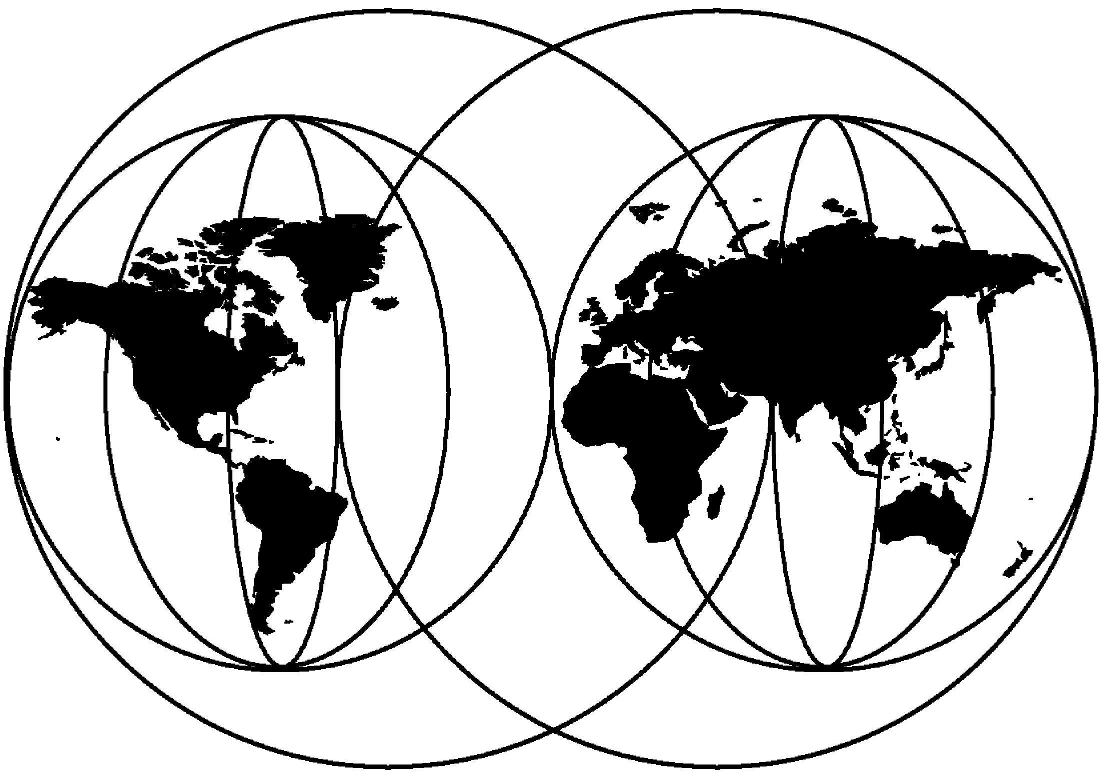
OS/390 Parallel Sysplex Configuration Volume 1: Overview

Frank Kyne, Luiz Fadel, Michael Ferguson, Rafael Garcia,
Keith George, Masayuki Kojima, Alan Murphy, Henrik Thorsen



International Technical Support Organization

www.redbooks.ibm.com

SG24-5637-00

International Technical Support Organization

SG24-5637-00

OS/390 Parallel Sysplex Configuration Volume 1: Overview

August 2000

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix A, "Special Notices" on page 233.

First Edition (August 2000)

This edition applies to

Program Name, Program Number	Version, Release Number
CICS/ESA, 5655-018	4.1
CICS TS for OS/390, 5655-147	1.3
DB2 for OS/390, 5655-DB2	5.1
DB2 UDB for OS/390, 5645-DB2	6.1
DB2 UDB for OS/390, 5675-DB2	7.1
DFSMS/MVS, 5695-DF1	1.5
IMS/ESA, 5695-176	5.1
IMS/ESA, 5655-158	6.1
IMS/ESA, 5655-B01	7.1
System Automation for OS/390, 5645-045	1.3

for use with:

Program Name	Version, Release Number
OS/390, 5645-001	1.3
OS/390, 5647-A01	2.4
OS/390, 5647-A01	2.5
OS/390, 5647-A01	2.6
OS/390, 5647-A01	2.7
OS/390, 5647-A01	2.8
OS/390, 5647-A01	2.9
OS/390, 5647-A01	2.10

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. HYJ Mail Station P099
2455 South Road
Poughkeepsie, New York 12601-5400

Contents

OS/390 Parallel Sysplex Configuration Volume 2: Cookbook

OS/390 Parallel Sysplex Configuration Volume 3: Connectivity

Figures

OS/390 Parallel Sysplex Configuration Volume 1: Overview

OS/390 Parallel Sysplex Configuration Volume 2: Cookbook

OS/390 Parallel Sysplex Configuration Volume 3: Connectivity

Tables

OS/390 Parallel Sysplex Configuration Volume 1: Overview

OS/390 Parallel Sysplex Configuration Volume 2: Cookbook

OS/390 Parallel Sysplex Configuration Volume 3: Connectivity

Preface

The Parallel Sysplex Configuration redbooks consist of three volumes that help you design and configure a Parallel Sysplex. These redbooks assist you in making the key hardware and software choices necessary to meet your performance, capacity, and availability requirements.

In addition to discussing many sysplex-related topics not documented elsewhere, these books contain, for the IBM large systems customer, practical information that has been collected from many sources and brought together into a single, comprehensive reference.

The redbooks provide rules of thumb for configuring a Parallel Sysplex. They discuss central processing complex sizing, CF sizing, the Sysplex Timer, connectivity guidelines, the implications of Parallel Sysplex for IBM software such as OS/390 V2R8 and related subsystem configurations, and useful tools and services. Network, systems management and availability considerations are also discussed.

These redbooks are a starting point for those involved in designing and configuring a Parallel Sysplex. They also may be used by those already having a Parallel Sysplex when further exploitation is being considered. The books refer to other relevant publications that cover specific areas in more depth.

These books are an update to an existing set of books:

- OS/390 MVS Parallel Sysplex Configuration Volume 1: Overview, SG24-2075
- OS/390 MVS Parallel Sysplex Configuration Volume 2: Cookbook, SG24-2076
- OS/390 MVS Parallel Sysplex Configuration Volume 3: Connectivity, SG24-2077

New information since the last books were published (January 1998) has been added, and information pertaining to back levels of hardware and software has been removed. The old *books* continue to be orderable for customers using those hardware or software levels.

The Team That Wrote These Redbooks

These projects were designed, managed and partly documented by:

Henrik Thorsen International Technical Support Organization, Poughkeepsie
Frank Kyne International Technical Support Organization, Poughkeepsie

The authors of these documents are:

Erik Bach	IBM Denmark	**Kuno Kern**	IBM Germany
Chris Barwise	IBM UK	**Masayuki Kojima**	IBM Japan
Frank Byrne	IBM UK	**Frank Kyne**	IBM Ireland
Dave Clitherow	IBM UK	**Geoff Leong**	IBM Canada
Richard Cwiakala	IBM USA	**Jock Marcks**	IBM Australia
Noshir Dhondy	IBM USA	**Dennis Moore**	IBM USA
Luiz Fadel	IBM Brazil	**Alan Murphy**	IBM Ireland
Linda Ferdinand	IBM USA	**Rachel Pickering**	IBM UK
Michael Ferguson	IBM Australia	**Alvaro Salla**	IBM Brazil
Per Fremstad	IBM Norway	**Alain Servais**	IBM France
Rafael Garcia	IBM USA	**Keld Teglgaard**	IBM Denmark
Keith George	IBM UK	**Henrik Thorsen**	IBM Denmark
Dick Jorna	IBM Netherlands	**Ian Waite**	IBM UK
Kazumasa Kawaguchi	IBM Japan		

These publications are the result of a series of residencies conducted at the International Technical Support Organization, Poughkeepsie Center. However, much valuable information and assistance was received from various organizations across IBM. The production of these redbooks is a team effort. The organizations that participated include:

- **Washington System Center, US M&S**
- **S/390 Parallel Center, S/390 Division**
- **S/390 Development, S/390 Division**

Also, thanks to the following people for reviewing these publications, providing material and offering invaluable advice and guidance. Their contributions have made these books eminently more accurate and more readable than they otherwise might have been.

Bill Abbot	Computer Associates
Robert M. Abrams	IBM USA
Riaz Ahmad	IBM USA
Michelle Ahearn	Computer Associates
Jay Aiken	IBM USA
Maria Sueli Almeida	ITSO Almaden
Steve Anania	IBM USA
Jaime Anaya	IBM USA
Stuart Arnell	IBM UK
John Arwe	IBM USA
Alan Ashcroft	IBM UK
Judi Bank	IBM USA
Paola Bari	IBM USA
Harley Beier	IBM USA
Ed Berkel	IBM USA
Evanne Bernardo	IBM USA
Connie Beuselinck	IBM USA
Page Borchetta	IBM USA
Mark Brooks	IBM USA
Jane Brockbank	IBM UK
Paolo Bruni	ITSO Almaden
Fred Bucci	IBM USA
Charlie Burger	IBM USA
Chuck Calio	IBM USA
John Campbell	IBM USA
Walt Caprice	IBM USA
Bernice Casey	IBM USA
Jose Castano	IBM USA
Alfred Christensen	IBM USA
Rich Conway	IBM USA
Roy Cornford	IBM UK
Curt Cotner	IBM USA
Mike Cox	IBM USA
Mary Crisman	IBM USA
Vijay Dalal	IBM USA
Jerry Dearing	IBM USA
Mac Devine	IBM USA
George Dillard	IBM USA
Kristin Donceel	IBM Belgium
Terry Draper	IBM UK
James Doyle	IBM USA
Joann Duguid	IBM USA
Gregory Dunlap	IBM USA
Mary Duval	IBM USA
Greg Dyck	IBM USA
Scott Fagen	IBM USA
Bob Fantom	IBM USA
Erika Fenn	Software AG
Martin Ferrier	IBM UK
John Fitch	IBM USA
Jonathan Follows	IBM ITSO Raleigh
Jim Fyffe	IBM USA
David Fox	IBM USA
Peter Garvin	IBM USA
Rob Geiner	IBM USA
Tony Giaccone	IBM USA
Rusty Godwin	IBM USA
Dionne Graff	IBM USA
Steve Grabartis	IBM USA
Jennifer Green	IBM USA
Susan Greenlee	IBM USA
Bob Haimowitz	IBM USA
Jim Hall	IBM USA
Dick Hannan	IBM USA
Ricardo Haragutchi	IBM Brazil
Johnathan Harter	IBM USA
Evan Haruta	IBM USA
Steve Heinbuch	IBM Canada
Audrey Helffrich	IBM USA
Gary Henricks	IBM USA
Steffen Herzog	IBM Germany
Ken Hewitt	IBM Australia
Juergen Holtz	IBM Germany
Janice Hui	IBM Australia
Patricia Jakubik	IBM USA
Dave Jeffries	IBM France
Curt Jews	IBM USA
Kim Johnson	IBM USA
Paul Johnson	IBM UK
Michael Jordan	IBM USA
Jeff Josten	IBM USA
Ken Kapulka	IBM USA
Gary King	IBM USA
Carl Klitscher	IBM New Zealand
Raimo Korhonen	CompMeas, Finland
Jeff Kubala	IBM USA
Ravi Kumar	IBM USA
Paul Lekkas	IBM Germany
Dave Levish	IBM USA
Rich Z. Lewis	IBM USA
Patrick LiVecchi	IBM USA
Alan Little	IBM USA
Sherman Lockard	IBM USA
Nick Matsakis	IBM USA
John Matush	IBM USA
Geoff Meacock	IBM UK
Kershaw Mehta	IBM USA
Geoff Miller	IBM USA
Roy Moebus	IBM USA
Thomas Murphy	IBM USA
Stephen Nichols	IBM USA
Arthur Neil	IBM UK
Geoff Nicholls	IBM Australia
Jeff Nick	IBM USA
Madeline Nick	IBM USA
Ron Northrup	IBM USA
Marianne Orman	IBM UK
Ron Parrish	IBM USA
Betty Patterson	IBM USA
Anthony Pearson	IBM USA
Kris Perry	IBM USA
James Perlik	IBM USA
David B. Petersen	IBM USA
Silvio Podcameni	IBM Brazil
Paul Powell	IBM USA
Richard Prewitt	IBM USA
David Raften	IBM USA
Tom Rankin	IBM USA
Patrick Rausch	IBM USA
Frank Ricchio	IBM USA
Paul Rogers	IBM USA
Tom Russell	IBM Canada
Eugene Sale	IBM USA
Hany Salem	IBM USA
Chris Saul	IBM USA
Jonathan Scott	IBM UK
Joachim Schmalzried	IBM Germany
Martin Schulman	IBM USA
Tom Shaw	IBM USA
Joseph Shelak	IBM USA
Sim Schindel	IBM France
Carol Shanesy	IBM USA
Beth Somers	IBM USA
Orrin Stevens	Computer Associates
Nat Stevenson	IBM USA
Bill Stillwell	IBM USA
Jeff Stokes	IBM USA
David Surman	IBM USA
John Swift	IBM UK
Lisa Taylor	IBM USA
Clarisse Taaffe-Hedglin	IBM USA
Michael Teuffel	IBM Germany
Douglas Trimboli	IBM USA
Ken Trowell	IBM USA
John Tuccio	IBM USA
Søren Understrup	IBM Denmark
Robert Vaupel	IBM Germany
Ambrose Verdibello	IBM USA
Phil Wakelin	ITSO Almaden
Kathy Walsh	IBM USA
Sharon Wang	IBM USA
James Warnes	IBM USA
Vern Watts	IBM USA
Klaus Wegele	IBM Germany
Maryela Weihrauch	IBM USA
Suzie Wendler	IBM USA
Gail Whistance	IBM USA
David Whitney	IBM USA
Joseph Winkelbauer	IBM USA
Lynn Winkelbauer	IBM USA
David Yackel	IBM USA
Douglas Zobre	IBM USA

Further thanks to the staff and the editorial team of the International Technical Support Organization, Poughkeepsie.

To everyone else who contributed, our sincere thanks.

Comments welcome

Your comments are important to us!

We want our redbooks to be as helpful as possible. Please send us your comments about this or other redbooks in one of the following ways:

- Fax the evaluation form found in "IBM Redbooks evaluation" on page 277 to the fax number shown on the form.
- Use the online evaluation form found at `http://www.redbooks.ibm.com/`
- Send your comments in an Internet note to `redbook@us.ibm.com`

Chapter 1. Introduction to the Configuration of a Parallel Sysplex

These redbooks are designed as cookbooks. They contain a series of "recipes" you may use to design your new Parallel Sysplex or enhance your existing Parallel Sysplex. Which "recipes" you use and in what order you use them depends on what you want to achieve: the books are *not* intended to be read sequentially. These books should not be your only source of information. They are intended to consolidate other sources of information at a suitable level of detail.

For a list of the sources used, refer to:

- "Recommended Sources of Further Information" at the beginning of each chapter.
- The packages and books listed in Appendix B, "Related Publications" on page 237.

Recommended Sources of Further Information

The following sources provide support for the information in this chapter:

- *OS/390 MVS Setting Up a Sysplex*, GC28-1779
- *OS/390 Parallel Sysplex Overview: Introducing Data Sharing and Parallelism in a Sysplex*, GC28-1860
- The Parallel Sysplex task roadmap, available on the Web site: `http://www.s390.ibm.com/pso/task.html#roadmap`

1.1 How to Use These Redbooks - Read This First

This section contains a series of roadmaps to guide you through the books in a fast and efficient way. The overall roadmap is like a long-distance route planner to determine your overall path through the books. Each chapter then contains a more detailed "local roadmap" to guide you through the chapter efficiently.

Each chapter also contains a list of sources of relevant information recommended for further reading.

Welcome to the S/390 Parallel Sysplex Home Page!

In line with IBM's direction to provide more information via the Internet, there is now a Web site dedicated to the topic of Parallel Sysplex. The site is intended to provide up-to-date Parallel Sysplex-related information, and to act as a hub from which you can be directed to other relevant sites. To make the use of this site easier, you should add the address of the following Web site to your bookmarks in your Web browser:
`http://www.s390.ibm.com/pso/`

At the time of writing, this Web site was organized into the following sections:

- News and events related to Parallel Sysplex.
- Library and education pages. These contain information about Parallel Sysplex education offerings, information about the announced CF Levels, and a list of the available Parallel Sysplex-related white papers.
- Availability characteristics. These are short documents that discuss how OS/390 features can be used to improve the availability of your Parallel Sysplex.
- Announcement information for latest IBM S/390 processors and major software products.
- Parallel Sysplex customer success stories.
- IBM Parallel Sysplex-related service offerings. This includes information on the IBM certification program and the Geographically Dispersed Parallel Sysplex offering.
- Downloads. This provides a link to the OS/390 Integration Test Web site where sample Parmlib members and other system definitions can be downloaded.
- Support. This provides links to the OS/390 Enhanced HOLDDATA site, the Washington Systems Center FLASHes, and information about how VM/ESA can be used to create a test Parallel Sysplex environment.

A Few Words on the Terminology Used in These Books

To avoid confusion, you should read the following list of terms as used throughout these books. The terms that we use may be different than you have seen elsewhere, but they are used consistently within these books:

CF rather than Coupling Facility partition.

CF link rather than coupling link, ISC link, or Coupling Facility channel.

CFCC rather than Coupling Facility Control Code.

CP rather than CPU, engine, or processor.

CPC rather than CEC, processor, model, computer, server, or machine.

DASD rather than disk.

External link identifies any CF link that connects between different CPCs. This includes ISC (50 MB and 100 MB) links and ICB (copper) links.

Failure-Independent CF identifies any CF that resides in a CPC that does not contain any other images (CF or OS/390) from the same sysplex.

Fiber link rather than ISC or real link. Contrast fiber links to ICs (internal) and ICBs (copper) CF links.

GBP rather than group buffer pool.

IC rather than Internal Coupling Link.

ICB rather than Integrated Cluster Bus or Cluster Bus Link.

ICF rather than Internal Coupling Facility.

Image rather than OS/390 image.

LP rather than PR/SM LPAR partition, partition, logical partition, or LPAR.

MVS and *OS/390* are used in the same context.

Parallel Sysplex, rather than Parallel Sysplex cluster.

RACF and *OS/390 Security Server* are used in the same context.

Standalone CF rather than external CF or 9674.

S/390 Partners in Development, rather than solution developers, business partners, or independent software vendors (ISVs).

Storage rather than memory or RAM.

Sysplex rather than complex or systems complex.

VTAM, *OS/390 Communications Server*, *e-network Communications Server*, *SecureWay Communications Server*, and *Communications Server for OS/390* are all used in the same context.

For more information on the terms and acronyms used in this book, refer to the sections “Glossary” on page 249 and “List of Abbreviations” on page 269.

Table 1 is the “roadmap” that will help to guide your route through the book. The roadmap can help you read the right information in the right order.

Table 1 (Page 1 of 2). Parallel Sysplex Configuration Roadmap

You want to configure:	**If you are especially interested in:**	**Then refer to:**
Parallel Sysplex		
	An introduction to these books and to the task of configuring a Parallel Sysplex: How to use these books What is a Parallel Sysplex? What are the advantages of a Parallel Sysplex?	Chapter 1, “Introduction to the Configuration of a Parallel Sysplex” on page 1
	High level design concepts for Parallel Sysplex What are other “plexes”?	Chapter 2, “High-Level Design Concepts for Parallel Sysplex” on page 17
	Availability in Parallel Sysplex Software considerations Hardware considerations Network considerations Disaster recovery	Chapter 3, “Continuous Availability in Parallel Sysplex” on page 89
	Workloads and Parallel Sysplex e-business workloads Transaction management Database management Batch workloads TSO/E in a Parallel Sysplex	Chapter 4, “Workloads in Parallel Sysplex” on page 141
	CPC and CF configuration Technology choices CF options CPC capacities CF exploiters	1.0, “CPC and CF Configuration in Parallel Sysplex” on page 3 in Volume 2: Cookbook, SG24-5638
	Details of CF structures: What size should I define? Where should I place my structures? What about CF volatility? What about structure rebuild?	2.0, “A Structured View of CF Structures” on page 59 in Volume 2: Cookbook, SG24-5638
	Parallel Sysplex connectivity	1.0, “I/O Connectivity in Parallel Sysplex” on page 3 in Volume 3: Connectivity, SG24-5639
	Network connectivity General issues Gateway options	2.0, “Network Connectivity for Parallel Sysplex” on page 21 in Volume 3: Connectivity, SG24-5639
	Sysplex Timer considerations	3.0, “Sysplex Timer Considerations” on page 71 in Volume 3: Connectivity, SG24-5639
	How do I configure consoles in a Parallel Sysplex? HMC considerations	4.0, “Consoles and Parallel Sysplex” on page 103 in Volume 3: Connectivity, SG24-5639

Table 1 (Page 2 of 2). Parallel Sysplex Configuration Roadmap

You want to configure:	If you are especially interested in:	Then refer to:
	Systems management in Parallel Sysplex RACF Performance products Operations management SMP/E HCD	Appendix A, "Systems Management Products for Parallel Sysplex" on page 123 in Volume 3: Connectivity, SG24-5639
	Tools to assist in the process of designing Parallel Sysplex configurations	Appendix A, "Tools and Services Catalog" on page 161 in Volume 2: Cookbook, SG24-5638
	RMF reports related to Parallel Sysplex	Appendix B, "RMF Reporting in Parallel Sysplex" on page 189 in Volume 2: Cookbook, SG24-5638
	Tuning DB2 GBP, SCA and IRLM structures	Appendix D, "Tuning DB2 Structures" on page 225 in Volume 2: Cookbook, SG24-5638
	What hardware features are available on which IBM CPCs (for 9672 G4 and prior models, 9674, 9021-711, and 9121, see Table 2 on page 15 and Table 3 on page 16 in this chapter). Also see SG24-2075, SG24-2076, SG24-2077.	Appendix E, "Functional Differences between IBM 9672-Based CPCs" on page 233 in Volume 2: Cookbook, SG24-5638

1.2 The Purpose of These Books

These books have been written to help you configure a Parallel Sysplex. The emphasis in these books is on *high level configuration design.* By this we mean that you:

- Order the right hardware and software.
- Decide how it will work. For example:
  - Will all subsystems run on every image?
  - How will my subsystems work in normal operation?
  - How will I operate and manage the subsystems/sysplex?
  - What happens if a component fails?
  - How will systems management functions work across the sysplex?

The books are not designed to help you justify the use of a Parallel Sysplex, nor to implement it (install it and make it work). They are designed with two purposes in mind:

1. For new users of Parallel Sysplex.

   To help you make the *initial* design decisions so that implementation does not uncover additional hardware or software that is needed, or fundamental misconceptions about how it should work.

2. For existing users of Parallel Sysplex.

   To provide information regarding enhancing your Parallel Sysplex by utilizing the latest functions and facilities of new hardware and software.

For information to help you estimate the value of Parallel Sysplex to your business, ask your IBM representative for a copy of the BV390 Package from MKTTOOLS.

Initially you will probably only be interested in approximate sizings. Later you will need to be sure that the hardware and software ordered is both correct and complete. If you have not thoroughly considered the operational issues listed in these books, you may find that you have to make alterations at a later date.

You can expect to make several iterations through the books, at different levels of detail, as your Parallel Sysplex evolves from an idea to a firm decision.

These books bring together new information and information that is already available but scattered among many different sources. They contain the latest information *based on the experience at the time of writing.* As Parallel Sysplex continues to evolve, you should always check the latest information.

These books contain information about the environments that are capable of being configured now. These are primarily the DB2, IMS TM, IMS DB, CICS, and CICS/VSAM RLS transaction processing and data sharing environments. These books also contain information about other exploiters including OS/390 functions such as JES, RACF, VTAM, BatchPipes, system logger, shared tape, Enhanced Catalog Sharing, and GRS.

Recommendation to Check Background Information

The content of these books is based on many sources. These sources include information from relevant Web sites, FORUMS, whitepapers, other redbooks, product documentation and so forth.

For a deeper understanding of the background information, always check the detailed information.

1.3 Base Hardware Levels for These Books

The minimum hardware level required for all the facilities described in this book is the IBM 9672 G5 CPC. Many, but not all, of the facilities will work on previous levels of CPCs. Mixed levels of hardware may cause some functions of a Parallel Sysplex to work differently.

Information about the older levels of hardware—9021s, 9121s, 9674s, and pre-G5 9672s—was removed to make the document more readable. Also, at the time of writing many customers have moved to the G5 and later generations of 9672 CPCs, so it was felt that the old information was no longer required.

More information regarding facilities available on previous levels of CPCs can be found in 1.8, "Function Support Table for Pre-G5 9672 CPCs" on page 15. Also, for those customers that are still using these older levels, the previous level of these books is still orderable. The names and order numbers are:

- *OS/390 MVS Parallel Sysplex Configuration Volume 1: Overview*, SG24-2075
- *OS/390 MVS Parallel Sysplex Configuration Volume 2: Cookbook*, SG24-2076
- *OS/390 MVS Parallel Sysplex Configuration Volume 3: Connectivity*, SG24-2077

1.4 Main Reasons for Parallel Sysplex

The main reasons for moving to a Parallel Sysplex are briefly discussed here. They are covered more thoroughly in *OS/390 Parallel Sysplex Overview: Introducing Data Sharing and Parallelism in a Sysplex*, GC28-1860.

For a discussion of the key design points for Parallel Sysplex, refer to "S/390 Cluster Technology: Parallel Sysplex," in *IBM Systems Journal, Volume 36 No. 2*, G321-0127.

1.4.1 Continuous Application Availability

When you have a single copy of any system component, hardware, software, or data, you are inevitably exposed to system outages because of either failure of the component or because of planned changes to the component that require it to be taken offline.

One of the goals of Parallel Sysplex is to eliminate the impact that scheduled outages have on application availability, and minimize the effects of an unscheduled outage by allowing work to continue executing on the remaining systems in the sysplex. This requires, among other things, that the system be designed for redundancy within the Parallel Sysplex. Applications must be capable of running across multiple systems, with access to the data being possible from at least two systems. If at least two instances of a resource exist, your applications can continue to run even if one of the resources fails.

It is important to remember that Parallel Sysplex only provides the facilities for continuous availability. Parallel Sysplex on its own will not eliminate scheduled or unscheduled application outages; the application itself must also be designed for continuous availability. Sharing data is only one part of this design.

A classic example is where IMS databases are made unavailable to online users to allow for batch updates. These batch update programs could be re-written as Batch Message Processing (BMP) programs which can perform updates while still having the databases available for online users.

Another example is that data in a DB2 database may be unavailable while it is being reorganized (and this will occur no matter how many data sharing members you have). DB2 for OS/390 reduces the impact of database reorgs by the provision of the online reorg utility. This utility allows full access to the data for *most* of the time that the reorg is taking place.

Similar considerations apply to other IBM subsystems.

IBM also provides a facility called FlashCopy which allows data on DASD volumes to be copied while it is still being accessed by online users. FlashCopy is available in the 2105 Enterprise Storage Server. FlashCopy is completely self-contained within the 2105 and does not require any additional host software.

On the earlier Ramac Virtual Array (RVA) DASD subsystem, a similar function called SnapShot was available. SnapShot is a combination of hardware features in the RVA and host software features provided by IBM Extended Facilities Product (IXP). Both products interact with DFSMSdss to provide high speed dumps, drastically reducing the impact of data backups.

More discussion on continuous application availability is found in Chapter 3, “Continuous Availability in Parallel Sysplex” on page 89, and in the *Parallel Sysplex Availability Checklist* white paper, available via the Parallel Sysplex home page, or directly from:
`http://www.s390.ibm.com/ftp/marketing/position/availchk_parsys.pdf`

1.4.2 Workload Balancing

Without workload balancing, installations with multiple CPCs have often had to upgrade one CPC to provide more capacity, while another CPC may have had spare capacity. The alternative was to redistribute work manually, which is time-consuming and can only handle short-term imbalances. Manual intervention also has the potential to introduce errors and could not be used to dynamically and constantly balance workload across the installed CPCs.

With a Parallel Sysplex, the basic framework exists for workload balancing. There is a small cost for doing workload balancing in a Parallel Sysplex, but the cost by and large is not related to the number of systems, and the workload balancing advantages are significant. Many subsystems exploit this framework to redistribute work across systems, thus allowing the systems to be run at higher utilizations, and allowing spare capacity anywhere in the Parallel Sysplex to be used to satisfy the demands of the workload.

All the major IBM subsystems—CICS, DB2, IMS, MQSeries, TSO, VTAM, TCP/IP, and JES—contain support for Parallel Sysplex workload balancing. This is discussed further in 2.6, “Dynamic Workload Balancing in Parallel Sysplex” on page 83. More discussion on subsystem exploitation of workload balancing is found in Chapter 4, “Workloads in Parallel Sysplex” on page 141.

1.4.3 Nondisruptive Addition of Scalable CPC Capacity

If a part of your workload cannot be split between multiple images or has grown larger than the largest available single-image system, Parallel Sysplex may offer a solution to allow you to add additional capacity (both MIPS and CPC storage) without rewriting the applications or re-partitioning the databases. For example, it may allow multiple instances of the application to share the data across multiple images.

More discussion on how to share data in a Parallel Sysplex is found in 4.3, “Database Management in Parallel Sysplex” on page 194.

1.4.3.1 Granularity

Once a Parallel Sysplex has been set up, the increments in which capacity can be added can vary from small to large quite easily. Capacity is available on demand. You can match your investment in hardware more closely to the business needs, rather than being forced to take large capacity (and cost) increments at arbitrary points in your workload growth. This is particularly true if you have a number of different CPC types installed, and the flexibility to run your work on any of those CPCs. So if you require an additional 200 MIPS, you upgrade the CPC with 200 MIPS engines. If you just require an additional 100 MIPS, then you upgrade the CPC with 100 MIPS engines.

Another option is to add an additional CPC. The ability with Parallel Sysplex to run work on any image, coupled with the cloning facilities in OS/390, make it far simpler to add a new image than was the case in the past. The feasibility of this option is an important consideration when designing your Parallel Sysplex. A new CPC can be added quickly, providing ESCON director ports are available for access to the I/O subsystems, and extensive use of symbolic parameters permits the cloning of software.

For more information on CPCs participating in a Parallel Sysplex, refer to 1.0, "CPC and CF Configuration in Parallel Sysplex" on page 3 in Volume 2: Cookbook, SG24-5638.

For more information on connectivity in a Parallel Sysplex, refer to 1.0, "I/O Connectivity in Parallel Sysplex" on page 3 in Volume 3: Connectivity, SG24-5639.

1.4.3.2 Just-In-Time Nondisruptive Growth

With a Parallel Sysplex, more options exist for providing more capacity. The previous options were to upgrade or replace existing CPCs (vertical growth). The Parallel Sysplex allows new CPCs to be added alongside existing CPCs (horizontal growth). There is also the hybrid option of both vertical and horizontal growth.

Which option you choose will depend on several factors, including:

- Availability requirements
- Whether there is an upgrade path
- Whether there is a larger CPC available
- CP speed considerations
- Restrictions imposed by existing configuration
- Cost

A Parallel Sysplex may allow CPCs to be added or removed nondisruptively. Upgrades might be accommodated by taking a CPC out of the Parallel Sysplex and upgrading it, while continuing to run the workload on the remaining CPCs in the Parallel Sysplex. The upgraded CPC can then be reintroduced to the Parallel Sysplex when testing is complete.

Removing or adding a system requires certain steps to be executed so that the removal or addition happens nondisruptively to the other work within the sysplex. There are sysplex and multisystem application considerations that need to be taken into account.

Adding additional capacity by horizontal growth in a Parallel Sysplex will not immediately benefit a single-threaded application if it is unable to share data between multiple instances. However, other work in the system can be moved or directed away from the system running the single-threaded application.

Vertical Growth: This is the traditional option, and requires little discussion. Previously, this would almost certainly be disruptive, unless you are upgrading within a Parallel Sysplex, in which case you can keep the Parallel Sysplex workload running on other CPCs while the upgrade takes place.

However, the introduction of the Capacity Upgrade on Demand (CUoD) feature on the G5 and later 9672 CPCs provides the ability to (in most cases) nondisruptively add additional CPs to an existing processor. More information on CUoD is available on the Internet at:
`http://www.s390.ibm.com/pes/apg/`

Horizontal Growth: If CUoD is not available on the installed CPCs, adding CPCs to an existing Parallel Sysplex can be an easier way of adding processing capacity. The additional CPC can probably be installed during normal operation, without time pressure or risk to the existing service. Connection to the existing Parallel Sysplex can often be achieved *nondisruptively,* though there are some situations in which nondisruptive connection is not possible. Testing can then be performed as required. When the new CPC has been proven to everyone's satisfaction, work can then be gradually migrated onto it, with less critical work migrated first to allow further validation.

1.4.3.3 Scalability

With Parallel Sysplex, it is possible to have an application that can run without modification on the smallest IBM 9672 CPC or on a Parallel Sysplex of multiple IBM 9672s. The difference in capacity is of several orders of magnitude. For a data sharing application, this could allow serious sizing errors at the design stage to be overcome without an application rewrite (which would probably be much more costly than any additional hardware required).

Another consideration for scalability in a Parallel Sysplex is that it is not subject to the same "drop off" in benefit from adding more images as a tightly coupled multiprocessing (MP) CPC is when more CPs are added. As more images are added to the Parallel Sysplex, you achieve *almost linear growth.* This is shown graphically in Figure 1.

Some information on performance in a Parallel Sysplex is found in 1.0, "CPC and CF Configuration in Parallel Sysplex" on page 3 in Volume 2: Cookbook, SG24-5638.

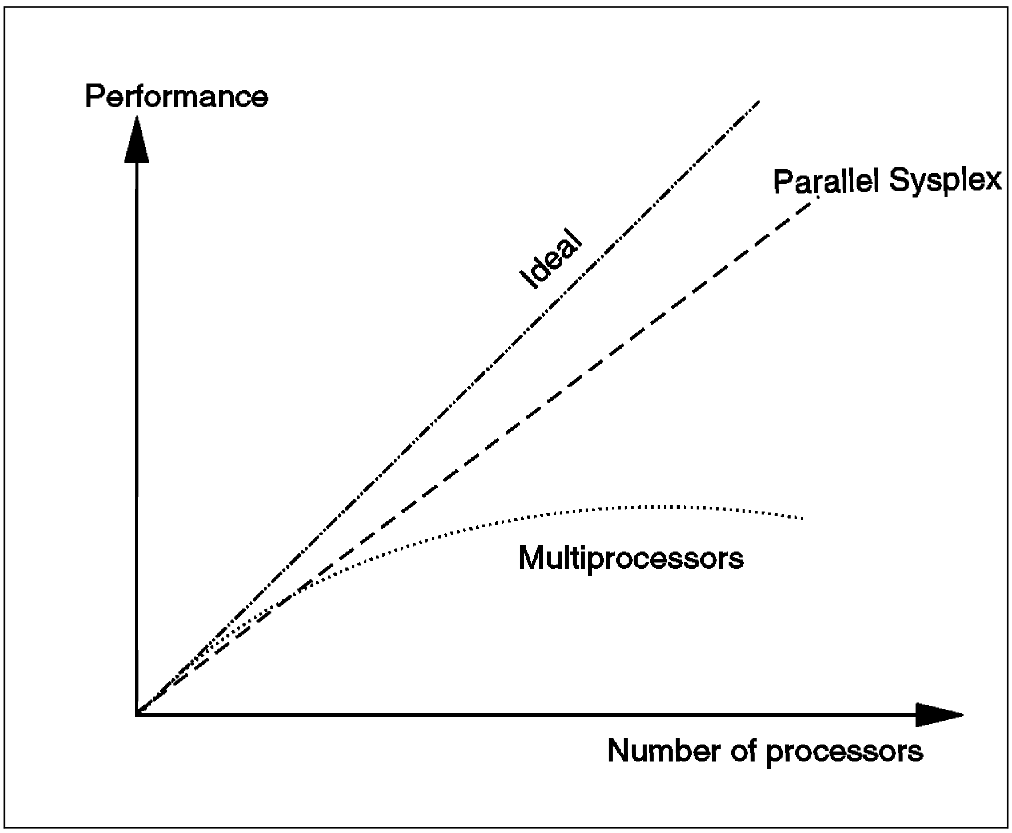


Figure 1. Scalability in Parallel Sysplex

1.4.4 Reduced Total Cost of Computing

Compared to other mainframe chip technologies, the CMOS-based IBM 9672, Parallel Sysplex may offer you a less expensive way to add processing capacity. Alternatively, if you have multiple CPCs today and more capacity is needed on one, but there is spare capacity on another, it may be less expensive to implement a Parallel Sysplex than to upgrade the system.

Even if it is not immediately less expensive, the benefits of a Parallel Sysplex and the flexibility to balance workload between the CPCs may make implementing a Parallel Sysplex attractive at this stage. Successive generations of IBM 9672 CPCs provide significant savings with respect to energy, facilities requirements, and maintenance costs.

1.4.4.1 Application Portfolio Investment Protection

Existing applications can often exploit the availability benefits of the Parallel Sysplex environment with few or no changes. The most significant common change that may be required is that you may want to alter the frequency of commits or checkpoints to reduce the contention for shared resources.

The majority of changes needed for applications to exploit a data sharing or dynamic workload balancing environment are implemented in the IBM subsystems such as CICS, DB2, IMS, MQSeries, and VTAM.

In essence, this means that Parallel Sysplex provides compatibility such that your existing applications will continue to run.

1.4.4.2 Software Pricing in Parallel Sysplex

The ability to aggregate the MSUs of the CPCs in your Parallel Sysplex provides the possibility of reduced software licenses when going to a Parallel Sysplex. Options such as Parallel Sysplex License Charges (PSLC) and Measured Usage License Charge (MULC) should be investigated to calculate the benefit.

Solution developers' (ISV) software licenses also need to be investigated to get the complete picture of software licenses in the Parallel Sysplex. There is some discussion about managing non-IBM S/390 software costs in a white paper entitled *How to Better Manage Your S/390 Costs from Independent Software Vendors* which can be found on the Web at:
`http://www.s390.ibm.com/marketing/gf225090.html`

There is also a mailing list that you can subscribe and contribute to, where customers can discuss concerns and information about their ISV costs. The mailing list can be accessed at:
`http://www.can.ibm.com/isvcosts`

There are also a number of specialist companies that offer contract review and negotiation services. Your IBM representative can provide a list of some these companies.

Another way that Parallel Sysplex can help you to reduce costs is to use it to route selected work to just certain images. For example, you may have a two-CPC Parallel Sysplex configuration: one CPC is 450 MIPS and the other is 200 MIPS. Regardless of how many or how few users each product may have, a software license is required for each CPC that it runs on. For example, if you have a product that is only using 20 MIPS, then it may prove cost effective to use Parallel Sysplex to route all work that uses that piece of software to the smaller

200 MIPS CPC. Thus a license for the product will only be required on the smaller CPC. It is much easier to do this in a Parallel Sysplex than would have been the case previously.

For information about IBM software licenses in the Parallel Sysplex, refer to:

- Announcement letters:
  - *Clarification of Parallel Sysplex Pricing Terms and Conditions Including Changes to the Parallel Sysplex Licence Charge Exhibit* (198-001)
  - PSLC Level C Price For S/390 Software Reduces the Cost of Growing Workloads (298-355)
  - *New Terms and Conditions for S/390 Software Products* (298-383)
  - *New S/390 Usage Pricing for S/390 Software Products* (298-357)
  - *S/390 New Application Growth Environment and S/390 New Applications Licence Charge* (298-371)
  - *PSLC Level D Price Introduction for S/390 Software Products* (299-052)
  - *Measured Usage License Charges* ZA94-0154 (294-155)
- PRICE application in DIALIBM/IBMLINK.
- IBM representatives may request the SWPRICER package from MKTTOOLS. SWPRICER S/390 PSLC Software Pricer (US Version) is an OS/2 tool designed to assist in pricing S/390 software using Parallel Sysplex License Charges. There is also a Windows version of the tool, SWPWIN, available from MKTTOOLS.
- Further information may be obtained from your IBM representative, who can request the following packages from MKTTOOLS:
  - G3260594 Package - S/390 Software Pricing Reference Guide
  - EPPG Package - S/390 Software Pricing Presentation Guide
  - S390USE Package - S/390 Usage Pricing Presentation Guide
  - S390USEQ Package - S/390 Usage Pricing Questions and Answers
  - FASTREF Package - S/390 Fast Pricer Reference Card
  - GF225065 Package - S/390 Enterprise Computing and Software Pricing Directions
  - SWPG Package - Software Pricing Announcement Guide
  - OS390CK Package - Assists with OS/390 Version 2 Pricing Analysis
  - SWLOSREF Package - Software Licensing and Offering Reference Sheet (1997)
  - 390LOS Package - S/390 Software Pricing Presentation Guide (1997)
  - PSLCUS package (US version) (1998)

1.4.5 Single System Image

Parallel Sysplex can potentially provide a logical single system image to users, applications, and the network. In addition, Parallel Sysplex provides the ability to have a single point of control for your systems operations staff.

Single system image is discussed further in:

- Chapter 2, “High-Level Design Concepts for Parallel Sysplex” on page 17.
- Chapter 3, “Continuous Availability in Parallel Sysplex” on page 89.
- 1.0, “I/O Connectivity in Parallel Sysplex” on page 3 in Volume 3: Connectivity, SG24-5639

- 4.0, “Consoles and Parallel Sysplex” on page 103 in Volume 3: Connectivity, SG24-5639.
- Appendix A, “Systems Management Products for Parallel Sysplex” on page 123 in Volume 3: Connectivity, SG24-5639.

1.5 Value of Parallel Sysplex - A Summary

The value of Parallel Sysplex is best described by its ability to deliver the functions which meet the ever-increasing requirements of today's businesses. These requirements include:

- Continuous availability of applications.
- Reduction or elimination of planned application outages
- Scalability to virtually unlimited capacity to meet the high transaction volumes and response times of today and tomorrow.
- Investment protection of existing applications by providing functions which allow these applications to operate in the e-business environment without a complete rewrite.
- A secure environment for existing and e-business transactions.
- A development environment which provides the tools and languages to develop new applications for today and tomorrow.
- A platform for server consolidation, to reduce the cost and complexity of having a large server farm to manage.
- A simple growth path delivered with low incremental and total cost of computing.

Parallel Sysplex is ideally suited to today's environment and is continually being developed to meet new and changing requirements.

1.6 The Distinction between Basic and Parallel Sysplex

Parallel Sysplex evolved from the basic sysplex. A general description of each follows.

1.6.1 Basic Sysplex

In September 1990, IBM introduced the SYStems comPLEX, or sysplex, to help solve the difficulties of managing multiple OS/390 systems. This established the groundwork for simplified multisystem management through the Cross-System Coupling Facility (XCF) component of OS/390. XCF services allow authorized applications on one system to communicate with applications on the same system or on other systems. In a base sysplex, connectivity and communication between images is provided by channel-to-channel (CTC) links. The couple data set, which is shared between all of the images, holds control information and provides a mechanism for monitoring the status of the images. When more than one CPC is involved, a Sysplex Timer synchronizes the time on all systems.

Basic Sysplex Definition

A basic sysplex is that set of systems that share a sysplex couple dataset and all of which have the same sysplex name.

1.6.2 Parallel Sysplex

The basic sysplex laid the foundations for communications between subsystems on the participating OS/390 images, but these were insufficient to provide the speed and integrity necessary for data sharing. To provide this capability, a concept called the Coupling Facility (CF) was introduced. This was implemented in a Logical Partition (LP) of a CPC.

The design of the CF ensures that requests are queued in a logical order that is not dependent on the speed of the connected CPCs. This removes the unpredictable response time that is inherent in conventional devices such as DASD and CTC. The use of the CF by subsystems, such as IMS, DB2, and CICS/VSAM RLS ensures the integrity and consistency of data throughout the sysplex. The capability of linking many systems and providing multisystem data sharing makes the sysplex platform ideal for parallel processing, particularly for online transaction processing (OLTP) and decision support.

Parallel Sysplex Definition

A Parallel Sysplex is that set of systems within a sysplex that all have access to the same one or more CFs.

While a basic sysplex is an actual entity, with a defined name (the sysplex name), a Parallel Sysplex is more conceptual. There is no member or couple dataset anywhere that contains a name for the Parallel Sysplex, or a list of the systems it contains. Rather, it is the superset of a number of other plexes (RACFplex, VTAMplex, and so on) that all share the same CF or set of CFs. For more information refer to 2.3.3, "What Different 'Plexes Are There?" on page 39.

The G3 9672s introduced the ability to configure CPCs with Integrated Coupling Facility (ICF) engines. These are specialized CPs which can only run CFCC. ICFs provide lower entry cost into Parallel Sysplex, and provide more configuration flexibility. When ordering a 9672, a decision needs to made on how many CPs will be used to run OS/390 LPs and how many ICFs are needed. The 9672 is then configured as requested and delivered with some CPs to run OS/390 and the remaining ones as ICFs to run CFCC. The capacity of an ICF is subtracted from the capacity of the CPC when working out the software license charge.

In short, a Parallel Sysplex builds on the base sysplex capability, and allows you to increase the number of CPCs and images that can directly share work. The CF allows high performance, multisystem data sharing across all the systems. In addition, workloads can be dynamically balanced across systems with the help of workload management functions. IMS/ESA V6 shared message queues, CICSPlex SM, and VTAM generic resources are some examples of workload balancing implementations in Parallel Sysplex.

1.7 How Do I Quickly Get to Parallel Sysplex?

IBM provides end-to-end services to implement a Parallel Sysplex for you. IBM project assistance is tailored to your environment to plan, configure, define, install, and operate a Parallel Sysplex. Tailored starter set definitions are used for system and application enablement.

The project may include enablement of selected applications running data sharing in your production environment.

For more information, contact your IBM representative, who can access the EPSO or EPSOEMEA (for EMEA) PACKAGEs on MKTTOOLS. Also, refer to A.1.19, "EPSO (Enhanced Parallel Sysplex Offering)" on page 174 in Volume 2: Cookbook, SG24-5638.

1.8 Function Support Table for Pre-G5 9672 CPCs

As mentioned previously, information about IBM CPCs prior to the G5 9672s has been removed from these books. However, as a quick reference, Table 2 contains information about which CF levels are supported on those older CPCs. More information about specific EC level requirements and the functions contained in each CFLEVEL can be obtained from the Web site: `http://www.s390.ibm.com/products/pso/cftable.html`

Table 2. Pre-9672 G5 CF Level Support

CPC Models	CF Level 8	CF Level 7	CF Level 6	CF Level 5	CF Level 4	CF Level 3	CF Level 2	CF Level 1	CF Level 0
9674 C05	yes	yes	yes	yes	yes	na	na	na	na
9674 C04	yes	yes	yes	yes	yes	yes	na	na	na
9674 C03	no	no	no	yes	yes	yes	yes	yes	na
9674 C02	no	no	no	yes	yes	yes	yes	yes	na
9674 C01	no	no	no	no	yes	yes	yes	yes	yes
9672 R5	yes	yes	yes	yes	yes	na	na	na	na
9672 R4	yes	yes	yes	yes	yes	yes	na	na	na
9672 R3	no	no	no	yes	yes	yes	yes	yes	na
9672 R2	no	no	no	yes	yes	yes	yes	yes	na
9672 R1/E/P	no	no	no	no	yes	yes	yes	yes	yes
9021 711-based	no	no	no	yes	yes	yes	yes	yes	yes
9121 511-based	no	no	no	no	yes	yes	yes	yes	yes

Table 3 shows which Parallel Sysplex-related features are supported on different models of the older CPCs.

Table 3. Pre-9672 G5 Parallel Sysplex-related Features

CPC Models	ICB	ICF	IC	HiperLink	Dynamic CF Dispatch	Dynamic ICF Expansion	CF in LPAR	ICMF
9674 C05	no	na	na	yes	yes	na	yes	na
9674 C04	no	na	na	yes	yes	na	yes	na
9674 C03	no	na	na	yes	yes	na	yes	na
9674 C02	no	na	na	yes	yes	na	yes	na
9674 C01	no	na	na	yes	yes	na	yes	na
9672 R5	no	yes	no	yes	yes	yes	yes	yes
9672 R4	no	yes	no	yes	yes	yes	yes	yes
9672 R3	no	no	no	no	yes	no	yes	yes
9672 R2	no	no	no	no	yes	no	yes	yes
9672 R1/E/P	no	no	no	no	yes	no	yes	yes
9021 711-based	no	no	no	no	yes	no	yes	yes
9121 511-based	no	no	no	no	no	no	yes	no

Chapter 2. High-Level Design Concepts for Parallel Sysplex

In this chapter, we look into some of the considerations that have to be taken into account when designing your Parallel Sysplex configuration.

Recommended Sources of Further Information

The following sources provide support for the information in this chapter:

- *Batch Processing in a Parallel Sysplex*, SG24-5329
- *CICS/TS for OS/390 V1.3: Implementation Guide*, SG24-5274
- *DB2 UDB for OS/390 V6 Data Sharing: Planning and Administration*, SC26-9007
- *DB2 UDB for OS/390 V6 Performance Topics*, SG24-5351
- *IMS/ESA Version 6 Shared Queues*, SG24-5088
- *IMS/ESA V6 Parallel Sysplex Migration Planning*, SG24-5461
- *OS/390 MVS Initialization and Tuning Reference*, SC28-1752
- *OS/390 MVS Planning: Global Resource Serialization*, GC28-1759
- *OS/390 MVS Setting Up a Sysplex*, GC28-1779
- *OS/390 Parallel Sysplex Application Migration*, GC28-1863
- *OS/390 Parallel Sysplex Hardware and Software Migration*, GC28-1862
- *OS/390 Parallel Sysplex Overview*, GC28-1860
- *OS/390 Parallel Sysplex Test Report*, GC28-1963
- *OS/390 Security Server (RACF) Planning: Installation and Migration*, GC28-1920
- *S/390 Parallel Sysplex Resource Sharing*, SG24-5666
- *SNA in a Parallel Sysplex Environment*, SG24-2113
- *TCP/IP in a Sysplex*, SG24-5235
- *Using VTAM Generic Resources with IMS*, SG24-5487

Refer to the roadmap for this chapter to locate the section for which you need recommendations.

Table 4 (Page 1 of 2). Parallel Sysplex Configuration Roadmap - High-Level Design

You want to configure:	**If you are especially interested in:**	**Then refer to:**
Parallel Sysplex from a high-level perspective		
	Is Parallel Sysplex right for me? Which major IBM subsystems exploit Parallel Sysplex? What about S/390 Partners in Development (ISV) software? Where can I find other customer references?	2.2, “Deciding If Parallel Sysplex Is Right for You” on page 18
	How many Parallel Sysplexes do I need? What test configurations should I configure? Does everything have to be symmetrical? What other 'plexes are there, and how do they map to the Parallel Sysplex? Can I share DASD among sysplexes?	2.3, “Parallel Sysplex High-Level Design” on page 26

Table 4 (Page 2 of 2). Parallel Sysplex Configuration Roadmap - High-Level Design		
You want to configure:	**If you are especially interested in:**	**Then refer to:**
	CF architecture What are list, lock, and cache structures? How are they used and by what functions and products?	2.4, "CF Architecture" on page 71
	Data integrity in the Parallel Sysplex This section provides step-by-step examples on how it works. What are false, real, and total lock contention?	2.5, "Data Integrity and Buffer Pool Consistency Considerations" on page 78
	Dynamic workload balancing What is it? Who are the exploiters?	2.6, "Dynamic Workload Balancing in Parallel Sysplex" on page 83

2.1.1 Important Rules of Thumb Information

This book contains several rules of thumb for configuring a Parallel Sysplex.

Rules of Thumb Disclaimer

Parallel Sysplex is a fundamental architecture, complementing an architecture that has matured over thirty-five years. In the various stages of Parallel Sysplex evolution and exploitation, you can expect to see changes in design and subsystem exploitation.

Rules of thumb do not cover every situation and may be misleading if they are not fully understood. They usually work when the assumptions under which they are arrived at originally are correct and maintained, but these assumptions are not always understood or passed on with the rule of thumb, so it is not always easy to notice when they are violated.

For this reason, you should always check the latest information available (ask your IBM representative) to see if any changes have recently been made.

IBM's current intentions and plans stated in this book are subject to review. Announcement of any product is based on IBM's business and technical judgement.

2.2 Deciding If Parallel Sysplex Is Right for You

When Parallel Sysplex was first announced, it was primarily seen as being of interest only to very large customers. Also, it was at a time when availability requirements were not as challenging as they are today. However, as time has passed, requirements have changed, and Parallel Sysplex technology has seen many improvements. As a result, Parallel Sysplex is now a technology that potentially applies to many more customers—and this trend will accelerate with future hardware and software enhancements. In this section, we discuss the things to consider when deciding if Parallel Sysplex is the right technology for your company.

Firstly, do you have application availability requirements that are proving difficult to achieve? One survey found that outage costs across 10 industries ranged from $25,000 (US) an hour up to $6,500,000 (US) an hour—and this was before the advent of e-business! If you plan to use Parallel Sysplex to improve your availability, you need to review all of your planned and unplanned outages over the past year, and identify which ones could have been avoided in a Parallel Sysplex environment. Are you already configured for maximum availability? There is less benefit in moving to Parallel Sysplex if you are going to leave other single points of failure in your configuration.

Maybe you wish to move in order to get better utilization from your installed CPCs; in this case you need to review your workloads to decide which ones you can split over the various systems, and then consider the following questions:

- Will the splitting of these workloads have a significant enough effect to balance out your CPC utilizations?
- Can the workloads be split?
- Do most of the products required by these workloads support data-sharing and workload balancing?

Or maybe your company is getting serious about e-business, and you want to know which is the best platform to host these new applications. There are a number of requirements if you wish to provide a successful e-business application:

- The application must be available 24 hours a day, 7 days a week—remember that your customers could now be in any country in the world, and will expect your application to be available at any time that suits *them*.
- You must be able to provide consistent, acceptable, response times, and react to abrupt changes in the rate of requests.
- You want your customers to be able to see current data, not data as it existed twelve hours ago. This means the e-business application has to have access to the live data, not a point-in-time copy.
- You need to be able to fully utilize all the installed MIPS, while at the same time being able to protect selected critical applications.

Most e-business applications either access host-resident data, or front-end existing host-based applications. Parallel Sysplex data sharing helps ensure that the data and applications are continuously available, and the Parallel Sysplex workload balancing features help ensure that incoming requests get routed to the Web server that is most able to provide the required service levels.

Perhaps your aim is for simplified operations and systems management. This is likely to be of interest to installations that already have a number of systems and are struggling with controlling them in a consistent manner.

For some installations, the aim may be to reduce software costs. Depending on your current configuration, implementing a Parallel Sysplex could have a significant impact on your software licence costs, depending on your product mix and current licence agreements.

The use of data mining applications has become common, but companies sometimes find that queries can run for unacceptably long times. Parallel Sysplex, together with WLM and DB2, can help reduce these run times by an order of magnitude, while still protecting the response and turnaround times of existing production work.

Whatever your reason for considering Parallel Sysplex, there will be some associated cost. It is said that there is no such thing as a "free lunch." To get from where you are, to the point of having a Parallel Sysplex up and running, is going to require an investment in both time and hardware. If you have done sufficient investigation, you should by now have an idea of the savings and additional business value that a Parallel Sysplex brings you. After consulting these three books, you are hopefully also in a position to accurately predict what these costs are.

For further information, *Why Implement an S/390 Operational Single Image (Sysplex)?*, available as the document OPSYSPLX package on MKTTOOLS, provides a discussion of the advantages to be gained from implementing a Parallel Sysplex prior to enabling data sharing mode. There is also a white paper, *Value of Resource Sharing*, available on the Parallel Sysplex home page, and a redbook, *OS/390 Parallel Sysplex Resource Sharing*, SG24-5666, that discuss the use of Parallel Sysplex for resource sharing.

VM/ESA can play an important role in Parallel Sysplex environments. For those installations that have not yet implemented a Parallel Sysplex, VM provides the ability to define and test a Parallel Sysplex configuration without requiring any specialized Parallel Sysplex hardware (CFs and CF links). For those customers that have already implemented Parallel Sysplex, there is still a potential role for VM. VM provides an ideal training environment for Parallel Sysplex, both for operators and system programmers, and allows destructive testing to be carried out without risk to the production environment.

The announced IBM transaction managers and database managers that exploit Parallel Sysplex, and their respective general availability (GA) dates, are shown in Table 5 on page 21, and Table 6 on page 22.

Table 5. IBM CICS and DFSMS Support of Parallel Sysplex				
Product	**Number**	**Support**	**Availability**	**More Information**
CICS TS for OS/390 R3	5655-147	• CICSPlex (R) System Manager enhancements • Resource definition online (RDO) for CICS temporary storage • Ability to externalize CICS ENQ/DEQ • Named Counters and Data Tables in CF • Enhancements to START and DPL requests	1Q 1999	4.2.2, "CICS Transactions in a Parallel Sysplex" on page 169
CICS TS for OS/390 R2	5655-147	• CICS extended system logger exploitation • CICS resources installation across multiple CICS	3Q 1997	4.2.2, "CICS Transactions in a Parallel Sysplex" on page 169
CICS TS for OS/390 R1	5655-147	• Shared CICS temporary storage • CICS exploitation of system logger • CICS/VSAM RLS	4Q 1996	4.2.2, "CICS Transactions in a Parallel Sysplex" on page 169
CICS/ESA V4.1	5655-018	• CICS MRO usage of XCF • CICS WLM Goal mode operation • CICS exploitation of VTAM GR • CICS ARM support	4Q 1994	4.2.2, "CICS Transactions in a Parallel Sysplex" on page 169
DFSMS/MVS V1.5	5695-DF1	• Enhanced sysplex catalog sharing • Multiple DFSMShsm subplexes within a sysplex • DFSMShsm Secondary host promotion. • Access to OAM objects from any system in the sysplex • HFS integrity improvements in a shared environments	2Q 1999	"SMSPLEX" on page 52
DFSMS/MVS V1.4	5695-DF1	• DFSMShsm CDS RLS and larger CDSs	3Q 1997	A.4.1, "Data and Storage Backup and Recovery" on page 141 in Volume 3: Connectivity, SG24-5639
DFSMS/MVS V1.3	5695-DF1	• CICS/VSAM RLS	4Q 1996	4.3.3, "CICS/VSAM Record Level Sharing Considerations" on page 207

Table 6. IBM IMS and DB2 Support of Parallel Sysplex				
Product	**Number**	**Support**	**General Availability**	**More Information**
IMS/ESA V7.1	5655-B01	• Rapid Network Recovery • Online Recovery Service	TBA	4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187, and 4.3.2.1, "IMS DB Data Sharing in a Parallel Sysplex" on page 203
IMS/ESA V6.1	5655-158	• IMS shared message queue • IMS exploitation of VTAM generic resources • Enhanced Parallel Sysplex workload balancing • Enhanced Parallel Sysplex database sharing • Other enhancements	4Q 1997	4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187, and 4.3.2.1, "IMS DB Data Sharing in a Parallel Sysplex" on page 203
IMS/ESA V5.1	5695-176	• IMS/ESA DB n-way data sharing • IMS/ESA exploitation of ARM	2Q 1995	4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187, and 4.3.2.1, "IMS DB Data Sharing in a Parallel Sysplex" on page 203
DB2 UDB for OS/390 V7	5675-DB2	• Restart Light option • DB2PM Support of sysplex	TBA	4.3.1, "DB2 Data Sharing Considerations" on page 194
DB2 UDB for OS/390 V6	5645-DB2	• GBP Duplexing • More Caching options for GBPs • Other enhancements	2Q 1999	4.3.1, "DB2 Data Sharing Considerations" on page 194
DB2 for OS/390 V5.1	5655-DB2	• GBP rebuild support • Other enhancements • GBP Duplexing	2Q 1997 2Q 1999	4.3.1, "DB2 Data Sharing Considerations" on page 194
DB2 V4.1	5695-DB2	• N-way data sharing • DB2 exploitation of ARM	4Q 1995	4.3.1, "DB2 Data Sharing Considerations" on page 194
MQSeries		• Message Queues in CF	TBA	2.18, "MQSeries Shared Queue Structures" on page 156 in Volume 2: Cookbook, SG24-5638

Refer to Chapter 4, "Workloads in Parallel Sysplex" on page 141 for more information on the subsystem exploitation of the Parallel Sysplex.

As can be seen in Table 5 on page 21, CICS was the first IBM transaction subsystem to deliver support for Parallel Sysplex. If you currently use CICS with one or more of IMS/DB, DB2, or VSAM, then you are in a strong position to derive benefit from Parallel Sysplex, possibly with few or no application changes. Each release of CICS since CICS/ESA 4.1 has provided additional support for

Parallel Sysplex, and reduced the number of issues that can constrain your ability to do workload balancing.

Starting with DB2 V5.1, users can avail themselves of DB2's support of VTAM generic resources and database sharing. Also, DB2 and Parallel Sysplex together provide the ability to manage very large databases: this is sometimes called a *Teraplex*. You can find more information about DB2 Parallel Sysplex support in general in *DB2 UDB for OS/390 Version 6 Data Sharing: Planning and Administration*, SC26-9007, and in the redbook *DB2 UDB for OS/390 Version 6 Performance Topics*, SG24-5351.

IMS/ESA V6.1 and later provide support for both VTAM generic resources and shared message queues, as well as support for sharing all of its database types (except MSDBs, which should be migrated to VSO DEDBs). Specific information about this support is available in *IMS/ESA V6 Release Planning Guide*, GC26-8744.

A future release of MQSeries for OS/390, previewed in April 2000, introduces the ability to place the message queues in a CF structure. Queue-sharing also leads to easy system management of MQSeries resources, and you can broadcast MQSeries commands to all queue managers in a queue sharing group.

2.2.1 S/390 Partners in Development Software

Several products from S/390 Partners in Development also support and exploit Parallel Sysplex. The following database managers have announced exploitation of Parallel Sysplex:

2.2.1.1 Adaplex+

In 1996, Software AG's leading DBMS product, Adabas, took advantage of IBM's Parallel Sysplex clustering technology to provide users with 24x7 support. Using the Parallel Sysplex technology, Adabas provides continuous availability of the DBMS by distributing the database load across multiple OS/390 images. Subsequent to its initial release, the Parallel Sysplex technology has made significant technological advancements. To stay abreast of this technology, Software AG will, in two consecutive phases, be introducing a new, improved version of Adaplex as a part of the Adabas V7.1 product.

Foremost with the new Adaplex is its multi-update support—any of the 1 to 32 Adaplex nuclei (that is, instances), running in the 1 to 32 OS/390 images, will have the capability to update the DBMS. This advancement alone greatly reduces networking traffic across the images by ensuring that all DBMS requests have an affinity to the Adaplex nucleus local to that user. Furthermore, with multi-update support, the possibility of an outage will be greatly reduced because:

- Loss of an update nucleus is no longer a single point of failure.
- OS/390 ARM support all but eliminates down-time of a nucleus.
- Sysplex System Managed Rebuild functionality ensures continuous availability of the DBMS during planned outages.
- Adabas's enhanced online Recovery mechanism automates the recovery of a failed nucleus or user in a timely manner.

Each of the multiple instances of the Adabas nuclei will use the Coupling Facility structures to maintain integrity and consistency of the shared data. Adaplex will ensure serialization of the individual Adabas instances across each of the local buffer pools using an optimistic serialization protocol. The actual serialization is achieved via a buffer validity test, a buffer invalidate process, and Coupling Facility conditional write function—a unique feature of the Parallel Sysplex coupling technology. Additionally, the Sysplex XES services will be used for optimistic locking to resolve any block contention during concurrent transactions and DBMS activities.

Through closer integration with the Parallel Sysplex Technology, for example, use of XCF facilities for enhanced communication services, support of Dynamic Transaction Routing, and other strategic enhancements, it is anticipated that there will be further performance improvements.

As for the DBMS itself, the following are some of the features inherent of the DBMS itself:

- Parallel compression/decompression
- Parallel format buffer translation
- Parallel sorting, retrieval, and searching
- Command/Protection Logging
- Universal Encoding Support
- Online Utilities

For more information about Adaplex+, refer to Software AG or see the Web site: `http://www.softwareag.com/adabas/adapalex/adabas%5Fsysplex.htm`

2.2.1.2 Datacom R9

The following text was kindly provided by Computer Associates:

"CA-Datacom Release 9.0 contains significant enhancements focused on meeting the needs for enhanced scalability and system manageability. Hardware and software capabilities are exploited. The key enhancements are:

- **Enhanced Symmetric Multi-Processing**—Exploiting the current mainframe architecture and improving throughput is critical in today's cost-conscious mainframe world. CA-Datacom 9.0 provides significant features to improve throughput with Symmetrical Multi-Processing support. These changes provide the ability to significantly increase the throughput of existing mainframe applications by using multiple processors. SMP is also implemented so that existing applications do not have to be changed. This capability is carefully crafted to support the new generation of CMOS-based processors where multiple smaller less-expensive processors are used to replace more expensive traditional mainframes.
- **Sysplex Exploitation**—It is also critical to exploit the new generation of mainframe hardware and software from IBM, the Parallel Sysplex environment. CA-Datacom 9.0 provides two key features to exploit this new environment:
  - **Application access to databases anywhere within the Sysplex**—Allows applications running on one MVS image within the Sysplex to access a database Multi-User Facility (MUF) running on a different MVS image within the Sysplex. The sysplex Coupling Facility is exploited for the cross-system communication.

- **Dynamic database connection**—Allows the database manager to select whether the application connection should use the standard SVC/Cross Memory processing or the Sysplex communication facilities to connect to the specified Multi-User Facility (MUF). The application does not need to know the location of the MUF within the Sysplex MVS images.

CA-Datacom 9.0 is a major step forward into the new world. Today's high levels of performance are extended through exploitation of the Parallel Sysplex environment and Symmetric Multi-Processing.

In the 10.0 release the data ownership concept will be augmented with the new option called data sharing. With data sharing a database will be shared by all images and owned by none. A combination of shared and unshared databases will be supported. It will still be possible for an application to connect to an unshared database image in a different box.

A data sharing database may be processed by multiple images at the same time. The data is shared among the images. Databases configured for data sharing are not partitioned into disjoint sets, so applications and shared databases can be linked in any configuration. Critical applications will be able to connect to a group of images and transactions from those applications can be processed by any image in the group. This provides load sharing across the images, but it also enhances reliability. If one image of the group fails, then transactions will continue to be processed by the remaining images."

For more information about Datacom, contact Computer Associates or refer to: `http://www.cai.com/solutions/os390/`

2.2.1.3 IDMS R14

Computer Associates has announced that IDMS R14 provides Parallel Sysplex support. In IDMS R14, the multitasking feature was extended to include the database engine, scratch-, security-, and storage managers. This mode of operation is especially beneficial for an update Central Version (CV). Because only one CV at a time can have a database area open for update, multitasking provides improved throughput for update transactions.

The IDMS R14 Dynamic Database Session Routing feature can be used in either environment (one or multiple images), but offers more opportunity for continuous operations and better workload balancing in an environment with multiple images on multiple CPCs. This environment is well suited to starting and stopping "cloned" CVs to meet the current workload for retrieval transactions.

The Shared Cache feature is useful in an environment where there are multiple IDMS CVs accessing the same database files. The CVs may be on the same CPC, or spread across multiple CPCs. Only one CV at a time may have an area open for update. This feature provides the opportunity for the dual benefit of improved throughput from reduced I/O and access to current data by retrieval CVs. This feature should be considered for use with the Dynamic Database Session Routing feature.

The CF is used to manage the workload balancing for dynamically routed retrieval database sessions. Two list structures are allocated in the CF for each DBGroup defined in the Parallel Sysplex. One list structure is used by front-end CVs to post a request soliciting a volunteer back-end CV in the DBGroup. The second list structure is used by the back-end CVs to indicate which CV is the volunteer for the database session.

The following text, describing CA-IDMS Release 15, was extracted from a white paper on the Computer Associates Web site at:
`http://www.cai.com/solutions/os390/roadmaps/idms.pdf`

"CA-IDMS 15.0 is the planned release that will fully exploit the IBM Parallel Sysplex data sharing capabilities by allowing concurrent access to the same database areas for update as well as retrieval by multiple Central Versions. Coupling Facility lock, list, and cache structures are exploited to deliver this functionality. Release 15.0 delivers enhanced data availability as well as improved scalability through this significant feature.

By allowing database areas to be updated by multiple Central Versions, true parallel processing of a workload by multiple CA-IDMS systems can be achieved. As workload increases, additional CA-IDMS systems can be started to handle the load. Increasing the number of systems is transparent to currently executing applications and users.

In Release 15.0, full data sharing will be supported within a Data Sharing Group. A Data Sharing Group is a collection of CA-IDMS systems that can share update access to data. Each CA-IDMS system is referred to as a member of the group. A database area may be concurrently updated by only one Data Sharing Group. Control structures within the Coupling Facility record information about the group and the areas and files to which it has access. Group management is provided through DCMT commands that can operate on all group members. In addition to database areas, other resources such as QUEUEs can also be shared across members of a data sharing group."

For more information about IDMS, contact Computer Associates or refer to:
`http://www.cai.com/solutions/os390/roadmaps/idms.pdf`

2.2.1.4 S/390 Partners in Development Information

If you are using a third-party database manager or transaction manager, check with the vendor for Parallel Sysplex support and SW license information.

You may also check the list of S/390 Partners in Development products that tolerate and exploit Parallel Sysplex under the "Applications by S/390 Technology" heading at the following URL:
`http://www.s390.ibm.com/products/s390da/applications/guide.html`

2.3 Parallel Sysplex High-Level Design

Do not skip this step and dive straight into configuring and sizing the hardware. *There are important considerations here that will affect the configuration.* It will not take long to review this section!

2.3.1 How Many Parallel Sysplexes Do I Need?

The number of Parallel Sysplexes needed in an installation will vary. For most installations however, two Parallel Sysplexes is likely to be the norm. This should cover both the current users' *production* environments and systems programmers' *testing* environments.

For some installations, a need will exist for additional Parallel Sysplexes because of availability, technical, or business reasons. Application development will normally be part of the production sysplex. As in the past, when new releases of software are brought in, these new releases will be put on the development images first.

For companies offering Facilities Management (outsourcing providers) services, for example, there might be specific business reasons why they would not want to include all their workloads in the same Parallel Sysplex.

Some of the specific things to consider when deciding how many Parallel Sysplexes to have are:

- Ease of operation and management

  It is easier to operate and manage a single large Parallel Sysplex than many smaller ones. If improved availability is one of your goals, this should not be overlooked - more outages are caused by human error than by hardware failures.

  Assuming that you have a test sysplex, which we recommend, the test sysplex should be used as learning environment to get familiar with the aspects of managing a sysplex and also as a place to test commands and procedures before they are used on the production system.

- Cost

  In nearly all cases, the cost will increase relative to the number of Parallel Sysplexes you create. Remember that CF partitions cannot be shared between different sysplexes, and that each Parallel Sysplex will need its own dedicated links between OS/390 images and CFs that are not in the same CPC. With G5 and later, it is possible to share an ICF CP between several CF LPs; however, remember that each CF LP must have its own dedicated CPC storage.

- Protecting your production systems

  There is a certain amount of risk associated with putting development systems into the same Parallel Sysplex as your production systems. The amount of this risk will vary from installation to installation. If your developers have a talent for bringing the development system down every other day, then it may be wiser to isolate them from the production systems. On the other hand, if you have a properly configured test Parallel Sysplex, and have thorough test suites, then the risk should be negligible.

  If you have not implemented a thorough test suite, then multi-system outages or problems may be encountered in the production environment rather than on the test systems. Some installations would therefore prefer to thoroughly test new software releases that have sysplex-wide effects (such as JES, XCF and so on), in a less availability-critical environment. In this situation, you may prefer to keep the development environment in a sysplex of its own.

  The system programmer test sysplex should always be kept separate from the production sysplex. The use of VM is one way of providing a self-contained environment where the configuration can be changed with just a few commands, and mistakes cannot impact a system outside the VM environment.

- The scope of work for batch, CICS and IMS transactions, shared data, physical location and shared DASD

  If there are systems that share nothing with other systems, maybe for business or security reasons, it may make more sense to place those systems in separate Parallel Sysplexes.

All of these factors, and more, have to be considered when trying to decide which systems should be in or out of a given Parallel Sysplex.

2.3.1.1 Which Environments Should Be Parallel Sysplexes?

Normally, a logical basis for the sysplex boundary is the JES2 MAS or the JES3 complex. In some instances, the CPCs that share DASD and tape devices become the logical basis for the sysplex. If you have complexes with different naming conventions, different job schedulers, or different security systems, there might be a need to map these into separate Parallel Sysplexes. Always evaluate the test environment and the software test/migration philosophy that is currently used.

2.3.1.2 System Software Level Considerations

To provide the maximum flexibility for introducing new levels of software into the Parallel Sysplex, IBM has committed to providing coexistence support for up to four consecutive releases of OS/390. The following text is extracted from the October 1998 announcement:

N and N+x Support

"Up through the announcement of OS/390 V2R4, IBM allowed three consecutive releases of OS/390 to coexist in a multisystem complex or Parallel Sysplex environment. To better meet customer requirements as they plan and prepare for improved migration compatibility and Year 2000 activity, IBM has further invested in its industry-leading OS/390 Integration Testing and has extended the coexistence policy so that it supports four consecutive releases of OS/390 in a multisystem complex or Parallel Sysplex environment. This extension to the OS/390 coexistence policy includes all OS/390 V1 releases, all currently announced OS/390 V2 releases, and all future releases of OS/390. The period for OS/390 coexistence now extends from 18 months to a maximum of two years based on the current six month release cycle. The coexistence extension from three consecutive releases to four consecutive may be withdrawn at IBM's discretion at a later date. In all cases, the current practice of providing at least 12 months' written notice prior to withdrawal of support will continue."

For a discussion about how frequently you should apply maintenance and how current your software should be for maximum availability, refer to the redbook *Parallel Sysplex - Managing Software for Availability*, SG24-5451, and a white paper entitled *S/390 Maintenance Suggestions to Help Improve Availability in a Parallel Sysplex Environment* available on the Web at:
`http://www.s390.ibm.com/marketing/psos390maint.html`

The Service Update Facility (SUF) is an important strategic tool that you should use to help you obtain service from IBM. SUF provides faster turnaround on PTF requests, and significantly reduces the volume of service shipped to you from IBM. For more information on SUF, as well as a facility to take a "test drive," see the SUF Web site at:
`http://www.S390.ibm.com/suf`

For a description of OS/390 coexistence considerations, refer to the paper *Planning Guide for Multisystem Customers: Coexistence and Planning Considerations Through the Year 2000* at the URL:
`http://www.S390.ibm.com/stories/year2000/coexist.html`

Note that for the subsystems, the coexistence considerations are different.

For DB2, you can have two consecutive releases if you have DB2 V4 (DB2 V4 and V5), and up to three consecutive releases if you have DB2 V7 (DB2 V5, V6, and V7) coexisting in a data sharing group.

Following the announcement of IMS V7, you can have up to three consecutive releases of IMS within a data sharing group at the same time, as long as all the releases are still supported.

The current CICS position is that any supported CICS release may coexist in a CICSplex with any other supported release; however, certain functions may be limited by the release level of the members of the CICSplex.

However, some restrictions exist when running a mixed release data sharing group that may reduce availability or operability. This support is intended for migration purposes only. Plan to maintain this situation for as short a period of time as possible and make sure that a fallback plan is in place before beginning actual implementation. Review *DB2 for OS/390 V5 Release Guide*, SC26-8965, *DB2 UDB for OS/390 V6 Release Guide*, SC26-9013, and *IMS/ESA V6 Release Planning Guide*, GC26-8744, and *IMS/ESA V7 Release Planning Guide*, GC26-9437, for more information.

Implementing this coexistence support may require the installation of “compatibility PTFs” on some or all systems in the sysplex before upgrading any system in the sysplex to the n+1 level.

2.3.1.3 Parallel Sysplex Test Environment

The environment where changes are tested before going into full production must be a Parallel Sysplex environment.

Recommendation for Test Environment in Parallel Sysplex

The current recommendation is that a test Parallel Sysplex should be *separate and distinct* from the production sysplex to ensure that any failures do not propagate to the production environment. This environment should also be as close in configuration (have at least the same *number* of CFs, but not necessarily the same type of CFs) to the production environment as is possible; however, the same capacity is *not* necessarily required.

Remember that the *more* the test environment matches the production workload, including stress testing, the *higher* the probability that potential problems will be discovered before they reach the production environment.

Also remember that there can be only one CF level active in a CPC at a time. So the only way to test a new CF level before moving it to production is to have the test CF in a different CPC to the production CPCs.

The systems programmers have the capability of testing new versions of software in a Parallel Sysplex environment. For more discussion on testing, refer to 4.8, “Test Considerations in Parallel Sysplex” on page 228.

Each separate Parallel Sysplex needs its own CF partitions. CF partitions *cannot* be shared between multiple sysplexes. An in-use CF partition is “owned” by one and only one sysplex, and no other sysplex is allowed to access it while it remains owned. A CF contains architected controls that reflect this sysplex ownership, and has several types of architected objects which relate to the CF as a whole, not to any particular structure within the CF. This plays into the design of the OS/390 CFRM sub-component, which was designed around managing a CF and its resources with a sysplex scope. CFs in a multi-CPC production sysplex are usually provided by either ICFs on CPCs or LPs on standalone CFs.

For reference, here is a complete list of CFs:

- An LP on IBM 9672- or 9021-711- or 511- or 2003- or 3000- or 7060-based CPCs running the Coupling Facility Control Code (CFCC) and using emulated CF links (ICMF).

  These options are, to a large extent, replaced by ICFs on selected 9672s using ICs.

- An LP on IBM 9674-, IBM 9672- or 9021-711-based CPCs running CFCC and using fiber CF links.

  These options are, in some cases, being replaced by ICFs on selected 9672s using a combination of ICs, ICBs and fiber CF links. Remember however, that a CF that is being used in conjunction with data sharing should be failure-independent from all the other CFs and OS/390s in that sysplex.

The Internal Coupling Facility (ICF) permits you to run a CF LP on a CPC without incurring the software costs that would be involved if these additional CPs were available to the operating system. The ICF LPs can operate with ICMF, ICs, ICBs, or fiber CF links (or a combination of ICs, ICBs or fiber CF links).

For a more complete discussion, refer to 1.4.1, “CF Topology Considerations” on page 28 in Volume 2: Cookbook, SG24-5638, the WSC Flash 98029, *Parallel Sysplex Configuration Planning for Availability*, and the CF Configuration Options white paper available on the Web at:
`http://www.s390.ibm.com/marketing/gf225042.pdf`

Figure 2 on page 31 shows a possible configuration that includes two Parallel Sysplexes, one aimed for *production* (non-shaded) and one aimed for *test* (shaded). The production Parallel Sysplex is in this case using standalone CFs.

At the time of writing, it is recommended you use failure-independent CFs in the production Parallel Sysplex if you are doing data sharing. If you are doing resource sharing, then you may opt for one or both of the standalone CFs being replaced by an ICF—either failure-independent or non-failure-independent. In case of a production Parallel Sysplex with two ICFs, then these two ICFs would be placed on separate CPCs.

Figure 3 on page 31 and Figure 4 on page 32 show combined Parallel Sysplex test and production environments using ICFs.

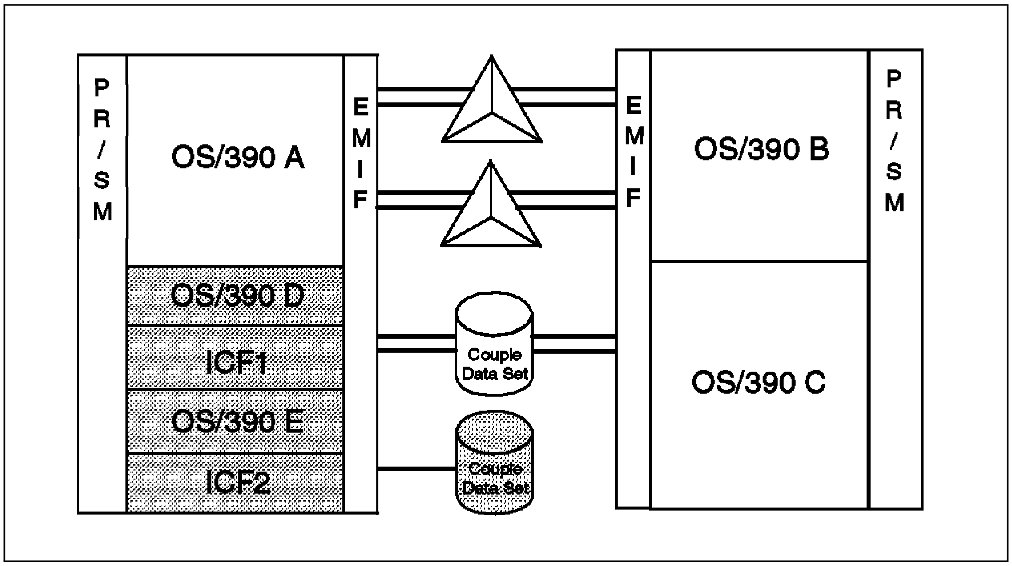


Figure 2. Combined Environments for Data Sharing

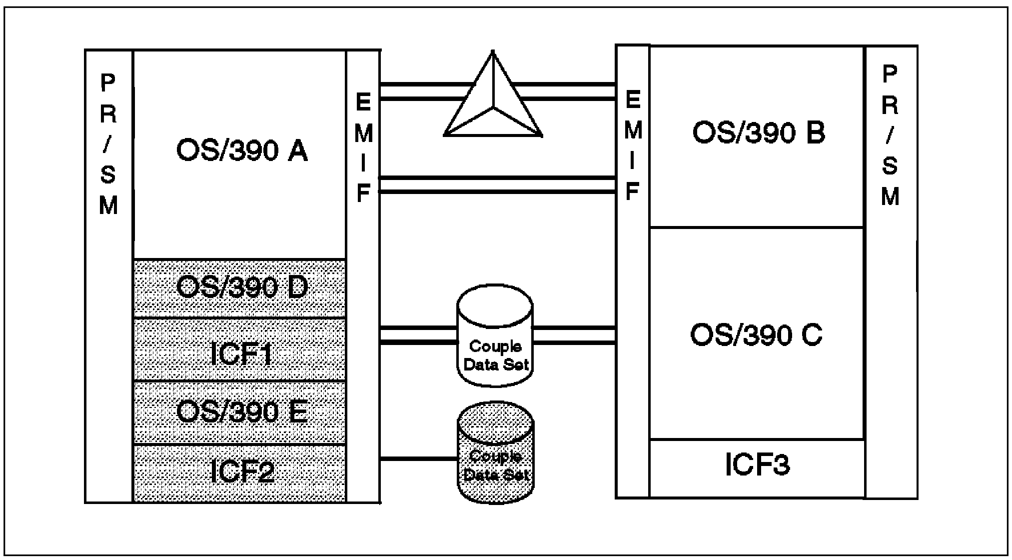


Figure 3. Combined Environments for Data Sharing (GBP Duplexing)

The configurations in Figure 2 and Figure 3 have an advantage over the configuration in Figure 4 on page 32 because it is possible to test new CF Levels in these configurations without having to upgrade the CF Level of either of the production CFs. In Figure 4 on page 32, upgrading the CF Level in use by test CFs ICF1 and ICF2 would also cause the CF Level used by production CF ICF4 to be upgraded at the same time.

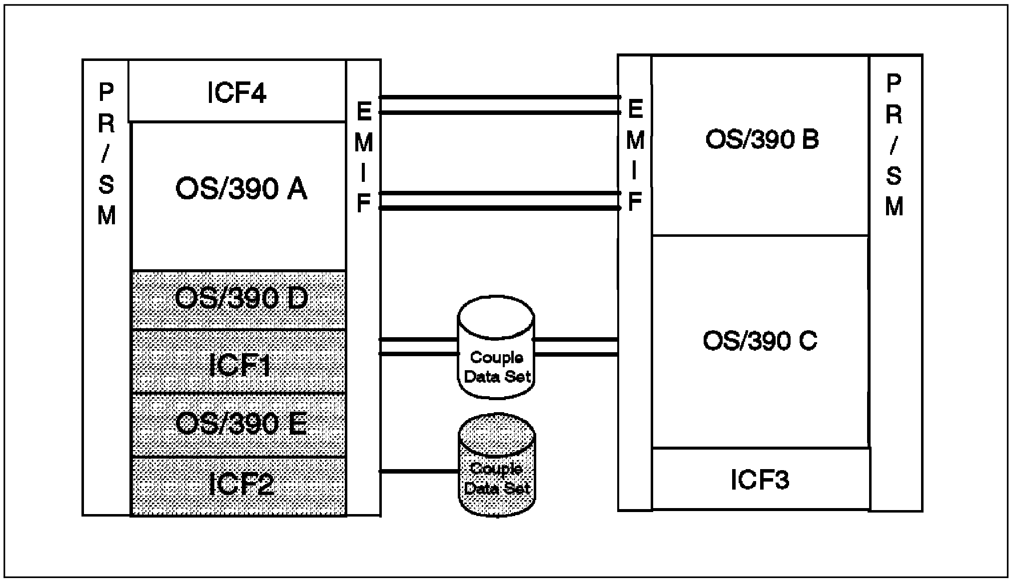


Figure 4. Combined Environments for Resource Sharing

Note that two LPs in the same CPC, but in different sysplexes, require separate physical CF links to attach to their respective CFs. (This is not shown in any of the figures). However, if two LPs in the same sysplex and in the same CPC attach to the same CF, then they can share physical CF links to make that attachment. This is because CF Sender (CFS) links can be shared between operating system LPs using EMIF, but CF Receiver (CFR) links cannot be shared between CF LPs.

2.3.1.4 Pre-Production Environment Cost Options

Sysplex and Parallel Sysplex impose restrictions on test environments as follows:

- A CF LP *cannot* be shared across Parallel Sysplexes. However, two CF LPs in a single CPC can belong to two separate Parallel Sysplexes.
- In general, DASD data sets *should not* be shared between sysplexes. This is discussed further in 2.3.1.5, "DASD Sharing" on page 35.
- As VM does not support external CF links, VM guests cannot currently participate in a multi-CPC Parallel Sysplex.

  A VM guest can, however, participate in a Parallel Sysplex within the one CPC.[1] Enhancements to VM and selected CPCs provide the ability to run the CFCC in a virtual machine (a CFVM), and VM will emulate the CF links between the CFVM and the OS/390 guests. This gives you the ability to run multiple sysplexes. Information about VM's support for Parallel Sysplex, and specifically how VM can assist in Parallel Sysplex testing, is available on the Web at:
  `http://www.vm.ibm.com/os390/pso/os390p1.htm`

  Note: If you decide to run OS/390 Parallel Sysplex Sysplex production in the guest environment on VM/ESA, you must understand that the VM/ESA operating system and the S/390 processor are both single points of failure. If these conditions are acceptable for your usage and environment, this could

---

1 This requires VM/ESA 2.3.0 and a 9672 G3, Multiprise 2000 or later CPC.

be a cost-effective way to take advantage of all the technical advantages available for Parallel Sysplex.

Prior to this enhancement, it was possible to run OS/390 guests in a base sysplex. However, VM does not support the Sysplex Timer, so any VM guests in a sysplex are confined to the single CPC.

It is imperative that you have a test sysplex that is separate and distinct from the production sysplex available for certain types of testing.

There are several alternatives:

1. A CF using ICFs on a CPC

   This alternative does not necessarily require external CF links for the CF function, but it does require storage on the CPC containing the LPs. Examples (shaded) are shown in Figure 2 on page 31, Figure 3 on page 31, and Figure 4 on page 32.

2. CF with external links

   These alternatives require some hardware in the form of external CF links and storage on the CFs, and external CF links on the CPCs housing the OS/390 images. Examples of three test Parallel Sysplexes are shown in Figure 5 on page 34.

3. A test Parallel Sysplex in a Multiprise 3000 (MP3000)

   The Multiprise 3000 processor provides support for a self-contained Parallel Sysplex. Like VM, the MP3000 does not support external CF links; however, it is possible (maybe even desirable) to define a complete test Parallel Sysplex within the one CPC.

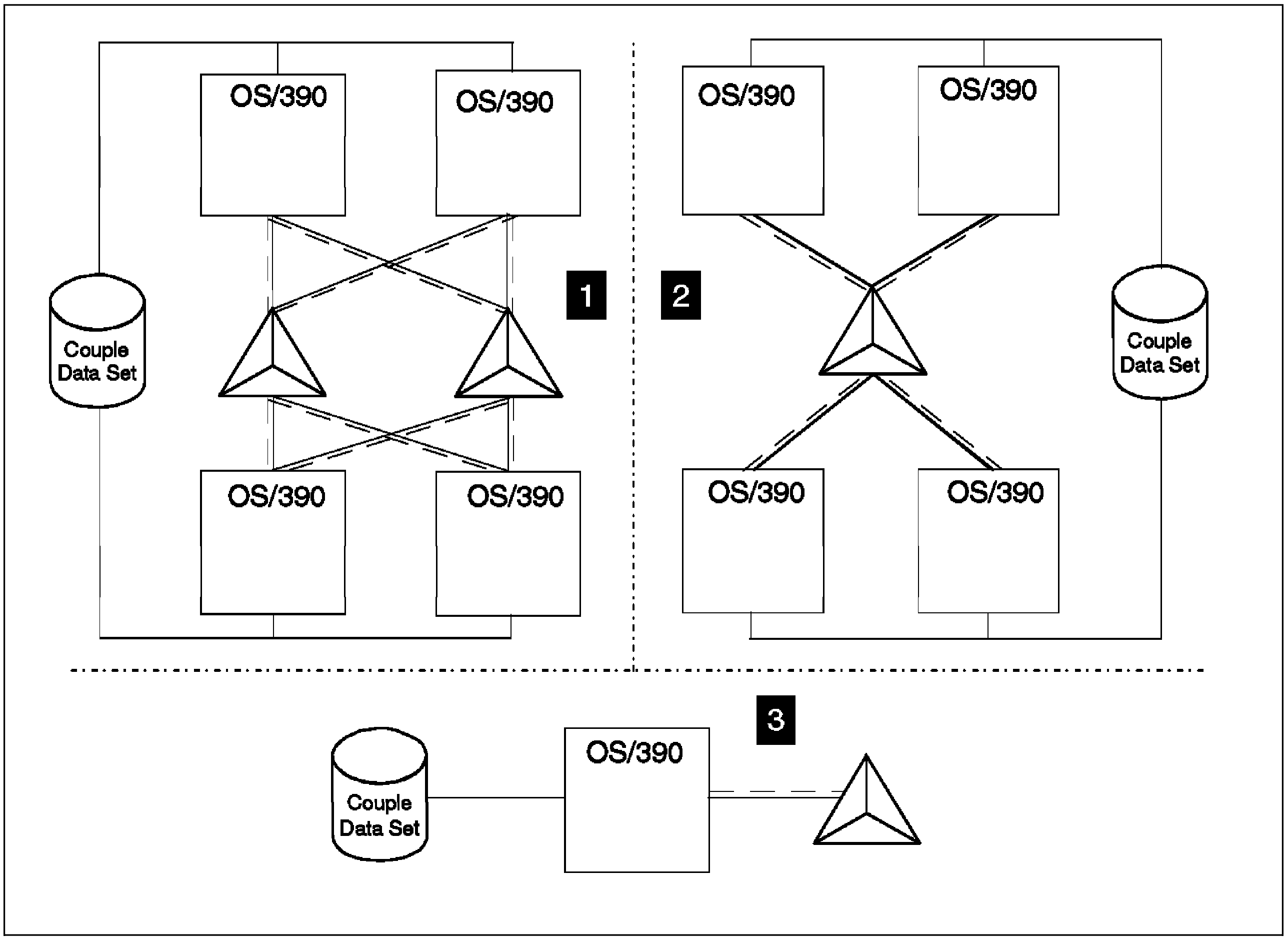


Figure 5. Examples of a Test Parallel Sysplex. Configurations with External CFs

The three Parallel Sysplex configurations using standalone CFs, as shown in Figure 5, can be mapped to corresponding ICFs in a CPC environment. The key point being discussed in the following is not whether you use standalone CFs or ICFs on your CPCs for your test Parallel Sysplex configuration; the aim is rather to discuss what level of testing can be accomplished in each of the three configuration types depicted in Figure 5 (irrespective of whether standalone CFs, ICFs, or a combination thereof is used).

Let us discuss the different Parallel Sysplex configurations and see what level of testing can be accomplished in each:

.1. Two or more OS/390 systems with two or more CFs.

This is the full baseline configuration, which is similar to the production environment. With this configuration there are no issues, since all recovery scenarios and Parallel Sysplex options can be tested.

A similar environment can be established based on using ICFs on one or more of the CPCs.

.2. Two or more OS/390 systems with only one CF.

This allows full testing in normal operation, but does not allow testing of recovery procedures (for example, structure rebuild because of CF or connectivity failures from an image).

A similar environment can be established based on an ICF on one of the CPCs.

3. One OS/390 system and one CF.

This can exercise the CF and CF links, but it does not allow multisystem data sharing and therefore does not test key data sharing functions. It may allow familiarity to be gained with CF operation, but full testing would still be required in another environment. A similar environment can be established based on using an ICF on the CPC.

2.3.1.5 DASD Sharing

In a multisystem environment, access to shared DASD needs serialization to maintain data integrity. Global Resource Serialization (GRS) is the component that manages resource serialization within and across systems. GRS has been available since MVS/SP V1.3.

Management of cross-system resources requires a fast communication method between each GRS subsystem. Prior to the announcement of OS/390 R2, the only way GRS communicated was to transfer a buffer called the *RSA message* around a ring architecture. This ring architecture was actually implemented through CTC or XCF links between each system that wished to share resources.

Since OS/390 R2, GRS has implemented a method to communicate called *GRS star*. GRS star uses a lock structure in the CF to store the status of shared resources. GRS still supports the ring architecture option; however, the GRS star architecture is the recommended option. Table 7 gives the main characteristics of each option.

Table 7. GRS Ring and Star Comparison (IBM Benchmark: SG24-4356)

GRS Characteristic	**Ring Topology**	**Star Topology**
ENQ response Time (ms) (see note)	10+ (increases with the number of systems)	<.04 (stable)
Real storage (frames) (see note)	1000+ (increases with the number of systems)	1500+ (stable)
System ITR (see note)	0.9	1.0
Availability and recovery	Ring reconstruction is complex and may take a long time. It is a function of the number of systems in the ring.	No operator action required if another system fails.

Note: These numbers (rounded) are extracted from *S/390 Parallel Sysplex Performance*, SG24-4356. They apply to a 4-way GRS complex running OS/390 R2. Refer to Chapter 13, “Ring/Star Special Study” in SG24-4356 for a complete description of the test environment.

To summarize:

- ENQ response time is significantly better in a star topology, even for sysplexes with as few as two systems. For batch jobs that process a large number of files, the reduced ENQ and DEQ times can have a significant impact on the elapsed time of the jobs.
- ITR may improve, mainly due to the removal of RSA message-passing overhead.

- Availability and recovery are better in star topology due to the way a lock structure is accessed.

GRS Star or Ring in a Parallel Sysplex?

The improvements brought by the star topology (speed, scalability, and recovery capabilities) make it the recommended choice for all Parallel Sysplexes. However, ring topology implementation may be dictated by specific installation needs (see 2.3.1.6, "DASD Sharing Outside Parallel Sysplex").

In a Parallel Sysplex, the following rules apply:

- GRS must be active in each system of the Parallel Sysplex.
- All systems in a Parallel Sysplex must be part of the same GRS complex.
- Systems in the sysplex, but not belonging to the Parallel Sysplex, must also be in the same GRS complex as those in the Parallel Sysplex. (In this situation, you obviously can't use GRS star mode.)
- Whereas the use of the GRS ring topology allows systems outside the sysplex to be part of the GRS complex, the use of the GRS star topology requires that the GRS complex exactly matches the Parallel Sysplex.
- If systems outside the sysplex are part of the GRS ring complex, they cannot be part of another sysplex.

2.3.1.6 DASD Sharing Outside Parallel Sysplex

If there is a need to share DASD outside the Parallel Sysplex or across Parallel Sysplexes, hardware reserve/release *must* be used to serialize OS/390 access to these devices. Figure 6 on page 37 depicts two Parallel Sysplexes and a standalone OS/390 sharing DASD. DASD in the middle of the picture can be shared by all systems if hardware reserve/release is used to serialize the access.

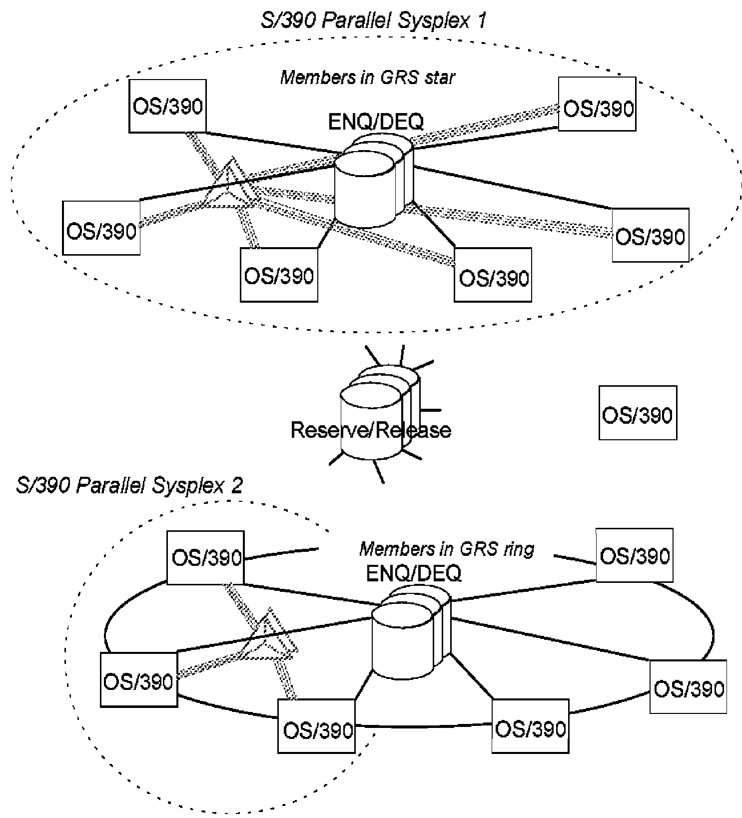


Figure 6. Sharing DASD between Multiple Sysplexes/GRS Complexes

Sharing DASD across sysplexes requires a strict naming convention. It is the responsibility of the function or program that is writing on shared DASD to issue the RESERVE macro.

How to Control Hardware Reserve/Release Requests

ISGGREX0 exit routine (member ISGGREXS of SYS1.SAMPLIB) can be used to ensure that all access to shared DASD outside the GRS complex is serialized with hardware reserve/release. It is recommended to activate this exit if you plan to share data across GRS complexes. It should be noted that this exit is *not* meant as a general purpose way of sharing DASD between sysplexes, and it has a number of restrictions.

Note: Because the multisystem data set serialization component of Multi Image Manager (MIM) from Computer Associates (CA) does not maintain a knowledge of sysplex, nor does it function as a ring, it does not have the same restrictions as GRS. Using this product, you are able to share DASD across multiple sysplexes without any integrity exposure and without having to depend on programs/functions issuing a reserve.

If you wish to use such a product, you should specify GRSRNL=EXCLUDE in your IEASYSxx member. This tells GRS that all ENQ, RESERVE, and DEQ macro requests with a scope of SYSTEMS are treated as though they had been found in the SYSTEMS exclusion RNL. Their scope is changed to SYSTEM and they are processed locally.

If an ENQ is issued with RNL=NO specified, the exclusion is bypassed, and the scope of SYSTEMS is retained. RNL=NO is used for special resources that must be processed by GRS, even if an equivalent product is installed.

If you already have MIM installed, it is possible to run MIM alongside GRS. You may use GRS for all global serialization except for those data sets you want to share between sysplexes. Include these data sets in the GRS exclusion list, and define them in the MIM inclusion list. This will be easier if you have strict naming conventions for these types of data sets.

There are, however, a number of ENQueues that are issued with a scope of SYSTEMS that are meant to provide serialization within a sysplex. Using a product such as MIM can cause *false enqueue contention* when a resource that is being serialized on a system in one sysplex is passed to a system that resides within another sysplex. The following are examples of this type of ENQueue (all of which are issued with RNL=NO specified):

- SYSZRACF (data sharing mode)
- SYSZRAC2
- SYSZLOGR
- SYSZAPPC
- SYSZDAE
- SYSZMCS

For more information on these topics, refer to:

- *OS/390 MVS Setting Up a Sysplex*, GC28-1779
- *OS/390 MVS Planning: Global Resource Serialization*, GC28-1759

Also, note that as of OS/390 R3, IBM provides a tool called the ENQ/DEQ/RESERVE Monitor to help monitor the number of ENQs and who is issuing them. For information on this tool, refer to Appendix A, "Tools and Services Catalog" on page 161 in Volume 2: Cookbook, SG24-5638. This tool is shipped in SYS1.SAMPLIB in OS/390 R3 and follow-on releases and is documented in Chapter 3 of *OS/390 MVS Planning: Global Resource Serialization*, GC28-1759.

IBM has accepted customer requirements to be able to share DASD between different sysplexes, and, at the time of writing, is investigating possible solutions that might be delivered in a future release of OS/390.

2.3.2 System Symmetry When Configuring a Parallel Sysplex

It is recommended that systems in a Parallel Sysplex are configured *symmetrically*. For more information on this, refer to *OS/390 Parallel Sysplex Systems Management*, GC28-1861.

Symmetry, in the context of a Parallel Sysplex discussion, refers to replicating or cloning the hardware and software configurations across the different physical CPCs in the Parallel Sysplex. That is, an application that is going to take advantage of parallel processing might have identical instances running on all images in the Parallel Sysplex. The hardware and software supporting these applications should also be configured identically (or as close to identical as possible) on most or all of the systems in the Parallel Sysplex, to reduce the amount of work required to define and support the environment.

This does not mean that every CPC must have the same amount of storage and the same number or type of CPs; rather, the *connectivity* should be symmetrical (for example, connections to devices, CFs, CTC and so on). A device should also have the same device number on every OS/390 system.

The concept of symmetry allows new systems to be easily introduced, and permits automatic workload distribution in the event of failure or when an individual system is scheduled for maintenance. Symmetry also significantly reduces the amount of work required by the systems programmer in setting up the environment. Systems programmers and operations personnel using the following will find it easier to operate a Parallel Sysplex where the concept of symmetry has been implemented to a large degree, as in the following:

- Consistent device numbering
- A single IODF, containing the definitions for all CPCs and devices in the installation
- Good naming conventions
- System symbols
- Single Point Of Control (SPOC)/Single System Image (SSI) support in OS/390 and subsystems (for example, using the enhanced ROUTE command)

These make planning, systems management, recovery, and many other aspects much simpler. A move toward Parallel Sysplex is a move toward an environment where any workload should be able to run anywhere, for availability and workload balancing. Asymmetry can often complicate planning, implementation, operation, and recovery.

There will, however, be some instances where asymmetric configurations may exist. At times this may even be desirable. For example, if you have a Parallel Sysplex environment that includes an application requiring a specific hardware resource, you may consider having that resource only on one (or a subset) of the CPCs in the Parallel Sysplex. An example is Advanced Function Printing (AFP) printers connecting to certain systems.

2.3.3 What Different 'Plexes Are There?

Within the Parallel Sysplex environment, there are many different 'plexes referred to. In this section, we give recommendations on how each of them relate to one another. Further, there is a brief overview of what constitutes each 'plex.

We would *strongly* recommend that, to the extent possible in your installation, you should try to have all the 'plexes line up with either the basic sysplex or the Parallel Sysplex, as is appropriate. If this is not done, the resulting confusion about which systems are in or out of which 'plex is nearly certain to lead to problems, especially in recovery situations. Having a single logical image both for units of work and for operations and management will lead to simpler, easier-to-understand operations, and also make systems management *far* easier.

The following lists many of the 'plexes commonly referred to today. This list is still growing, so if your favorite is not mentioned here, do not feel offended!

- Sysplex
- JESplex

- Pipeplex
- Batchplex
- GRSplex
- RACFplex
- SMSplex
- HSMplex
- BCSplex
- WLMplex
- OAMplex
- RMFplex
- OPCplex
- CICSplex
- VTAMplex
- Tapeplex
- HMCplex
- CECplex
- GDPSplex

Note: These terms are not always used in the same way in standard IBM documentation.

You may wonder why we have not listed DB2plexes or IMSplexes, or even VSAMRLSplexes or MQSeriesplexes. The reason is that these 'plexes are usually referred to as data sharing groups rather than as 'plexes. Also, they tend to be much more flexible and often contain just a subset of the systems in the Parallel Sysplex, and therefore do not have many of the attributes associated with the 'plexes we describe below.

In the following sections we define what is meant by each of the 'plexes listed above, and show each one's relationship to Parallel Sysplex, and in certain cases to other 'plexes as well.

Sysplex: This describes the set of one or more systems that is given the same XCF sysplex name, and in which the authorized programs in the systems can use XCF services. All systems in a sysplex must be connected to a shared sysplex couple data set. An XCFLOCAL sysplex does not allow a sysplex couple data set. None, one, some, or all of the systems can have CF connectivity. The sysplex is the basis for our relationships in the discussion that follows.

Parallel Sysplex This is the set of systems within a sysplex that all have connectivity to the same CF. The recommended configuration is for *all* systems in the sysplex to be connected to *all* the CFs in that sysplex.

If only a subset of your systems is connected to a given CF, that subset is (throughout these books) referred to as a Parallel Sysplex. In some documents, you may also find this subset with CF connectivity being referred to as a “CF Subplex.”

Using the latter definition of a Parallel Sysplex, you may have more than one Parallel Sysplex within a sysplex if all systems are not connected to each CF. This is *not* a recommended setup.

Note: Not all IBM literature agrees on a unique definition of what constitutes a Parallel Sysplex. For the purpose of this redbook and the discussion that follows, we will use the definition provided on page 14. Other documents mention that a Parallel Sysplex is composed of *two* or more systems, whereas we say that it can consist of just one system and a connected CF. Still other documents mention that a Parallel Sysplex may comprise systems that do *not* have CF connectivity - but this most certainly does not meet our definition.

You will also see the following terms used in IBM literature to describe different Parallel Sysplexes:

Resource Sharing Parallel Sysplex: This describes the non-data sharing implementation. S/390 resource sharing provides the following functionality:

- XCF Signaling - providing multi-system signaling
- GRS Star - multi-system resource serialization
- JES Checkpointing - multi-system checkpointing
- Shared Tape - multi-system tape sharing
- Merged Operations Log - multisystem log
- Merged LOGREC - multisystem log
- RACF - sysplex data sharing
- Shared Catalog - multi-system shared catalogs
- WLM - Multi-system enclave support

Note that there is no standalone CF requirement for a Resource Sharing Parallel Sysplex. A Resource Sharing Parallel Sysplex applies to both multi-CPC and single CPC environments.

Data Sharing Parallel Sysplex: This describes the extension into data sharing, and covers enhancements that are made over time into areas such as high availability and disaster recovery.

A Parallel Sysplex example is shown in Figure 7 on page 42. This example is used as the base for the additional figures in the 'plex discussion.

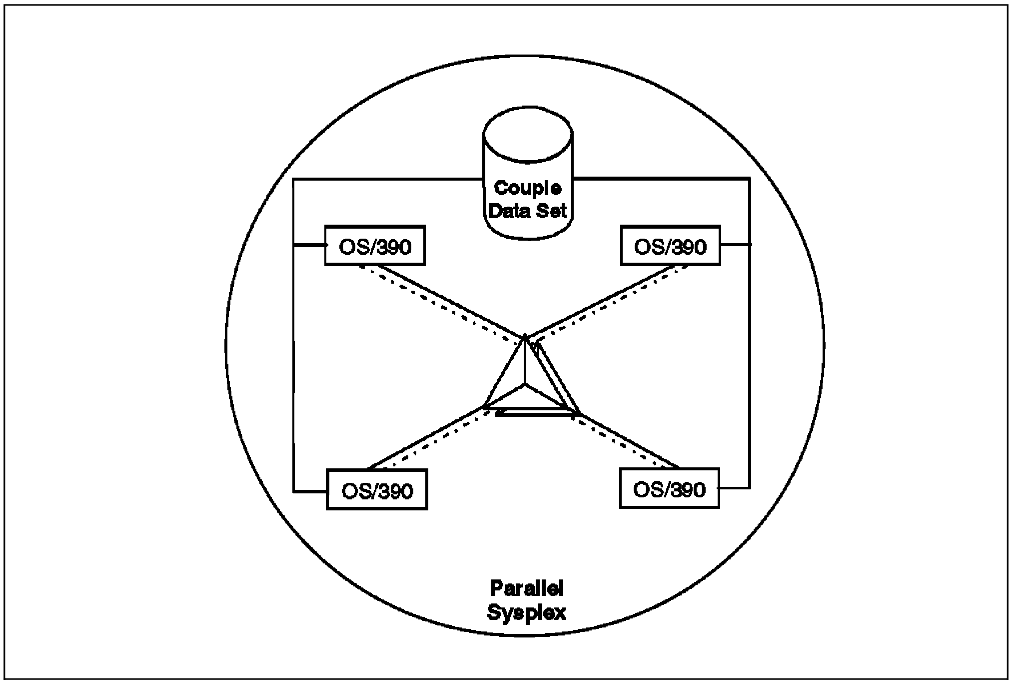


Figure 7. Parallel Sysplex Configuration

Before you read the discussion about how the sysplex, Parallel Sysplex, and all the other 'plexes match, note that it is normally recommended that you *match as many of your 'plexes as possible*. However, you may have valid business or technical reasons for not following the recommendation initially. An example may be that, in a migration phase, you have CPCs that are not yet equipped with CF links. Images on these CPCs may be in the same base sysplex as the systems in the Parallel Sysplex. However, as you move these systems into the Parallel Sysplex, you should plan on lining them up with all the other systems already in the Parallel Sysplex. In the following section, we look at some pros and cons of certain configurations and provide recommendations where possible.

2.3.4 How Do 'Plexes Relate to Each Other?

In order to simplify the operation and management costs in a multisystem environment, it is critical that the relationship between all the different components of that environment are thought out in advance. The redbook *Parallel Sysplex - Managing Software for Availability*, SG24-5451, contains a discussion about the different 'plexes and how they relate to each other, especially from a naming conventions perspective.

In this section, we discuss the capabilities, restrictions, and relationships between the following 'plexes:

- JESplex on page 44.

  For the JESplex, a distinction is made between:

  - A data sharing JESplex, where checkpointing is implemented using the CF.
  - A non-data sharing JESplex.

- Pipeplex on page 46.
- Batchplex on page 46.

- GRSplex on page 48.

  For the GRSplex, a distinction is made between:

  – A GRSplex based on star topology.
  – A GRSplex based on ring topology.

- RACFplex on page 50.

  For the RACFplex, a distinction is made between:

  – A data sharing RACFplex, where the RACF database is buffered in the CF.
  – A non-data sharing RACFplex.

  Further, a distinction is made between whether:

  – Sysplex communication is enabled.
  – Sysplex communication is not enabled.

- SMSplex on page 52.
- HSMplex on page 54.
- BCSplex on page 56.
- OAMplex on page 58.
- WLMplex on page 60.
- RMFplex on page 62.
- OPCplex on page 64.
- CICSplex on page 66.
- VTAMplex on page 68.
- Tapeplex on page 68.
- HMCplex on page 68.
- CECplex on page 68.
- GDPS is discussed further in 3.7.1, “Multi-Site Sysplexes” on page 127.

For a summary table with key information on how 'plexes relate, including recommended configurations, refer to Table 8 on page 70.

JESplex This describes the job entry subsystem configuration sharing the same spool and checkpoint data sets. For JES2 installations, this is the multi-access spool (MAS) environment; for JES3 installations, this is the JES3 global and locals configuration. *It is recommended that there be only one MAS in a sysplex, and that the MAS matches the sysplex boundary.* The same is true for a JES3 global and locals configuration. There are a number of installations that run more than one JES2 MAS in a sysplex, or more than one JES3 complex in a sysplex, for any number of reasons.

Beginning with V5 releases of both JES2 and JES3, the JES MAS or complex must be within a sysplex; they cannot span sysplexes. If you want to run both JES2 and JES3 in the same sysplex, this can be done. Here it is unlikely that either the JES2 or JES3 'plex will match the sysplex. Figure 8 on page 45 shows recommended and valid configurations for the JESplex.

The top configuration in Figure 8 on page 45 shows the recommended configuration where the JESplex matches the Parallel Sysplex. The JESplex should also match the GRSplex, SMSplex, BCSplex, RACFplex and possibly some of the other 'plexes discussed in the following sections.

The center configuration in Figure 8 on page 45 shows that several JESplexes can coexist in the same Parallel Sysplex. Several JES2plexes, JES3plexes, and a combination of these can coexist in a Parallel Sysplex. JES2plexes may even overlap (have systems in common with) JES3plexes. Note that Figure 8 on page 45 is not meant to imply that a JESplex must use the CF. However, there are certain advantages to JES2 using the CF for checkpointing.

The bottom configuration illustrates that systems outside the Parallel Sysplex, but within the same sysplex, can belong to a JESplex. Therefore, if the sysplex is larger than the Parallel Sysplex, a JESplex can also be larger than the Parallel Sysplex (as shown in the figure). This is true for both JES2 and JES3 starting from V5.1 due to the introduction of JESXCF.

If the JES members of a JESplex are V4, then the JESplex can span a sysplex; it can include systems outside a sysplex. JES releases prior to V5, do not require XCF; so a JESplex does not need to be confined in a sysplex. When using CF structures for JES2 checkpointing, the JESplex must not exceed the Parallel Sysplex.

For more information, refer to 4.4.1, “JES2 Considerations in Parallel Sysplex” on page 213, and 4.4.2, “JES3 Considerations in Parallel Sysplex” on page 214.

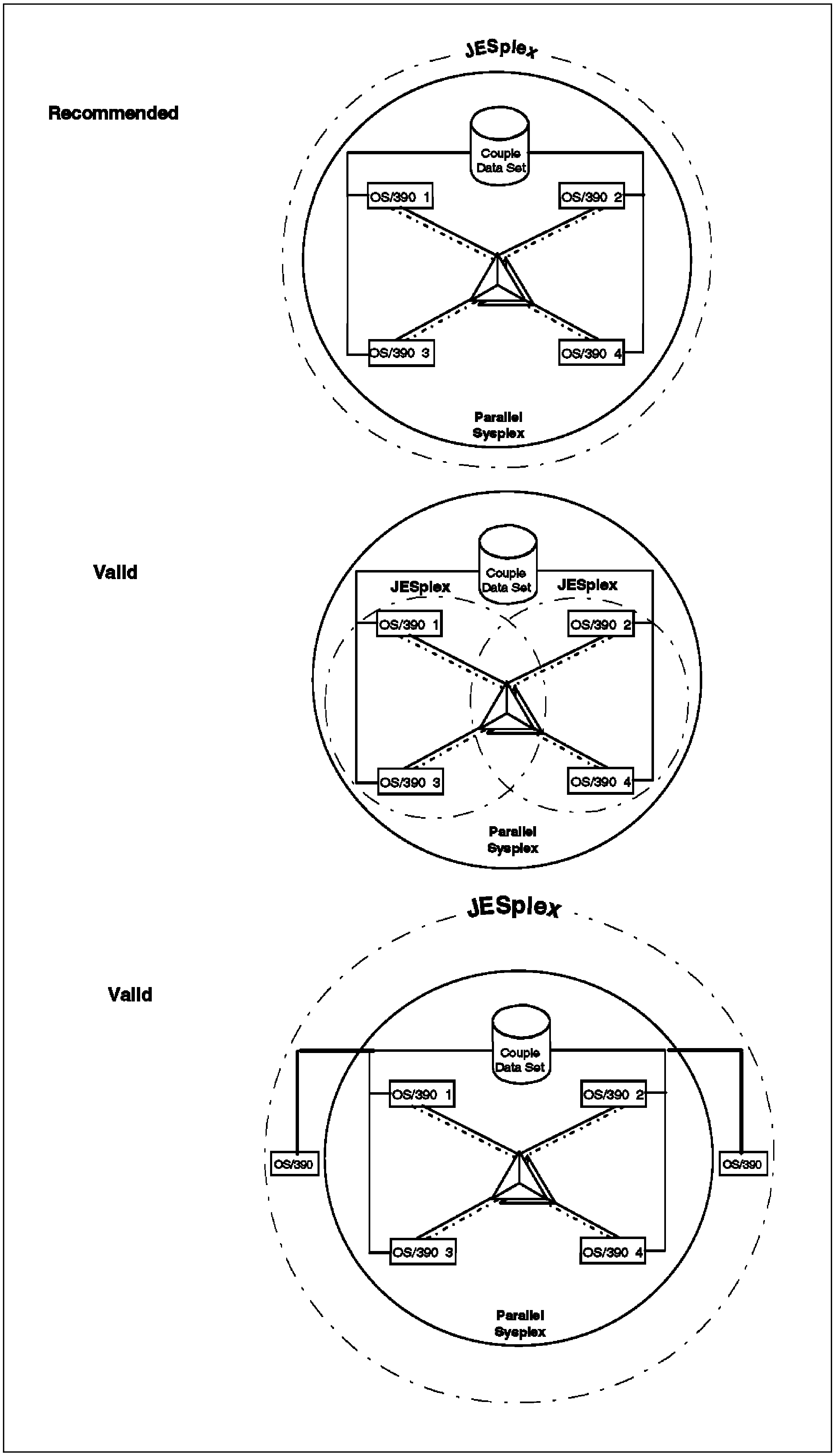


Figure 8. JESplex Configurations

Pipeplex This describes those systems that enable BatchPipes piping of data between jobs or steps running on different systems.

Since Pipeplex uses a structure in the CF, the scope of the Pipeplex can be smaller or equal to the Parallel Sysplex, but not larger. Figure 9 on page 47 shows recommended, valid, and not valid configurations for Pipeplex. Note that you can run multiple Pipeplexes within a Parallel Sysplex. However, every subsystem in a given Pipeplex must have the same subsystem name. Therefore it follows that if you have more than one BatchPipes subsystem on one OS/390 image, then each of those subsystems must be in a different Pipeplex. Each Pipeplex has its own CF structure. An OS/390 system can be in more than one Pipeplex.

You may have a single system Pipeplex. In this configuration, data is piped through memory rather than through a CF structure.

Batchplex SmartBatch requires definition of a Batchplex to enable Batch Accelerator scheduling of job steps across systems. Batch Accelerator is a component of SmartBatch.

Since the Batchplex uses XCF for communication, the scope of a Batchplex is less or equal to the sysplex and can thus be larger than the Parallel Sysplex (*this is not true if you also have a Pipeplex*).

Note that, although most installations will use only one, more than one Batchplex can be defined in the same sysplex. This can be useful if you want to target only two images in your test environment.

Notice that a Pipeplex and Batchplex must contain the same number of OS/390 images.

For more information, refer to 4.4.5, "IBM BatchPipes for OS/390" on page 215.

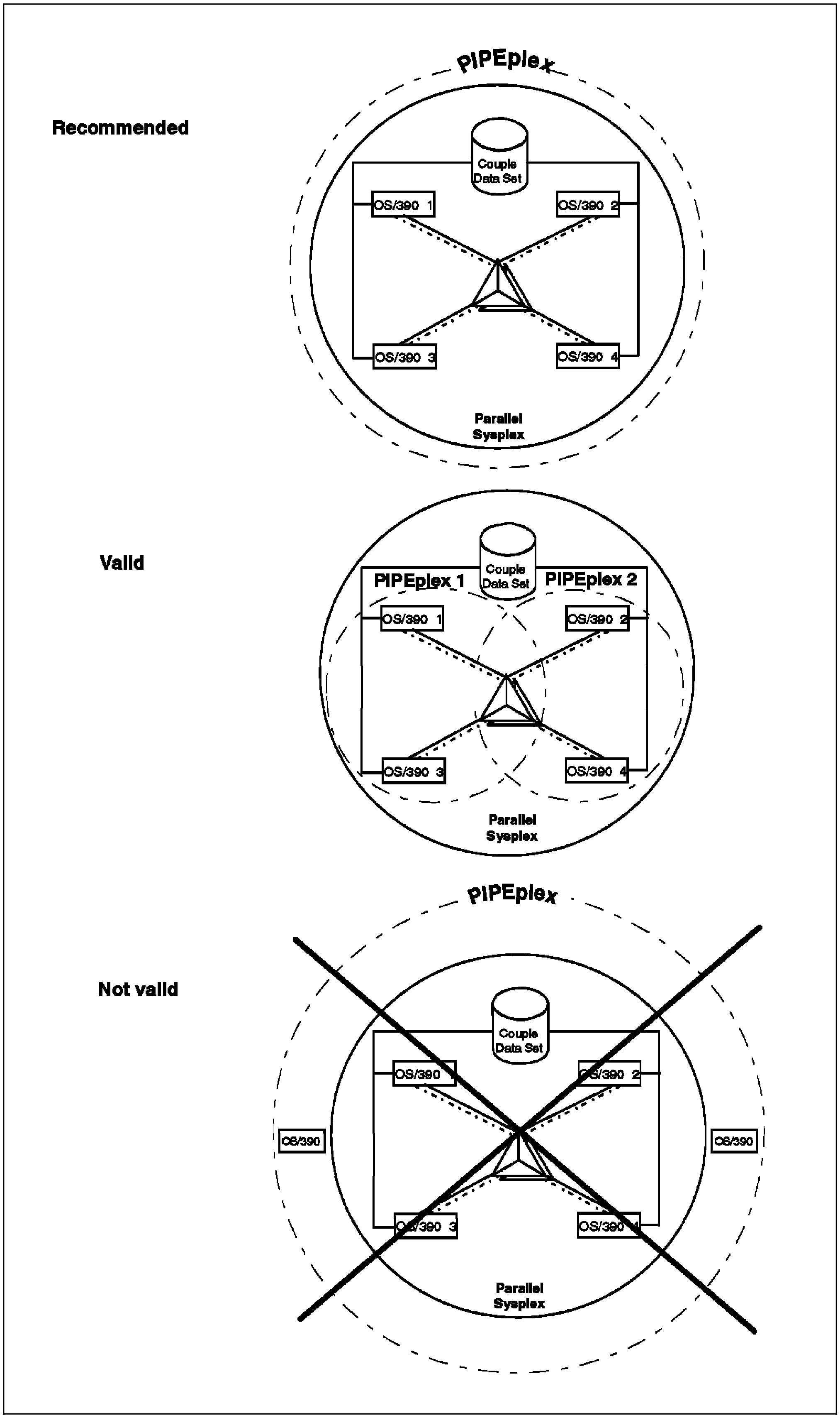


Figure 9. Pipeplex Configurations

GRSplex GRSplex describes one or more OS/390 systems that use global resource serialization to serialize access to shared resources (such as data sets on shared DASD volumes). A GRSplex describes all systems that are part of either a GRS ring or GRS star configuration.

In a GRS ring configuration, systems outside a sysplex may share DASD with those systems in the sysplex. It is possible for a GRSplex based on ring topology to be larger than the sysplex, but additional systems cannot belong to another sysplex. This is also called a *mixed GRS complex.* In a mixed GRS complex, systems within the sysplex would automatically use XCF to communicate (regardless of what is specified in the GRSCNF member), whereas GRS would need its own dedicated CTCs to communicate with any systems outside the sysplex.

With the GRS star topology, however, the GRSplex *must* match the Parallel Sysplex. Also, every system in a Parallel Sysplex must be a member of the same GRSplex and the Parallel Sysplex must match the sysplex. In other words, in a GRS star topology, the GRSplex, the Parallel Sysplex and the sysplex must match completely.

No matter whether GRSplexes are based on star or ring topology, device serialization *between* multiple Parallel Sysplexes must be achieved by the reserve/release mechanism.

Figure 10 on page 49 shows recommended, valid, and not valid configurations for the GRSplex.

The top configuration in Figure 10 on page 49 shows the recommended configuration where the GRSplex matches the Parallel Sysplex.

The center configuration in Figure 10 on page 49 shows that the GRSplex (based on GRS ring topology) may include systems outside the sysplex. If your GRSplex is larger than the sysplex, there are some operational aspects to be considered:

- Lack of automatic GRS ring rebuild in a recovery situation.
- No automatic adjustment of the residency time value (RESMIL).
- No dynamic changes to the Resource Name Lists (RNLs).

Note that the figure is not meant to imply that a GRSplex must use the CF. However, a GRSplex based on the star topology must use the CF and, therefore, cannot include systems outside the Parallel Sysplex.

The bottom configuration illustrates that several GRSplexes cannot coexist in the same Parallel Sysplex.

For more information related to the GRSplex configuration in the Parallel Sysplex environment, refer to *OS/390 MVS Planning: Global Resource Serialization*, GC28-1759. Also refer to 2.3.1.5, "DASD Sharing" on page 35 for a discussion on MIM and GRS considerations.

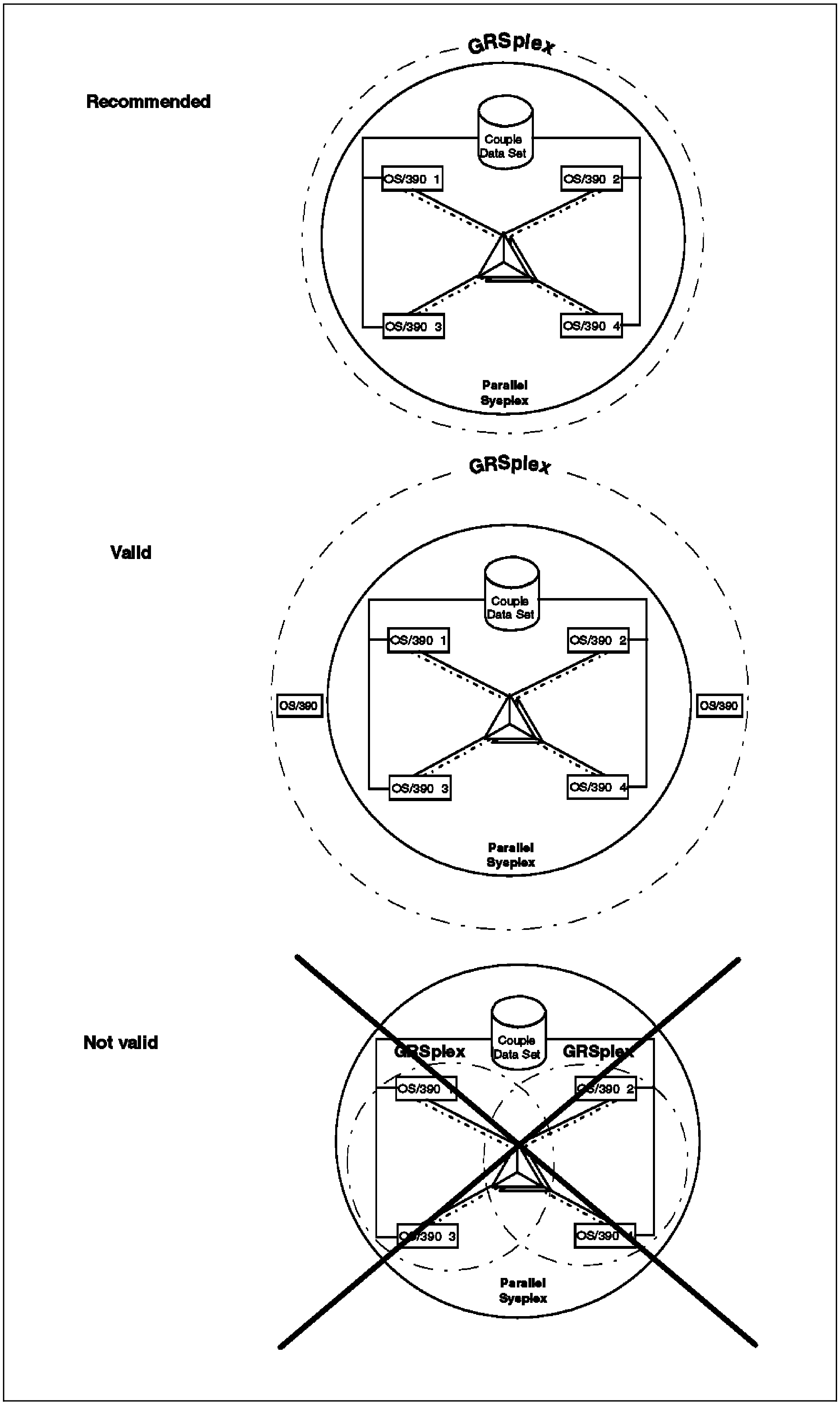


Figure 10. GRSplex Configurations

RACFplex This describes the systems that share the same RACF database. If the systems sharing the RACF database are all members of the same sysplex, RACF can be enabled for RACF sysplex communication. This implies RACF will use GLOBAL ENQ instead of reserve/release for serialization. In addition, commands are propagated to all systems, and updates made on one system are available to all.

If the systems sharing the RACF database are all members of the same Parallel Sysplex, RACF can also be enabled for sysplex data sharing. This implies that RACF will exploit the CF. There can be only one RACF data sharing group within a Parallel Sysplex. The installation cannot alter the RACF XCF group name (IRRXCF00), so it is not possible to have more than one RACF sysplex data sharing group in the sysplex. *It is recommended that you make the RACFplex match the Parallel Sysplex and enable it for sysplex data sharing.*

When the RACFplex matches the Parallel Sysplex, RACF can be configured with all systems sharing the same RACF database and one of the following:

- All systems not enabled for RACF sysplex communication (in this case, you should not do RACF sysplex data sharing).
- All systems enabled for RACF sysplex communication.
- All systems enabled for RACF sysplex communication and RACF sysplex data sharing.

After RACF sysplex communication has been enabled on all systems, RACF sysplex data sharing mode can be enabled or disabled. When sysplex communication is enabled, the second and subsequent system IPLed into the sysplex enters the RACF mode that currently exists, regardless of the data sharing bit setting in the Data Set Name Table (DSNT). After IPL, the RVARY command can be used to switch from sysplex data sharing to non-data sharing mode and vice versa.

You can share the RACF database with systems outside the Parallel Sysplex boundary. However, in that case there are operational aspects to be considered:

- RACF sysplex data sharing cannot be enabled.
- Reserve/release is used for serialization.
- If the system is outside the sysplex, then it should not be part of another sysplex.
- Commands will not be propagated to systems outside the sysplex boundary.

You may have more than one RACFplex in the Parallel Sysplex, but only one of the RACFplexes can be enabled for RACF sysplex data sharing. The other RACFplexes will not be able to use the CF for buffering the RACF database.

If you want to exploit RACF sysplex data sharing, the participating systems all have to be within the Parallel Sysplex. Figure 11 on page 51 shows recommended, valid, and not valid RACF configurations. For more information about RACFplexes, refer to *OS/390 MVS Security Server (RACF) System Programmer's Guide*, SC28-1913, and the redbook *OS/390 Parallel Sysplex Resource Sharing*, SG24-5666. Also refer to A.2, "OS/390 Security Server (RACF) Sysplex Exploitation" on page 126 in Volume 3: Connectivity, SG24-5639.

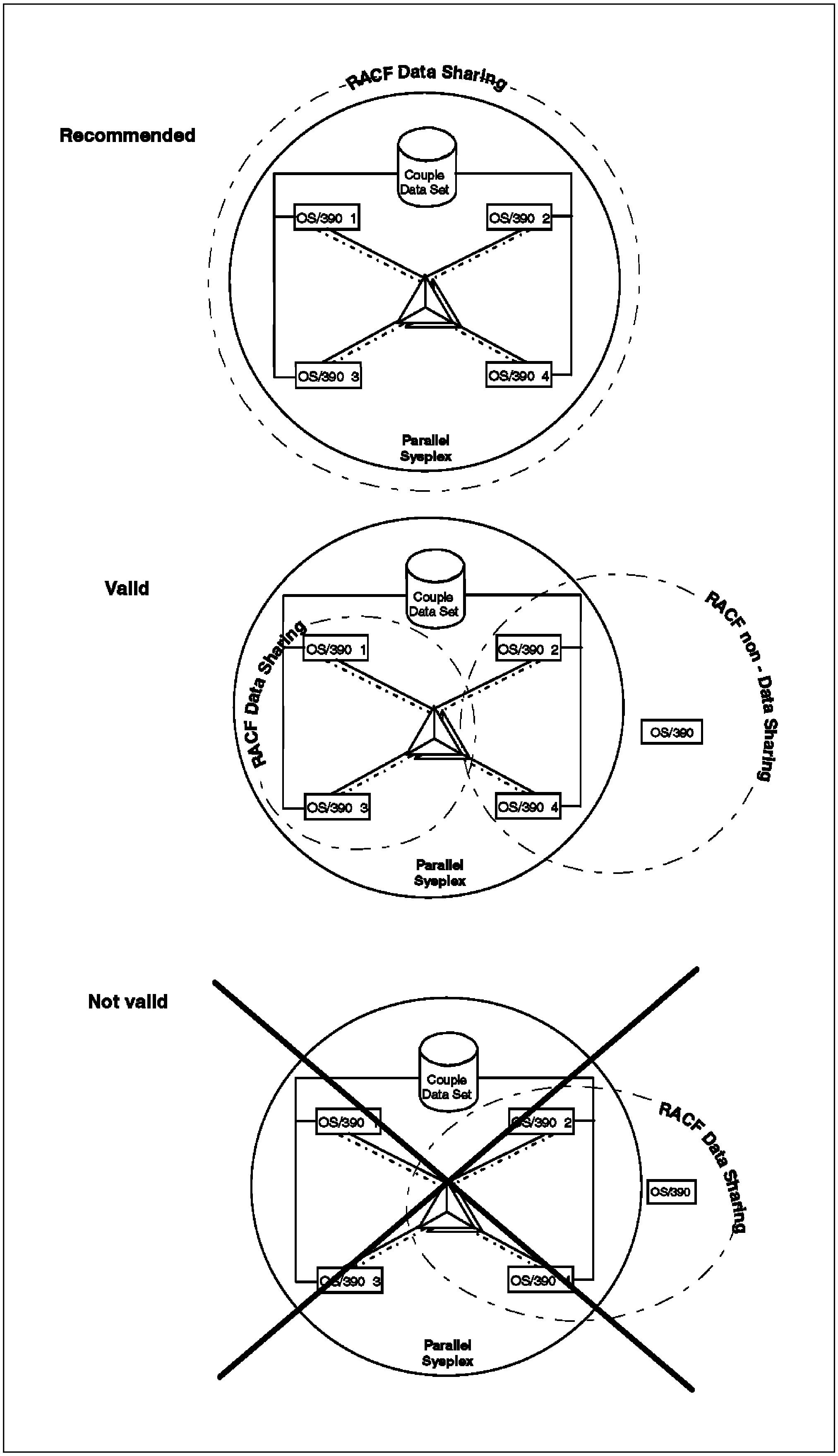


Figure 11. RACFplex Configurations

SMSplex This describes those systems that use the same SMS base configuration stored in the active control data set (ACDS), and the same communications data set (COMMDS). While it is possible for the SMSplex to include systems outside the sysplex, *it is recommended that, if possible, the SMS complex, or SMSplex, should match the sysplex*.

Figure 12 on page 53 shows recommended and valid configurations for the SMSplex. The top configuration shows the recommended configuration where the SMSplex matches the Parallel Sysplex. The SMSplex should also match the JESplex, GRSplex, RACFplex, and possibly some of the other 'plexes discussed.

The center configuration in Figure 12 on page 53 shows that the SMSplex may include systems outside the Parallel Sysplex. If your SMSplex is larger than the Parallel Sysplex, then it is likely that your SMSplex matches the sysplex and the GRSplex. The SMSplex may span several sysplexes and Parallel Sysplexes, in which case it is not possible to match the GRSplex. In this configuration, care should be taken with respect to the performance of the catalogs, ACDS and COMMDS data sets. For a related discussion, refer to 2.3.1.5, "DASD Sharing" on page 35.

The bottom configuration illustrates that the SMSplexes may be smaller than the Parallel Sysplex. Several SMSplexes may coexist in the Parallel Sysplex. For more information about SMS in a Parallel Sysplex environment, refer to *DFSMS/MVS V1R5 Planning for Installation*, SC26-4919, and to A.4.2.2, "SMSplex" on page 143 in Volume 3: Connectivity, SG24-5639.

The Enhanced Catalog Sharing (ECS) protocol, introduced with DFSMS/MVS V1.5, uses a CF cache structure to hold change information for shared catalogs. This eliminates catalog-related I/O to the VVDS, resulting in better performance for both user and master catalog requests. All systems sharing a catalog that is being used in ECS mode *must* have connectivity to the same CF, and *must* be in the same GRSplex. So, if you are using ECS, the middle configuration in Figure 12 on page 53 would not be valid.

SMS uses XCF communication for a number of components, including:

- OAM. This is described in the OAMplex section on page 58.
- DFSMShsm. This is described in the HSMplex section on page 58.
- VSAM/RLS. If you are using the VSAM Record Level Sharing support for the DFSMShsm CDS in DFSMS V1.4, then all participating systems must have connectivity to the CF and therefore be in the same Parallel Sysplex.
- PDSE Sharing. If you are sharing PDSEs across systems, then XCF services are used to communicate status information between systems.

  There are APARs available for DFSMS 1.4 and 1.5 (and supporting products) that remove the restriction that HFS and PDSE files must reside on SMS-managed volumes. More information about this support is available in WSC Flash 10007, *HFS and PDSE Support on non-Managed SMS Volumes PTFs Available*.

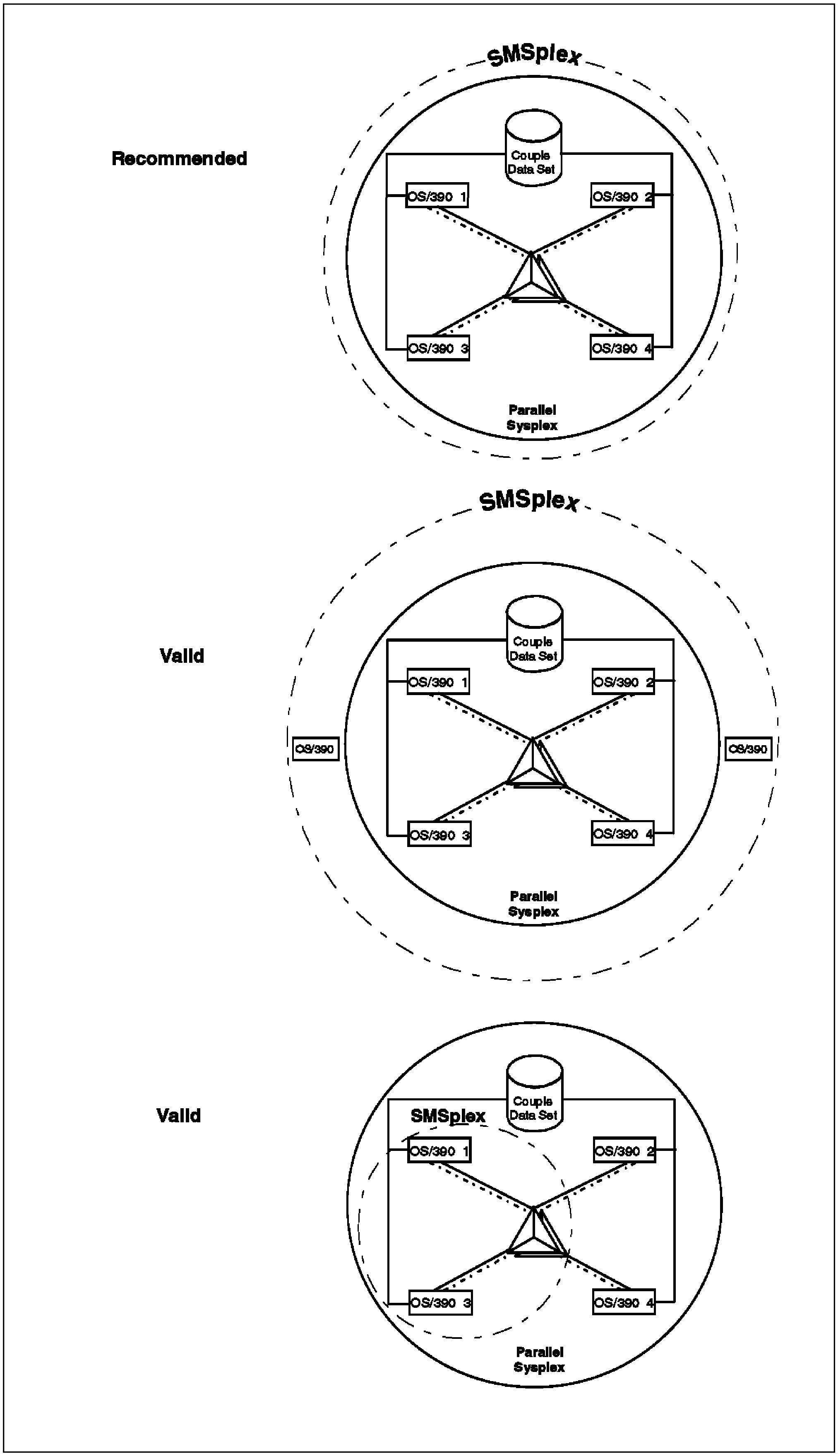


Figure 12. SMSplex Configurations

HSMplex Closely related to the SMSplex is the HSMplex, which is the set of DFSMShsm systems that share the same HSM journal and control data sets.

DFSMS 1.5 introduced three new Parallel Sysplex-related facilities:

- Secondary Host Promotion
- “Single GRSplex”
- Large CDS support (this was retrofit to DFSMS 1.4 via APAR OW33226)

All of these impact the systems that constitute the HSMplex.

Secondary Host Promotion provides the ability for a HSM to take over unique responsibilities from another HSM Primary or Secondary Space Management (SSM) host should that host, or the system it is running on, fail. This ability is controlled by the HSM SETSYS PRIMARYHOST and SETSYS SSM commands, and uses XCF services to communicate the status of the Primary and SSM hosts to the other HSMs in the HSMplex. In order for a host to be eligible to take over, it must be in the same HSMplex as the failed host, as defined by the new PLEXNAME parameter. There is one DFSMShsm XCF group per HSMplex. The XCF group name is the HSMplex name. There is one XCF group member for each DFSMShsm host in the HSMplex.

The next new function is known as “Single GRSplex” support. Prior to DFSMS 1.5, when HSM executed a given function, it would issue an ENQ against a fixed name—this is to protect the integrity of the control data sets. The problem with this mechanism is that if you have two independent HSMplexes within the same GRSplex, you would get false contention if the two HSMplexes tried to execute the same function (but against a different set of DASD) at the same time.

“Single GRSplex” gives you the option of changing HSM so that it issues a unique ENQ, utilizing the data set name of the associated HSM CDS. This effectively creates an HSMplex, but the scope of this HSMplex is based on the data set names of the CDSs used by the member HSMs. To enable this new support, you must specify a new parameter, RNAMEDSN=Y, in the HSM startup JCL.

The third new function (also available on DFSMS 1.4) is DFSMShsm support for the use of RLS for its own Control Data Sets. The MCDS, BCDS and OCDS can be optionally be defined as EA VSAM data sets, which greatly increases the capacity beyond 16 GB for the MCDS and BCDS, and beyond 4 GB for the OCDS. All the HSM systems accessing a given set of CDSs must have access to the same CF. The HSMplex must be the same as, or smaller than, the RLSplex.

One additional thing to consider in relation to the scope of a HSMplex is if you use the MVS Logger. When you are using Logger, all systems that are connected to a LOGR structure get notified when any of the logstreams within that structure exceed the user-specified threshold. Any of those systems are eligible to offload the logstream data to an offload data set on DASD. Therefore, all the systems that are connected to a given LOGR structure should be in the same HSMplex. Figure 13 on page 55 shows recommended and valid configurations for the HSMplex.

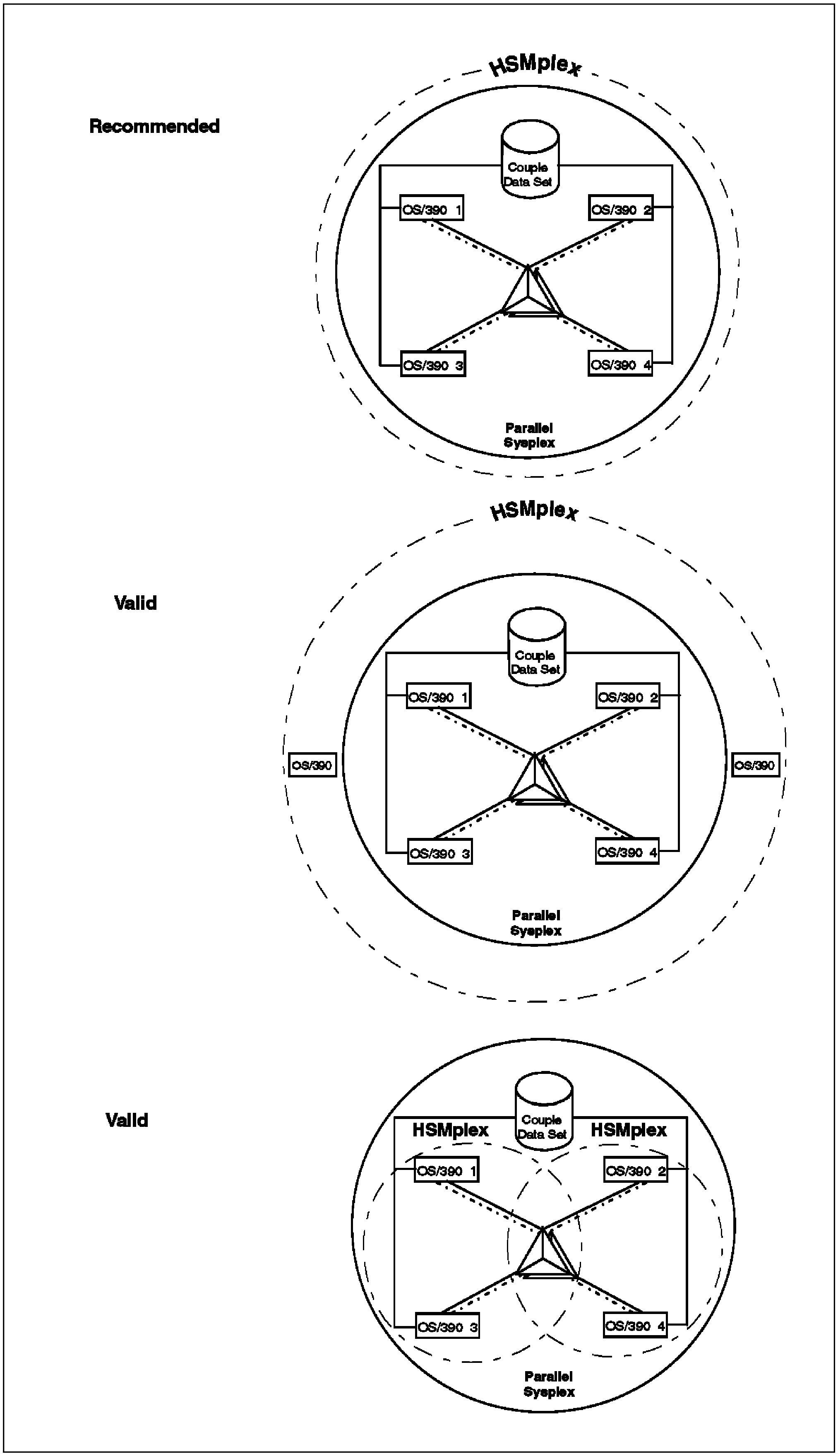


Figure 13. HSMplex Configurations

BCSplex The BCSplex is the set of systems sharing a set of catalogs in Enhanced Catalog Sharing (ECS) mode. Once a catalog is open in ECS mode, it cannot be concurrently opened by a system that does not have ECS support, or that does not have access to the ECS structure in the CF. A new ENQ is used to indicate the fact that a catalog is open in ECS mode, so all the systems sharing a catalog that *could potentially* be opened in ECS mode *must* be in the same GRSplex. Failure to adhere to this restriction will almost definitely result in a damaged catalog.

ECS mode is supported on DFSMS 1.5 and higher systems. Toleration service is provided for lower level systems so that they will check for the ENQ, and understand its meaning.

Figure 14 on page 57 shows both valid and invalid configurations for the BCSplex. The top configuration is valid because all the systems in the BCSplex are within the Parallel Sysplex, and thus have access to the ECS structure in the CF. The middle configuration is also valid. As long as *all* the systems accessing any catalog in the sysplex (that is eligible to be opened in ECS mode) have access to the same ECS CF structure, that is a valid configuration. As systems 1 and 3 both have access to the same CF, that is a valid configuration. The bottom configuration is not valid because the systems outside the Parallel Sysplex cannot access the ECS structure in the CF and therefore cannot open a catalog that is open in ECS mode by a system within the Parallel Sysplex.

For more information, refer to the redbook *Enhanced Catalog Sharing and Management*, SG24-5594.

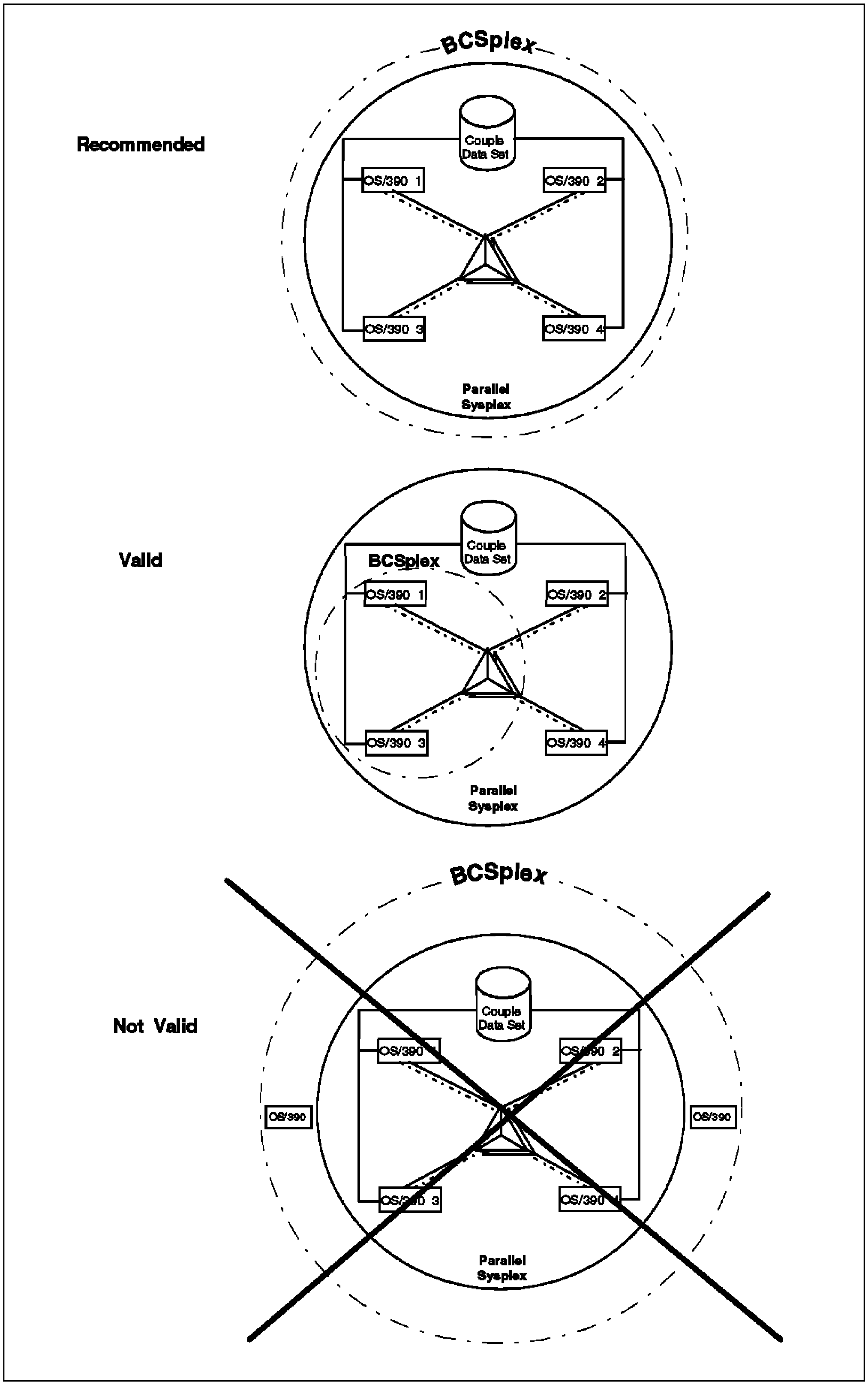


Figure 14. BCSplex Configurations

OAMplex DFSMS 1.5 introduced the ability for multiple OAM instances, on multiple OS/390 images, to have access to shared OAM objects. The OAM instances must all be in XCF mode, in the same XCF group, and the associated DB2 subsystems must all be in the same data sharing group. An OAMplex, therefore, is the set of OAM instances that have access to the shared OAM objects.

In an OAMplex, each OAM instance can directly access OAM objects resident on DASD, or on tape if the tape is not currently in use by another OAM instance. If the object is on a 3995, or on a tape that is currently being used by another OAM instance, the requesting OAM instance will route the request to the OAM instance that owns the 3995 or is currently using the required tape volume. These requests are routed using XCF services.

If the system that owns the 3995 fails, the 3995 must be manually brought online to another system. As soon as the 3995 comes online to the other system, that OAM will take ownership of it and use XCF to notify the other OAM instances in the OAMplex that requests for objects in that 3995 should now be routed to that OAM instance.

Figure 15 on page 59 shows both valid and invalid configurations for the OAMplex. The top two configurations are valid because all the systems in the OAMplex are within the Parallel Sysplex, and thus have access to the DB2 structures in the CF. The bottom configuration is not valid because the systems outside the Parallel Sysplex cannot be in the DB2 data sharing group, and therefore cannot be in the OAMplex. Even if they are in the same sysplex (and thus can communicate using XCF) they still cannot be in the OAMplex—they must also be in the same DB2 data sharing group as all the other systems in the OAMplex.

For more information, refer to *DFSMS/MVS V1R5 Planning for Installation*, SC26-4919, and *Object Access Method Planning, Installation, and Storage Administration Guide for Object Support*, SC26-4918.

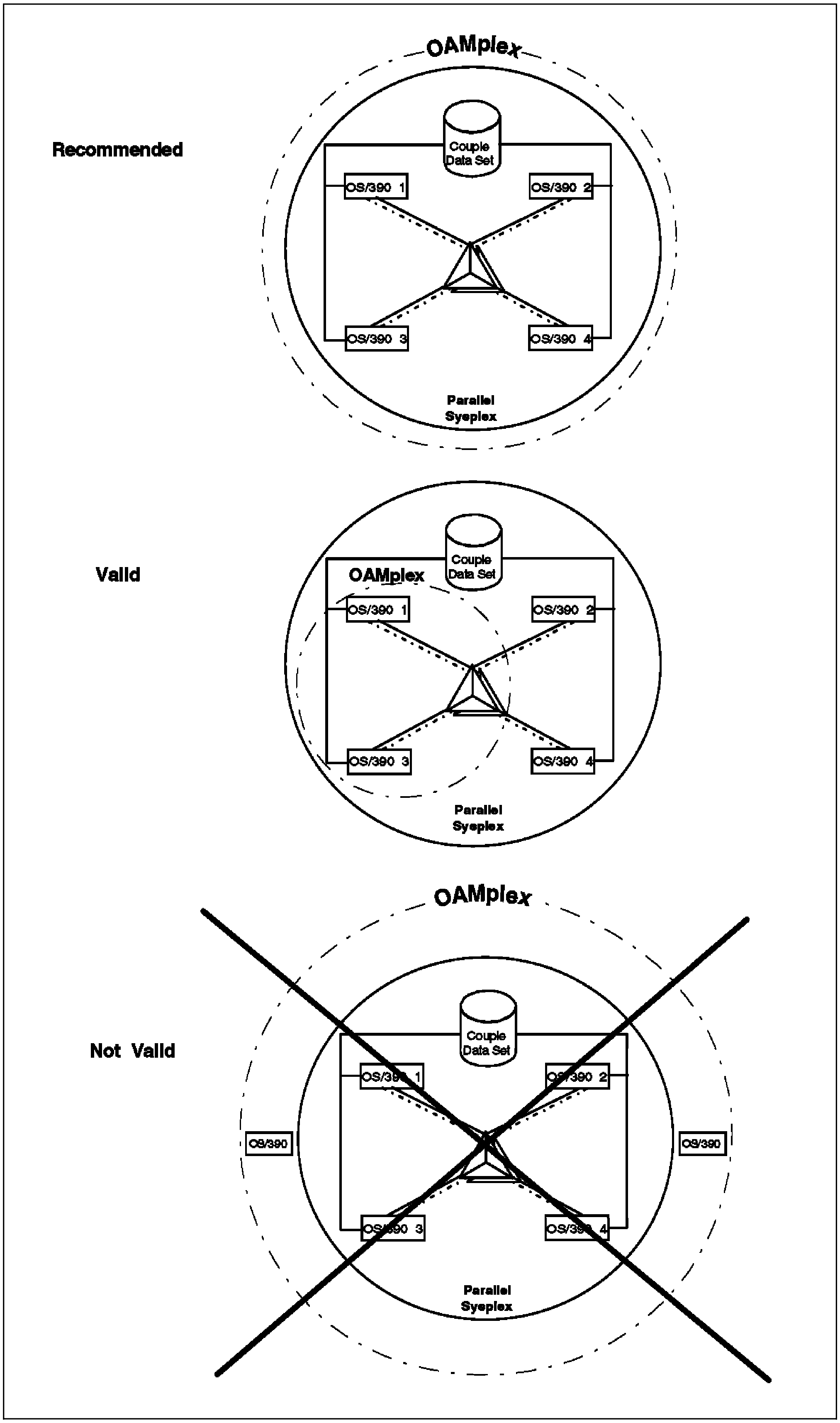


Figure 15. OAMplex Configurations

WLMplex This describes those systems within the sysplex that share a WLM control data set. The systems may or may not be in goal mode, or some may be in goal mode and some in compatibility mode. Even systems that are not in goal mode use the WLM service definition for classification of enclaves, for reporting of CICS and IMS transactions prior to migrating to goal mode, and for scheduling environments. *If all systems in a sysplex operate in goal mode, then the WLMplex matches the sysplex*. All systems in the same WLMplex must be in the same sysplex.

The WLM couple data sets (primary and alternate) contain information about the sysplex they belong to (sysplex name). The WLMplex can, however, be larger than a Parallel Sysplex.

Figure 16 on page 61 shows both recommended and valid configurations for the WLMplex. The top configuration in Figure 16 on page 61 shows the recommended configuration where the WLMplex matches the Parallel Sysplex.

The center configuration in Figure 16 on page 61 shows that some systems can operate in WLM goal mode, whereas others operate in compatibility mode. Only one WLM Policy can be active at a time within a given sysplex.

The bottom configuration illustrates that systems outside the Parallel Sysplex but belonging to the same sysplex (they share the same couple data sets) can belong to the same WLMplex. If systems outside a Parallel Sysplex belong to the WLMplex, then these systems must have XCF connectivity in form of CTC links.

The configurations shown are not meant to suggest that CF structures are used for anything other than XCF signalling in the WLMplex.

For more information related to WLM in the Parallel Sysplex environment, refer to *OS/390 MVS Planning: Workload Management*, GC28-1761. See also A.3.1, “Workload Manager (WLM) in the Parallel Sysplex Environment” on page 131 in Volume 3: Connectivity, SG24-5639.

There is also a new type of WLMplex, introduced in OS/390 V2R9. This is related to the use of Multisystem enclaves. Multisystem enclaves provides the ability for enclaves to run in address spaces on multiple systems within a Parallel Sysplex cluster. Work begun on one system, for instance, can be split into parallel tasks that can then be exported to other systems for parallel processing. Multisystem enclaves provide the capability to manage and report on work requests that are executed in a parallel environment on multiple OS/390 images as single entities. Enclaves can be managed across the Parallel Sysplex cluster with this new function. At the time of writing, this support is exploited by the DB2 Intelligent Miner for Data product.

Information about the enclaves is stored in a new CF structure, which means that any systems running an element of the multisystem enclave must have access to the CF containing the structure.

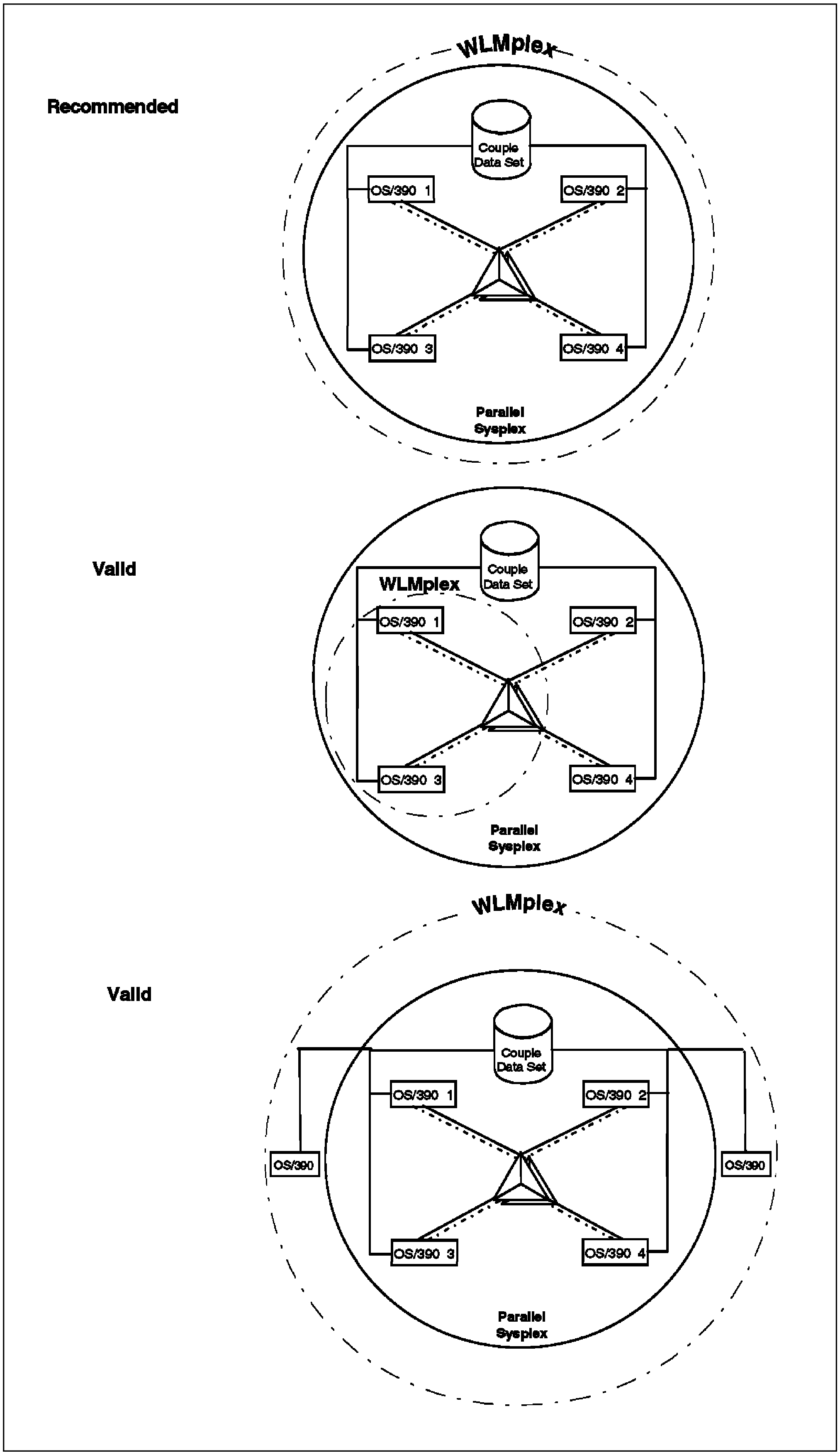


Figure 16. WLMplex Configurations

RMFplex This describes those systems that are connected by their use of the RMF Sysplex data server. The data server uses an in-storage wrap around buffer (data space) to hold SMF and RMF data. Since RMF uses XCF for communication, *the RMFplex can be smaller but not larger than the sysplex*. Note, however, that if the in-storage buffer is not defined, data from a system is not available interactively to other systems for interrogation using RMF Monitor II and Monitor III.

Monitor III data sets may be offloaded and serve as input to Monitor III sessions set up remotely. This implies that systems outside the sysplex may be used for "post-incident analysis." If you do sysplex analysis (or any sort of multi-system analysis), then all the corresponding Monitor III data sets must be offloaded. Monitor III only works in stop mode if working on preallocated data sets.

Figure 17 on page 63 shows both recommended and valid configurations for the RMFplex. While RMF reports on the use of CF structures, it does not use any CF structures itself for anything other than XCF signalling.

For more information about RMF in the Parallel Sysplex environment, refer to the *OS/390 RMF User's Guide*, SC28-1949. See also Appendix B, "RMF Reporting in Parallel Sysplex" on page 189 in Volume 2: Cookbook, SG24-5638.

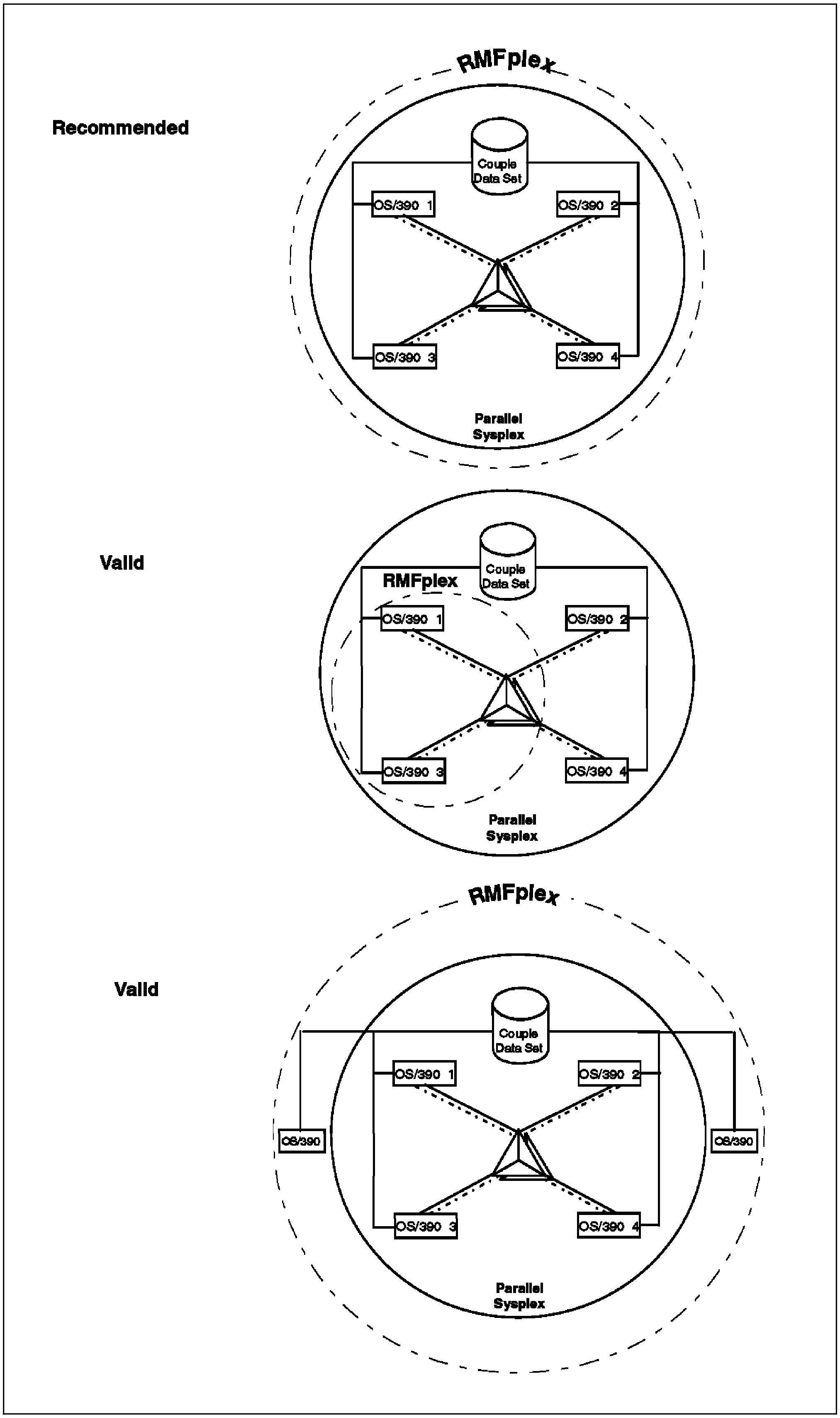


Figure 17. RMFplex Configurations

OPCplex This describes all the systems on which the Tivoli OPC controller has "trackers" running. *This can be a greater or lesser number of systems than that in the sysplex,* and may include other non-OS/390 platforms where OPC tracker agents are available.

Tivoli OPC allows you to start up multiple controllers using the PLEX initialization parameter. This will cause the first controller to start in a sysplex to be the controller. All other controllers that start after the first controller will be started in "standby" mode. Should the first controller fail, or the OS/390 system that it is running on fail, the standby controllers are notified via XCF. One of the standby controllers will then take over functions from the failed controller automatically. For the standby function to work, all controllers must be in the same sysplex. The trackers can be outside the sysplex.

The takeover function can also be manually initiated by an operator for a normal shutdown of the controlling system.

Figure 18 on page 65 shows valid configurations for the OPCplex. The configurations shown are not meant to suggest that CF structures are used for anything other than XCF signalling in the OPCplex.

For more information about OPC, refer to *TME 10 OPC Installation Guide*, SH19-4379.

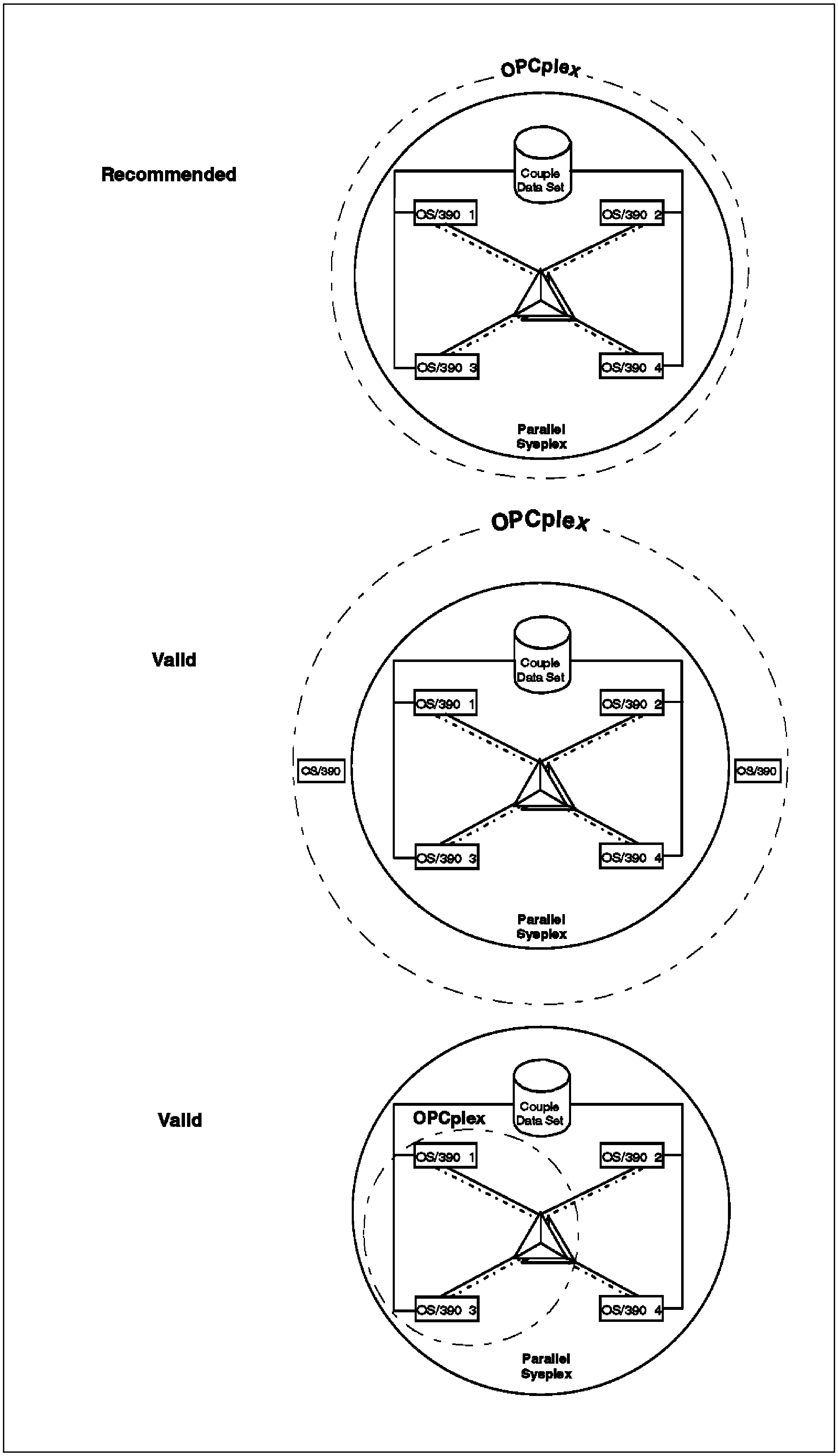


Figure 18. OPCplex Configurations

CICSplex Unfortunately, there are a number of different configurations that can be termed a CICSplex. Some of these are:

- The set of CICS regions that share a common VTAM generic resource name.
- The set of CICS regions that share non-recoverable temporary storage in a CF.
- The set of CICS regions that have access to shared data tables in a CF.
- The set of CICS regions that all access the same shared data (IMS, DB2, or VSAM).
- The set of CICS regions that all write to the same logger structure in a CF.
- The set of CICS regions that use XCF for MRO communication.
- A CICSPlex SM term for a group of CICS regions that it is to monitor and control as a logical entity.

Of these, the CICSplex as defined to CICSPlex SM can include CICS systems outside the sysplex, even on non-S/390 platforms, and therefore will be excluded from the following discussion.

If you plan to use XCF for MRO communication, then all the participating CICS regions need to be in the same base sysplex.

For the other four CICSplexes, all systems in each CICSplex need to communicate with the same CF. To make recovery easier, it is *highly* recommended that all four 'plexes line up with each other and with the Parallel Sysplex. Figure 19 on page 67 shows both recommended and valid configurations for the CICSplex. Note that the second figure is only valid for the CICSPlex SM type of CICSplex - all the others need to at least be in the same base sysplex.

Refer to *CICS TS for OS/390 Migration Guide*, GC34-5353, and *CICS TS for OS/390 Installation Guide*, GC33-1681, for a much more detailed explanation of the role of the CF in a CICS TS environment.

CICS TS R2 removed the requirement for CF for CICS logging and journals, using the DASD Logger in OS/390 V2R4 and subsequent releases instead. However, this configuration is only valid in a single system environment. It does *not* support data sharing environments.

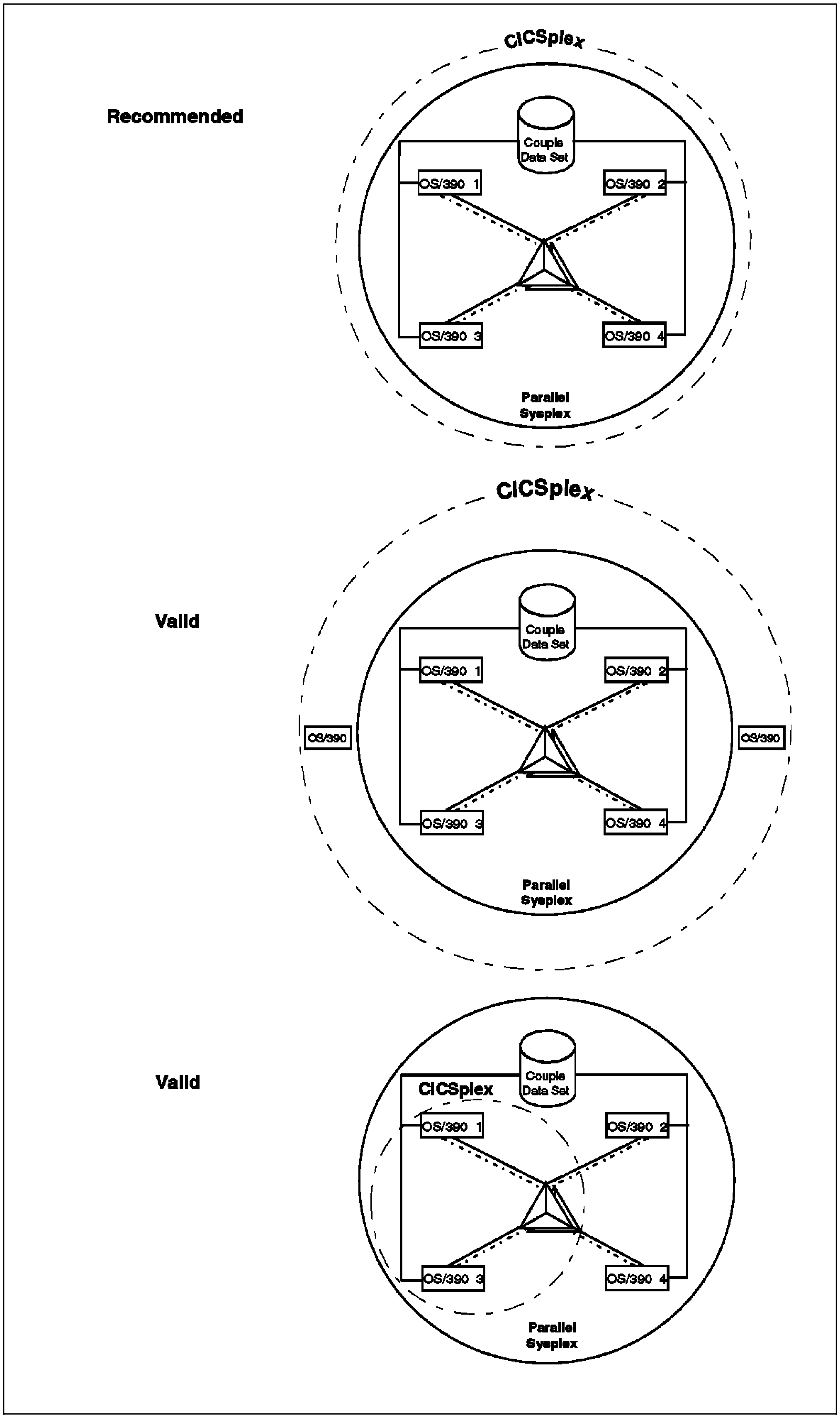


Figure 19. CICSplex Configurations

VTAMplex This may be used to describe the OS/390 systems that share VTAM generic resources. Several generic resource groups may exist within a Parallel Sysplex. All the systems with applications within a generic resource group must have connectivity to the same CF. Also, systems with applications using MNPS must have connectivity to a CF, as session data is stored in the CF to permit session recovery on another system.

Tapeplex Automatic tape switching uses a CF structure (IEFAUTOS) to keep track of tape allocation status and therefore an Auto Switchable (AS) tape drive can only be shared among systems within one Parallel Sysplex. If you have several Parallel Sysplexes, you need several AS tape pools, one for each Parallel Sysplex.

You may vary a tape drive offline on all systems in one Parallel Sysplex followed by a vary online on all systems in another Parallel Sysplex and vice versa, to move a tape drive from one pool to another, just as you would in a pre-Parallel Sysplex JES2 environment.

HMCplex This describes the largest set of CPCs that is monitored by one or more HMCs. Each HMC can manage up to 32 CPCs. To achieve maximum availability, the recommendation is to have at least two HMCs. The HMCplex can span only part of one sysplex, or it can span one sysplex, or it can span more than one sysplex. However, it is recommended that it match the sysplex, if possible. An HMC can manage IBM 9672 and 9674 systems in multiple sysplexes, provided they are all in the same HMC management domain. Even though a maximum of 4 HMCs can be ordered with each CPC, up to 16 HMCs can participate in the same management domain.

CECplex This name is sometimes used for the *hardware* construct of putting multiple CPCs in a single 9672-Exx/Pxx CPC. This applied to the first generations of IBM 9672 CPCs.

It can also be used to define a group of CPCs that are related either by physical location, or by the systems they support. The name of the CECplex is not defined anywhere.

2.3.5 'Plex Summary

Table 8 on page 70 summarizes the information about the various 'plexes. The table shows that the recommendation is almost always to match the 'plex with the Parallel Sysplex.

The columns in the table indicate:

- **'plex = Parallel Sysplex = sysplex**
  The 'plex matches the Parallel Sysplex and all systems are connected to each CF. In other words, the same number of images belong to the 'plex, the Parallel Sysplex, and the sysplex.
- **CF required**
  The 'plex requires a CF. Certain 'plexes have a mandatory CF requirement.
- **Data sharing using a CF**
  The 'plex uses the CF for data sharing among members of a data sharing group.

- **< Parallel Sysplex**
  The scope of the 'plex is smaller than the Parallel Sysplex. In other words, fewer systems may belong to the 'plex than to the Parallel Sysplex.
- **> Parallel Sysplex**
  The scope of the 'plex is greater than the Parallel Sysplex. In other words, a larger number of systems may belong to the 'plex than to the Parallel Sysplex.

  Care should be exercised at this point. In certain cases, systems outside the sysplex cannot be part of other sysplexes, or other restrictions may apply.
- **> 1 per Parallel Sysplex**
  Several 'plexes can coexist in the Parallel Sysplex. More than one 'plex may exist concurrently within the Parallel Sysplex. In certain cases the 'plexes may partly overlap (for example, a system may belong to two JESplexes in that it may be in both a JES2plex and a JES3plex).
- **> 1 Parallel Sysplex per 'plex**
  Several Parallel Sysplexes can coexist in the 'plex. Or put the other way around: the 'plex may comprise more than one Parallel Sysplex. Having a √ in this column does *not* always have an identical meaning. For a more complete discussion on this, refer to the last column in Table 8 on page 70 which contains a pointer to a discussion pertaining to particular 'plexes.
- **Non-S/390 Extension**
  S/390 can provide a single point of control for non-S/390 instances of the application.

Table 8. 'Plex Summary								
'plex	**'plex = Parallel Sysplex = sysplex**	**CF required**	**< Parallel Sysplex**	**> Parallel Sysplex**	**> 1 per Parallel Sysplex**	**> 1 Parallel Sysplex per 'plex**	**Non-S/390 Extension**	**More info on page**
Sysplex	√ √			√		√ (1)		39
JESplex (chkpt in CF)	√ √	√	√ (2)		√			44
JESplex (chkpt not in CF)	√		√	√ (3)	√	√		44
Pipeplex	√ √ (6)	√	√		√			46
Batchplex	√ √ (6)		√	√	√			46
GRSplex (star)	√ √ √	√						48
GRSplex (ring)	√			√ (4)				48
RACFplex (sysplex data sharing)	√ √	√	√ (5)					50
RACFplex (non-sysplex sharing)	√		√	√	√			50
SMSplex	√	√ (9)	√	√	√	√		52
HSMplex (using VSAM/RLS CDS)	√ √	√	√		√ (8)			54
HSMplex (non-data sharing)	√		√	√	√ (8)	√		54
BCSplex	√ √ √	√	√					56
OAMplex	√ √	√	√		√			58
WLMplex	√ √		√	√				60
RMFplex	√ √		√	√				62
OPCplex	√		√	√	√ (7)	√	√	64
CICSplex	√	√ (10)	√	√	√	√	√	66
VTAMplex (GR)	√ √	√	√		√			68
VTAMplex (MNPS)	√ √	√	√		√			68

Note:

(1): This configuration is not recommended.
(2): This configuration is possible if the JES2 MAS does not contain all the systems in the Parallel Sysplex.
(3): For pre-V5 JES2 and JES3 releases, there is no restriction on which systems are in the JESplex. For V5 and above JES2 and JES3, the JESplex can be greater than the Parallel Sysplex, but it cannot extend beyond the base sysplex.
(4): Systems outside the Parallel Sysplex may not be part of other sysplexes.
(5): If only a subset of the systems in the Parallel Sysplex are sharing the RACF database.
(6): Batchplex must match the Pipeplex.
(7): For the OPC standby function to work, all controllers must be in the same sysplex.
(8): This option is available starting at DFSMS/MVS V1.5.
(9): If using ECS, all sharing systems must be in same Parallel Sysplex and same GRSplex.
(10): CF required if using temporary storage, data tables, or named counter in CF support.

Entries with:

- Triple checkmarks (√ √ √) indicate a mandatory configuration.
- Double checkmarks (√ √) indicate a recommended configuration.
- Single checkmark (√) indicate a possible configuration. Potential drawbacks are discussed in the section referred to in the last column.

2.4 CF Architecture

The Parallel Sysplex exploits CF technology. The CF architecture uses hardware, specialized licensed internal code (LIC), and enhanced OS/390 and subsystem code. All these elements are part of the Parallel Sysplex configuration. This section outlines the CF architecture.

The CF consists of hardware and specialized microcode (control code) that provides services for the systems in a sysplex. These services include common storage and messaging in support of data integrity and systems management by high speed processing of signals on links to and from the systems in the sysplex. The CF specialized microcode, called Coupling Facility Control Code (CFCC) runs in an LP in a standalone CF or a CPC. Areas of CF storage are allocated for the specific use of CF exploiters. These areas are called *structures.* There are three types of structures:

1. *Lock*, for serialization of data with high granularity. Locks are, for example, used by IRLM for IMS DB and DB2 databases, by CICS for VSAM RLS, and by GRS star for managing global resource allocation.
2. *Cache*, for storing data and maintaining local buffer pool coherency information. Caches are, for example, used by DFSMS for catalog sharing, RACF databases, DB2 databases, VSAM and OSAM databases for IMS, and by CICS/VSAM RLS. Caches contain both directory entries and optional data entries.
3. *List*, for shared queues and shared status information. Lists are, for example, used by VTAM generic resources, VTAM Multi-Node Persistent Sessions, IMS shared message queues, system logger, the JES2 checkpoint data set, tape drive sharing, CICS temporary storage, and XCF group members for signalling.

For more information on structures, refer to 2.0, "A Structured View of CF Structures" on page 59 in Volume 2: Cookbook, SG24-5638.

There is also an area in the storage of the CF called *dump space*. This is used by an SVC dump routine to quickly capture serialized structure control information. After a dump, the dump space is copied into OS/390 CPC storage and then to the SVC dump data set. The dump space can contain data for several structures.

The CF has CF links but no I/O channels. The CF link normally refers to the connectivity medium, while CFS and CFR channels refer to *sender* and *receiver* (as defined through the IOCP). The CF has only CFR channels.

CF links provide high bandwidth communication with copper or fiber optic cables between the CF and the connected CPCs. Currently there are several types of links available:

- Internal Coupling Channel (IC), 6 Gb/sec, in this case the IC emulates a CF link between sysplex images *within the same CPC*.
- Integrated Cluster Bus (ICB) copper cable, 280 MB/sec, maximum cable distance up to 10 m.
- Multimode fiber optics, 50 MB/sec, maximum distance up to 1 km.
- Single-mode fiber optics, 100 MB/sec, maximum distance up to 3 km.

  Note: There is an RPQ available to extend the single-mode fiber to 40 km.

The CFCC has the following characteristics:

- CFCC runs only under LPAR, or in a Coupling Facility Virtual Machine (CFVM) under VM. When running in an LP, the CPs serving the CF LP can be shared or dedicated.
- CFCC is loaded from the support element (SE) hard disk (for IBM 967x CPCs). When running in a CFVM, VM issues a call to the S/390 processor when the CF machine is logged on and uploads the CFCC into the CF virtual machine and IPLs it. at LP activation. LPs are listed in IOCP along with their correspondence to a LP number. CF LP definitions with respect to storage, number and type of CPs are defined independent of HCD or IOCP. CF link definitions are done in HCD or IOCP.
- The major CFCC functions are:
  - Storage management.
  - Dispatcher (with MP support).
  - Support for CF links.
  - Console services (HMC).
  - Trace, logout, and recovery functions.
  - "Model" code that provides the list, cache, and lock structure support.

  CFCC code is usually in a continuous loop (known as an *active wait*) searching for work. Dynamic CF Dispatching (DCFD) causes the dispatching algorithms to enable CFCC to share CPs in LPAR mode with other active workloads and only take CPC resource when needed. When there is no work to do, a CFCC partition using DCFD takes less than 2% of the shared CPs. Refer to 1.2.4.1, "Dynamic CF Dispatching" on page 11 in Volume 2: Cookbook, SG24-5638 and WSC FLASH 9731, *MVS/ESA Parallel Sysplex Performance Dynamic CF Dispatch*, for more information on the Dynamic CF Dispatching enhancement.
- CFCC is not "interrupt-driven." For example:
  - There are no inbound CF interrupts (the CF receiver link does not generate interrupts to CF logical CPs).
  - There are no outbound interrupts. The completion of asynchronous CF operations is detected by OS/390 using a polling technique.

The CF is accessed through a privileged instruction issued by an OS/390 component called cross system extended services (XES). The instruction refers to a subchannel. The instruction is executed in *synchronous* or *asynchronous* mode. These modes are described as follows:

- **Synchronous:** The CP in the image waits for the completion of the request. There are two types of synchronous request:
  1. Synchronous Immediate (such as Lock requests).
  2. Synchronous Non-Immediate.

  OS/390 sometimes converts Synchronous Non-Immediate requests to Asynchronous requests; that is, it effectively modifies the XES macro issued by the requester. This is reported by RMF as "changed" requests. OS/390 never converts Synchronous Immediate requests.

OS/390 will transform a non-immediate request (also referred to as a “synchronous list” or “cache request”) to execute asynchronously if:

– A necessary resource, such as a subchannel, is not available at the time of the request.
– A multiple-buffer transfer of data was requested, regardless of the amount of data transferred.
– It is expected that the operation will take a long time to execute.

If OS/390 is running on shared CPs on the same CPC as a CF with shared non-ICF CPs, then PR/SM may perform an “under the covers” conversion to asynchronous for both Immediate and Non-Immediate requests. OS/390 is not aware of this conversion and it is not reported by RMF. The reason for the conversion is to prevent a potential deadlock from occurring. If OS/390 can determine such a deadlock will not happen, it uses a parameter on the XES macro to suppress the conversion by PR/SM. (For example, the OS/390 system may be issuing the request to an external CF).

- **Asynchronous:** The CP in the image issues the request, but the CP does not wait for completion of the request. XES will either return a return code to the requester, who may continue processing other work in parallel with the execution of the operation at the CF, or XES will suspend the requester. XES recognizes the completion of the asynchronous request through a dispatcher polling mechanism and notifies the requester of the completion through the requested mechanism. The completion of the request can be communicated through vector bits in the hardware system area (HSA) of the CPC where the image is running.

2.4.1 XES Providing Software Access to the CF

All the accesses to CFs are handled by the OS/390 XES component. XES application programming interfaces (APIs)[2] are the only means by which programs and subsystems can access structures in the CF. An example of how structures are accessed is discussed in 2.5.2, “Data Integrity in Parallel Sysplex” on page 80.

For a comprehensive discussion of the XES application interface, refer to *OS/390 MVS Programming: Sysplex Services Guide*, GC28-1771.

2.4.1.1 XCF and XES Considerations

Now we take a look at XCF and XES and what considerations have to be made when configuring a Parallel Sysplex.

XCF allows three methods for passing messages between members of XCF groups:

- Memory-to-memory: between members of the same XCF group in a single image
- Channel-to-channel (CTC) connections: between members of the same group on different images
- CF list structure: between members of the same group on different images

[2] Using the macros IXL...

The method chosen is an internal decision made by XCF based upon the target of the message and the available signalling paths. In this discussion, we only examine CTC and CF structure options. This is because they are the methods *most* relevant for consideration when configuring the Parallel Sysplex.

To ensure acceptable performance, it is vital that the CFs have sufficient capacity in terms of storage, links, and processing power. The CF configurations are set using the coupling facility resource management (CFRM) policy definitions. One of the considerations is the structure size. For a discussion on this topic, refer to 2.2.1, "Initial Structure Sizes" on page 71 in Volume 2: Cookbook, SG24-5638. For details on other aspects of the CFRM policy, refer to *OS/390 MVS Setting Up a Sysplex*, GC28-1779. Sample Parallel Sysplex policies may be downloaded from the OS/390 Integration Test Web site at: `http://www.s390.ibm.com/os390/support/os390tst/samplel.htm`

To create your own policies, we recommend that you use the Parallel Sysplex Configuration Assistant, available on the Web at: `http://www.s390.ibm.com:80/pso/psotool/ps_intro.html`

The use of the tool to define the CF structures and policies is described in the redbook *Parallel Sysplex Resource Sharing*, SG24-5666.

The other thing to consider is the size and placement of the various couple data sets. In A.5.7.1, "Couple Data Set Sizing" on page 164 in Volume 3: Connectivity, SG24-5639, we take a look at couple data sets. There are several types of couple data sets used in Parallel Sysplex. Their size and availability is critical for the continued operation of Parallel Sysplex.

2.4.1.2 XCF Signalling Path Considerations

Before we move on to the advantages of CF structures versus CTCs for signalling, look over the following terms, which are XCF-related terminology.

XCF Group: A group is the set of related members defined to XCF by a multisystem application in which members of the group can communicate (send and receive data) between OS/390 systems with other members of the same group. A group can span one or more of the systems in a sysplex and represents a complete logical entity to XCF.

Multisystem Application: A multisystem application is a program that has various functions distributed across OS/390 systems in a multisystem environment. Examples of multisystem applications are:

- CICS
- Global resource serialization (GRS)
- Resource Measurement Facility (RMF)
- Workload manager (WLM)

You can set up a multisystem application as more than one group, but the logical entity for XCF is the group.

Member: A member is a specific function (one or more routines) of a multisystem application that is defined to XCF and assigned to a group by the multisystem application. A member resides on one system in the sysplex and can use XCF services to communicate (send and receive data) with other members of the same group. However, a member is not a particular task and is not a particular routine. The member concept applies to all authorized routines running in the address space in which the member was defined. The entire address space could act as that member. All tasks and SRBs in that address space can request services on behalf of the member. Members of XCF groups are unique within the sysplex. In Figure 20, the association between groups and members is portrayed.

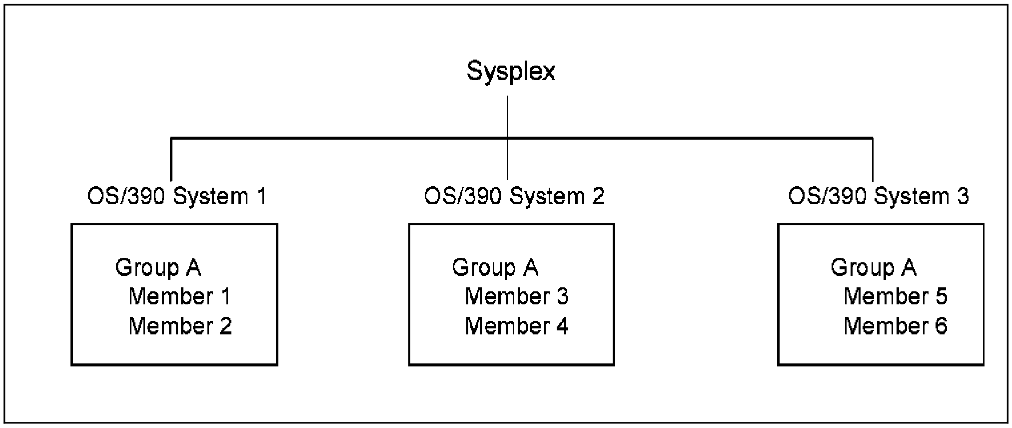


Figure 20. Systems, Groups, and Members in an XCF Sysplex

The two types of signalling paths of interest are:

- CTC connections
- CF list structures

When planning the signalling paths for XCF to use, remember to ensure that there is *redundancy* in whatever you configure.

For example, if you use CTCs for signalling, then you should define at least *two paths* into each system in the Parallel Sysplex and *two paths* out. This ensures that, if one path should fail, there will always be an alternative path available.

When using CF list structures exclusively for XCF signalling (no CTC connectivity between systems exists), at least two signalling structures should be defined and they should be allocated in at least two CFs.

Recommendation for CF Signalling Structures

If two CF signaling structures are used and placed in two CFs, then you should define at least *two paths into* each system in the Parallel Sysplex and *two paths out*. Note that this refers to paths, rather than links.

Multiple paths can, and should, be defined on each physical link so that if you lose a link, you will still have a path-in and a path-out available on the other link. As a consequence, connectivity is maintained in case of problems with one structure or one link.

XCF supports the structure rebuild function of XES, but it will take some time if there is no alternative signalling connectivity (either CTCs or a second CF structure) available. XCF will use, as a last resort, the sysplex couple data set to accomplish the rebuild process synchronization. The XCF-internal signals used to coordinate the rebuild process generate heavy traffic to the couple data set. As a result, normal user-specified signals (for example, an instance of CICS sending a signal to another instance of CICS) may be delayed during rebuild of the last or only signalling structure in the sysplex.

Another good reason for having *multiple signalling paths,* either CTC or structure, is for *throughput* considerations. If XCF has multiple paths, it will not try to put a request onto an already busy path. Alternatively, you can decide to route messages for specific XCF groups along specific paths. This can be useful if one particular group is known to have specific signalling requirements, such as it always uses long messages. The systems programmer could assign XCF groups to *transport classes* and then associate the transport class with a specific signalling path or paths.

For detailed information, refer to *OS/390 MVS Setting Up a Sysplex*, GC28-1779.

Recommendation about XCF Message Sizes and Signalling Paths

When using XCF for the first time, assign all signalling paths to the *default* transport class. Then examine RMF XCF Activity reports to determine if a larger message buffer than that of the default transport class is needed, based on the percentage of messages that are too big for the default class length. This is indicated in the %BIG column. Refer to WSC Flash 10011 (a replacement for Flash 9723) *MVS/ESA Parallel Sysplex Performance: XCF Performance Considerations*, for more information on XCF performance considerations.

If a message buffer is too *small* for a particular message request, the request incurs additional overhead because XCF must format a larger message buffer. If a message buffer is too *large* for a particular message request, central storage is wasted. XCF adjusts the size of the buffers in a particular transport class on a system-by-system basis based on the actual message traffic. Overhead is indicated in the %OVR column.

For example, in a given RMF interval, if 50,000 messages are sent from one system to another and the %SML=0, %FIT=90, %BIG=10 and %OVR=95, this means that 5000 messages were too big to fit in the existing message buffer and that 95% of the 5000 (big) messages incurred overhead from XCF buffer tuning functions. See Appendix B, “RMF Reporting in Parallel Sysplex” on page 189 in Volume 2: Cookbook, SG24-5638 for a sample of the RMF XCF Activity report.

If the amount of overhead from XCF buffer tuning is unacceptable, consider defining another transport class for the larger messages and assigning a signalling path or paths to that transport class to handle the larger message traffic. The following three transport classes work well in many cases and are, at the time of writing, recommended by the OS/390 integration test team:

```
CLASSDEF CLASS(DEFAULT) CLASSLEN(20412) MAXMSG(750) GROUP(UNDESIG)
CLASSDEF CLASS(DEF8K) CLASSLEN(8124) MAXMSG(1000) GROUP(UNDESIG)
CLASSDEF CLASS(DEFSMALL) CLASSLEN(956) MAXMSG(750) GROUP(UNDESIG)
```

Note: It is generally not necessary to eliminate all buffer expansion and overhead when tuning XCF transport classes.

There are a few known common causers of large (64 KB) signals. WLM will occasionally send them. GRS GQSCAN processing can send them, and RMF Sysplex Dataserver processing can send them as well.

If the signalling paths are via XCF structures, then there are no means to figure out who is sending them. If you are using CTCs, then a GTF CCW trace may reveal some information. When you specify the value for CLASSLEN, subtract 68 bytes for XCF control information. XCF will round to the next 4 K increment.

Defining XCF transport classes with GROUP(UNDESIG) allows XCF to select which transport class (and their assigned signalling paths) messages should be transferred over based on message size.

Note: See APAR OW16903 for an enhancement to the XCF buffer adjustment algorithm.

The right number of signalling paths is indicated by the column BUSY in the RMF XCF Activity Report. If the busy count rises, the number of paths or the message buffer space for the outbound path should be increased. Experience shows that if all paths are made available to any transport class that needs them, the workload is spread evenly, and workload spikes are handled well.

Assigning groups such as GRS or RMF to a specific transport class is *not recommended* unless there is a proven need to do so.

Generally speaking, it is acceptable to mix the different transport classes on a single link. However, if you have tens of thousands of XCF messages per second, there may be a benefit to dedicating a transport class to a link, and have separate links for PATHIN and PATHOUT. Also, if you have very large XCF messages, you will see a greater benefit from the use of high-bandwidth links (such as IC links) than if your messages are small.

2.4.1.3 What Type of Signalling Path Should I Use?

There are three possible options for the signalling between systems in a Parallel Sysplex:

- CTCs only
- CF structures only
- A combination of CTCs and CF structures

For low XCF rates of activity, there is little performance difference between CTC and CF signalling paths. At high rates, performance may be improved by using CTCs, especially if no HiPerLinks, ICBs or ICs are available. If they are available, however, either structures or a mix of structures and CTCs should be used.

At the time of writing, XCF does not attempt to balance the PATHOUT activity. XCF does not try to balance path utilization. XCF will send test signals down the link and watch the response times for different paths. XCF then places the paths into either a “best” stack or a “good” stack. XCF will use the path at the top of the “best” stack until its test signal response time indicates the XCF signalling performance is impacted at which time it will use the next path on the “best” stack. When the “best” stack is empty, XCF will attempt to use the paths on the “good” stack. Only when both “best” and “good” stacks are empty, will XCF resort to a round robin selection.

To net it out, XCF path selection is based on a set of dynamic response times which are managed by XCF, so even utilization of the paths is not necessarily expected. Other traffic on the CHPID can also influence XCF since other non-XCF traffic can impact the response times of the XCF test signals.

Note: APAR OW40631 should be applied to ensure optimum path usage in environments with high XCF message rates. This APAR also enhances the IXC356I message by adding response time information for each path.

XCF will always use the fastest path for signalling, when there is a choice. The current recommendation therefore is to configure both CF structures and CTCs and allow XCF to determine the best path. For a more detailed discussion of this, refer to *OS/390 MVS Setting Up a Sysplex*, GC28-1779.

The HiPerLinks available for G3 and follow-on CPCs and corresponding CFs significantly improve the performance for CF signalling paths. ICBs and ICs, available for G5 and follow-on CPCs further improve the performance for CF signalling paths. For sysplexes consisting of more than two CPCs, the performance is comparable with, or in some cases better than, CTC connections. Therefore, with HiPerLinks, ICBs and ICs, it is possible to use only CF XCF signalling paths and dispense with the added complexity and expense of CTCs

For more specific details of HiPerLinks, ICBs and ICs, refer to 1.2.5.3, “High Performance Links (HiPerLinks)” on page 15 in Volume 2: Cookbook, SG24-5638.

CTC Configuration Guidelines

Full CTC connectivity, redundancy, and extensibility requires four ESCON channels between each CPC. If you have multiple CPCs, this really mandates the use of ESCON directors.

2.5 Data Integrity and Buffer Pool Consistency Considerations

Data integrity is an important issue. First we discuss data integrity prior to Parallel Sysplex (both in one system and in a multisystem environment). Then we discuss data integrity in a Parallel Sysplex.

2.5.1 Data Integrity before Parallel Sysplex

When only one OS/390 system has access to database records, current S/390 data management products are able to address data integrity for thousands of active users. They apply a *database manager locking* technique to ensure data is valid for multiple tasks that may be accessing the same data. They use in-storage buffering to improve transaction performance, holding frequently used data and recently used data in buffer pools to avoid the overhead of reading data from DASD whenever possible.

You may have different reasons to grow from one OS/390 to multiple OS/390 systems, for example:

- A demand for continuous availability.
- The workload has grown too large for one OS/390 system.
- A desire to lower overall cost of computing.

In a multisystem environment, it is more complex for the systems to do locking and buffering than in a single system environment. When two or more systems need to share data, communication between those systems is essential to ensure that data is consistent and that each system has access to the data. One way that this communication can occur is through *message passing*.

Let us assume two images (S1 and S2) share the same database, as pictured in Figure 21 on page 79. When we start, the same record, ABC, is stored in each local buffer pool in each image.

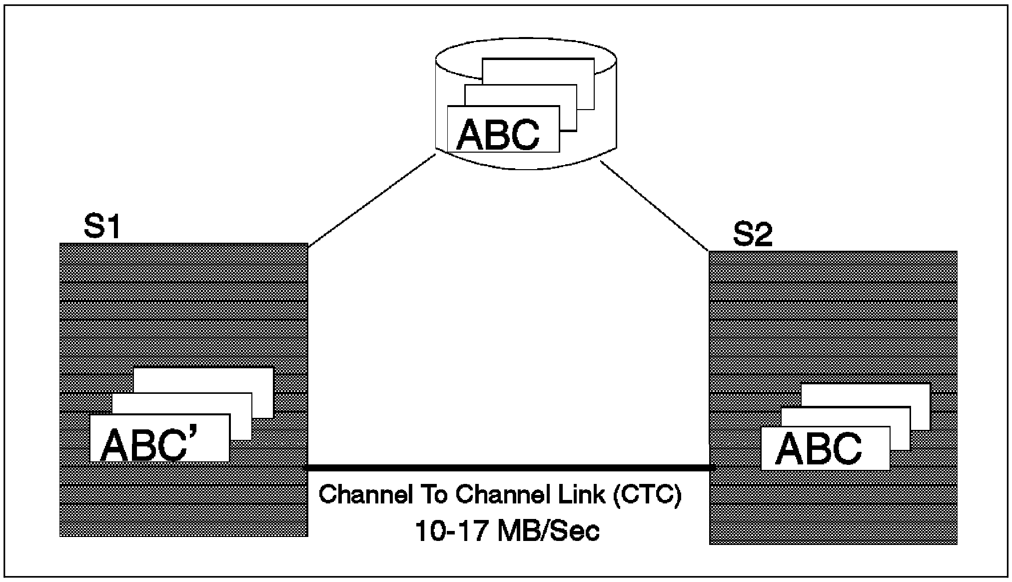


Figure 21. Multisystem Data Sharing

Let us assume that S1 needs to update the ABC record. S1 sends a message to obtain a lock across a channel-to-channel (CTC) link to S2, requesting access to the record. If S2 is not using the record, it sends a message back to confirm that S1 can have the access it requires. Consider two things about what was just described:

1. Both S1 and S2 applications had to stop what they were doing to communicate with each other.
2. If there are three or more systems, this scheme becomes even more complex. The *cost* of managing the sharing in this method is *directly proportional* to the number of systems involved.

This problem is even more complicated than just described. Suppose that S2 wants to read the ABC data and that the most recent copy of the record ABC is not yet written on the DASD file that both systems are sharing. It might be in an internal buffer of S1 (record ABC'). In this case, S2 must first be informed that the copy on the disk is not valid. S1 must then write it to DASD before S2 can have access to it.

The mechanism that tells S2 that the copy of the data in its buffer pool is the most recent one is called *maintaining buffer pool coherency*.

Some database managers offered such mechanisms before the Parallel Sysplex became available. For example, IMS DB provided two-way data sharing through IRLM (using VTAM CTCs for communication).

2.5.2 Data Integrity in Parallel Sysplex

Parallel Sysplex provides, among other functions, the ability to manage database integrity (locking and buffering) for up to 32 images in a much simpler way than in a non-Parallel Sysplex multisystem configuration. We need to allow the sharing of data among *many systems* without causing the problems described in the prior example. Data must be shared in such a way that adding systems does not increase the complexity or overhead, and still preserves data integrity across updates. To solve this complex problem, a sharing technique is needed. This is implemented through the CF technology that manages the *communication* and *serialization* necessary for *sharing* of data. The CF function works with *structures* in the CFs that help manage the serialization of access to data and the coherency of the data.

OS/390 provides a set of XES services that enable subsystems and authorized applications to use CF structures. These services provide buffer pool coherency and locking serialization, which allow data sharing to occur between many systems. As an example of how this works, look at Figure 22, which shows an example of how cache structures are used when two or more systems are sharing data.

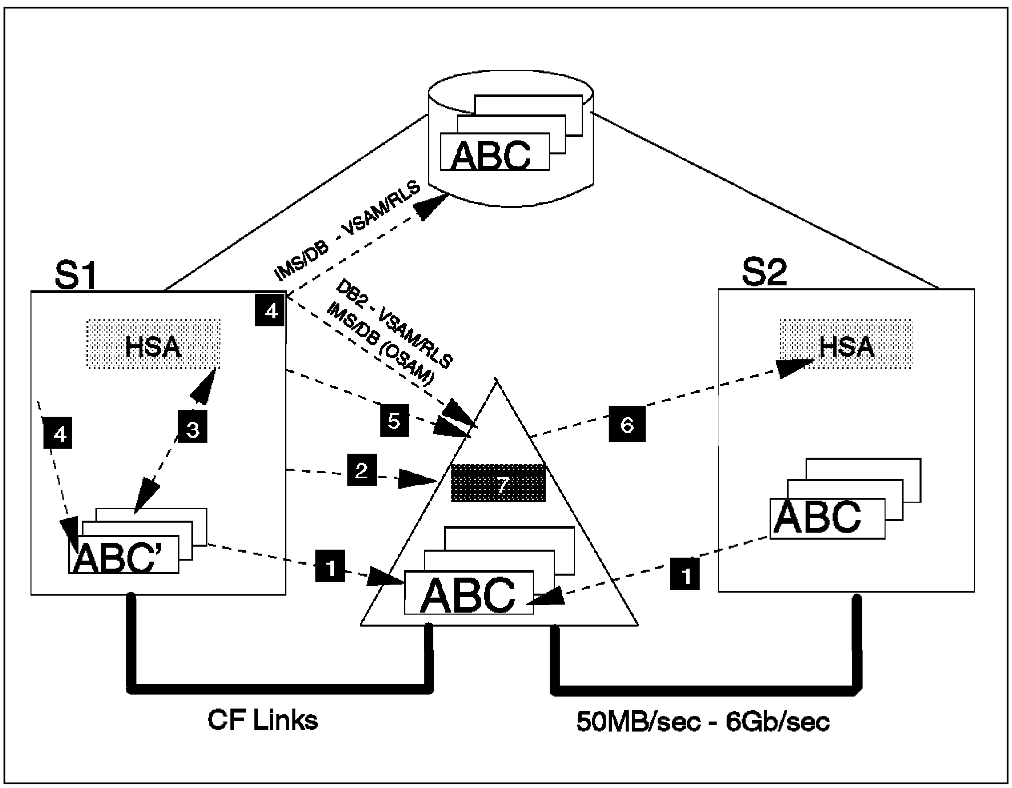


Figure 22. Multisystem Data Sharing in Parallel Sysplex

Two systems are sharing database records. Looking at the local buffer pool for each system, you can see some records (including the ABC record). Assume both systems have read record ABC in the past, and a valid current copy exists in each buffer pool.

.1. Both systems have registered their interest in record ABC with the CF. Now an application in S1 (on the left) needs to update this ABC record to ABC'.

.**2**. The DBM in S1 invokes an OS/390 service that calls the CF to obtain an exclusive lock for the update. Depending on the subsystem implementation, the lock is obtained for one or more records.

Assume that the lock is granted.

.**3**. Now S1 looks in its HSA vector bit table associated with the buffer to see if the record in the buffer is a valid copy.

Assume it is valid, so S1 does not have to go to DASD or CF to obtain a copy.

.**4**. S1 changes the record to ABC'. It is changed in S1's local buffer, and depending on the subsystem protocol, it is written to:

- A cache structure of the CF(s). An example is IMS DB V6 (OSAM using a store-in technique) or DB2. On a per-tablespace basis, DB2 provide several options for how data is cached in the structure:
  - All data.
  - Changed data only.
  - Part of the data (space map pages for LOBs).

  For duplexed DB2 cache structures, the data is also written to the secondary structure, but *asynchronously*.
- A cache structure in the CF *and* to DASD. Examples are VSAM/RLS and IMS DB V5 (OSAM) using a store-through technique.
- DASD only. An example is IMS DB V5 (VSAM) using a directory-only technique or DB2 using the “no cache” option.

.**5**. A signal is sent by the database manager to the CF to show that this record has been updated. The CF has a list, stored in a cache structure directory, of every system in the Parallel Sysplex that has a copy of the record ABC, which has now been changed to ABC'. A directory entry is used by the CF to determine where to send cross-invalidation signals when a page of data is changed or when that directory entry must be reused.

.**6**. Without interrupting any of the other systems in the Parallel Sysplex, the CF invalidates all of the appropriate local buffers by changing the bit setting in the HSA vector associated with the record to show that the record ABC in the local buffer is down level. This operation is called *cross-invalidation*. Buffer invalidation may also occur under other circumstances, such as when a contention situation is resolved through directory reclaim.

.**7**. At this point, the serialization on the data block is released when the update operation is complete. This is done with the use of global buffer directories, placed in the CF, that keep the names of the systems that have an interest in the data that reside in the CF. This is sometimes referred as the *unlock operation*.

In a store-in operation, the database manager intermittently initiates a cast-out process that off-loads the data from the CF to the DASD.

The next time the images involved in data sharing need record ABC, they know they must get a fresh copy (apart from S1, which still has a valid copy in its local buffer). The systems are not interrupted during the buffer invalidation. When the lock is freed, all the systems correctly reflect the status of the buffer contents.

If S1 does not have a valid copy in the local buffer (as was assumed in step 3.), it must read the record from either the cache structure in the CF or from DASD. The CF only sends messages to systems that have a registered interest in the data, and it does not generate an interrupt to tell each system that the buffer is now invalid.

The example described above details the steps necessary every time a record is referenced. Note that the shared locking also takes place if a system reads a record without ever doing an update. This guarantees that no other systems update the record at the same time. The example discussed here can easily be expanded to more than two systems.

A summarization of the event sequence needed to access a database with integrity in a Parallel Sysplex includes:

1. An element, or list of elements, is locked before it is updated.
2. An element is written to DASD and/or the CF no later than sync point or commit time.
3. Buffer invalidation occurs after the record is written.
4. Locks are released during sync point or commit processing, and after all writes and buffer invalidations have occurred. The locks may be released in batches.

2.5.3 Locking in Parallel Sysplex

An important concept when designing a locking structure is the *granularity* of the database element to be serialized. Making this element very small (such as a record or a row) reduces lock contention, but increases the number of locks needed. Making the element bigger (such as a set of tables) causes the opposite effect. There are options in IMS DB, CICS/VSAM RLS, and DB2 where the installation can define this granularity. Refer to the following sections for more information:

- 4.3.2.1, "IMS DB Data Sharing in a Parallel Sysplex" on page 203.
- 4.3.3, "CICS/VSAM Record Level Sharing Considerations" on page 207.
- 4.3.1, "DB2 Data Sharing Considerations" on page 194.

To avoid massive database lock structures in the CF, a *hashing technique* is used. Every element in the database to be locked has an identifier (ID). This ID is hashed, and the hash result is an index into the lock structure. Because there are more database elements than lock entries, there is a probability of *synonyms*. A synonym is when two or more distinct elements point to the same lock entry. When this situation happens, all the subsystems from different images that are using this lock structure need to perform lock resource negotiation. After this negotiation, they decide if the contention was *false* (just a synonym) or *real* (two subsystems trying to get access to the same lock).

Note that the lock requester may or may not be suspended until it is determined that there is no real lock contention on the resource. For example, the requester may be processing other work in parallel while XES makes the determination. Contention can be a performance problem for workloads that have a high level of sharing (locking) activity.

Total contention is the sum of false and real contention and should be kept low.

Recommendation for Locking Contention

Total locking contention should be kept to less than 1.0% of the total number of requests. If possible, try to keep false contention to less than 50 percent of total global lock contention. If total global lock contention is a very low value, it might not be as important to reduce false contention.

Real Contention: Real contention is a characteristic of the workload (which is the set of locks that are obtained, lock hold time, and the nature of workload concurrency), and is not easy to reduce by tuning the Parallel Sysplex configuration. It is tuned by *tuning the workload itself*. For example, you can try to avoid running long batch jobs concurrent with online work, avoid running concurrent batch jobs that access the same data, increase the frequency of checkpoints, accept "dirty reads," and so forth.

Real contention is also reduced by controlling the degree of data sharing. CP/SM and WLR may, for example, help to make sure that transactions needing certain data are run on certain systems.

False Contention: False contention has to do with the hashing algorithm that the subsystem is using, and the size of the lock table in the lock structure. It is true that increasing the size of the lock table in the lock structure reduces false contention. Carrying this to extremes, by making the lock table arbitrarily large, one can make the false contention become as small as one wants. This has practical limits, since, for IRLM, the only way to increase the size of the lock table is to increase the size of the lock structure as a whole. To minimize the effect of false contention, there are three options:

- Increase the lock structure size in the CF.
- Decrease the granularity of locking; or in other words, reduce the number of locks. However, by decreasing the granularity of locking, you increase the probability of real contention at expense of the false contention. Therefore, this option might not always be considered a solution.
- Decrease the number of users connecting to the lock structure, thus allowing more lock entries within a given structure size.

For more information on lock contention and false lock contention, refer to 2.0, "A Structured View of CF Structures" on page 59 in Volume 2: Cookbook, SG24-5638.

2.6 Dynamic Workload Balancing in Parallel Sysplex

With parallel processing, the workload can be distributed and balanced across a subset, or across all the CPCs in the Parallel Sysplex. Dynamic workload balancing aims to run each arriving transaction in the "best" system. IBM's two major transaction management systems, IMS and CICS, due to their different structures, implement workload balancing in different fashions, as the following describes:

- IMS/ESA V6 offers workload balancing by implementing algorithms that make it possible to *pull* messages from a sysplex-wide shared message queue. In IMS/ESA V6, you can also perform workload balancing using VTAM GR. Using VTAM GR is complementary to using shared queues. Generic resources distributes user logons across multiple IMS subsystems, while shared queues distributes the task of executing the transactions across multiple message processing regions associated with multiple IMS subsystems.

IMS support for the OS/390 Workload Manager (WLM) helps OS/390 balance the workload in accordance with business objectives.

- For IMS/ESA V5, the IMS/ESA Workload Router implements workload balancing by distributing an IMS workload among two or more IMS online systems interconnected through MSC communication links. IMS WLR uses user exits to distribute the workload. It detects and reroutes transactions around inactive resources, reconfigures workload distribution among active resources, and reinstates work when resources become available. IMS WLR also assists in managing peak workloads by allowing work to be directed to systems with available capacity. Refer to 4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187.
- CICS implements workload management by *pushing* transactions to different AORs. These AORs are in the same or on different images.

  In CICS, there are two aspects to workload management:

  1. Workload separation—for example, using TORs, AORs, and FORs.
  2. Workload balancing—distributing logon requests and transactions across more than one CICS region.

  Both may be in operation within a CICSplex. CICS may use three workload balancing techniques:

  – Goal mode algorithm, using CP/SM.
  – Shortest queue algorithm, using CP/SM.
  – Dynamic transaction routing program (either supplied by CICSPlex SM or by your installation).

  Starting with CICS TS for OS/390 R1, it was also possible to share CICS non-recoverable temporary storage sysplex-wide. When the temporary storage queues are shared by all images in the Parallel Sysplex, there are fewer affinities tying a transaction to a particular CICS region.
- DB2 V4 and subsequent releases provides automatic and dynamic work balancing for certain work.
- At the time of writing, MQSeries for OS/390 V2.2 has previewed the fact that it will support placement of Message Queues within a CF structure, thus allowing servers on multiple OS/390 images to process messages from a single queue.
- OS/390 R4 implements workload balancing for batch jobs. The workload manager can dynamically manage jobs that are goal-oriented by distributing them to multiple systems in a sysplex, reducing operational demands and improving their total response time.
- OS/390 R3 implements workload balancing for TSO/E using generic resources; for further descriptions refer to 4.7, "TSO/E and Parallel Sysplex" on page 227.

Another helpful aspect of workload balancing is availability. If an OS/390 failure occurs, only a portion of the logged-on users will be impacted, and subsequent transactions can be routed to another OS/390 system that is unaffected by the failure.

One of the characteristics of dynamic workload management is an increase in the rate of cross-invalidation of records in the local buffer pools. Total flexibility increases the probability that the same record is referenced and updated in more than one system, raising the number of cross-invalidations in the local buffer pools, and increasing the overhead.

2.6.1 Dynamic Workload Balancing Exploiters

Currently, the exploiters of dynamic workload balancing in a Parallel Sysplex are CICS/ESA V4.1, IMS/ESA V6.1, DB2 V4.1, VTAM V4.2, WLM for OS/390 V2R4, the JES2 component of OS/390 V2R4 (for batch), the JES3 component of OS/390 V2R8 (for batch), the TSO component of OS/390 R3, TCP/IP V3R2, and follow-on releases of all these. In addition, WLR, while not using any Parallel Sysplex facilities, can be used to balance IMS transactions over a number of TM regions, even in IMS/ESA V5.1 which does not support VTAM generic resources. The listed exploiters are described as follows:

CICS TS for OS/390 R3: CICS extends its dynamic routing facility to provide mechanisms for dynamic program link (DPL) requests, and a subset of START commands. The routing mechanisms allow workload balancing to be managed by CICSPlex SM enabling workload balancing for EXCI clients, CICS Clients, and started tasks.

It also provides the ability to externalize CICS ENQ and DEQ commands, so a CICS resource can be ENQ'd in one CICS region and DEQ'd from another. Support is also provided for placing CICS data tables in a CF—this can be used to replace the use of CSA or CWA to pass data between transactions.

CICS TS for OS/390 R1 added support for placing CICS temporary storage queues in the CF. This allows you to remove affinities between programs that share temporary storage queue entries.

CICS/ESA V4.1: In CICS, the terminal owning region (TOR) can route the arriving transaction to an application owning region (AOR) in the same OS/390 image, in another image in the Parallel Sysplex, or to any image in the network. This routing can be done through CICSPlex SM, or through installation-written exit routines.

Starting with CICS/ESA V4.1, CICS may obtain the goal from WLM (if in goal mode). If your system is in goal mode and you have CICSPlex Systems Manager (CICSPlex SM) installed, CP/SM will route the transaction to the AOR that historically presented a response time as close as possible to the goal of the transaction. For more information, refer to 4.2.2, "CICS Transactions in a Parallel Sysplex" on page 169.

IMS/ESA V7.1: IMS Transaction Manager Version 7 introduces a new facility called Rapid Network Recovery. This facility utilizes VTAM MNPS to enable faster recovery of IMS sessions following a failure of the IMS subsystem that owned those sessions. For more information, see *IMS/ESA V7 Release Planning Guide*, GC26-9437.

IMS/ESA V6.1: IMS Transaction Manager supports VTAM Generic Resources. This support allows VTAM to balance the network workload by routing session requests across multiple IMS TM images in the Parallel Sysplex using selection criterions. IMS TM also provides shared message queues support that allows up to 32 IMS DB/DC subsystems to have access to a common workload. Refer to 4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187 for more information. Also, 4.2.5.2, "IMS VTAM Generic Resources" on page 187 contains more information on the selection criteria used by VTAM.

DB2 V5.1: DB2 provides automatic work balancing through its sysplex query parallelism function.

DB2 V5.1 also introduced WLM-managed Stored Procedures. This facility allows you to let WLM manage the number of stored procedure address spaces, based on the importance of the stored procedure and the available capacity on the system.

DB2 V4.1: DB2 can distribute Distributed Data Facilities (DDF) requests received from a gateway across servers on different images, using information from the OS/390 workload manager about the CPC capacity available to those images. For more information, refer to 4.3.1, "DB2 Data Sharing Considerations" on page 194.

Additionally, DB2 V4.1 uses VTAM generic resources to route DRDA session initialization requests across to DB2 subsystems in the data sharing group.

VTAM V4.4: VTAM provides Multi-Node Persistent Session (MNPS). This support allows sessions to survive a VTAM or OS/390 and OS/390 failure thus going one step further to continuous end-to-end availability. This support is provided in an APPN/HPR network. VTAM V4.4 also supports APPN/HPR link definition through XCF and then provides fast communication paths for all VTAM to VTAM traffic across images in a sysplex.

VTAM V4.4 extends the VTAM generic resource support to TSO/E; this provides the capability to balance session distribution across TSO systems in the Parallel Sysplex using the CF. Refer to 4.5, "Network Workload Balancing Capabilities" on page 219 for more information. Also see the ITSO redbook *VTAM in a Parallel Sysplex Environment*, SG24-2113.

VTAM V4.3: VTAM allows you to dynamically add applications. In a sysplex environment, model application program definitions can be contained in a single VTAMLST accessible to all VTAMs in the sysplex. For example, you can define a model application with a VTAM APPLID of SERVLU? on every system. The APPLID does not get activated because VTAM knows this is only a model. However, when application SERVLU1 starts, VTAM will then activate the resource.

An application program can open its access method control block (ACB) on any VTAM in the sysplex. In addition, an application program can be moved to any VTAM in the sysplex by closing its ACB on one VTAM and opening it on another.

VTAM V4.2: VTAM provides a generic resource function. This function is exploited by CICS/ESA V4.1 to distribute CICS logons across different TORs in OS/390 systems in the Parallel Sysplex. DB2 V4.1 exploits the VTAM generic resource function to balance DRDA sessions across DB2 regions. For more information, refer to 4.5.1.2, "VTAM Workload Balancing for Generic Resources" on page 221.

JES3 in OS/390 V2R8 In OS/390 V2R8 and above, JES3 will work with WLM so that WLM can control the stopping and starting of initiators. WLM uses the goals defined for the batch jobs to decide if more initiators are required to meet the goals, or if it can meet the goals with fewer initiators.

JES2 in OS/390 V2R4 In OS/390 V2R4 and above, JES2 will work with WLM so that WLM can control the stopping and starting of initiators. This occurs both within a single system and also across multiple systems in a JES2 MAS. WLM uses the goals defined for the batch jobs to decide if more initiators are required to meet the goals, or if it can meet the goals with fewer initiators.

TCP/IP in OS/390 V2R8 In OS/390 V2R8, the TCP/IP stack will automatically register and deregister with the Automatic Restart Manager (ARM) during initialization and normal termination. If TCP/IP has successfully registered with ARM and then terminates abnormally, TCP/IP will be automatically restarted on the MVS image on which it was originally started.

Another new feature in OS/390 V2R8 is Dynamic Virtual IP Addressing. Virtual IP addresses (VIPAs) are not tied to any particular stack and are accessible outside the owning stack (host) via any physical interface to that stack. In previous releases, if that stack failed, you could manually move the application (server) workload and activate a VIPA on another stack via the VARY TCPIP,OBEY command. With Dynamic VIPA support, the VIPA will be automatically taken over by another stack within the same sysplex. New VIPAs are dynamically created. Furthermore, VIPAs can become the address of an application server in the sysplex.

TCP/IP in OS/390 V2R7 In OS/390 V2R7, TCP introduced Dynamic XCF, in which new TCP stacks in a sysplex configured for Dynamic XCF automatically gain IP connectivity to all existing stacks also configured for Dynamic XCF without having to change the configuration on the existing stacks. This support makes it much easier to add a new stack to the sysplex, and removes the requirement for dedicated connections between the systems.

TCP/IP in OS/390 V2R4 TCP/IP uses a technique called Connection Optimization. This feature uses the Domain Name Server (supplied with the TCP/IP stack for OS/390 UNIX Services) and WLM to balance IP connections and workload across a sysplex domain. This feature is available with OS/390 V2R4 and also depends on support from TCP/IP server applications. With OS/390 V2R4, the servers providing the necessary support are TN3270 and DB2. Additional servers support this function in OS/390 V2R5. See 2.3.2.3, “TCP/IP Connection Optimization” on page 49 in Volume 3: Connectivity, SG24-5639 for more information on this topic.

TSO in OS/390 R3 In OS/390 R3 and above, it is possible to define a generic resource for TSO. If this is done, VTAM in conjunction with WLM, will route TSO logon requests to the system with the most available capacity. Additional support is provided so that LOGON RECONNECT requests are routed to the correct system.

Network Dispatcher The network dispatcher is software that runs in a router device. The objective of the network dispatcher is similar to that of DNS and WLM—namely, to select the most appropriate server for each new TCP/IP connection request.

Using the network dispatcher, clients reference a single, special IP address, known as a *cluster* address. This address is shared between a set of hosts in the network and the network dispatcher. The network dispatcher selects the most appropriate server at the time the connection request arrives, and sends the packet to that server. The next request that arrives from the client, using the same IP address, could potentially get routed to a different server.

As stated above, the objective of the network dispatcher and the new DNS support is similar, but each is more suited to different types of applications. The use of these two features is discussed in detail in the redbook *TCP/IP in a Sysplex*, SG24-5235.

APPC in OS/390 R3 Starting with OS/390 R3, APPC/MVS can associate a VTAM generic resources name with APPC/MVS LUs, to improve availability of APPC/MVS resources, and to balance sessions among APPC/MVS LUs.

Chapter 3. Continuous Availability in Parallel Sysplex

Continuous availability is one of the major advantages of migrating to a Parallel Sysplex. This chapter discusses how some of the configuration choices affect system availability in a Parallel Sysplex.

Recommended Sources of Further Information

The following sources provide support for the information in this chapter:

- *DB2 on the MVS Platform: Data Sharing Recovery*, SG24-2218
- *DB2 UDB for OS/390 and Continuous Availability*, SG24-5486
- *DB2 UDB for OS/390 V6 Data Sharing: Planning and Administration*, SC26-9007
- *DB2 UDB for OS/390 V6 Release Guide*, SC26-9013
- *DB2 Universal Database Server for OS/390 V7 What's New?*, GC26-9017
- *HCD and Dynamic I/O Reconfiguration Primer*, SG24-4037
- *IBM Fiber Saver (2029) Implementation Guide*, SG24-5608
- *IBM Global Services High Availability Services*, available on the Web at: `http://www.ibm.com/services/tsm/`
- *Implementing ESS Copy Services on S/390*, SG24-5680
- *IMS/ESA Sysplex Data Sharing: An Implementation Case Study*, SG24-4831
- *IMS/ESA Data Sharing in a Parallel Sysplex*, SG24-4303
- *IMS/ESA Multiple Systems Coupling in a Parallel Sysplex*, SG24-4750
- *IMS/ESA V6 Administration Guide: System*, SC26-8730
- *IMS/ESA V6 Release Planning Guide*, GC26-8744
- *IMS/ESA V7 Release Planning Guide*, GC26-9437
- *OS/390 Integration Test Site*, `http://www.s390.ibm.com/os390/support/os390tst`
- *OS/390 MVS Parallel Sysplex Hardware and Software Migration*, GC28-1862
- *OS/390 Parallel Sysplex Test Report*, GC28-1963
- *OS/390 Setting Up a Sysplex*, GC28-1779
- *Parallel Sysplex Availability Checklist* available on the Web at: `http://www.s390.ibm.com/ftp/marketing/position/availchk_parsys.pdf`
- *Parallel Sysplex Configuration Planning for Availability*, WSC Flash 9829
- *Parallel Sysplex - Managing Software for Availability*, SG24-5451
- *Parallel Sysplex Recovery*, GA22-7286
- *RAMAC Virtual Array: Implementing Peer-to-Peer Remote Copy*, SG24-5338
- *S/390 MVS Parallel Sysplex Continuous Availability Presentation Guide*, SG24-4502
- *S/390 MVS Parallel Sysplex Continuous Availability SE Guide*, SG24-4503
- *SFM Functions and the Impact of OW30814 for Parallel Sysplex*, WSC Flash 9825
- *SNA in a Parallel Sysplex Environment*, SG24-2113
- *TCP/IP in a Sysplex*, SG24-5235
- *XCF Service Recommendations to Maximize Sysplex Availability*, WSC Flash 9838

The following roadmap guides you through the chapter, providing a quick reference to help you find the appropriate section.

Table 9. Continuous Availability in Parallel Sysplex Roadmap

You want to configure:	**If you are especially interested in:**	**Then refer to:**
Continuous Availability in Parallel Sysplex		
	Is availability important? • How can Parallel Sysplex help me minimize the effect of planned and unplanned outages? • How do I limit the scope of an outage?	3.1, "Why Availability Is Important" on page 91
	What availability considerations are related to SW? • What is the concept of "cloning"? • What SW resources are key? • What about JES3? • What about subsystems?	3.2, "Software Considerations for Availability" on page 97
	How can networking functions contribute to continuous availability in Parallel Sysplex?	3.3, "Network Considerations for Availability" on page 104
	What availability considerations are related to hardware? • How is hardware redundancy helping availability? • What are some of the hardware elements that are redundant in Parallel Sysplex? • How does redundancy permit CPC and CF upgrades?	3.4, "Hardware Considerations for Availability" on page 111
	Are there limitations to achieving continuous availability in Parallel Sysplex?	3.5, "Limitations to Continuous Availability" on page 116
	Recovery in Parallel Sysplex: What can SFM and ARM do for me?	3.6, "Recovery Considerations for Availability" on page 117
	Disaster recovery and Parallel Sysplex: • Is a geographically dispersed Parallel Sysplex viable? • What remote copy techniques are available? • CICS disaster recovery in Parallel Sysplex. • DB2 disaster recovery in Parallel Sysplex. • IMS disaster recovery in Parallel Sysplex. • Where can I find more information?	3.7, "Disaster Recovery Considerations in Parallel Sysplex" on page 126

3.1 Why Availability Is Important

There are real costs associated with system outages. In business terms, an outage costs real dollars and can seriously impact the ability to respond to changing business needs.

A *planned outage* is inconvenient at best. Some companies simply cannot tolerate any unplanned down time, and so will schedule a planned outage for a time (for example, 3 a.m. Sunday morning) when the fewest users are affected. This means that necessary system work is postponed until a time when the applications are not available. It also means that system programmers must work inconvenient hours, and help may be unavailable if something goes wrong.

An *unplanned outage* can, of course, happen during the busiest time of the day and therefore have a serious impact. Also, because of the panic nature of such an outage, the potential exists for additional harm to be caused by errors that are introduced as people rush to get the system restored.

Availability Definitions

Levels of availability are defined as follows:

High availability: A system that delivers an acceptable or agreed level of service during scheduled periods

Continuous operation: A system that operates 7 days a week, 24 hours a day, with no scheduled outages

Continuous availability: A system that delivers an acceptable or agreed level of service 7 days a week, 24 hours a day

Continuous availability is the *combination* of high availability and continuous operation.

Source: *S/390 MVS Parallel Sysplex Continuous Availability Presentation Guide*, SG24-4502.

For many reasons, a continuously available system is becoming a requirement today. In the past, you most likely concentrated on making each part of a computer system as fail-safe as possible and as fault-tolerant as the technology would allow. Now it is recognized that it is impossible to achieve true continuous availability without the careful *management of redundancy.*

3.1.1 Parallel Sysplex Is Designed to Allow Management of Redundancy

One of the advantages of Parallel Sysplex is that it allows a configuration of multiple redundant parts, each of which is doing real work. In the past, we had the concept of one system acting as backup for another, doing low priority work or essentially very little work unless the primary system failed. This is costly and difficult to manage.

With Parallel Sysplex, many systems are actively doing work and are peers to each other. The implementation of this capability requires multiple access paths through the network, multiple application processing regions and multiple database managers sharing common data, with mechanisms for workload balancing.

In case of a failure, or a need to do preventive maintenance on one of the systems, the remaining systems can assume the extra work without interruption, though users connected through the network to a failed or removed system (for example, a system containing a CICS TOR) will need to reconnect. No intervention is required by the remaining systems.

Several levels of redundancy are provided in Parallel Sysplex. In addition to "system redundancy," there can also be redundancy at, for example, the subsystem level, such as having several AORs per OS/390 image.

The old adage says that you should not put all your eggs in one basket. With one very large single system, that is just what you are doing. This is binary availability, so either all of it runs, or none of it runs.

If, instead, the processing power is broken up into a number of smaller pieces, much of the total capacity remains if one of these pieces break. For this reason, many installations today run multiple stand-alone systems. The workload is partitioned so that a single system outage will not affect the entire installation, although one of the workloads will not run. To manage installations in this manner, the systems programmer must manage several systems, Sysres packs, master catalogs, and parmlib members, all of which are different.

Parallel Sysplex allows the system programmer to manage several copies of one *single system image*. Each of the systems in Parallel Sysplex can be a clone of the others, sharing master catalogs, Sysres packs and parmlib members. In addition, because of the fact that each individual system can access the data equally, if one system is lost, the work is shifted to another system in the complex and continues to run.

Parallel Sysplex- the Instant Solution?

You should not be under the illusion that implementing Parallel Sysplex alone will immediately produce continuous availability. The configuration of Parallel Sysplex has to conform to the same requirements for addressing continuous availability as in a non-sysplex environment. For example, you must have:

- Adequate redundancy of critical hardware and software.
- No single points of failure.
- Well-defined procedures/practices for:
  - Change management.
  - Problem management.
  - Operations management.

Thus the design concept of Parallel Sysplex provides a platform that enables continuous availability to be achieved in a way that is not possible in a non-sysplex environment, but it cannot be achieved without consideration of the basic rules of configuring for continuous availability.

Also, the full exploitation of data sharing and workload balancing available through the Parallel Sysplex, and the level of software that will support these functions, will ultimately help provide continuous availability.

When continuous availability is discussed, it is often centered on hardware reliability and redundancy. Much emphasis is put on ensuring that single points of failure are eliminated in channel configurations, DASD configurations, environmental support, and so on. These are important, but far short of the whole picture.

A survey was done to determine what activities were being carried out in installations during *planned* outages. The numbers quoted are not necessarily definitive, but do serve to highlight a point. The percentages show the amount of time spent on each activity:

- Database backups 52%
- Software maintenance 13%
- Network 10%
- Application 8%
- Hardware maintenance 8%
- Other 9% (for example, environmental maintenance)

The key message from this survey is that during the time allocated for planned outages, nearly 75% of the activity is directed at *software-related issues*. Therefore, while the hardware element is important when planning for continuous availability, equal if not more consideration must be given to the software element.

Recommendation on Availability Analysis

We recommend that an analysis of *all* outages that occurred over the previous 12-24 months be carried out and compared with an estimate of what would have happened in Parallel Sysplex. In many cases, the analysis may need to be qualified by statements such as: "for systems which have implemented data sharing," or "provided transaction affinities have been resolved." In addition, we recommend that a formal availability review be done for hardware, software, and the environment, for both failures and changes. Parallel Sysplex with data sharing and load balancing is a *prerequisite* for continuously available OS/390 systems, but so is a *well-designed* environment.

A white paper that provides a list of things to check for, to help you obtain the highest possible application availability is available on the Web at: `http://www.s390.ibm.com/ftp/marketing/position/availchk_parsys.pdf`

3.1.2 Planned Outages

Planned outages constitute by far the largest amount of system down time. According to recent IBM studies, roughly 90% of all outages are planned outages. For an installation achieving 99.5% planned availability, this would equate to roughly 400 hours of planned down time a year. Based on the industry figures quoted in 2.2, "Deciding If Parallel Sysplex Is Right for You" on page 18, this equates to a cost of between $10,000,000 (US) and $2,600,000,000 (US) per annum per installation. Even allowing that planned outages will be during off-peak hours when the impact will be minimized, it can be seen that the real cost is substantial.

Restarting an image to add new software or maintenance, to make configuration changes, and so on, is costly. Much progress has already been made in this area to avoid IPLs. For example, OS/390 has many parameters that can be changed dynamically, I/O configurations can be changed dynamically, and, by using ESCON, new peripheral devices can be added dynamically.

Dynamic I/O reconfiguration allows a channel path definition to be changed from converter (CVC) to ESCON (CNC). However, changing a channel from parallel to ESCON, even on a 9672, will require a POR. Note that dynamic I/O reconfiguration requires specification of HSA expansion factors, which in turn require a POR to be set initially.

In Parallel Sysplex, HCD supports cross-system activation of the new I/O configuration. This is discussed in detail in the IBM Redbook *HCD and Dynamic I/O Reconfiguration Primer*, SG24-4037.

Many parts of the hardware allow concurrent repair or maintenance, resulting in fewer outages due to this work.

Capacity Upgrade on Demand (CUoD) on G5 and later processors allows processor upgrades to add processors for OS/390 or ICF upgrades concurrent with normal operation by utilizing a currently spare PU. In basic mode, with current maintenance applied, the new CP can be brought online without a re-IPL of OS/390. In LPAR mode, the CP can be added to the shared pool of CPs, or it can be used as a dedicated CP by an LP that uses dedicated CPs—in either case, no POR or image reset is required. More information about CUoD can be obtained on the Web site:
`http://www.s390.ibm.com/pes/apg`

The Capacity Backup Option, CBU, is a disaster recovery option which exploits CUoD to non-disruptively bring additional processor capacity online. A typical use would be in a remote site where a 9672-Z17 could be upgraded via CBU to a 9672-ZZ7 if the primary site is lost in a disaster. Special Terms and Conditions control the use of CBU. The *CBU Users Guide* contains more information about the use of CBU.

Concurrent I/O conditioning pre-installs support cards (such as FIBBs and channel driver cards) so that channel cards (such as ESCON, FICON and ICBs) can later be added concurrently. Because FICON and ICBs can only occupy certain I/O slots, installation of this feature may involve moving existing channel cards and CHPID number changes.

Some upgrades are still non-concurrent, for example storage upgrades.

Parallel Sysplex can make planned outages less disruptive.

A system can be removed from the sysplex and work will continue to be routed to the remaining systems. From the end-user perspective, the application continues to be available, (although end users may have to log on again to reestablish the session), even if one of the systems is down.

While a system is removed from the sysplex, maintenance or new software levels can be applied. When the system is reintroduced to the sysplex, it can coexist at the new software level with the other systems. New function might not be available until the software is upgraded on all systems, but existing functions are not affected.

In order for this strategy to work, new levels of software must be written to allow compatibility with previous levels. No software maintenance should require a *sysplex-wide* IPL. See 2.3.1.2, "System Software Level Considerations" on page 28 for more information on the IBM strategy for this coexistence.

3.1.3 Unplanned Outages

For many years IBM has continued to improve the fault-tolerance and non-disruptive upgradability of the hardware.

The processor module on 9672 processors contains identical Processing Units (PUs), which can function as OS/390 CPs, System Assist Processors (SAPs), or ICF CPs, depending on microcode load. Not all processors have spare PUs because they may all be in operation in normal use—for example, a 9672-ZZ7 has none.

Sparing involves the utilization of a spare PU in the processor module to replace a failing PU (which could be an OS/390 CP, a SAP or an ICF CP).

With improvements in sparing on G5 and later CPCs, if a spare PU is available, the failure of a SAP, ICF, or OS/390 CP does not cause an outage in most cases.

For an overview of these and other CPC hardware features related to availability, refer to the RAS section of Table 36 on page 233 in Volume 2: Cookbook, SG24-5638.

Parallel Sysplex has superior availability characteristics in *unplanned*, as well as in planned, outages, because work continues to be routed to existing systems when one system is removed. In addition, when a system fails or loses connectivity to the rest of the sysplex, the remaining systems can immediately take steps to make sure that the system in question cannot contaminate any existing work; they do this by cutting the failed system off from the running sysplex. The sysplex components, working together, also begin to recover the failing system and any transactions that were in progress when the failure occurred.

The installation-established recovery policies determine what happens when failures occur, including how much is done automatically and how much requires operator intervention.

3.1.4 Scope of an Outage

An additional consideration for planning for availability is to limit the scope of the effects outages will have. For example, two smaller, physically separate CPCs have better availability characteristics than one CPC twice as powerful, because a catastrophic hardware failure causes only half the available capacity to be lost. In a CICS environment, splitting an application region (AOR) into two AORs in Parallel Sysplex allows work to flow through one region if another has failed. Putting the two AORs on separate physical hardware devices gives even more availability. Figure 23 on page 96 shows how separating elements can help limit the scope of failure.

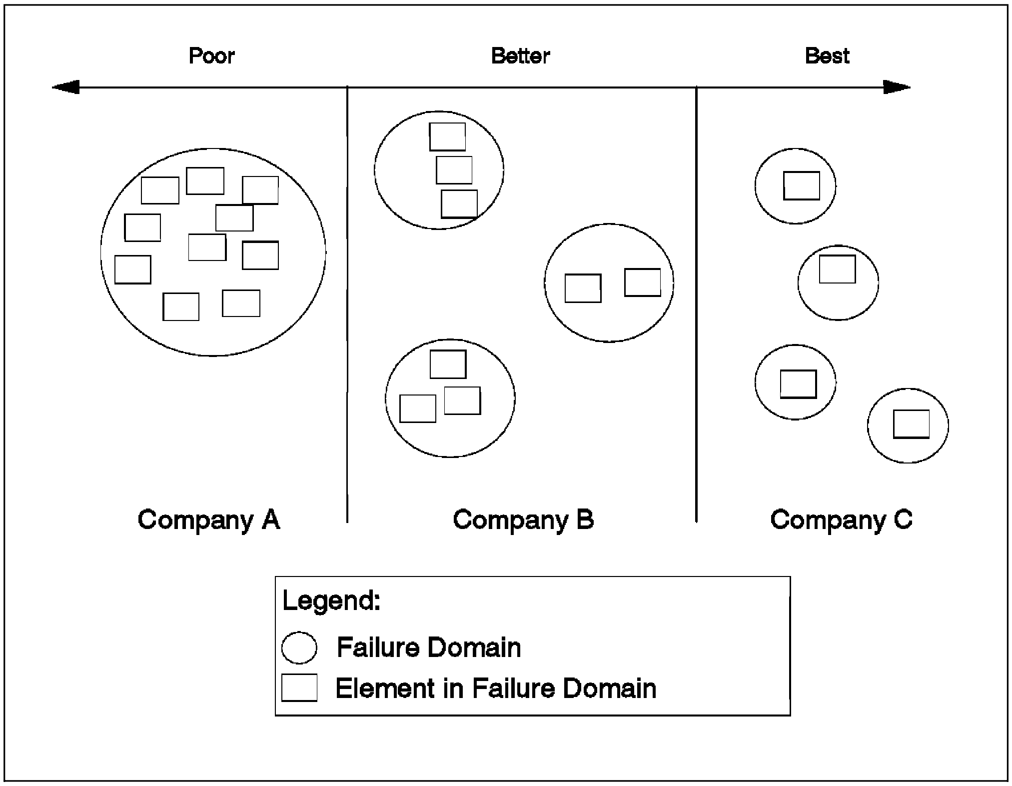


Figure 23. Failure Domain Example

The more separate pieces, or domains, there are, the less total capacity there is to lose when a domain fails. CICS is an example of how this concept applies. Other examples to create "small" failure domains include:

- Separate power connections to hardware where possible.
- Physically separate channels and CF links that attach the same pieces of hardware. Care should also be taken to spread the channels and CF links over as many 967x cards as possible. The guidelines in the 9672 Systems Assurance Product Review (SAPR) Guide will help you achieve this.
- Place the CFs and the OS/390 LPs they are coupled to in separate CPCs.
- Define backup NCPs to allow minimal or fast recovery.
- Put backup or spare copies of critical data sets on volumes belonging to different DASD subsystems. An example of this is primary and alternate couple data sets.
- Use disk mirroring or dual copy for all important data sets such as shared Sysres, all catalogs, other vital system data sets, and, obviously, all critical user data.
- Have a mixture of CF structures and CTCs for XCF signaling.
- Use XCF for inter-VTAM and inter-TCP/IP communication.
- Spread CF structures across multiple CFs according to the current guidelines, and do rebuild on separate CFs.
- Choose Sysplex Timer redundancy features.
- Distribute ESCON channels, CTCs, and control units across several ESCON directors.

The following sections briefly touch on some of these examples. For a more complete discussion, refer to the IBM Redbooks *Continuous Availability S/390 Technology Guide*, SG24-2086, and *S/390 MVS Parallel Sysplex Continuous Availability Presentation Guide*, SG24-4502.

Various methods exist to assist in determining failure domains or single points of failure. One such method is Component Failure Impact Analysis (CFIA). A CFIA study can be performed by your IBM Services specialist.

3.2 Software Considerations for Availability

When managing multiple systems, it is much easier to manage multiple copies of a single image than to manage many different images. For that reason, it is sensible to create OS/390 systems that are clones of each other. They can share the master catalog, Sysres volumes, and Parmlibs and Proclibs. There is additional information in Appendix A, “Systems Management Products for Parallel Sysplex” on page 123 in Volume 3: Connectivity, SG24-5639, about cloned OS/390 systems.

However, there are decisions to be made about striking a balance between the benefits of sharing and the risks of creating single points of failure. This is discussed in more detail in the IBM Redbook *Parallel Sysplex - Managing Software for Availability*, SG24-5451.

3.2.1 Data Set Redundancy

Some data sets need to be redundant in Parallel Sysplex in order to achieve continuous availability.

3.2.1.1 System Data Sets

In a non-sysplex environment, certain system data sets are key to high availability. These data sets become even more crucial in Parallel Sysplex because they are now shared by all the systems in the sysplex and as such become potential single points of failure. Consideration should be given to the use of disk mirroring techniques and/or RAID 5 technology to ensure a live backup at all times. Examples of such resources are:

- **Master Catalog**

  Refer to A.5.5, “Shared Master Catalog” on page 157 in Volume 3: Connectivity, SG24-5639 to get more information on the redundancy required for the master catalog.

- **Couple Data Sets**

  Refer to A.5.7, “Considerations for Couple Data Sets” on page 160 in Volume 3: Connectivity, SG24-5639 to get more information on the redundancy required for the couple data sets. Place the alternate couple data set on a separate control unit from the primary. For the couple data sets, rely on XCF to maintain two copies of the data set (primary and alternate) rather than using a hardware solution such as dual copy or mirroring.

- **Sysres**

  Refer to A.5.6, “Shared System Residence (SYSRES) Volume” on page 158 in Volume 3: Connectivity, SG24-5639 to get more information on the redundancy required for the Sysres volumes.

3.2.1.2 Subsystem Data Sets

In a non-sysplex environment, certain *subsystem* data sets are key to high availability. These data sets become even more crucial in Parallel Sysplex because they are now shared by several instances of the subsystems. Some examples of such data sets are:

- DB2 catalog
- DB2 directory
- IMS RECON data set

Using disk mirroring or a RAID-capable DASD controller for these data sets is highly recommended.

Chapter 4, "Workloads in Parallel Sysplex" on page 141 provides further detail on subsystems within a Parallel Sysplex.

3.2.2 Couple Data Set Considerations

When formatting the sysplex couple data set, specifying the GROUP and MEMBER keywords in the format utility determines the number of XCF groups and the number of members per group that can be supported in the sysplex. These values should not be over-specified, as this can lead to elongated recovery or IPL times.

These values cannot be *decreased* non-disruptively using an alternate couple data set and the SETXCF COUPLE,PSWITCH command. A sysplex-wide IPL is required. However, this technique can be used to *increase* the values for GROUP and MEMBER non-disruptively.

For further couple data set considerations see A.5.7, "Considerations for Couple Data Sets" on page 160 in Volume 3: Connectivity, SG24-5639, and for general details on couple data sets see *OS/390 MVS Setting Up a Sysplex*, GC28-1779.

3.2.3 JES3 and Continuous Availability

There are some specific considerations for JES3 in a Parallel Sysplex.

3.2.3.1 JES3 Continuous Availability Considerations

There are a number of areas that require consideration for high availability:

- Some initialization deck changes require a warm start.
- Up until OS/390 V2R9, JES3-managed devices were ineligible for dynamic I/O reconfiguration. This is particularly important for DASD devices, for which we recommend SMS management. Operationally, this lead to some new situations. For example, operators will not be able to use commands such as *D,D=devno, or *I S,V=volser.

  JES3 provides support for tape sharing. This could be replaced by the tape sharing support in OS/390. However, in general, DASD is more likely to be installed or de-installed than tape, so lack of dynamic reconfiguration is less important in this case. The operational impact of moving to OS/390 tape management would be greater. For example, jobs requiring tape drives could go into execution with no available drives and then go into allocation recovery. We expect some JES3 users will retain JES3 management of tape.

  OS/390 V2R9 introduced JES3 support for dynamic I/O reconfiguration. Prior to this release, changes to the JES3-managed devices required a sysplex-wide IPL in order to rebuild the information it keeps in storage for all JES3-managed devices.

Further discussion of JES3 in Parallel Sysplex can be found in 4.4.2, "JES3 Considerations in Parallel Sysplex" on page 214, and from a continuous availability point of view in the IBM Redbook *S/390 MVS Parallel Sysplex Continuous Availability SE Guide*, SG24-4503.

OS/390 V2R4 Enhancements: OS/390 V2R4 provides some enhancements to JES3 to reduce outages needed to make certain configuration changes. These include:

- Dynamic update support for adding or modifying certain SNA RJP and VTAM FSS initialization statements.
- A start type called HOTSTART with REFRESH, which allows for the reading of the initialization stream during a hotstart without an IPL.
- JES3 restart processing is enhanced to reduce the time it takes the JES3 global to reinitialize.

OS/390 V2R6 Enhancements: Further enhancements were made to JES3 in OS/390 V2R6 to improve its availability characteristics and usability:

- LPA modules can be changed without an IPL and CLPA being performed. A new level of JES3 LPA module is activated when a JES3 hot or local start (without IPL) is performed. JES3 automatically picks up the new levels of the module during JES3 initialization.
- JES3 now supports WLM scheduling environments. This support improves your ability to ensure that jobs get executed on whichever system has the particular resources required by that job. If a system becomes unavailable, the unique resources that were previously only available on that system can be enabled on another system, and JES3 will ensure that jobs requiring those resources will be routed to that system.

OS/390 V2R8 Enhancements: OS/390 V2R8 added JES3 support for the WLM Managed Initiator function that was added to JES2 in OS/390 V2R4. The use of WLM-managed initiators means that the amount of batch work processed by a system can be increased or decreased depending on its relative importance compared to other work running on that system.

OS/390 V2R9 Enhancements: OS/390 V2R9 added JES3 support for Dynamic I/O Reconfiguration for JES3-managed devices. As a result, it is no longer necessary to do a sysplex-wide IPL, or *any* IPL to make changes to a JES3-managed device. This support introduces a new type of JES3 start—the hot start with refresh. More information about this support is available in *OS/390 JES3 Migration*, GC28-1799.

3.2.3.2 XES External Hang Detect

Sometimes, problems can arise in a sysplex where one application in the sysplex issues a request to another member, but never gets a reply. An example might be a subsystem that is using a CF structure and wishes to rebuild the structure to another CF. The subsystem initiating the rebuild will contact all its peers on the other OS/390 images in the sysplex, requesting them to take part in the rebuild. If one of those peers does not reply, the rebuild process will hang. Unfortunately, it is not easy to identify which process in the sysplex is the one that is causing the hang.

To address this situation, a new function called *XES External Hang Detect* was introduced in OS/390 V2R8. With this new support, XES on the system to which an event is delivered for processing by an instance of the structure exploiter will start a timer every time it receives such an event. If the task does not respond within two minutes, XES will issue an IXL040E/IXL041E message; identify the system, jobname, and ASID of the task that is not responding. While this does not alleviate the hang situation, it does identify the causer of the hang, enabling the operators or automation to respond appropriately, thus freeing up the hang.

3.2.3.3 System Managed Rebuild

In the early days of Parallel Sysplex, MVS provided basic services to users of the CF. These services provided the ability to carry out any necessary tasks, however it was the responsibility of the exploiter to make most of the decisions and drive most of the processing. Conceptually, this was like DASD storage management prior to System Managed Storage—you could do just about anything you wanted, but you needed an amount of knowledge to get the end result that you desired.

As a result, there was inconsistent use of the CF. Some products supported structure rebuild and others did not. Some supported seamless recovery from a CF failure, and others did not. And some products used MVS interfaces (the SETXCF command) to manage their structures, while others (like JES2) provided an interface of their own for such activities.

As the start of a move to make it easier for additional products and functions to use the CF, and for installations to manage the CF, OS/390 V2R8 introduced a new capability known as System Managed Rebuild. When this capability is used, MVS will take responsibility for the process of rebuilding a structure, providing a much simpler interface to the exploiter. Eventually, this will lead to a more consistent use of the CF functions, and a more consistent user interface across the various CF exploiters.

This function requires specific support in the products that use CF structures. As a result, it will take some time before it is used by all the current CF exploiters. In OS/390 V2R8, JES2 announced support for System Managed Rebuild. Since then, WLM and MQSeries have also announced that they will use System Managed Rebuild for their CF structures.

There are some things that must be considered, however, before you start using System Managed Rebuild:

1. Any CF that is likely to contain a structure that will be managed by System Managed Rebuild *must* be at CFLEVEL=8 or higher.
2. The CFRM couple data set must be formatted with a new keyword. ITEM NAME(SMREBLD) NUMBER(1) indicates that System Managed Rebuild may be used in this Parallel Sysplex.
3. Any system that is going to use a CFRM CDS formatted for System Managed Rebuild support *must* be running OS/390 V2R8 or later.
4. Once a CFRM CDS containing these new keywords is activated, it is not possible to nondisruptively fall back to a CDS that does not support System Managed Rebuild. If you have to fall back to such a CDS, a sysplex-wide IPL is required.
5. System Managed Rebuild is designed to support planned configuration changes. However, it does not support rebuild in case of an error situation,

such as a connectivity or CF failure. For either of these situations, other procedures must be used to rebuild the structures.

Two subsequent new functions, furthering the goal of making CF structures easier to use and manage, are *Structure Full Monitoring*, announced in OS/390 V2R9, and *Auto Alter*, announced in OS/390 V2R10. Structure Full Monitoring is described in more detail in 2.2.1.1, "Structure Full Monitoring" on page 75 in Volume 2: Cookbook, SG24-5638 and Auto Alter is described in 2.2.1.2, "Auto Alter Support" on page 76 in Volume 2: Cookbook, SG24-5638. Both functions are also described in the OS/390 V2R10 or later level of *Setting Up A Sysplex*, GC28-1779.

3.2.4 Subsystem Considerations

This section provides pointers to information about subsystems and availability issues.

3.2.4.1 CICS

The IBM Redbook *Planning for CICS Continuous Availability in an MVS/ESA Environment*, SG24-4593 has a full discussion of the issues relating to CICS availability. See the "Unavailability Cause and Solution Checklist" in Chapter 1 of that publication for an analysis of causes of unavailability and strategies for circumventing them.

3.2.4.2 DB2

DB2 UDB for OS/390 V6 builds on the high availability features of previous versions of DB2. Most important, is the ability to duplex the DB2 Group Buffer Pools (GBPs). This effectively removes the GBPs as a single point of failure, and dramatically reduces recovery time following a CF, CF link, or structure failure.

GBP duplexing can be more of an asset if used with an Integrated Coupling Facility (ICF). By placing the primary copy of the GBP in the ICF, you can enjoy the performance aspects of an ICF and still rely upon the availability characteristics of a duplexed structure. This new facility has also been retrofitted to DB2 for OS/390 V5 with APAR PQ17797. (You also require CFLEVEL=5 or later and OS/390 APAR OW28460.) For high availability configurations, it is not recommended to place the DB2 SCA or Lock structures in an ICF that is in the same failure domain as any of the members of the DB2 data sharing group.

Another enhancement to DB2 is a new option on the -STO DB2 command. The STO DB2,CASTOUT(NO) command tells DB2 not to do castout processing for pagesets for which this member is the last updater. The result is that the shutdown is faster, avoiding the time that would otherwise be required to write all those changed pages out to DASD.

When shutting down all members of a data sharing group, if CASTOUT(NO) is specified for one or more members, then there is likely to be some changed data that still resides in the group buffer pools after all members are shut down. This means that the database data on DASD probably will not have all of the most recent data. So if you are shutting down all members to get a consistent copy of the databases on DASD that can be copied and sent offsite, then CASTOUT(NO) should not be used.

Given the fact that this option can result in pages that have not yet been written to DASD remaining in the CF, it really should only be used if DB2 is being stopped just for a short time to allow the application of maintenance. For longer duration stops, or stops that involve components other than just DB2, a normal stop should be used.

Several other enhancements have also been introduced. Improved partition rebalancing lets you redistribute partitioned data with minimal impact to data availability. One REORG of a range of partitions both reorganizes and rebalances the partitions. You can alter VARCHAR column lengths easily while your data remains available.

The DB2 UDB for OS/390 V6 utilities have been enhanced to provide better performance and improved data availability:

- Faster backup and recovery with COPY and RECOVER which can process a list of image copies in parallel; and RECOVER can also rebuild indexes and table spaces at the same time as using image copies.
- Parallel index build reduces the elapsed time of LOAD and REORG jobs of table spaces, or partitions of table spaces, that have more than one index; the elapsed time of REBUILD INDEX jobs is also reduced.
- Inline statistics embeds statistics collection into utility jobs, making table spaces available sooner.
- Discard during REORG lets you select rows to discard, and you can perform a faster unload of data in external format.
- You can determine when to run REORG by specifying threshold limits for relevant statistics from the DB2 catalog.

A number of the DB2 V6.1 enhancements, in addition to duplex GBPs, have been made available for DB2 V5.1. These build on the GBP structure rebuild support, automatic recovery of GBP data, and concurrent database reorganization features that were already included in DB2 V5.1. For more information on these enhancements see 2.8, “DB2 Structures” on page 106 in Volume 2: Cookbook, SG24-5638 and 4.3.1, “DB2 Data Sharing Considerations” on page 194 of this book, and the DB2 manual, *DB2 UDB for OS/390 V6 Release Guide*, SC26-9013. For information on the enhancements made to DB2 V5.1, refer to the IBM Redbook *DB2 Server for OS/390 Version 5 Recent Enhancements - Reference Guide*, SG24-5421.

In another enhancement designed to further improve DB2 availability, DB2 UDB for OS/390 V7 introduced a new startup option, to speed up recovery in a data sharing environment. “Restart Light” allows you to do a cross-system restart of a DB2 member using a smaller DB2 storage footprint and optimized for retained lock freeing (DB2 terminates after releasing as many retained locks as it can). This improves availability in that the resources protected by retained locks can get freed quicker.

3.2.4.3 IMS

IMS/ESA V7, the most recent version of IMS at the time of writing, introduces further exploitation of the facilities of Parallel Sysplex to improve IMS availability and restart. The Rapid Network Recovery feature uses the facilities of VTAM MNPS to dramatically reduce the elapsed time required to reconnect all user sessions following a failure of the IMS subsystem. There are other enhancements in IMS V7, and more information will be available closer to the general availability date which, at the time of writing, has not been announced.

IMS/ESA V6 introduced a number of enhancements that significantly improve the potential availability of IMS TM and DB in a data sharing environment. Among them are:

- Shared message queues
- Shared VSO DEDBs
- Shared Fast Path DEDBs with SDEPs
- Daylight saving time support
- VTAM generic resource support
- Fast database recovery
- Fast Path DEDB online change

VTAM GR support allows you to access multiple IMS subsystems using a single GR name, offering a single-system image while using the resources of many IMS subsystems. In general, if one IMS subsystem fails, you can log onto another IMS subsystem in that GR group. This is discussed in more detail in the IBM Redbook *Using VTAM Generic Resources with IMS*, SG24-5487, and in the -11 level of the *OS/390 Parallel Sysplex Test Report*.

The Fast Database Recovery (FDBR) feature, introduced in IMS/ESA V6, provides quick access to shared database resources in a sysplex environment by releasing records that might otherwise be locked by a failed IMS until the failed system is restarted. This can significantly reduce the impact of an IMS failure in a data sharing environment, especially if you are using shared message queues.

These features are discussed in 4.2.5, “IMS Transaction Manager in Parallel Sysplex” on page 187. For further information on all of the enhancements in IMS/ESA V6, refer to *IMS/ESA V6 Release Planning Guide*, GC26-8744, and the IBM Redbook *IMS/ESA V6 Parallel Sysplex Migration Planning Guide for IMS TM and DBCTL*, SG24-5461.

It should be noted that IMS/ESA V5.1 does not support data sharing for the following types of databases:

- MSDBs
- DEBDs with SDEPs
- DEBDs using virtual storage option (VSO)

Customers that wish to share these types of databases are recommended to upgrade to IMS/ESA V6 or later.

Refer to *IMS/ESA V5 Administration Guide: Database Manager*, SC26-8012, *IMS/ESA V5 Administration Guide: System*, SC26-8013, and *IMS/ESA V5 Administration Guide: Transaction Manager*, SC26-8014 for information on IMS/ESA V5 in a data sharing environment.

IMS/ESA V5 uses local times for log timestamps, which means that twice a year it has to be shut down. When the clocks are moved back, then all IMS systems have to be shut down for at least one hour to avoid duplicate time stamps. When the clocks are moved forward, then all IMS systems have to be shut down and restarted.

IMS/ESA V6.1 has addressed this problem by using Universal Time Coordinates (UTC) for its internal timestamps. However, this is only a solution for the IMS code, and it cannot address problems with time changes in applications. The IBM Redbook *S/390 MVS Parallel Sysplex Continuous Availability SE Guide*, SG24-4503, contains a discussion on time changes in a Parallel Sysplex and procedures required for IMS.

3.2.4.4 Subsystem Software Management

For each of the subsystems—CICS, DB2, and IMS—you need to have a strategy that will allow you to apply maintenance or install a new release without impacting application availability. You must be able to introduce new levels of code onto one system at a time. You also need to be able to keep the remaining subsystems up and running while you stop one to make a change to it. And you need to have the ability to quickly and easily back out any changes, to allow a subsystem fall back to a prior software level as quickly as possible. All these requirements, and a suggested way of addressing them, are addressed in the IBM Redbook *Parallel Sysplex - Managing Software for Availability*, SG24-5451.

3.3 Network Considerations for Availability

The Communications Server for OS/390 provides several facilities that can increase the availability of both SNA and TCP/IP applications in a Parallel Sysplex. An enhancement to the APPN architecture called High Performance Routing (HPR) can increase network availability for SNA applications. TCP/IP has a number of functions that can enhance availability of TCP applications in a Parallel Sysplex. Each of these functions is discussed in this section.

3.3.1 VTAM Generic Resources Function

The ability to view the Parallel Sysplex as one unit of many interchangeable parts is useful from a network perspective. For workload balancing, it is nice to be able to have users simply log on to a generic name, and then route that logon to a system with available capacity, keeping the routing transparent to the user. The user therefore views the system as one unit, while the workload balancing under the covers knows that it really consists of multiple parts. This capability was introduced by VTAM V4.2. The function is called *generic resources*. It allows the installation to specify a single generic name to represent a collection of VTAM applications in the Parallel Sysplex. You may define several generic names to represent different workload collections.

The generic resources function also provides a level of availability in the Parallel Sysplex. If a system fails while a user is logged on, he must log on again, but simply to the same generic name as before. The new session is then established with another subsystem in the Parallel Sysplex. Therefore, the end user no longer has to know the name of a backup system and can get back onto the application faster.

For more information about the VTAM generic resources function and how it works, refer to 4.5.1, “VTAM Generic Resources Function” on page 219. For more information about subsystems exploiting the VTAM generic resources function, refer to 4.5.1.3, “Generic Resource Subsystem Support” on page 224.

3.3.2 Persistent Sessions

An availability enhancement introduced in VTAM V3.4.1 and exploited by several subsystems is *persistent session* support. This function provides the ability for sessions to be maintained when a subsystem fails. The subsystem can be restarted on the same VTAM, and the end-user sessions are still active when the new subsystem becomes available. VTAM keeps the session state information that is required to support this function in a private data space.

The subsystems that support persistent sessions specify how long they are willing to wait from the time of the failure to the re-opening of the access method control blocks (ACBs). This tells VTAM how long the session state information is to be kept in VTAM storage. This timeout value is specified in the PSTIMER value on the SETLOGON macro, which the subsystems use after opening their ACBs.

The IBM subsystems that support persistent sessions are:

- CICS V4.1 and higher
- APPC/MVS for MVS/ESA SP V4.2.2 and higher

3.3.2.1 CICS Persistent Session Support

Persistent session support improves the availability of CICS in the following manner. If CICS fails, the user perception is that CICS is "hanging." What is on the screen at the time of the failure remains until persistent session recovery is complete. After a successful CICS emergency restart, the recovery options defined for the terminals or sessions take effect. Depending on the CICS setup, all you may have to do is clear the screen, sign on again, and continue to enter CICS transactions.

For more information on CICS exploitation of persistent session support, refer to *CICS TS for OS/390 Release Guide*, GC33-1570.

3.3.2.2 APPC/MVS Persistent Session Support

A number of applications make use of the services of APPC/MVS for certain functions. These applications may or may not support end users. Examples include:

- IMS V4.1 and higher
- OPC/ESA
- RACF V2.2
- MERVA/ESA
- Resource Recovery Services (RRS) for OS/390 R3 (distributed synch point services).

In the future, it is expected that more applications will use APPC/MVS.

3.3.3 Multi-Node Persistent Sessions

In a Parallel Sysplex environment, the availability benefits of persistent sessions are enhanced by allowing the failed subsystem to be restarted on a different VTAM system. This allows recovery from hardware failures or failures in system software such as OS/390 or VTAM. This capability is called *Multi-Node Persistent Session* support.

Multi-Node Persistent Session support extends the persistent session support provided in VTAM V3.4.1 for application programs. Multi-Node Persistent Sessions enable an application to re-establish its sessions after a VTAM, hardware, or operating system failure. Sessions are preserved across application outages because session information is maintained in the CF.

3.3.3.1 High Performance Routing

The Multi-Node Persistent Session function uses *High Performance Routing* (HPR), which is an extension to the APPN architecture. HPR connections are established to carry either some or all of the session from the application in the Parallel Sysplex to the remote end user. For more information on how HPR connections operate, refer to 2.1.3, "High Performance Routing (HPR)" on page 29 in Volume 3: Connectivity, SG24-5639.

If there is a failure of a system in the Parallel Sysplex, the endpoint of the HPR connection is moved to a different system. HPR support introduced in VTAM V4.3 can only provide support for routing around failures on intermediate nodes or links (as opposed to failures on the HPR endpoints). The ability to reroute the endpoint of an HPR connection in a Parallel Sysplex is a special case, and this function is provided in VTAM V4.4 as the Multi-Node Persistent Session support.

It is possible to support Multi-Node Persistent Sessions even if HPR is used only inside the Parallel Sysplex. In this case, an HPR connection is established between the end node where the application is running and one of the network nodes in the Parallel Sysplex. For more information on APPN node types, refer to 2.1.1, "APPN Overview" on page 24 in Volume 3: Connectivity, SG24-5639.

Figure 5 on page 30 in Volume 3: Connectivity, SG24-5639, shows the effects of placing HPR functionality at different places in the network. Using end-to-end HPR will give the maximum benefit to the end user, in terms of availability. If any component in the network fails, such as an intermediate routing node or a telecommunications line, HPR will route around that failure. If there is a failure in the hardware or system software on the Parallel Sysplex system where the end user's application is running, HPR and the Multi-Node Persistent Session function will move the endpoint to another system. For a list of hardware and software that provide HPR support, see 2.1.3.1, "HPR Implementations" on page 31 in Volume 3: Connectivity, SG24-5639.

Where Do You Use HPR?

To achieve maximum availability, you should consider moving the HPR function as far out into the network as possible.

3.3.3.2 APPN Configuration

When using the Multi-Node Persistent Session function, sessions are moved from a subsystem on one VTAM to a restarted subsystem on another VTAM. This support requires both VTAM systems to be configured as APPN end nodes in the same sysplex.

VTAM Configuration for Multi-Node Persistent Session

VTAM systems that wish to use Multi-Node Persistent Sessions should be configured as APPN end nodes (or migration data hosts).

For more information on APPN node types, refer to 2.1.1, "APPN Overview" on page 24 in Volume 3: Connectivity, SG24-5639.

3.3.3.3 Subsystem Support of Multi-Node Persistent Sessions

The IBM subsystems that support Multi-Node Persistent Sessions at the time of writing are:

- CICS V4.1 (with PTF) and higher.
- APPC/MVS for OS/390 R3 and higher.
- IMS/ESA V7 uses the facilities of MNPS to provide faster session recovery following a failure of the IMS subsystem.

The implications of Multi-Node Persistent Session support are the same as in the single node case. For example, in the case of a CICS recovery with a Multi-Node Persistent Session, the end user will have to sign on again if he was signed on at the time of the failure.

3.3.4 High Performance Routing

As mentioned previously, HPR is a prerequisite for the Multi-Node Persistent Session support. However, even for subsystems that do not support Multi-Node Persistent Sessions, there are availability benefits for the end user in implementing HPR on the network. This is because HPR provides the ability to reroute sessions around failures in the network, through the feature known as *nondisruptive path switching*. The ability to reroute sessions implies that an alternate route is available for the automatic rerouting to work. Note that HPR also can hold sessions until the failing link/node recovers (assuming the failing component recovers in time), which means that an alternate route is not required.

Using HPR nondisruptive path switching in the network does not require any specific Parallel Sysplex features. However, to get the maximum availability benefits, you should try to use HPR for as much of the end user's session as possible. For example, by providing the HPR endpoint function in VTAM inside the Parallel Sysplex, and also at the end user's workstation, with all intermediate nodes providing HPR routing function, the end user will get the best availability from the network and the Parallel Sysplex.

Figure 5 on page 30 in Volume 3: Connectivity, SG24-5639 shows the effects of placing the HPR functionality at different places in the network. For a list of hardware and software that provide HPR support, see 2.1.3.1, "HPR Implementations" on page 31 in Volume 3: Connectivity, SG24-5639.

3.3.5 VTAM Systems Management

A number of systems management facilities supported by VTAM can increase overall availability.

3.3.5.1 Automatic Restart Manager Support

Support for the Automatic Restart Manager (ARM) facility of OS/390 was introduced in VTAM V4.3. ARM exploitation reduces the impact on end users when a VTAM failure occurs. VTAM will register with ARM at startup time. ARM can then automatically restart VTAM after a failure, as long as the ARM function is still active. Other ARM participants are automatically notified after the VTAM recovery. For more information on ARM, refer to 3.6.2, "Automatic Restart Management (ARM)" on page 123.

3.3.5.2 Cloning

VTAM V4.3 was the first release to support the cloning facilities in OS/390. VTAM allows you to dynamically add applications in the Parallel Sysplex environment. Model application definitions can be used, so that a single set of VTAMLST members can be used for all VTAM application major nodes.

VTAM can also be an OS/390 cloned application itself, using OS/390 symbols in all VTAMLST members. It is possible to use a single set of VTAMLST members for all the VTAM systems. It is also possible to use OS/390 symbols in VTAM commands.

If the VTAM systems are defined as APPN nodes (either end nodes or network nodes), it is much easier to make full use of the OS/390 cloning facilities. VTAM APPN systems require fewer system definitions than VTAM subarea systems. The definitions that are required are simpler and lend themselves readily to the use of the OS/390 symbols.

VTAM's Use of Cloning

VTAM systems that use APPN rather than subarea protocols can make optimum use of the OS/390 cloning facilities.

The cloning of applications and the cloning of VTAM itself make it much easier to replicate systems in a Parallel Sysplex environment.

3.3.5.3 APPN Dynamics

APPN dynamics offer many advantages that may enhance availability. APPN features greater distributed network control that avoids critical hierarchical dependencies, thereby isolating the effects of single points of failure; dynamic exchange of network topology information to foster ease of connection, reconfiguration, and adaptive route selection; dynamic definition of network resources; and automated resource registration and directory lookup. APPN extends the LU 6.2 peer orientation for end-user services to network control and supports multiple LU types, including LU0, LU1, LU2, LU3, and LU6.2.

VTAM V4.4 provides the ability to use XCF services to create “logical” APPN S/390 server-to-S/390 server connections, thus eliminating the need for VTAM CTCs. See the IBM Redbook *SNA In a Parallel Sysplex Environment*, SG24-2113, for a more detailed description of this facility.

3.3.6 TCP/IP Virtual IP Addressing

There is an enhancement to TCP/IP called *Virtual IP Addressing (VIPA)*, which can improve availability in a sysplex environment. The VIPA function is used with the routing daemon (RouteD or OMPROUTE) to provide fault-tolerant network connections to a TCP/IP for OS/390 system. The RouteD daemon is a socket application program that enables TCP/IP for OS/390 to participate in dynamic route updates, using the Routing Information Protocol V1 or V2 (RIP-1). OMPROUTE supports the Open Shortest Path First (OSPF) and RIP routing protocol.

A TCP/IP for OS/390 system may have many physical interfaces. Each physical interface is assigned an IP address. The availability of an IP address that is assigned to a physical interface depends on the availability of the hardware components that constitute the physical interface, such as:

- An adapter on a 3172
- An adapter in a 3745 or 3746
- An OSA adapter
- An adapter on a 2216

For example, if the adapter in a 3172 or the 3172 itself becomes unavailable, the IP addresses of the interfaces that are associated with the failing hardware components become unavailable.

The VIPA functions allow the installation to define a virtual interface that is not associated with any hardware components and thus cannot fail. The IP address that is defined for a virtual interface is therefore always available.

Remote IP S/390 servers connect to the VIPA address through one of the physical interfaces of the TCP/IP for OS/390 system. Name servers must be configured to return the VIPA address of the TCP/IP for OS/390 system and not the IP addresses of the physical interfaces. If a physical interface fails, dynamic route updates will be sent out over the other physical interfaces, and downstream IP routers or S/390 servers will update their IP routing tables to use an alternate path to the VIPA address. The effect is that TCP connections will not be broken, but will recover nondisruptively through the remaining physical interfaces of a TCP/IP for OS/390 system.

Figure 24 shows an example of TCP/IP VIPA recovery of an inbound connection.

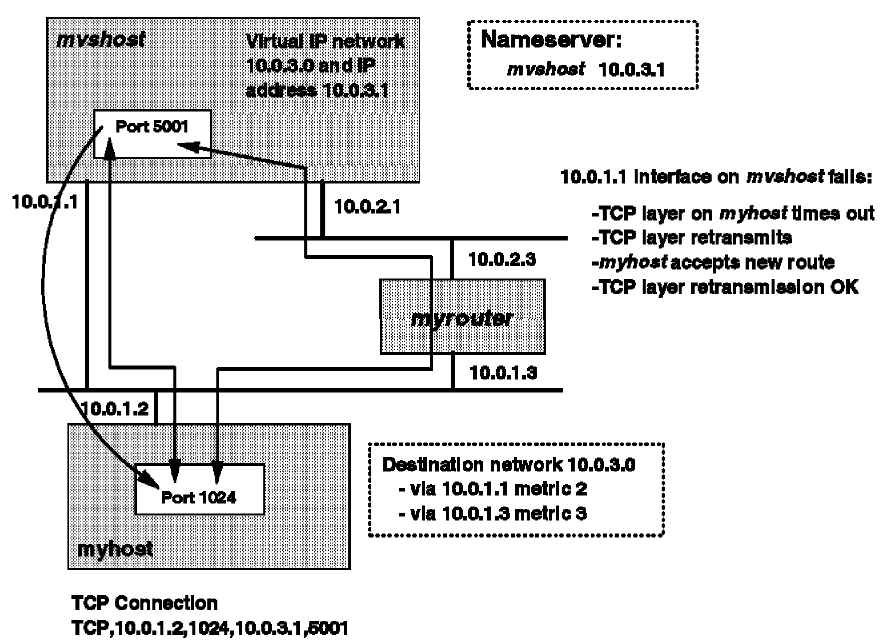


Figure 24. TCP/IP VIPA Example

If a complete TCP/IP for OS/390 system becomes unavailable, connections with clients will be broken. A standby TCP/IP for OS/390 system can, in that situation, be dynamically modified to take over the VIPA identity of the failing TCP/IP for OS/390 system. After such a takeover action, remote clients can re-connect to the VIPA address of the failing system, and they are connected to a backup system. If the backup system supports the same set of server functions as the failing system supported (such as FTP services, Telnet services, and so on), clients will be able to manually recover by re-establishing their FTP session or Telnet session.

The takeover action can be initiated through normal NetView/390 message automation functions in releases prior to OS/390 V2R8. In OS/390 V2R8 and later, Dynamic VIPA Takeover can be used to automatically activate the VIPA on another host.

VIPA may be used with static route definitions, but in case of an interface failure, a manual update of routing tables in both TCP/IP for OS/390 and downstream routers is required. This is impractical in any but the smallest network.

For more information, refer to the IBM Redbook *TCP/IP In a Sysplex*, SG24-5235.

3.3.7 Dynamic VIPA Takeover

A Virtual IP Address can be associated with either a complete IP stack, or with a specific IP application. OS/390 V2R8 introduced the capability to have the VIPA of a failed stack or a failed application automatically taken over by an alternate stack or by restarting the failed application.

For example, if your users are connecting to a Web server or to TSO in a Parallel Sysplex, it is transparent to them which system in the sysplex they are using. In this case, you would associate the VIPA with the stack, so that if the stack fails, another stack would accept requests destined for the failed stack, and route the requests to the Web server or TSO on that system.

On the other hand, if your users are accessing a unique application, such as IMS in a non-data sharing environment, all requests must be routed to that specific application. In this case, you would assign the VIPA to the IMS system, and define the VIPA as being dynamic. Now, if IMS fails, when it is restarted (on the same or another system), it will connect to the same VIPA when it is coming up, and requests will once again be routed to that application. You could use automation or ARM to automate the restart of IMS. The reconnection and activation of the VIPA will happen automatically as long as the correct definitions have been put in place.

More information about Dynamic VIPA, and how to set it up, is contained in the IBM Redbook *TCP/IP in a Sysplex*, SG24-5235.

3.3.8 Sysplex Distributor

The latest enhancement to TCP's workload balancing capabilities, is a new feature called "Sysplex Distributor." Sysplex Distributor builds on the VIPA and Dynamic VIPA Takeover support to effectively let you have a cloned application, on a number of OS/390 systems, with each instance having the same IP address. It provides the benefits of WLM/DNS without requiring WLM/DNS support in the application, nor requiring the client to abide by the Time To Live value specified by the server. Sysplex Distributor also adds the ability to non-disruptively move an application back to its original location after Dynamic VIPA Takeover has been used to move the application after a failure.

Sysplex Distributor is described in more detail in 2.3.2.8, "Sysplex Distributor" on page 54 in Volume 3: Connectivity, SG24-5639.

3.4 Hardware Considerations for Availability

The amount of redundancy to configure is a very important consideration. To a large extent, this will depend on the amount of availability required, and how much cost your business can afford. In most cases however, the redundant hardware should be used as a normal part of the configuration, rather than having to sit idle until a problem arises.

Generally speaking, and from a pure availability perspective, we recommend at least two of everything. For example, it is strongly suggested that an installation configure at least two CFs, with two CF links to each CPC that contains an image that is communicating with that CF. The system normally uses all CF links, but if one fails, the others have enough capacity to carry the full load.

The *Parallel Sysplex Availability Checklist*, available on the Parallel Sysplex home page, contains a detailed list of things to consider when configuring hardware for availability.

3.4.1 Number of CPCs in Parallel Sysplex

When planning for continuous availability and a *guaranteed service level,* you might want to configure one more CPC than the number you estimated for performance. In this way, if you suffer a hardware outage, the redistribution of the load to the running CPCs will not cause a performance degradation. Alternately, you can choose to provide extra *capacity* in the CPCs without providing an additional CPC.

3.4.2 Redundant Power

Both 9672 and 9674 CPCs have dual power cords that should be connected to separate power sources. In addition, a local UPS is available that provides a greater level of protection against power outages.

The local UPS (machine type 9910) can keep a 9672/9674 fully operational for 6 to 12 minutes, depending on configuration. If it is providing support to a 9674 CF or a CF LP in a 9672 in *power save state*, then the CF storage is preserved for 8 to 16 hours, maintaining the structures as they were just before the power drop.

This can significantly decrease recovery time for the CF exploiters, because they will not have to allocate and build the structures again. For a further discussion of UPS and power save mode/state, refer to the IBM Redbook *S/390 MVS Parallel Sysplex Continuous Availability SE Guide*, SG24-4503.

3.4.2.1 CF Volatility/Nonvolatility

The presence of the Internal Battery Feature or UPS determines the *volatility* of the CF; that is, it determines whether the contents of storage are maintained across an external power outage. Certain CF exploiters "react" to the volatility, or change of volatility of the CF. Details of the volatility attribute and the reaction of exploiters to it are provided in 2.0, "A Structured View of CF

Structures" on page 59 in Volume 2: Cookbook, SG24-5638. It is important to consider the volatility of the CF when configuring for continuous availability.

If you have an external UPS, you need to tell the CF that its status is nonvolatile. You do this with the CF MODE NONVOLATILE command that is entered at the CF console.

3.4.3 Isolate the CF

There are a number of reasons why you may want to isolate some or all of your CFs onto CPCs of their own. The first and most important is for recovery. Depending on the structures residing in the CF, if a CPC containing both a CF LP, and an OS/390 LP that is connected to the CF LP were to fail, the impact ranges from being minimal to your whole Parallel Sysplex being adversely affected. The exact implications depend on which structures were in the CF that failed.

The WSC Flash W98029, *Parallel Sysplex Configuration Planning for Availability*, categorizes the different users of CF into those that do not have to be failure-isolated through to those that do. When deciding whether you should have all ICFs, all standalone CFs, or some combination in between, you should consult that document to identify the most appropriate configuration for your environment.

Another reason for isolating the CF is that some structures are specifically designed not to reside in the same failure domain as the operating systems they are connected to. Forcing such structures into such a CF could have undesirable performance side effects.

An additional reason is discussed in 2.4, "CF Architecture" on page 71, and relates to the conversion of synchronous CF requests into asynchronous requests if the OS/390 LP is sharing CPs with a CF LP. This could have undesirable performance effects; however, as it does not affect ICF LPs, this should only be a concern if you plan to run a CF in an LP using CP rather than ICF engines.

Generally speaking, if you are doing data sharing with Parallel Sysplex, the data sharing-related structures should reside in a failure-independent CF. If you are only doing resource sharing, it may be acceptable to use CFs that are not failure-independent. 2.3.3, "What Different 'Plexes Are There?" on page 39 contains a list of the Parallel Sysplex exploiters that are regarded as resource sharers rather than data sharers. There is also information about the requirements of the different resource sharing exploiters in the IBM Redbook *S/390 Parallel Sysplex: Resource Sharing*, SG24-5666.

3.4.3.1 CF Configuration Options

There are a number of options regarding where the CF could be run. For a production CF, the viable options are:

- Use a standalone CF with dedicated or shared CPs. This is the preferred option for a production CF involved in data sharing.
- Use the ICF facility to run a CF in a CPC that is failure-isolated from the production sysplex OS/390 images.

- Use the ICF facility to run a CF in a CPC that is *not* failure-isolated from the production sysplex. This is really only acceptable for non-data sharing Parallel Sysplexes.

ICF Options

The 9672 G5 and later processors provide great flexibility as to how you can run a CF in a CPC. While it is still possible to run a CF in an LP that uses non-ICF CPs (that is, a CPs that can be used by an operating system), ICFs are more popular because they are lower cost than non-ICF CPs, and they are not taken into account when calculating the software costs for that CPC.

The possible configurations for an ICF LP are as follows:

- You can have an ICF LP with one or more dedicated CPs.
- You can have an ICF LP with one or more dedicated CPs, plus the ability to expand into the pool of shared operating system CPs.
- You can have an ICF LP with one or more dedicated CPs, plus the ability to expand into a shared ICF CP.
- You can have an ICF LP with just shared ICF CPs.

Which one is the most appropriate for you depends on which structures you are placing in the CF, the number of MIPS required for the CF, the speed of the CPs in your CPC, and whether the CF is being used for test or production.

For a test CF, you have more flexibility, mainly due to the reduced performance and recovery requirements for a test sysplex. The options for a test CF are:

- Run in an LP on a standalone CF, possibly sharing CPs with a production CF, or with other test CFs. If the CPs are shared between the LPs, you should run the production CF in active wait mode, and run the test CF with Dynamic CF Dispatching enabled. This will ensure acceptable performance for the production Parallel Sysplex, while still having the required functionality on the test Parallel Sysplex.
- Use the ICF facility to place the CF on a CPC. If this really is a test sysplex, it may be acceptable to use a CF that is not failure-isolated, even if you are testing data sharing. IC links could be used to connect to the test OS/390 LP if it is running on the same CPC. If the test OS/390 is running on another CPC, external CF links must be used.
- Have a CF LP that shares CPs with operating system LPs. If you have a very low MIPS requirement for the test CF, this may be a viable option. To control the amount of resource being consumed by the test CF, Dynamic CF Dispatching should be used. As this is only a test CF, the performance should be acceptable. Again, IC or external CF links could be used to communicate with the OS/390 LPs in the test Parallel Sysplex.

The different CF configurations are discussed further in 1.4.1.4, “Comparing CF Configuration Options” on page 31 in Volume 2: Cookbook, SG24-5638.

Continuous Availability CF Configuration

If your applications truly require continuous availability, we recommend placing critical CF structures in a standalone CF (9674 or 9672-R06) rather than in a CPC with operating system images. This is because it is possible that the CPC may need to be taken down to apply maintenance that is not CF-related, and your CF suffers an outage that it might otherwise avoid. Also, there is the remote possibility that an error in an operating system LP could cause an unplanned outage of the whole CPC.

Finally, if you are really serious about continuous availability, you should configure a third CF that can be used to provide fallback during periods when one of the CFs is down for a planned outage. If you have only two CFs, you are exposed during this period if there is any sort of problem on the one remaining CF. The following section discusses how you might achieve this.

3.4.3.2 Third CF

Although the recommended configuration of two CFs, each with sufficient resource to accommodate all the allocated structures, can provide industry-leading levels of availability, there may be installations where anything less than 100% is unacceptable.

If your installation falls into this category, you may wish to consider providing a third CF. This can be implemented as either a full production CF, or as a hot standby, ready to be called into use only when necessary.

At the time of writing, there is a small, but growing, number of customers that have three CFs in their production sysplex. If you wish to use this configuration, there are no special considerations, other than remembering that all three CFs should be specified in the preference lists, and all three of the CFs should have sufficient MIPS, CF link capacity, and storage to at least support all the critical structures in case of a failure of the other two CFs. Rather than splitting your CF load over two CFs, you would split it over three.

If you are using DB2 GBP duplexing, you should especially make sure to specify all three CFs in the preference list for the GBP structures. If one of the DB2 duplexed GBPs is lost, DB2 will revert back to simplex mode for that GBP. However, if a third CF is available, and is included in the structure's preference list, DB2 will automatically reduplex into the third CF.

However, if you wish to implement the third CF as a hot standby that only gets used when one of the other CFs is unavailable, the considerations are different.

For a hot standby to be effective, you ideally want it to use little or no resource when it is not being used, but still be able to give production-level response times when it is called into action.

The most likely way that you would implement a hot standby CF is in an LP, using shared CPs, with Dynamic CF Dispatching enabled, and a high weight so that it can quickly get the CPU it needs when called into action. For details on the Dynamic CF Dispatching feature, see 1.2.4.1, “Dynamic CF Dispatching” on page 11 in Volume 2: Cookbook, SG24-5638.

A detailed discussion about hot standby CFs and how best to configure them is contained in 1.4.3, “Hot Standby CFs” on page 43 in Volume 2: Cookbook, SG24-5638. For now, suffice to say that it *is* possible to configure for the level of availability that you require, without necessarily having to make significant additional hardware investments.

3.4.4 Additional CF Links

For high availability, redundant CF links should be configured between CPCs and CFs. This configuration removes a potential single point of failure in the Parallel Sysplex environment. In most customers, a single CF link should be sufficient to provide acceptable response time, so two CF links would be sufficient to provide redundancy in case of a CF link failure.

In a configuration with more than 16 CPCs, redundant CF links to a single CF cannot be maintained because the maximum number of external links to a CF is 32.

Recommendation for CF Links

It is recommended that you have multiple CF links between images and CFs. Where multiple CF links connect an image to a CF, they should be routed physically separate from each other. This procedure prevents having elements damaged by an external physical occurrence.

For performance reasons, it is also desirable to avoid the intensive use of a few physical CF links serving many OS/390 LPs in the same CPC. Also, if one physical CF link is down, many LPs could be affected. The recommendation is therefore to avoid or minimize the intensive sharing of CF links.

CF links are reconfigurable between CF LPs on G3 and later 9672s and 9674s. For more information on configuring CF links, refer to 1.5, “CF Links” on page 45 in Volume 2: Cookbook, SG24-5638.

3.4.5 I/O Configuration Redundancy

Normal channel configuration techniques should be applied for FICON, ESCON, and any parallel channels in a Parallel Sysplex. Channels should be spread across channel cards/IBB domains and ESCON directors to maintain channel path availability. The Systems Assurance Product Review (SAPR) Guide guide for the relevant CPC should be used so that all devices are attached for maximum availability.

Refer to 1.0, “I/O Connectivity in Parallel Sysplex” on page 3 in Volume 3: Connectivity, SG24-5639 for information on I/O redundancy.

Further enhancements in new models of the ESCON director hardware improve the ability to support continuous availability configurations. For example, the 9032 Model 3 and Model 5 have:

- Optional hardware component redundancy
- Concurrent LIC upgrade capability
- Dynamic addition of ports (hot plug)

For further information on the 9032 and other ESCON director features see 1.2.3, "ESCON Director (ESCD) Configuration Guidelines" on page 10 in Volume 3: Connectivity, SG24-5639.

3.4.6 Sysplex Timer Redundancy

In any sysplex environment, whether basic or parallel, the Sysplex Timer is a critical device, and can be a single point of failure for the entire sysplex if it is not configured for redundancy. Refer to 3.2.1, "Planning for Availability of Sysplex Timer" on page 77 in Volume 3: Connectivity, SG24-5639 to get information on Sysplex Timer redundancy.

3.4.7 RAS Features on IBM CPCs

For an overview of reliability, availability and serviceability (RAS) features on IBM CPCs, refer to Table 36 on page 233 in Volume 2: Cookbook, SG24-5638.

3.5 Limitations to Continuous Availability

There are situations in non-sysplex and Parallel Sysplex environments that limit the achievement of continuous availability. It is anticipated that many of these limitations will be fixed over time in further releases of LIC and software. Some of the traditional limitations still exist, but have been addressed to some extent:

- Depending on the types of data you have, and your backup tools, backing up your data may require applications to be stopped so that a coherent, point-in-time backup can be achieved. DB2 and IMS both provide backup facilities that allow the database to be backed up without having to take the data offline. For VSAM, the Backup While Open facility can be used to get a coherent backup without removing access to the data. For other file types, or for full volume dumps, facilities such as the ESS FlashCopy function can be used to minimize the outage time.
- Database reorganizations are disruptive. This can be offset to some extent through the use of partitioned databases. In this case, the disruption can be reduced to just that part of the database that is being reorganized. DB2 supports partitioned databases, and IMS/ESA V7 introduces support for partitioned Full Function databases.

  DB2 has an online reorg capability, whereby the data remains during most (but not all) of the reorg process. Online reorg capability is also available for Fast Path DEDB databases in IMS.
- Some upgrade/maintenance activities on the CF are nonconcurrent. With no rebuild support for certain CF structures, this necessitates an outage on the affected subsystem. The number of changes requiring an outage can be reduced through the use of the Capacity Upgrade on Demand feature. Also, the System Managed Rebuild facility introduced in OS/390 V2R8 is exploited by JES2 and WLM multisystem enclaves and will be expanded to other products in the future.
- Daylight saving time issues up to and including IMS/ESA V5. (This issue is addressed in IMS/ESA V6 for the IMS internal code.)

Some other inhibitors remain for the time being:

- Some CICS resources do not support data sharing in a Parallel Sysplex environment, for example recoverable temporary storage queues. These resources that must be shared require techniques such as function shipping to a single CICS region. This creates a single point of failure.

- VSAM does not support online reorganization of VSAM files.
- Coexistence support for n, n+1 is not fully implemented for all software products.

With careful planning, it is possible to minimize the disruption caused by these limitations.

3.6 Recovery Considerations for Availability

When failures occur in a Parallel Sysplex, work can continue on the remaining elements of the system. Also, many features have been added that will make recovery of the failed system easier.

If a failure occurs in a CF, structures may be automatically recovered on an alternate CF if the preference list for the structure contains the alternate CF name. This structure rebuild is initiated by each subsystem that has structures on the CF when the subsystem is informed that a failure has occurred. At present, not all subsystems support dynamic structure rebuild. See 2.0, “A Structured View of CF Structures” on page 59 in Volume 2: Cookbook, SG24-5638 for more detail.

3.6.1 Sysplex Failure Management (SFM)

To enhance the availability of a sysplex, XES provides the Sysplex Failure Management (SFM) facility. A recent analysis of multisystem outages indicated that the single change that would have eliminated the largest number of system outages would be the implementation of SFM, using ISOLATETIME(0) to automate the removal of disabled systems. The use of SFM is described in the white paper *Improve Your Availability with Sysplex Failure Management*, which is available on the Web at:
`http://www.s390.ibm.com/products/pso/availability.html`

To use SFM, you must define and activate an SFM policy. The SFM couple data set contains one or more SFM policies that allow an installation to predefine the actions OS/390 is to take when certain types of failures occur in the sysplex.

Each system in the sysplex periodically updates its own status and monitors the status of other systems in the sysplex. The status of the systems is maintained in a couple data set (CDS) on DASD. A status update missing condition occurs when a system in the sysplex does not update its status information in either the primary or alternate couple data set within the failure detection interval, specified on the INTERVAL keyword in COUPLExx, and appears dormant.

SFM allows you to specify how a system is to respond to this condition. System isolation allows a system to be removed from the sysplex as a result of the status update missing condition, without operator intervention, thus ensuring that the data integrity in the sysplex is preserved. Specifically, system isolation uses special channel subsystem microcode in the target CPC to cut off the target LP from all I/O and Coupling Facility accesses. This results in the target LP loading a non-restartable wait state, thus ensuring that the system is unable to corrupt shared resources.

For example, if one system loses signalling connectivity to the sysplex, SFM provides the method for reconfiguring the sysplex so that operations can continue. The “isolation” of a failed system can be automatic if SFM is enabled. Figure 25 on page 119 shows the process followed if a system fails to update its status. In this example, OS/390 A fails to update its status (.1.). This is detected by OS/390 B (.2.), which in turn requests, through the CF (.3.), the isolation of OS/390 A. Having received the request from OS/390 B, the CF sends the isolate signal to the CPC that OS/390 A is running on (.4.). The channel microcode in that CPC then stops all I/O and CF communication for the LP that OS/390 A is running in.

Note: This requires that OS/390 A is still in communication with the CF.

The timings related to the isolate function are dependent on a number of parameters coded in the SFM policy and the COUPLExx member of parmlib. For an explanation of the interaction between these parameters, refer to the IBM Redbook *Parallel Sysplex Continuous Availability Guide*, SG24-4503.

Only one SFM policy can be active in the sysplex at any one time, but installations may require different recovery actions at different times, for example during the online day and during the batch window. Different SFM policies can be defined in the SFM couple data set and activated using the SETXCF operator command. This function can be automated using an appropriate automation tool.

SFM allows you to assign a relative value WEIGHT to each system in the Parallel Sysplex. The weight is used for two purposes:

- If there is a partial loss of connectivity between the systems in the sysplex, SFM uses the weight determine which systems should be isolated in order to let the most important systems continue processing.
- The weight is also used to decide whether a structure should be rebuilt after a system loses connectivity to a CF.

In this way, the most important systems in the Parallel Sysplex can continue *without operator intervention* in the event of a failure.

As a simple example, assume there are three systems in a sysplex. OS/390 A has WEIGHT(10), OS/390 B has WEIGHT(10), and OS/390 C has WEIGHT(30). There is an XCF connectivity failure between OS/390 B and OS/390 C. The alternatives are to continue with OS/390 A and OS/390 B (total WEIGHT=20), or to continue with OS/390 A and OS/390 C (total WEIGHT=40). The latter configuration is used; that is, OS/390 B is isolated from the sysplex. If all systems in this example had the same weight (that is, if the weight of OS/390 C was 10), it would be unpredictable which two systems would be kept in the Parallel Sysplex.

Note: The defined weights are not used when reacting to a “status update missing” condition. In this case, the system that has not updated its status is partitioned, regardless of its weight.

Weights are attributed to systems based on a number of factors determined by the installation. Examples might be:

- Importance of the work running on a system.
- The role of the system—for example, the system that owns the network might be deemed to be high importance, even if it does not actually run any applications.

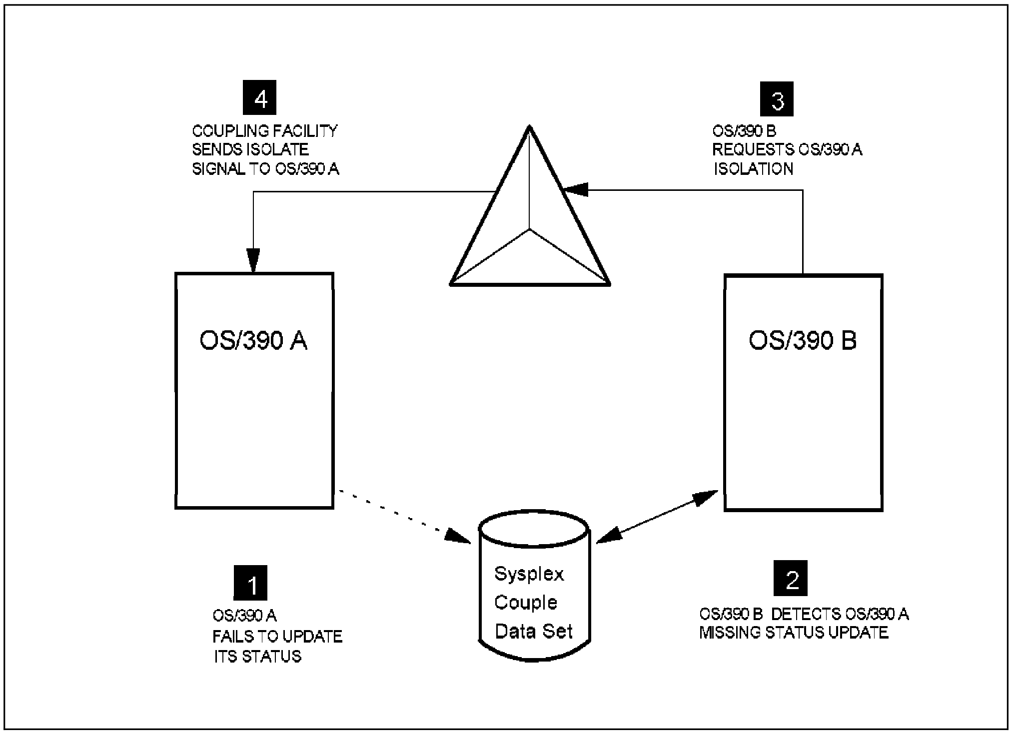


Figure 25. Isolating a Failing OS/390

- ITRs of the systems in the sysplex.
- Configuration dependencies, such as unique features or I/O connected to a particular system in the sysplex.

Weight is any value from 1 to 9999. Specifying no weight is equivalent to specifying WEIGHT(1). If you do not specify a WEIGHT parameter in the SFM policy, every system is given the same importance when it comes to partitioning.

The SFM policy, in conjunction with the CFRM policy, is also used to determine the rebuild activity for those CF structures that support it. The relationship between SFM and CFRM in the area of rebuild is that REBUILDPERCENT (CFRM) is calculated using the WEIGHT values as defined to SFM. Refer to 2.0, "A Structured View of CF Structures" on page 59 in Volume 2: Cookbook, SG24-5638 for details of rebuild characteristics of the various CF exploiters.

Availability Enhancement in OS/390 R2: In OS/390 R2, the CF selection process was enhanced. During structure allocation, the WEIGHT values specified in the SFM policy are used. Also, the structure rebuild process was improved by insuring that, in case of a loss of connectivity, the rebuilt structure has better connectivity to the systems in the sysplex than the old structure.

In this process, the WEIGHT values and REBUILDPERCENT value specified in the CFRM policy are used. Prior to OS/390 R2, if CONNFAIL(NO) is specified in the SFM policy, WEIGHT values were not saved, therefore APAR OW19718 was required for those systems to save the WEIGHT value regardless of what is specified for the CONNFAIL keyword. This APAR is available for systems starting with MVS/ESA SP V5.1 up to OS/390 R1. Refer to Table 10 on page 121 for a summary of rebuild characteristics. Also, refer to WSC FLASH 9825, *SFM Functions and the Impact of OW30814 for Parallel Sysplex.*

Availability Enhancement in APAR OW26649: This APAR increases the probability that XCF can recover from system status update I/O completion errors. Prior to this APAR, XCF depended on notification from the Missing Interrupt Handler (MIH) to know if an I/O operation to update its status was having a problem. If MIH informed XCF that the I/O had not completed within the specified MIH interval, XCF would attempt to update the alternate CDS instead.

However, if XCF was not aware that the I/O had not completed, it would not redrive the I/O, and the other systems would eventually detect a status update missing condition. This can happen if the XCF I/O is not the first I/O queued to the device. MIH will only redrive the actual I/O that is failing. Other I/Os will just wait, with no notification of there being a problem.

APAR OW26649 changed XCF to exploit the I/O Timing facility. XCF will set a time value for completion of each I/O. This is called *request-based I/O timing*. This value will be 15 seconds, or one-third of the system's failure detection interval, which is specified in INTERVAL keyword of COUPLExx parmlib member, whichever is smaller. If the I/O does not complete within this time, XCF will be notified and can redrive the I/O to the alternate CDS. This is a more precise method of ensuring that I/O completes than the MIH facility used earlier. This function is integrated in OS/390 V2R5, and is available as an APAR for systems starting with MVS/ESA V5.2 up to OS/390 V2R4. Also, refer to WSC FLASH 9838, *XCF Service Recommendations to Maximize Sysplex Availability*.

Availability Enhancement in APAR OW30814: When the loss of connectivity between a CF structure and its exploiter occurs, and an SFM policy is active in the sysplex, the structure that has suffered a connectivity loss may be rebuilt in another available CF. This allows the exploiter to connect to the structure in the new location and continue processing. For some CF structures, the product that allocated the structure will realize that connectivity to the structure has been impacted, and that product will automatically initiate a rebuild.

However, for some other structures, such a rebuild must be initiated by the operating system. For those structures, if an SFM policy is not active, a structure rebuild will not be initiated for any structure that supports the REBUILDPERCENT keyword unless *all* connectors lose connectivity. APAR OW30814 changes loss of connectivity processing so that if there is no active SFM policy and there is *any* loss of CF connectivity, a structure rebuild will be initiated by OS/390 R3 and higher systems as long as the structure users support the structure rebuild protocol. This APAR is available for systems starting with OS/390 R3 up to OS/390 V2R5. Refer to Table 10 on page 121 and 2.0, "A Structured View of CF Structures" on page 59 in Volume 2: Cookbook, SG24-5638 for details of rebuild characteristics of the various CF exploiters. Also, refer to WSC FLASH 9825, *SFM Functions and the Impact of OW30814 for Parallel Sysplex*.

Although this function provides some of the benefits of SFM to sysplexes that do not have SFM active, this does not mean that SFM is no longer required. This function will cause a rebuild if *any* of the systems in the sysplex lose connectivity to the structure. However, you might prefer *not* to have a rebuild take place if certain systems, a development system for example, lose connectivity. The use of SFM provides you with much more granular control over what actions take place.

Availability Enhancement in APAR OW30926: Every system in the sysplex updates its system status in the sysplex couple data set at some interval. Prior to this APAR, checking that status is the only way for the other systems to determine whether that system is alive or not. This APAR changes the way SFM reacts to a status missing condition. Now, instead of immediately partitioning the affected system out of the sysplex, XCF will also check to see if there are any XCF signals coming from the system that appears to be dead. If there are, XCF issues messages IXC427A and IXC426D, asking the operator whether the affected system should be partitioned out of the sysplex or not. This new support should ensure that healthy systems will not be erroneously reset, and should provide the confidence to use the ISOLATETIME(0) and no operator PROMPT parameters in the SFM policy to immediately partition out systems that are no longer alive.

This new function is integrated into OS/390 V2R7 and is available via APAR OW30926 for OS/390 R2 up to OS/390 V2R6.

Table 10 shows the summary of the relations between structure rebuild and APARs OW19718 and OW30814 in case of connectivity failure.

Table 10. Rebuild Characteristics in Connectivity Failure

OS	APAR	Save WEIGHT	OS/390 Initiated Rebuild	
			SFM Not Active	**SFM Active**
Prior to OS/390 R2	w/o OW19718	No/Yes [1]	No [2]	Yes (CONNFAIL=YES) No (CONNFAIL=NO)
	w/ OW19718	Yes	No [2,3]	Yes [3]
OS/390 R2	-	Yes	No [2,3]	
OS/390 R3, R4 and R5	w/o OW30814	Yes	No [2,3]	
	w/ OW30814	Yes	Yes	
OS/390 V2R6	-	Yes	Yes	

Note:

- (1) - WEIGHT value is saved if CONNFAIL(YES) is specified in active SFM policy.
- (2) - Most structures will initiate rebuild even if there is no active SFM policy.
- (3) - The structure rebuild is based on REBUILDPERCENT and WEIGHT value.

Availability Enhancement in APAR OW33615: APAR OW33615 provides a number of significant availability and manageability improvements. Perhaps most importantly, it provides improved recovery from a double failure. Prior to this APAR, if a CPC containing a CF and a connected OS/390 system were to fail, the other systems in the sysplex would realize that they had lost connectivity to the structures in the CF. However, because the OS/390 which resided on the same CPC (and is now dead) had not reported that *it* had lost connectivity, the other systems would not realize that it was a 100% loss of connectivity situation, and would happily wait for the OS/390 on the failed CPC to initiate recovery. They would continue waiting until a “status update missing” condition was raised for the dead OS/390. At that point, that OS/390 would be partitioned out of the sysplex and recovery could finally begin.

With this APAR applied, if connectivity to a CF is lost and is reported by all systems that are failure-independent from that CF (that is, all the systems in the sysplex that are not in the same CPC as the failed CF), but is not reported by any systems that are not failure-independent from that CF (that is, those systems that are in the same CPC as the failed CF), the system will assume that any systems that are not failure-independent have lost connectivity as well, but are unable to report it. The LOSSCONN percentage will be calculated at 100%, thus enabling recovery to occur in a more timely manner. For this reason, this APAR is highly recommended for any sysplex using ICFs, *especially* those sysplexes doing data sharing.

This APAR also introduces a new IXCMIAPU keyword, ENFORCEORDER, which can be used to force XCF to allocate structures in CFs in the order specified on the preference list.

Availability Enhancement in APAR OW41959: If REBUILDPERCENT is not specified for a particular structure in the active CFRM policy, and there is an active SFM policy in the sysplex, the system will use a default value of 100 for REBUILDPERCENT. This causes OS/390 to initiate a rebuild only if *every* system with active connections to the structure loses connectivity to the Coupling Facility in which the structure resides (if structure rebuild is supported).

To improve availability, this APAR changes the default REBUILDPERCENT value to be 1, meaning that a rebuild will be initiated if there is *any* loss of connectivity to the structure. This is similar to the effect of APAR OW30814; however, that APAR only applied to cases where there was *not* an SFM policy active.

APAR OW41959 is available for OS/390 R3 through OS/390 V2R9, and is integrated into OS/390 V2R10.

Other SFM Considerations: You can configure a CF and OS/390 on a single CPC, using ICF for example, both of which are in the same sysplex. In this configuration, if the CPC fails, both the CF and the OS/390 in the same CPC fail. This is called a *double failure*. To avoid the impact of a single point of failure, most components using a CF structure have implemented structure rebuild support, so that the structure can be automatically recovered. But for a double failure situation, some components may take a long time to recover or require manual intervention during the recovery process. Especially in this double failure situation, the most important thing for high availability is to isolate the failing system from the sysplex as quickly as possible so that recovery processing can proceed on the remaining systems in the sysplex. Therefore, it is imperative to:

- Ensure that the INTERVAL value is not too long.
- Ensure that the ISOLATETIME value is very low (ideally zero). This implies that PROMPT is *not* specified in the SFM policy.
- Ensure that operational procedures are in place to quickly detect the issuance of IXC102A and respond to the WTOR. There is specific Parallel Sysplex support added to the product System Automation for OS/390 by APAR OW39485 that deals with this, and other, Parallel Sysplex-related messages.
- Ensure that operational procedures are in place to monitor the recovery process.
- Ensure that all exploiters that require rapid recovery are not allowed to have their structures placed in a CF which can allow a double failure to occur.

For more detailed information, refer to WSC FLASH 9829, *Parallel Sysplex Configuration Planning for Availability*, and an excellent document on the OS/390 Integration Test Web site at:

http://www.s390.ibm.com/os390/support/os390tst/icfmst.htm

Recommendation for SFM

If the key impetus for moving to Parallel Sysplex is a business need for continuous availability, then minimizing operator intervention and ensuring a consistent and predefined approach to recovery from failure of elements within the sysplex is paramount. Having SFM enabled is highly recommended to achieve this requirement.

SFM is discussed in detail in the IBM Redbook *Parallel Sysplex Continuous Availability Guide*, SG24-4503, Chapter 2, and in *OS/390 MVS Setting Up a Sysplex*, GC28-1779.

3.6.2 Automatic Restart Management (ARM)

The purpose of automatic restart management (ARM) is to provide fast, efficient restarts for critical applications. These can be in the form of a batch job or started task (STC). ARM is used to restart them automatically whether the outage is the result of an abend, system failure, or the removal of a system from the sysplex.

When a system, subsystem, or application fails, it may hold database locks that cannot be recovered until the task is restarted. Therefore, a certain portion of the shared data is unavailable to the rest of the systems until recovery has taken place. The faster the failed component can be restarted, the faster the data can be made available again.

ARM is a function in support of integrated sysplex recovery and interacts with:

- Sysplex Failure Management (SFM)
- Workload manager (WLM)

ARM also integrates with existing functions in both automation (SA for OS/390) and production control (Tivoli OPC) products. However, care needs to be taken when planning and implementing ARM to ensure that multiple products (Tivoli OPC and SA for OS/390, for example) are not trying to restart the same elements. AOC/MVS V1.4 introduced support to enable proper tracking of ARM-enabled elements to ensure that multiple restarts are avoided. For further information, see A.4.3.1, "Considerations for System Automation in Parallel Sysplex" on page 144 in Volume 3: Connectivity, SG24-5639.

For more information on WLM, see A.3.1, "Workload Manager (WLM) in the Parallel Sysplex Environment" on page 131 in Volume 3: Connectivity, SG24-5639.

3.6.2.1 ARM Characteristics

ARM was introduced in MVS/ESA V5.2. ARM requires a couple data set to contain policy information, as well as status information, for registered elements. Both JES2 and JES3 environments are supported.

The following describe the main functional characteristics:

- ARM provides only job and STC restart. Transaction or database recovery are the responsibility of the restarted applications.

- ARM does not provide initial starting of applications (first or subsequent IPLs). Automation or production control products provide this function. Interface points are provided through exits, event notifications (ENFs), and macros.
- The system or sysplex should have sufficient spare capacity to guarantee a successful restart.
- To be eligible for ARM processing, elements (Jobs/STCs) must be registered with ARM. This is achieved through the IXCARM macro. Some subsystems come with this support built in. For example, CICS registers with ARM at startup time. For products that do *not* register with ARM, there is a program available, called ARMWRAP, that can be inserted into an existing job or STC, and this can be used to do the ARM registration. Your IBM representative can obtain the ARMWRAP package for you from the MKTTOOLS disk in IBM.
- A registered element that terminates unexpectedly is restarted on the same system.
- Registered elements on a system that fails are restarted on another system. Related elements are restarted on the same system (for example, DB2 and its corresponding IRLM address space).
- The exploiters of the ARM function are the jobs and STCs of certain strategic transaction and resource managers, such as the following:
  - CICS/ESA
  - CP/SM
  - DB2
  - IMS TM
  - IMS/DBCTL
  - ACF/VTAM
  - TCP/IP (as of OS/390 V2R8)
  - Tivoli Netview for OS/390
  - IRLM (with APARs PQ06465 and OW28526)

  These products, at the correct level, already have the capability to exploit ARM. When they detect that ARM has been enabled, they register an element with ARM to request a restart if a failure occurs.

3.6.2.2 ARM and Subsystems

When a subsystem such as CICS, IMS, or DB2 fails in a Parallel Sysplex, it impacts other instances of the subsystem in the sysplex due to such things as retained locks and so on. It is therefore necessary to restart these subsystems as soon as possible after failure to enable recovery actions to be started, and thus keep disruption across the sysplex to a minimum.

Using ARM to provide the restart mechanism ensures that the subsystem is restarted in a pre-planned manner without waiting for human intervention. Thus disruption due to retained locks or partly completed transactions is kept to a minimum.

It is recommended therefore that ARM be implemented to restart major subsystems in the event of failure of the subsystem or the system on which it was executing within the Parallel Sysplex.

A discussion on how ARM supports the various subsystems is provided in the IBM Redbook *S/390 MVS Parallel Sysplex Continuous Availability SE Guide*, SG24-4503. Further information on ARM function is contained in an ARM presentation available on the Redpapers section of the ITSO Web site: `http://www.redbooks.ibm.com`

Recommendation for Recovery in the Sysplex

OS/390 Parallel Sysplex Recovery, GA22-7286 contains information about:

- Hardware recovery
- CF recovery
- Subsystem recovery

It is recommended that you take a look at this document when planning for the recovery aspects of the Parallel Sysplex.

3.7 Disaster Recovery Considerations in Parallel Sysplex

Parallel Sysplex Is an Excellent Framework for Disaster Recovery

Parallel Sysplex can be useful in a disaster recovery strategy for the following reasons:

- The elements of a Parallel Sysplex may be physically spread over up to 40 kilometers.
- Parallel Sysplex allows you to configure almost every element redundantly, thus reducing the risk that a disaster could render your entire Parallel Sysplex inoperable.

  In addition to traditional disaster recovery configurations based on either cold or hot standby CPCs, Parallel Sysplex CPCs may be fully utilized at all times.
- Parallel Sysplex provides the ability to maximize the benefits of DASD remote copy. Following a complete failure of one site, the remaining members of the Parallel Sysplex in the alternate site can continue processing using the remote copies of the data from the failed site.

At a SHARE session in Anaheim, California, the Automated Remote Site Recovery Task Force presented a scheme consisting of six tiers of recoverability from disaster scenarios. These are as follows:

- Tier 0 - No DR plan

  No DR plan: all data is lost and recovery is not possible
- Tier 1 - Pick up truck access method

  Pickup Truck Access Method (PTAM): the system, the subsystem, and the application infrastructure along with application data is dumped to tape and transported to a secure facility. All backup data, such as image copies and archived logs, still onsite are lost in the event of a disaster (typically up to 24 to 48 hours). DR involves securing a DR site, installing IT equipment, transporting backup tapes from the secure facility to the DR site, restoring the system, the subsystem, and application infrastructure along with data, and restarting the workload (typically more than 48 hours). Cost factors include creating the backup copy of data, backup tape transportation, and backup tape storage.
- Tier 2- PTAM and hot site

  PTAM and hot site: same as Tier 1 except the enterprise has secured a DR facility. Data loss is up to 24 to 48 hours and the recovery window will be 24 to 48 hours. Cost factors include owning a second IT facility or a DR facility subscription fee in addition to the Tier 1 cost factors.
- Tier 3 - Electronic vaulting

  Electronic vaulting: same as Tier 2 except that the enterprise dumps the backup data to a remotely attached tape library subsystem. Data loss will be up to 24 hours or less (depending upon when the last backup was created) and the recovery window will be 24 hours or less. Cost factors include telecommunication lines to transmit the backup data and dedicated tape library subsystem at the remote site in addition to the Tier 2 cost factors.
- Tier 4 - Active secondary site (electronic remote journaling)

Active secondary site: same as Tier 3 except that transaction manager and database management system updates are remotely journaled to the DR site. Data loss will be seconds and the recovery window will be 24 hours or less (the recovery window could be reduced to 2 hours or less if updates are continuously applied to a shadow secondary database image). Cost factors include a system to receive the updates and disk to store the updates in addition to the Tier 3 cost factors.

- Tier 5 - Two-site two-phase commit

  Two-site two-phase commit: same as Tier 4, with the applications performing two-phase commit processing between two sites. Data loss will be seconds and the recovery window will be 2 hours or less. Cost factors include maintaining the application in addition to the Tier 4 cost factors. Performance at the primary site can be affected by performance at the secondary site

- Tier 6 - Zero data loss (remote copy)

  Zero Data Loss (remote copy): the system, the subsystem, and application infrastructure along with application data is continuously mirrored from the production site to a DR site. Theoretically there is no data loss if using a synchronous remote copy, and only seconds worth of changes lost if using an asynchronous remote copy. The recovery window is the time required to restart the environment using the secondary disks if they are data consistent (typically less than 2 hours). The synchronous solution conceptually allows you to reach zero data loss, but performance may be impacted and care must be taken when considering the rolling disaster which will leave inconsistent data at the secondary site.

Many installations are currently Tier 2; that is, backups are kept offsite with hardware installed for recovery in the event of a disaster.

Some installations are now looking at Tier 6; that is, a backup site is maintained and there is minimal-to-zero data loss. This approach requires the use of facilities such as IBM Peer-to-Peer Remote Copy.

Finally, there is a solution available that might be called Tier 6+; that is, a Tier 6 approach using Parallel Sysplex facilities across geographically dispersed sites. This solution is called a Geographically Dispersed Parallel Sysplex (GDPS).

3.7.1 Multi-Site Sysplexes

Many installations maintain two or more separate sites, with each site providing disaster recovery capability to the others.

ESCON distances to DASD can extend to 103 km (with the Enterprise Storage Server). CF links can cover a distance up to 20 km with single-mode fibers. The performance impact beyond 20 km makes active use of a Parallel Sysplex less practical. For a related Sysplex Timer discussion, refer to 3.2.7, “9037 Distance

Limitations” on page 90 in Volume 3: Connectivity, SG24-5639.

The IBM 2029 Dense Wavelength Division Multiplexer (usually called the Fiber Saver) enables up to 64 channels to be transmitted over a pair of fibers at distances up to 200 km. The actual maximum distance is limited by the devices you wish to connect. You can use the 2029 to extend:

- FICON channels
- ESCON channels
- Sysplex Timers
- Coupling Links (ISC)
- Gigabit Ethernet
- ATM
- FDDI links
- Fiber Fast Ethernet

However CF links and Sysplex Timer links are still restricted to their respective maximum distances.

The 2029 supports two configurations—a high availability configuration, and a base configuration. In the high availability configuration, the incoming channel signals are duplexed across the two links connecting the two 2029s. In this case, you are limited to 32 channels. In the base configuration, the signals are not duplexed, so you can use connect up to 64 channels.

For more details, refer to 1.2.6, “IBM 2029 Dense Wavelength Division Multiplexer” on page 15 in Volume 3: Connectivity, SG24-5639. The use of the 2029 may enable significant cost reduction in establishing a GDPS.

Be aware that no matter what technology is used, extended distances will degrade the response time of whatever device is at the other end. See 1.5, “CF Links” on page 45 in Volume 2: Cookbook, SG24-5638 for details of CF link degradation with distance.

Some installations may use two sites within these distances for disaster recovery purposes. At the time of writing, a number of customers in the Financial Services industry have announced an intention to use GDPS to help them achieve the availability and resilience their businesses require.

It may be technically advantageous to run the two sites as a Parallel Sysplex. Sysplex capability can provide benefits in other (non-disaster) recovery situations, such as a complete, planned, site shutdown for critical maintenance. At the time of writing, a GDPS cannot provide 100% continuous application availability in the event of a disaster affecting one site, however it can help limit the impact of such a disaster to less than two hours.

Figure 26 on page 129 represents a Geographically Dispersed Parallel Sysplex that could facilitate a number of recovery scenarios.

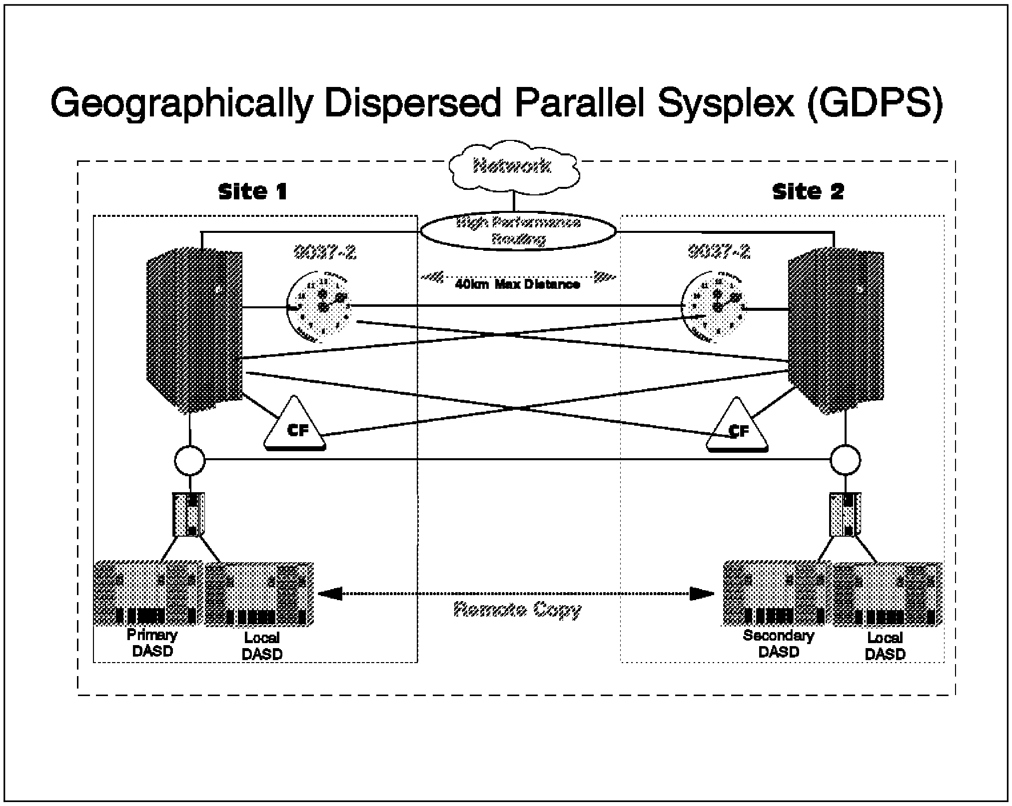


Figure 26. Sample GDPS Configuration

Site 1 and Site 2 can operate as a single Parallel Sysplex. However, the distance between the sites may dictate that spreading production workload across the complete Parallel Sysplex is not viable, due to the response time degradation caused by the distance. Assume then that Site 1 performs high priority work in its own subplex, and Site 2 performs low priority work in its subplex.

In the event of a loss of a CPC in Site 1, it would be possible for the workload associated with it to be taken over by LP X, for example, until the failed CPC is restored.

As Site 2 performs low priority work, it has no backup CF. In the event of a failure, therefore, spare CF capacity could be used in Site 1.

Site 2 can also act as a disaster recovery site for Site 1. By using remote copy techniques, copies of key system data and databases can be maintained at Site 2 for fast recovery of production systems in the event of a complete loss of production at Site 1. As Site 1 is a data sharing environment, the recovery site is required to provide a similar environment.

Each site can also provide additional capacity for the other for scheduled outages. For example, provided all structures support rebuild, the 9674 in Site 2 can be upgraded without interrupting the applications. If remote copy is used, whole applications can be moved between sites with minimal interruption to applications using techniques such as peer-to-peer dynamic address switching (P/DAS). This might be useful if a complete power down is scheduled in either site.

See 3.7.2, “Disaster Recovery Data” on page 130 for a discussion of remote copy techniques, and the remaining sections of this chapter for specific subsystem disaster recovery considerations.

3.7.2 Disaster Recovery Data

A key element of any disaster recovery solution is having critical data available at the recovery site as soon as possible and as up-to-date as possible. The simplest method of achieving this is with offsite backups of such data. However, the currency of this data is dependent on when the outage happened relative to the last backup.

A number of options are available that allow for the electronic transfer of data to a remote site. Two techniques, electronic remote journaling and remote DASD mirroring, are discussed further.

3.7.2.1 Electronic Remote Journaling

Electronic remote journaling requires an active CPC at the remote site, with appropriate DASD and tape subsystems. The transaction manager and database manager updates are transmitted to the remote site and journaled. Image copies of databases and copies of the subsystem infrastructure are required at the recovery site. In the event of an outage at the primary site, the subsystem infrastructure is restored, the journaled data is reformatted to the subsystem log format, and recovery of the subsystem (TM/DB) is initiated. The only *data lost* is that in the remote journal buffers which was not hardened to DASD, and that which was being transmitted at the time of the outage. This is likely to amount to only a few seconds worth of data. The quantity of data lost is proportional to the distance between sites.

Electronic remote journaling can be implemented using the Remote Site Recovery feature of IMS, or the Tracker Site feature of DB2. Another option is the Remote Recovery Data Facility (RRDF) product supplied by E-Net Corporation. More information about RRDF can be found on the Web site:
`http://www.ubiquity.com.au/content/products/suppliers/enet/rrdf01.htm`

3.7.2.2 Remote DASD Mirroring

IBM current provide two options for maintaining remote copies of data. Both address the problem of data made out-of-date by the time interval between the last safe backup and the time of failure. These options are:

- Peer-to-peer remote copy (PPRC)
- Extended remote copy (XRC)

Peer-to-Peer Remote Copy: PPRC provides a mechanism for *synchronous* copying of data to the remote site, which means that no data is lost in the time between the last backup at the application system and the recovery at the remote site. The impact on performance must be evaluated, since an application write to the primary subsystem is not considered complete until the data has been transferred to the remote subsystem.

Figure 27 on page 131 shows a sample peer-to-peer remote copy configuration.

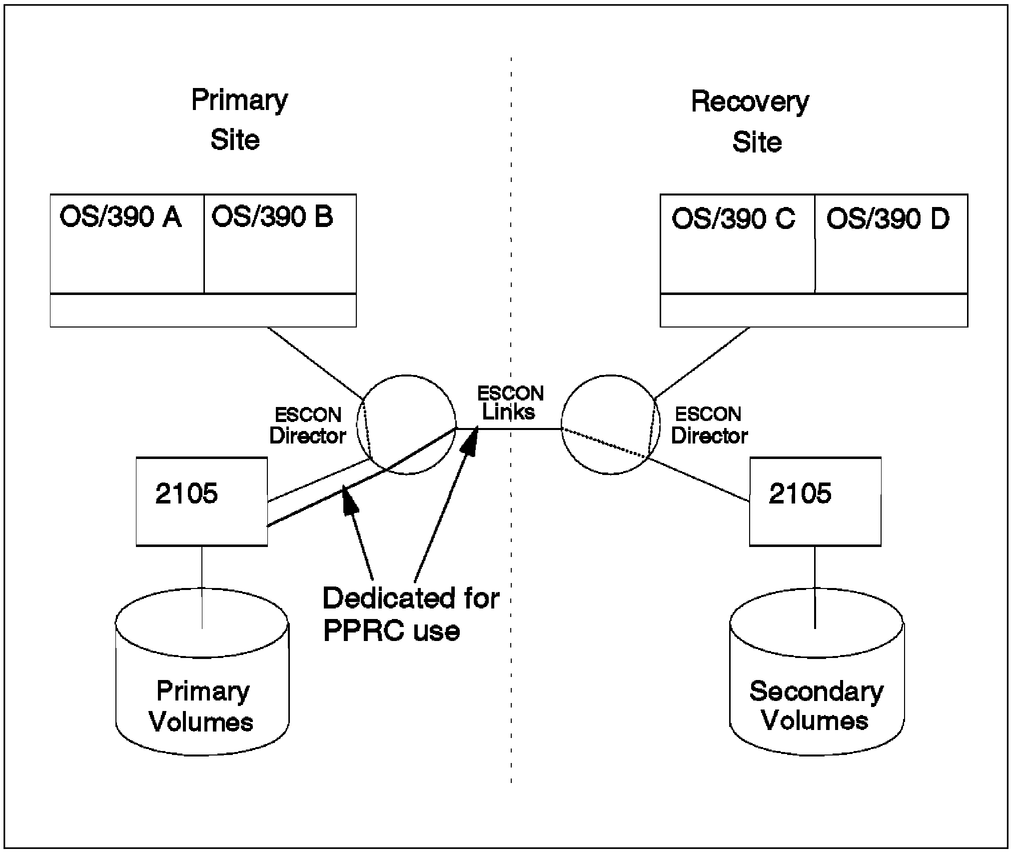


Figure 27. Peer-to-Peer Remote Copy Configuration

The peer-to-peer remote copy implementation requires ESCON links between the primary site DASD controller and the remote (recovery) site DASD controller. These links are provided via:

- Direct ESCON connection between controllers
- Multimode ESCON director connection, as shown in Figure 27
- XDF ESCON director connection
- 9036 ESCON extender connection

The mode of connection determines the distance limit of the secondary controller. Up to 103 km is available with the Enterprise Storage Server. At the time of writing, Enterprise Storage Server support for PPRC is planned to be available in the Fall of 2000.

Peer-to-peer dynamic address switching (P/DAS) provides the mechanism to allow switching from the primary to the secondary controller.

Extended Remote Copy: XRC provides a mechanism for *asynchronous* copying of data to the remote site. Only data that is in transit between the failed application system and the recovery site is lost in the event of a failure at the primary site. Note that the delay in transmitting the data from the primary subsystem to the recovery subsystem is usually measured in seconds.

Figure 28 on page 132 shows a sample extended remote copy configuration.

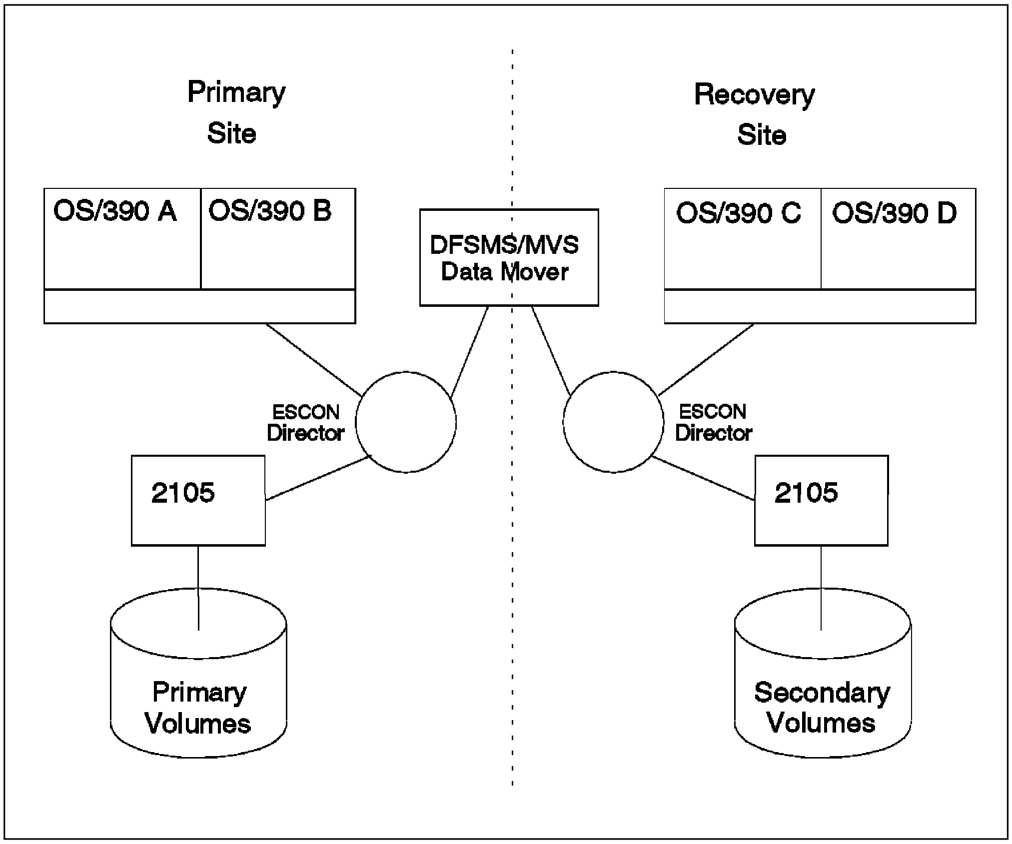


Figure 28. Extended Remote Copy Configuration

The extended remote copy implementation involves the transfer of data between the primary subsystem and the recovery subsystem. The recovery subsystem must be attached to another OS/390 image running the required level of DFSMS and can exist at the primary site, at the recovery site, or anywhere in between. Figure 28 shows an independently sited data mover at ESCON distance limits. However, XRC sites can be separated by distances greater than those supported by ESCON, with the use of channel extenders and high speed data links.

Mixing Remote Copy Solutions

Each remote copy solution uniquely addresses data sequence consistency for the secondary volumes. Combining the techniques may lead to unpredictable results, exposing you to data integrity problems. This situation applies to PPRC, XRC, and specific database solutions such as IMS RSR.

See 3.7.5, "IMS Disaster Recovery Considerations" on page 134 for an overview of IMS RSR.

The DASD Remote Copy solutions are data-independent; that is, beyond the performance considerations, there is no restriction on the data that is mirrored at a remote site using these solutions.

A complete description of PPRC and XRC functions, configuration, and implementation can be found in the IBM Redbooks *Planning for IBM Remote Copy*, SG24-2595, and *Implementing ESS Copy Services on S/390*, SG24-5680.

3.7.3 CICS Disaster Recovery Considerations

The way in which recovery data is provided to the secondary site determines what is required in terms of CICS infrastructure copies. For example, if recovery data is provided from offsite backup data, backup copies of VSAM files and a copy of the CICS infrastructure are required. If full remote DASD mirroring is in place, CICS journaling must be duplexed to the CF and LOGR staging data sets; the CICS/VSAM RLS data sharing group must be identical to that of the production site. VSAM files, system logger infrastructure, CICS/VSAM RLS infrastructure, and CICS infrastructure must also be copied to the secondary site.

For a full discussion on the options available for CICS disaster recovery, refer to the IBM Redbook *Planning for CICS Continuous Availability in an MVS/ESA Environment*, SG24-4593.

3.7.4 DB2 Disaster Recovery Considerations

Be aware that if you implement a DB2 data sharing group, your disaster recovery site *must* be able to support the same DB2 data sharing group configuration as your main site. It must have the same group name, the same number of members, and the names of the members must be the same. Additionally, the structure names in the CFRM policy must be the same (although the sizes can be different).

The hardware configuration, however, *can* be different. For example, your main site could be a multisystem data sharing group spread among several CPCs, with CFs and Sysplex Timers. Your disaster recovery site could be a large single OS/390 image, which could run *all* of the DB2 subsystems in the data sharing group.

Since some of the benefits of the Parallel Sysplex are lost by bringing the subsystems under one OS/390, all but one of the members could be stopped once the data sharing group has been started on the recovery site. Figure 29 on page 134 shows this configuration.

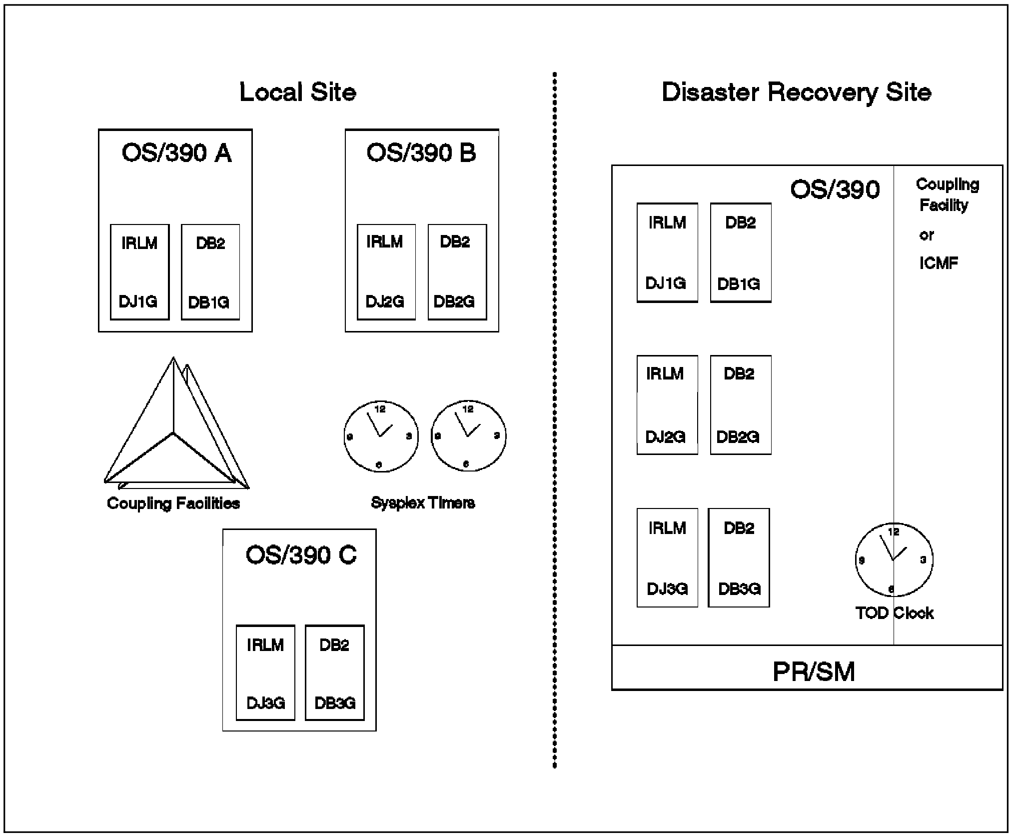


Figure 29. Example Configurations for DB2 Data Sharing Groups

The Remote Tracker facility was introduced in DB2 V5.1 by APAR PQ09972. It provides the ability to have a remote DB2 (tracker site) which is a close-to-current shadow copy of the DB2 at the local site: The remote DB2 is kept synchronized with the changes to the local DB2 by applying logs to copies of the catalog, directory, and user data. More information about the Remote Tracker facility can be found in the IBM Redbook *DB2 UDB for OS/390 and Continuous Availability*, SG24-5486.

Additional details about disaster recovery considerations for DB2 data sharing can be found in: *DB2 UDB for OS/390 V6 Data Sharing: Planning and Administration*, SC26-9007.

3.7.5 IMS Disaster Recovery Considerations

IMS was the first IBM S/390 program product to exploit *hot standby*. Using the extended recovery facility (XRF), one IMS system shadows another and takes over within seconds if the active system fails. With XRF, a number of IMS TM installations worldwide have achieved continuous service availability for more than 1000 days.

XRF is not generally considered a disaster recovery function. When data sharing is used, an IMS XRF active and its alternate must be in the same Parallel Sysplex. XRF is a recovery function for failures of IMS, OS/390, and CPCs.

With the introduction of IMS 5.1, the concept of a standby system was extended from a local perspective to remote site recovery (RSR), which provides system-managed transfer and application of log data to shadow databases at the remote site. A key feature of RSR is that the remote databases are assured of consistency and integrity. In the event of a site failure, the remote system can be restarted as the production system within minutes.

RSR offers two levels of support, selected on an individual database basis:

- Database Level Tracking

  With this level of support, the database is shadowed at the remote site, thus eliminating database recovery in the event of a primary site outage.

- Recovery Level Tracking

  No database shadowing is provided with this option. The logs are transmitted electronically to the remote site and are used to recover the databases by applying forward recovery to image copies in the event of an outage at the primary site.

RSR supports the recovery of IMS full-function databases, Fast Path DEDBs, IMS message queues, and the telecommunications network. Some of the facilities introduced in IMS/ESA V6 do have an impact on XRF and RSR usage. Note that:

- IMS XRF systems cannot participate as members of a generic resource group. However, XRF systems and generic resource members can connect to the same shared message queue group.
- IMS RSR systems cannot participate as members of a generic resource group. However, RSR systems and generic resource members can connect to the same shared message queues group.
- Subsystems that use XRF cannot use Fast DB Recovery.
- A DBCTL system with an alternate standby configuration cannot use Fast DB Recovery.

Additional details are found in *IMS/ESA V6 Administration Guide: System*, SC26-8730, and *IMS/ESA V5 Administration Guide: System*, SC26-8013.

As with DB2, if you have implemented data sharing at the primary site, then the remote recovery site must also be a Parallel Sysplex environment, and the IMS data sharing group must be identical.

3.7.6 Geographically Dispersed Parallel Sysplex Offering

GDPS is a multisite management facility that is a combination of system code and automation that utilizes the capabilities of Parallel Sysplex technology, storage subsystem mirroring and databases to manage processors, storage and network resources. It is designed to minimize and potentially eliminate the impact of a disaster or planned site outage. It provides the ability to perform a controlled site switch for both planned and unplanned site outages, with no data loss, maintaining full data integrity across multiple volumes and storage subsystems and the ability to perform a normal DBMS restart (not DBMS recovery) at the opposite site.

At one installation a simulated site disaster at one site caused no data loss and the recovery window to the other site was reduced from 12 hours to 22 minutes. Additionally, a user-defined planned site switch from one of the sites to the second site took 42 minutes.

3.7.6.1 What Is GDPS

GDPS is a multisite application availability solution which provides the capability to manage the remote copy configuration and storage subsystem(s), and automates Parallel Sysplex operational tasks, and failure recovery from a single point of control, thereby improving application availability.

GDPS provides the ability to perform a controlled site switch for both planned and unplanned site outages, with no data loss, thus providing full data integrity. The amount of data loss with traditional Disaster Recovery (D/R) methods (volume dumping, ABARS) is typically 24 to 48 hours. GDPS is application-independent and, therefore, covers the installation's complete application environment. Note that solutions such as IMS Remote Site Recovery are very effective, but applicable to IMS applications only. No other D/R solution is as complete as GDPS.

GDPS supports all transaction managers (e.g., CICS TS and IMS) and database managers (e.g., DB2, IMS, and VSAM), and is enabled by means of key IBM technologies:

- Parallel Sysplex
- Systems Automation for OS/390
- RAMAC Virtual Array DASD
- PPRC (Peer-to-Peer Remote Copy)
- Dense Wavelength Division Multiplexer

The physical topology of a GDPS consists of a base or Parallel Sysplex cluster spread across two sites known as Site 1 and Site 2 separated by up to 40 kilometers (approximately 25 miles) with one or more OS/390 systems at each site. The multisite Parallel Sysplex cluster must be configured with redundant hardware (for example, a Coupling Facility and a Sysplex Timer in each site) and the cross-site connections must be redundant. Figure 30 on page 137 shows an example of a GDPS.

All critical data resides on storage subsystem(s) in Site 1 (the primary copy of data) and is mirrored to Site 2 (the secondary copy of data) via PPRC synchronous remote copy.

GDPS consists of production systems, standby systems and controlling systems. The production systems execute the mission-critical workload. The standby systems normally run expendable work which will be displaced to provide processing resources when a production system or a site is unavailable. There must be sufficient processing resource capacity available, such as processor capacity, central and expanded storage, and channel paths, that can quickly be brought online to restart a system's or site's critical workload (typically by terminating one or more systems executing expendable (non-critical) work and acquiring its processing resource). The controlling system coordinates GDPS processing. By convention, all GDPS functions are initiated and coordinated by one controlling system.

All GDPS systems run GDPS automation based upon Tivoli NetView for OS/390 and System Automation for OS/390. Each system will monitor the Parallel Sysplex cluster, Coupling Facilities, and storage subsystems and maintain GDPS status. GDPS automation can coexist with an enterprise's existing automation product.

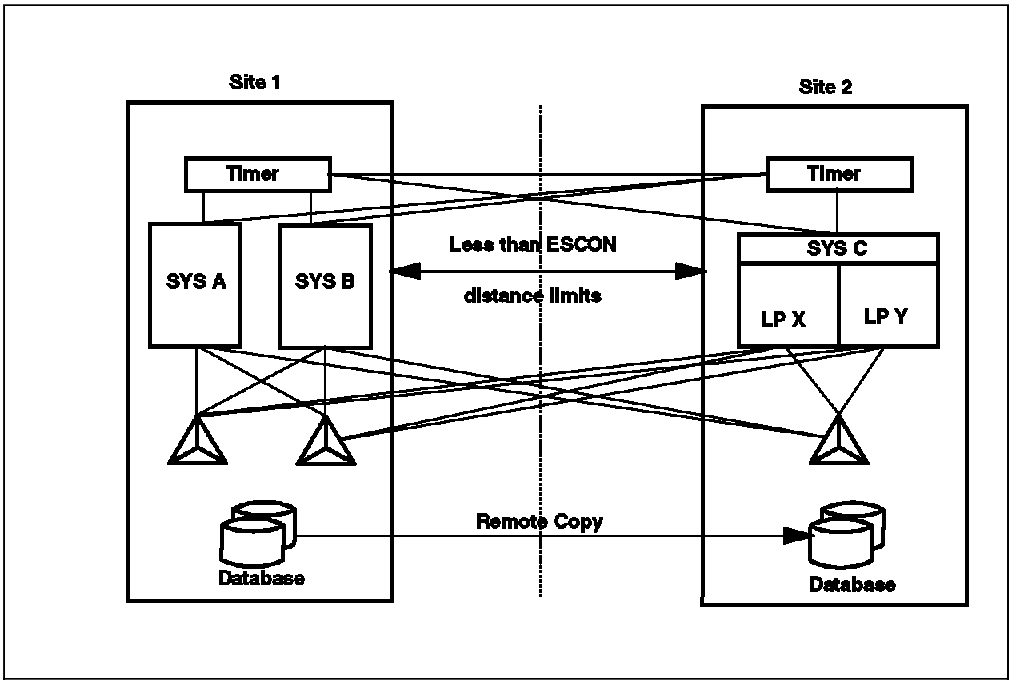


Figure 30. Example of Geographically Dispersed Parallel Sysplex (GDPS)

Planned Reconfigurations: GDPS planned reconfiguration support automates procedures performed by an operations center. These include standard actions to:

- Quiesce a system's workload and remove the system from the Parallel Sysplex cluster (e.g., stop the system prior to a change window);
- IPL a system (e.g., start the system after a change window); and
- Quiesce a system's workload, remove the system from the Parallel Sysplex cluster, and re-IPL the system (e.g., recycle a system to pick up software maintenance).

The standard actions can be initiated against a single system or a group of systems. Additionally, user-defined actions are supported (e.g., planned site switch in which workload is switched from processors in Site 1 to processors in Site 2).

Unplanned Reconfigurations: GDPS unplanned reconfiguration support not only automates procedures to handle site failures, but will also minimize the impact and potentially mask an OS/390 system or processor failure based upon the GDPS policy. If an OS/390 system fails, the failed system will automatically be removed from the Parallel Sysplex cluster, re-IPLed in the same location, and the workload restarted. If a processor fails, the failed system(s) will be removed from the Parallel Sysplex cluster, re-IPLed on another processor, and the workload restarted.

3.7.6.2 IBM Global Services Offerings

There are three GDPS services offered by IBM Global Services:

1. Remote Copy Management Facility (RCMF)

   The RCMF code and panels to manage the remote copy infrastructure will be installed, customized, and verified along with providing operational education for the enterprise.

2. Parallel Sysplex Management Facility (PSMF)

   The PSMF automation to manage Parallel Sysplex operations at a site will be installed, the automation policy customized, and the automation verified along with providing operational education for the enterprise.

3. GDPS

   Initially a planning session will be held to understand the enterprise's application availability and disaster recovery objectives. Once the objectives are understood, GDPS automation to manage the remote copy infrastructure, perform routine tasks, and recover from failures will be installed, the automation policy customized, and the automation verified along with providing operational education.

3.7.6.3 Pre-Requisites

For IBM to perform these services, you must have the following elements:

- OS/390 V1 or later
- System Automation for OS/390 1.2 or later
- Tivoli NetView for OS/390
- Site Monitor 3.18 (including HMC Bridge), Site Manager 2.14 (including HMC Bridge), or Availability and Operations Manager Multisite sysplex (Base or Parallel)
- Storage subsystems with PPRC

Note: IBM plans to add XRC support of the asynchronous XRC environment. GDPS/XRC, based upon the open IBM Extended Remote Copy (XRC) Architecture, is intended to address the requirements of enterprises that cannot use GDPS/PPRC—that is, GDPS/XRC is targeted for enterprises that:

- Can't tolerate the PPRC overhead, can't get dedicated fiber connections between their sites, or have two sites separated by greater more than 40 km.
- Can tolerate some data loss, and
- Can tolerate a failover instead of near-transparent disaster recovery

XRC can support unlimited distances, but in this case, you have a sysplex in Site 1 and a different sysplex in Site 2. The sysplex does not span sites (due to the current sysplex distance limitation of 40 km).

3.7.7 More Information on GDPS

For IBM Installation Services for Geographically Dispersed Parallel Sysplex and a more detailed GDPS white paper refer to:
`http://www.as.ibm.com/asww/offerings/mww62b1.htm`

For an overview of how planning, implementation and proper exploitation of S/390 Parallel Sysplex clustering can enable your business to achieve near continuous availability refer to "Five Nines / Five Minutes Achieving Near Continuous Availability" at:
`http://www.s390.ibm.com/marketing/390avail.html`

3.7.8 Recommended Sources of Disaster Recovery Information

The following is a list of books that provide valuable information for many aspects of disaster recovery.

- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Configuration Guidelines*, SG24-4927
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Presentation Guide*, SG24-4911
- *DB2 UDB For OS/390 and Continuous Availability*, SG24-5486

Chapter 4. Workloads in Parallel Sysplex

In this chapter, we review the important workload considerations for the Parallel Sysplex environment. We examine the most common workload types and look at the characteristics of their exploitation of the Parallel Sysplex. Also, where possible, any *changes to your configuration* caused or recommended by the implementation of a Parallel Sysplex are highlighted.

If you are running a Parallel Sysplex environment consisting of differing CPC types (for example, IBM 9672 G4 and IBM 9672 G6), consideration should also be given to workloads that require specific hardware features. See Table 36 on page 233 in Volume 2: Cookbook, SG24-5638 for a complete list of functional differences. Care must be taken to ensure that workloads dependent on these features are executed only on hardware platforms that support them.

Sizing CF Structures

A Web-based tool to help you determine the size for all your CF structures is available at:
`http://www.s390.ibm.com/cfsizer/`

The tool is called the CF Sizer and is designed to help you quickly determine appropriate sizes for any of your CF structures. The tool has been deliberately kept simple, to make it as easy to use as possible, and to minimize the amount of data you need to be able to use it. In addition to providing recommended structure sizes, the tool also provides IXCMIAPU statements that can be tailored to your configuration and used to help set up the structures.

Recommended Sources of Further Information

The following sources provide support for the information in this chapter:

- *A Performance Study of Web Access to CICS*, SG24-5748
- *Batch Processing in a Parallel Sysplex*, SG24-5329
- *CICS and VSAM RLS: Implementation Guide*, SG24-4766
- *CICS and VSAM RLS: Planning Guide*, SG24-4765
- *CICS and VSAM RLS: Recovery Considerations*, SG24-4768
- *CICS TS for OS/390: V1.3 Implementation Guide*, SG24-5274
- *CICS TS for OS/390 Intercommunication Guide*, SC33-1695
- *CICS TS for OS/390 Migration Guide*, GC34-5353
- *CICS TS for OS/390 Release Guide*, GC33-1570
- *CICS TS for OS/390 and Parallel Sysplex*, GC33-1180
- *CICS TS: Web Interface and 3270 Bridge*, SG24-5243
- *CICSPlex Systems Manager Business Application Services: A New Solution to CICS Resource Management*, SG24-5267
- *CICS Transaction Server for OS/390 Planning for Installation*, GC33-1789
- *CICS Workload Management Using CICSPlex SM and the MVS/ESA Workload Manager*, GG24-4286
- *CICSPlex Systems Manager V1 Concepts and Planning*, GC33-0786
- *DB2 on the MVS Platform: Data Sharing Recovery*, SG24-2218
- *DB2 UDB for OS/390 V6 Data Sharing: Planning and Administration*, SC26-9007
- *DB2 UDB for OS/390 V6 Release Guide*, SC26-9013
- *DB2 UDB for OS/390 V7 Release Guide* (available on the Web)
- *Host on-Demand 4.0*, SG24-5237
- *IMS/ESA Shared Queues: A Planning Guide*, SG24-5257
- *IMS/ESA V6 Shared Queues*, SG24-5088
- *IMS/ESA Sysplex Data Sharing: An Implementation Case Study*, SG24-4831
- *IMS/ESA Data Sharing in a Parallel Sysplex*, SG24-4303
- *IMS/ESA Multiple Systems Coupling in a Parallel Sysplex*, SG24-4750
- *IMS/ESA V6 Parallel Sysplex Migration Planning Guide*, SG24-5461
- *IMS/ESA V6 Administration Guide: System*, SC26-8730
- *IMS Version 7 Release Guide*, SG24-5753
- *IMS e-business Connect Using the IMS Connectors*, SG24-5427
- *JES2 Multi-Access Spool in a Sysplex Environment*, GG66-3263
- *JES3 in a Parallel Sysplex*, SG24-4776
- *OS/390 MVS Multisystem Consoles Implementing Sysplex Operations*, SG24-4626
- *OS/390 MVS Parallel Sysplex Test Report*, GC28-1963
- *OS/390 MVS Planning: APPC/MVS Management*, GC28-1807
- *OS/390 MVS Parallel Sysplex Application Migration*, GC28-1863
- *OS/390 R4 Implementation*, SG24-2089
- *OS/390 R5 Implementation*, SG24-5151
- *Revealed! CICS Transaction Gateway and More CICS Clients Unmasked*, SG24-5277
- *Revealed! Architecting Web Access to CICS*, SG24-5466
- *Securing Web Access to CICS*, SG24-5756
- *S/390 MVS Parallel Sysplex Migration Paths*, SG24-2502
- *SNA in a Parallel Sysplex*, SG24-2113
- *TCP/IP in a Sysplex*, SG24-5235
- *Web-to-Host Integration Solutions*, SG24-5237

Table 11. Workloads in Parallel Sysplex Roadmap			
You want to configure:	**If you are especially interested in:**	**Subtopic of interest:**	**Then refer to:**
Workloads in a Parallel Sysplex			
	e-business workloads		4.1, "e-business and Parallel Sysplex" on page 143
		HTML Serving	4.1.2, "HTML Serving" on page 147
		3270 emulation	4.1.3, "3270 Emulation" on page 147
		CICS	4.1.4, "CICS" on page 148
		DB2	4.1.5, "DB2" on page 156
		IMS	4.1.6, "IMS" on page 160
	Transaction management		4.2, "Transaction Management in Parallel Sysplex" on page 169
		CICS	4.2.2, "CICS Transactions in a Parallel Sysplex" on page 169
		IMS TM	4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187
	Database management		4.3, "Database Management in Parallel Sysplex" on page 194
		DB2	4.3.1, "DB2 Data Sharing Considerations" on page 194
		IMS DB	4.3.2.1, "IMS DB Data Sharing in a Parallel Sysplex" on page 203
		VSAM RLS	4.3.3, "CICS/VSAM Record Level Sharing Considerations" on page 207
	Batch processing		4.4, "Batch Workload Considerations" on page 213
		JES2	4.4.1, "JES2 Considerations in Parallel Sysplex" on page 213
		JES3	4.4.2, "JES3 Considerations in Parallel Sysplex" on page 214
		Workload balancing	4.4.4, "Batch Workload Balancing and Parallel Sysplex" on page 215
	Network access		4.5, "Network Workload Balancing Capabilities" on page 219
		SNA applications	4.5.1.2, "VTAM Workload Balancing for Generic Resources" on page 221
		TCP applications	4.5.2, "TCP Workload Balancing" on page 225
	Other workloads		
		Testing	4.8, "Test Considerations in Parallel Sysplex" on page 228
	Applications		4.9, "How to Select Applications to Exploit Parallel Sysplex" on page 231

4.1 e-business and Parallel Sysplex

My, how quickly things are changing in our industry. When the first version of this book was written, few people had heard of the World Wide Web, and even fewer had access to it. Now, www.something is all-pervasive. You hardly see an advertisement that does not have the company's URL at the bottom of the ad, and you take it as a given that you can sit at a Web browser, enter "www.widgets.com," and be taken to the Widgets company Web site.

The role that Parallel Sysplex plays in e-business is also rapidly evolving. When the Internet started merging with the business world, the majority of Web interactions involved downloading static HTML pages. In that environment, and given the level of HFS sharing available at the time, Parallel Sysplex was not seen as a big player in the Internet world.

Since then, things have changed dramatically. No longer is the Web used simply as a marketing medium; it is now the way companies actually transact business. Interactions might still include HTML pages, but they also include a lot more, like talking to backend CICS, DB2, and IMS systems. Companies are leveraging their existing applications and data to enable real e-business, in a much more efficient manner, and with a worldwide customer base.

Together with this move to e-business comes a requirement for much higher application availability. If your company cannot respond to a customer request, that customer can now be in your competitor's “shop” about 10 seconds later—meaning that lack of availability really does mean quantifiable lost business.

For these reasons, customers doing serious e-business are looking more and more at Parallel Sysplex as a way of meeting the divergent requirements of:

- Maintaining near-continuous application availability.
- Still having to process the traditional workloads, including batch jobs and data backups.
- Being able to keep the system software up to date, in order to keep pace with competitors.

The software in this area is evolving so rapidly that it is just about impossible to document it in a static publication like this. Hardly a month goes by without a product announcement that adds new or more flexible features. Therefore, in this section, we will only briefly discuss the options currently available, and refer you to documents that discuss these options in far greater detail.

4.1.1 Web Server

For most Web transactions, the primary means of communication will be via the Web Server, either running on OS/390 or on another platform. Starting with the replacement for the Lotus Domino Go Webserver (DGW), the Web Server was actually broken into two components. The Java servlet support was moved into a product called WebSphere Application Server (WAS). The remaining functions from DGW were provided by a new product called IBM HTTP Server of OS/390.

Table 12 contains a summary of the different generations of OS/390 Web Servers and the related OS/390 releases.

Table 12 (Page 1 of 2). OS/390 Web Servers

Server product name	**Related OS/390 release**
IBM Internet Connection Server (ICS) for MVS/ESA	MVS/ESA 5.2.2-OS/390 1.1
IBM Internet Connection Secure Server (ICSS) for OS/390 V2R1	OS/390 1.2
IBM Internet Connection Secure Server (ICSS) for OS/390 V2R2	OS/390 1.3
Lotus Domino Go Webserver (DGW) for OS/390 Release 4.6.0	OS/390 2.4

Table 12 (Page 2 of 2). OS/390 Web Servers	
Server product name	**Related OS/390 release**
Lotus Domino Go Webserver (DGW) for OS/390 Release 4.6.1	OS/390 2.5
Lotus Domino Go Webserver (DGW) for OS/390 Release 5.0	OS/390 2.6
IBM HTTP Server (IHS) 5.1 for OS/390 R7	OS/390 2.7
IBM HTTP Server (IHS) 5.2 for OS/390 R8	OS/390 2.8-2.10
Note: Fixes may be available to allow these products run on earlier levels of OS/390 than indicated in this table.	

The latest release of WAS available at the time of writing introduces support for storing Java servlet session state data in DB2. In a Parallel Sysplex environment, this means that the session could actually move from one Web Server to another, with each server being able to access the session state information; this provides improved flexibility and availability for Java servlet applications running in a Parallel Sysplex.

Among many other features, the Web Server provides support for a feature known as Scalable Web Serving. This support was added as part of ICSS 2.2. Using this support, one Web Server will take responsibility for accepting all incoming requests. Based on the number of requests awaiting execution and whether the Web Server is meeting its performance objectives, WLM may start additional Web Server or WAS address spaces. Figure 31 contains a sample configuration showing the use of Network Dispatcher with the Scalable Web Server. Network Dispatcher is discussed in 2.3.2.6, "Interactive Network Dispatcher" on page 52 in Volume 3: Connectivity, SG24-5639.

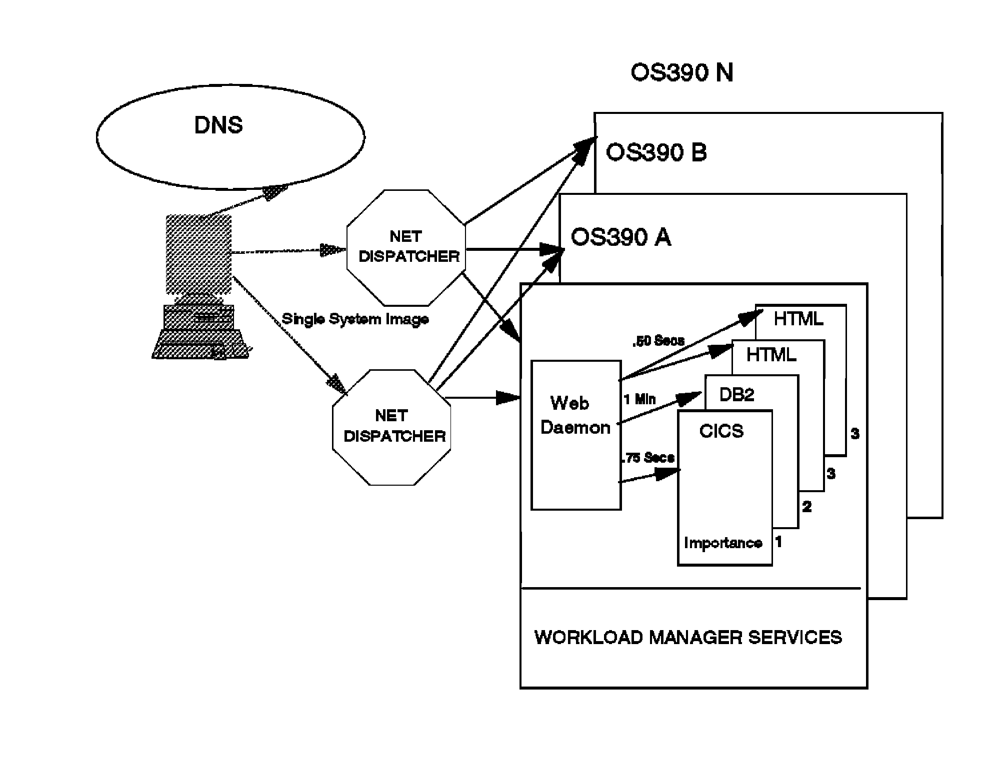


Figure 31. Sample Scalable Web Server Configuration

The combination of Network Dispatcher and the Scalable Web Server not only improves performance for the Web requests, it also provides improved availability by letting you spread the processing of Web requests across multiple server address spaces in multiple images.

More information about how to set up the Scalable Web Server can be found in the IBM Redbook *OS/390 e-business Infrastructure: IBM HTTP Server 5.1 - Customization and Usage*, SG24-5603.

The Web Server can be used to:

- Serve HTML pages
- Communicate with CICS
- Communicate with DB2
- Communicate with IMS

One of the first things to consider when selecting a Web-enabling solution is what business process or processes you wish to make available to the Web. This will help to determine the required data sources and the preferred Web solution architecture.

If the requirement is simply to make existing 3270 applications available to Web users, then a Java 3270 emulator may be the quickest and best solution. If the requirement is to enhance the user interface, then there are several options that we will list that should be assessed to determine if they meet the user's needs.

If an entirely new application is required in order to re-engineer existing business processes, the WebSphere development environment would be the best choice, with VisualAge for Java and the various Connectors for access to applications and data.

To help you identify the most appropriate choices, you should use the "OS/390 Planning Assistant for e-business" located at:
`http://www.s390.ibm.com/os390/wizards/ebiz/`

The next thing to consider, from a Parallel Sysplex point of view, is the availability characteristics of each option.

- If the solution you choose includes a Web Server or some other intermediate address space (which we will generically call the front-end), that address space should support workload balancing.
- You also have to consider if it is necessary for the front-end to reside on the same system as your backend application manager (CICS, DB2, or IMS).
- You need to decide on the location of the Web Server. Are you going to have a two-tier or a three-tier architecture?
- You have to consider the profile of the transactions. Is data carried from one transaction to another? If so, you either have to make sure that every part of the transaction is executed on the same server(s), or you need to use a solution that will make this transient data available across multiple servers.
- Finally, you should check whether the front-end can communicate with any instance of the application manager, or if it relies on a hardcoded applid or subsystem name.

In the following sections, we provide this information for each of the listed solutions.

4.1.2 HTML Serving

In the past, a possible inhibitor to exploiting a Parallel Sysplex as a Web Server was the lack of shared read/write support for the Hierarchical File System (HFS). Prior to OS/390 V2R9, if you wished to share a HFS between more than one OS/390 image, *all* of the images had to have the HFS mounted in read-only mode. This made it difficult to schedule updates to the HTML pages stored in the HFS while still maintaining the service.

This inhibitor has been removed in OS/390 V2R9. OS/390 V2R9 provides support for multiple OS/390 systems to share a HFS and still have the ability for any of the systems to make updates to the HFS.

As a result, it is now possible to maintain the availability of the Internet service running across a number of OS/390 images (for improved availability) and still be able to make any required changes to the HFS content.

In order to share a file in this manner across systems, all the systems that are sharing the HFS in question must be running OS/390 V2R9 or a later release and must be in the same sysplex. Information on setting up a shared HFS is available in *OS/390 UNIX System Services Planning*, SC28-1890 and in the WSC Flash 10020, entitled *UNIX System Services Performance: Shared HFS*.

To improve the performance for serving HTML pages, a new feature known as Fast Response Cache Accelerator was introduced in OS/390 V2R7. The Fast Response Cache Accelerator uses Adaptive Fast Path Architecture (AFPA) to speed up the processing of static Web pages by the WebSphere Application Server. Web pages are cached within the TCP/IP stack, and requests are handled without traversing the entire kernel or entering the user space. From the perspective of the TCP/IP stack, the Cache Accelerator consists of a set of exits that are executed during stack processing in order to support transient, in-kernel “quick” connections. As a result, the Cache Accelerator executes as an extension of the TCP/IP stack rather than as a functional layer above the TCP/IP stack.

The technology developed for the High Speed Web Access facility, introduced in OS/390 V2R5 to improve performance for Web serving, was integrated into the level of TCP/IP delivered with OS/390 V2R6.

4.1.3 3270 Emulation

If your requirement is to give Internet users access to existing 3270 applications, the easiest alternative might be a simple 3270 emulator. IBM's SecureWay Host On-Demand Version 4 (V4) gives you secure access to your host applications and data, using a simple, Java-enabled Web browser.

To improve usability and increase user productivity, SecureWay Host On-Demand V4 adds several enhancements to earlier versions of this product:

- 3270 Host Graphics
- Hot Standby Load Balancing support (SLP)
- SSL Client Authentication for increased security
- ActiveX Controls

- Ability to share and distribute Host On-Demand session configurations
- Support for LDAP storage and access to user/group information
- Conversion of IBM Personal Communication's host connection settings

For more information about Host on-Demand, refer to:
`http://www.software.ibm.com/network/hostondemand`

If you wish to provide something a little more appealing than a “green screen,” you may be able to use SecureWay Screen Customizer for Host Integration V1 (5801-AAR). This product works with SecureWay Host On-Demand or SecureWay Personal Communications. It automatically converts “green screens” into customizable GUIs without programming.

The default GUI is automatically presented for each host screen at the client and features automatic menu buttons, defined input fields, and hot spots. However, SecureWay Screen Customizer can also combine information from multiple screens into a single GUI.

The administrator can also customize screens by:

- Adding new objects
- Changing fonts or colors
- Hiding host fields
- Reordering work flow
- Combining data
- Automatically navigating the user through screens

All features added to the host application are done through simple drag and drop techniques. This allows you to produce more user-friendly screens without tying up valuable programming resources.

The Host on-Demand program is downloaded from a Web Server; however, the Web Server does not necessarily have to reside on the same system as the target application, so the client could potentially download the code from any of the available Web Servers. Host on-Demand communicates with a TN3270 server, and from there to the target application using SNA. One of the TCP/IP workload balancing mechanisms can be used to distribute the client connections across several TN3270 servers. For communication between the TN3270 server and the target SNA application, VTAM GR can be used to route the client sessions to any available member of the generic resource group.

4.1.4 CICS

If you are an existing CICS user, and have CICS applications that you would like to exploit as part of a Web transaction, there are many options open to you. The considerations and restrictions relating to each are too complex to discuss in detail here (and they would probably have changed by the time you read this anyway!), so we will just list the strategic options and provide pointers to sources of more detailed information.

When selecting a mechanism that most closely meets your business needs, you should bear two attributes in mind—flexibility and availability. Depending on how your current applications are written, you may have a greater or lesser choice of options. One of the basic principles of modular programming in CICS is that the application logic should be separated from the presentation logic. This is shown in Figure 32 on page 149.

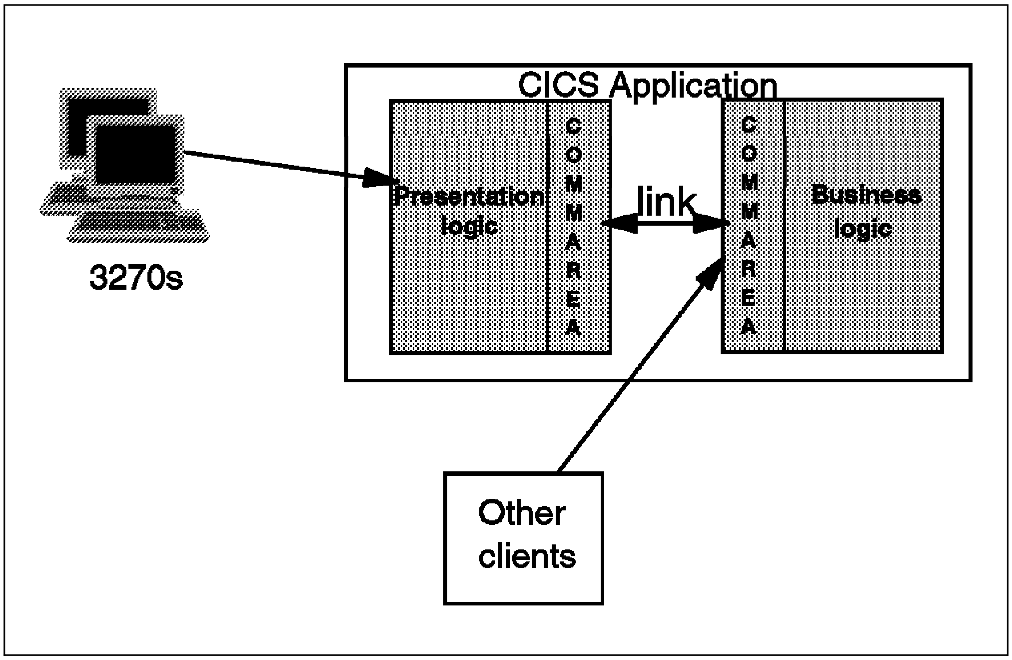


Figure 32. Separation of Application and Presentation Logic in CICS

If your applications are designed in this manner, it should be possible to easily plug in an alternative presentation logic that will use the COMMAREA interface. If you have this flexibility, you then select the mechanism that provides the level of availability you require.

If your applications do not separate the application and presentation logic, and you do not wish to change the applications, then you must use a mechanism that supports the 3270 interface.

If you are designing a brand-new CICS application, you may wish to use the CICS CORBA client support.

Table 13 on page 150 contains a summary of the mechanisms that are available for each of the three different connection options (Distributed Web Server, Web Server on S/390, and direct connection (no Web Server)).

Table 13. CICS Web Mechanisms		
Connection method	**Application interface**	**Applicable mechanisms**
Distributed Web server	COMMAREA	ECI Interface of CICS Transaction Gateway
	3270	EPI Interface of CICS Transaction Gateway, or Host on-Demand
	CORBA	CICS CORBA client support
OS/390 Web Server	COMMAREA	ECI Interface of CICS Transaction Gateway or CICS WebServer Plugin
	3270	CICS WebServer Plugin with 3270 bridge
	CORBA	CICS CORBA client support
Direct connection to CICS	COMMAREA	CICS Web support
	3270	CICS Web support with 3270 bridge
	CORBA	CICS CORBA client support

Figure 33 contains a (simplified) diagrammatic representation of these options.

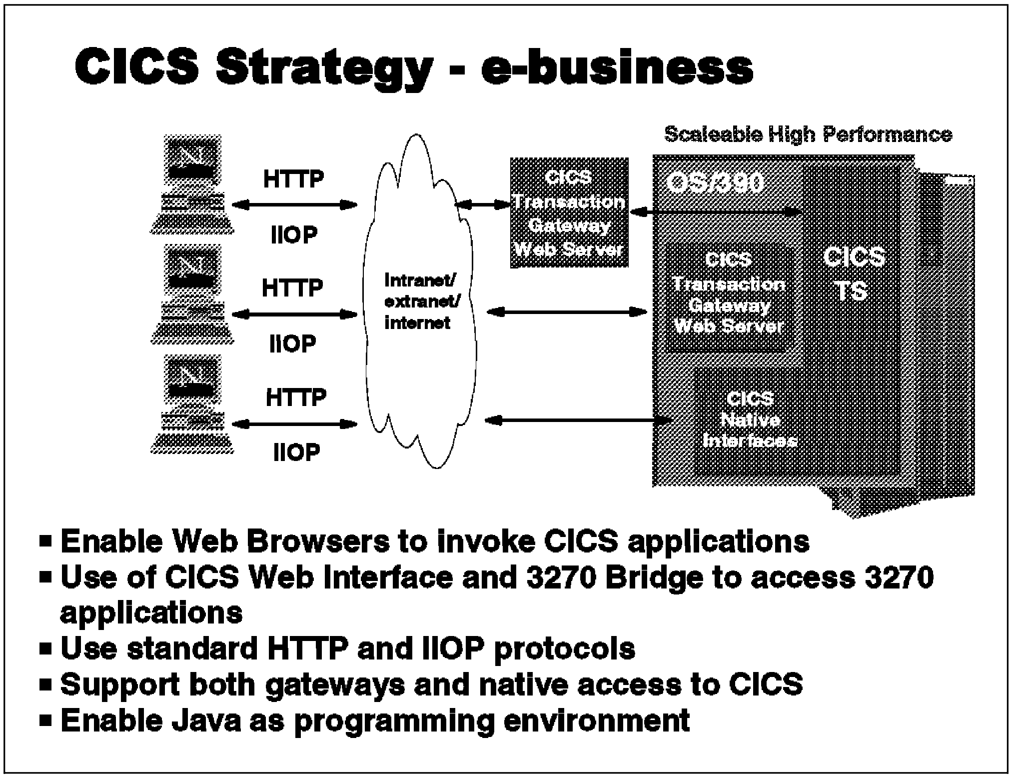


Figure 33. CICS e-business Connectivity Options

There are a number of ways that the actual connection into CICS can be achieved:

EXCI The External CICS Interface (EXCI) is used by the CICS Transaction Gateway (CTG) and the OS/390 Web Server to communicate with CICS transactions using the COMMAREA or 3270 interfaces when CTG or the Web Server is running in the same sysplex as the target CICS region.

With EXCI, it is not necessary for the program issuing the EXCI call to be in the same system as the CICS region that the call is being made to; however, they must both be in the same sysplex. This is because XCF is used for the communication, and XCF communications are limited to the scope of the sysplex. When CICS comes up, it registers with XCF; as a result, when a task looks for a specific CICS region, XCF knows which system it resides on, and passes the request to that region.

The disadvantage of EXCI is that a specific CICS applid must be specified on the EXCI call. This means that if that CICS region is unavailable, the request will fail. To get around this limitation, you can provide code for the DFHXCURM user replaceable module. The supplied sample allows you to distribute the requests to multiple regions in the same OS/390 system. The e-business Integration Team has modified the supplied sample to add the ability to route the request to one of multiple CICS regions on *multiple* systems. If the selected CICS region is not available, the request is re-driven to another region.

The DFHXCURM provides a crude form of WLM, and can at best do round robin (it has no knowledge of the load on different systems, or the response times being achieved by the various CICS regions). The supplied sample is less than ideal since it removes a region from the round robin table as soon as any failure occurs and doesn't retry until all the regions in the table are marked as having failed. If you have CICSPlex/SM available, it is much better to use that product to do any routing and just use DFHXCURM to provide failover by balancing across two CICS regions.

More information about the e-business Integration Team is available on the Web at:
`http://www.s390.ibm.com:80/nc/techinfo/ebitintr.html`

ECI/EPI The External Call Interface (ECI) and External Presentation Interface (EPI), which are utilized by CTG, use LU6.2 to communicate from the calling routine to the target CICS region. This connection supports the use of VTAM Generic Resources (described in 4.5.1, “VTAM Generic Resources Function” on page 219) in that the initial connection request can get routed to a member of the generic resource group. Subsequent requests will get routed to the same region until that connection is broken, at which point the next request will set up a new connection with the same or a different member of the group. This provides only limited workload balancing, but it does provide improved availability as there is no affinity to a particular CICS region. The LU6.2 connection will remain in place until the last session is terminated.

If an SNA network is not available between CTG and the OS/390 system, TCP62 can be used. TCP62 is a feature of Anynet that allows LU6.2 communications to be carried over a TCP/IP network. The use of VTAM GR is unaffected by whether you use an SNA or an IP network to carry the LU6.2 session.

TCP/IP A third option is to connect directly to CICS using TCP/IP. This option uses a Sockets listener transaction in CICS (not to be confused with the CICS TCP/IP Sockets Interface which provides an application level socket interface to CICS applications). With this mechanism, you can use one of the TCP/IP workload balancing methods to provide independence from the availability of a particular CICS region.

The TCP/IP workload balancing mechanisms are discussed in 2.3.2, "Workload Balancing Enhancements" on page 48 in Volume 3: Connectivity, SG24-5639.

3270 Emulation Host on-Demand is discussed in 4.1.3, "3270 Emulation" on page 147. For access to CICS applications from the Web, Host on Demand provides a 3270 emulator, giving the user access to traditional "green screen" CICS screens with some facilities for tailoring the screens. The latest releases of Host on Demand introduces facilities for transforming the traditional 3270 screens into something more closely resembling a standard HTML screen.

Using a combination of VTAM GR and TCP/IP workload balancing, Host on-Demand supports the ability to distribute user requests across multiple CICS regions in multiple systems.

As CICS was one of the first S/390 subsystems to support connection to the Web, it is probably the one with the largest number of options. As the technology advances, some of the early options get superseded by later ones. In this section we concentrate on the four strategic options available at the time of writing and discuss the availability characteristics of each.

All of these options are discussed in far greater detail in the IBM Redbooks *Revealed! Architecting Web Access to CICS*, SG24-5466, and *Securing Web Access to CICS*, SG24-5756. These books contain very clear descriptions and comparisons of the options and are a very good starting place for those that are new to this area.

Another excellent paper that discusses the pros and cons of the available options is available on the Web at:
`http://www.ibm.com/software/ts/cics/library/whitepapers/cicsweb/`

4.1.4.1 CICS Transaction Gateway (CTG)

The first option is known as CICS Transaction Gateway (CTG), formerly known as the CICS Gateway for Java. CTG provides a comprehensive set of Java-based Web Server facilities for accessing CICS applications from a Web browser. The Java program can be an applet, a servlet, or a custom Java application. However, access to transactions using the 3270 interface is only available if CTG is running on a distributed server.

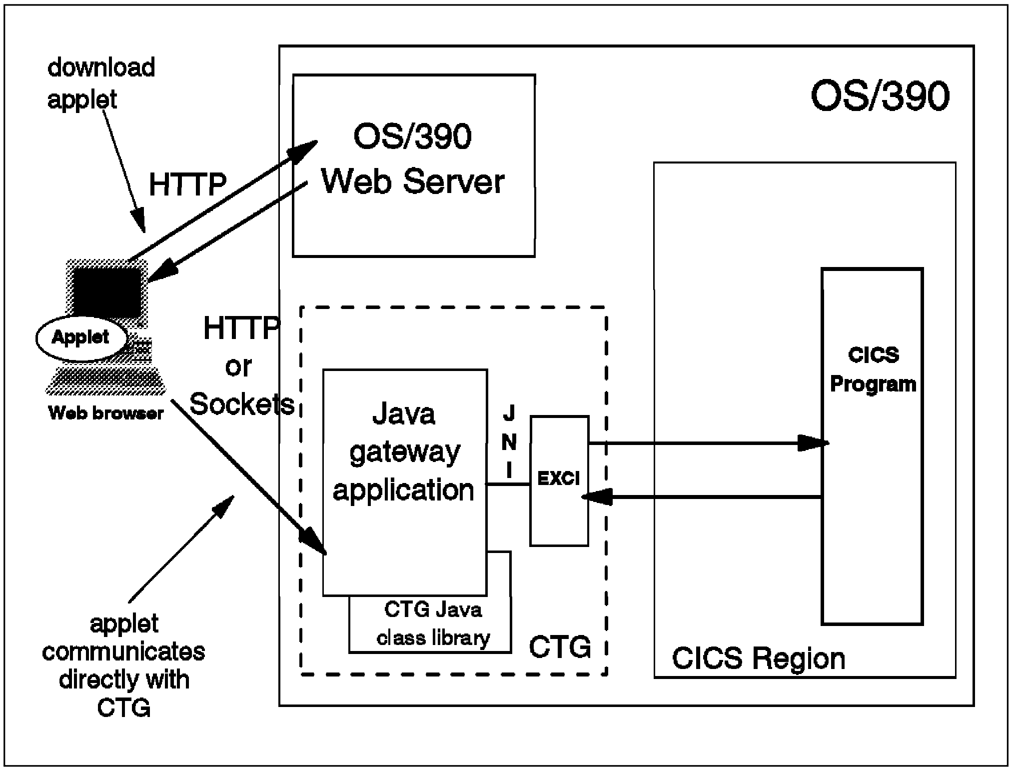


Figure 34. CTG with Java Applet

If you are using a Java applet, the Java program runs on the client and communicates with the CTG address space via HTTP or the sockets interface. CTG in turn communicates with the CICS address space using EXCI. There is an affinity between CTG and the target CICS region for the underlying EXCI request, unless you modify the DFHXCURM exit as described above to route the request to one of several CICS regions.

Similarly, as HTTP or sockets are used to communicate between the client and CTG, one of the TCP/IP workload balancing mechanisms (described in 2.3.2, "Workload Balancing Enhancements" on page 48 in Volume 3: Connectivity, SG24-5639) can be used to distribute the transactions across a number of CTG address spaces. This option is shown in Figure 34 on page 153.

If you are using a Java servlet, the Java program actually runs on the server. In this case, the Java program runs under the control of WebSphere Application Server. The CTG Java gateway application is not usually required when using servlets, since the CTG methods can invoke the CTG local protocol, which uses the JNI to pass the requests onto the underlying EXCI. The EXCI passes the request onto the CICS region, which is not aware that the request comes from a Java application. Once again, as EXCI is being used, the CTG can actually potentially communicate with any CICS region in the CICSplex.

Similarly, as the client communicates with the servlet using HTTP, the client request could be connected to any of a number of WebSphere Application Servers using the provided TCP/IP workload balancing options. This option is shown in Figure 35 on page 154.

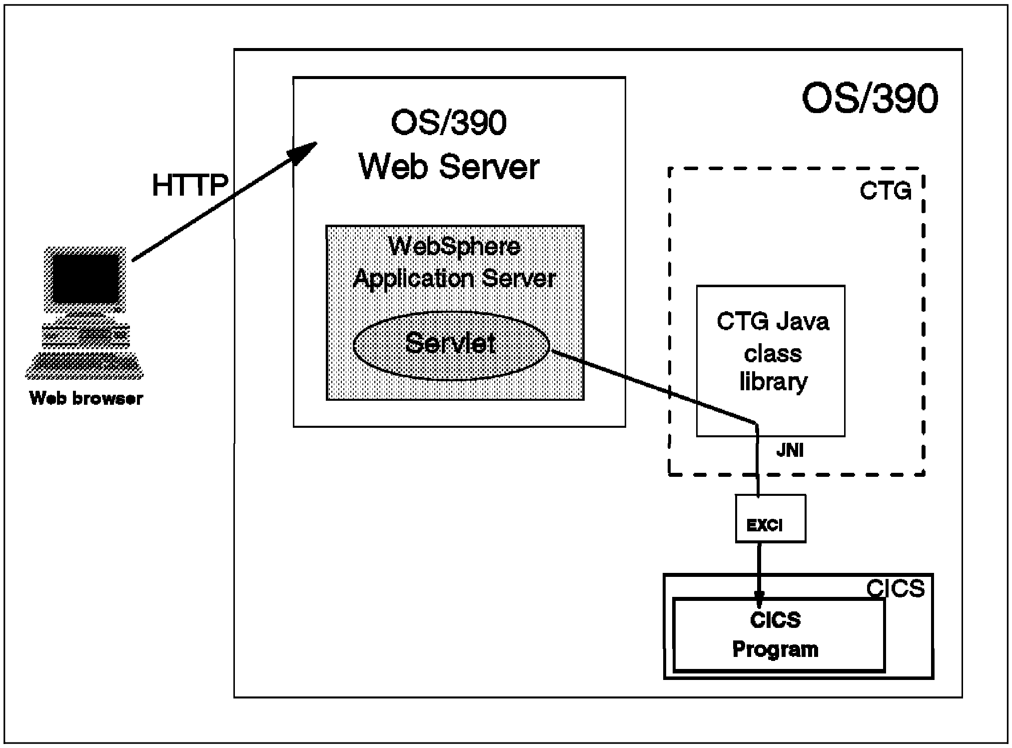


Figure 35. CTG with Java Servlet

Another option, which we do not illustrate here, is to have CTG running on a non-S/390 server. In this case, you can access transactions with both COMMAREA and 3270 interfaces. ECI and EPI are used respectively to communicate between CTG and the target CICS region. CTG uses LU6.2 to communicate with CICS on OS/390, but as noted earlier, either SNA or TCP/IP can be used to carry the communication. This conversation is able to us VTAM GR to target the request to one of a group of CICS regions. As long as the connection is open, all requests from a single client will be routed to the same CICS region. However, if that CICS were to go away, the next request would just get routed to a different member of the generic resource group.

For access to transactions with a 3270 interface, CTG provides a pre-coded servlet called the "terminal servlet." This servlet automatically translates the 3270 screens into a browser equivalent on a one-for-one basis. It is also possible to tailor the interaction with the 3270 transactions, so that a single Web request may initiate more than one CICS transaction, or the output from multiple transactions can be merged to form a single browser screen.

More information about CTG is available in *CICS Transaction Gateway for Administration Guide*, SC34-5528, and in the IBM Redbook *Revealed! CICS Transaction Gateway with More CICS Clients Unmasked*, SG24-5277.

4.1.4.2 CICS Web Support (CWS)

CICS Web support (CWS) is a set of resources supplied with CICS TS V1.3 that provide CICS with some functionality similar to a real Web Server. These provide the ability for a Web browser to access CICS directly, without having to use a separate Web Server. A summary of how a CICS application can be Web-enabled using CWS is shown in Figure 36 on page 155. In CICS Transaction Server V1.3, the CICS Web functionality, previously known as the

CICS Web Interface (CWI), was split into the listener support for TCP/IP and the protocol support for HTTP.

To invoke a 3270 transaction, the facilities of the 3270 bridge are used. The transaction remains unchanged and the 3270 output is converted to HTML.

To invoke a transaction using the COMMAREA, some new application presentation logic must be written. The application is responsible for extracting the required information from the HTTP data stream and building a HTTP data stream upon the completion of the transaction. CICS provides facilities to help with this task.

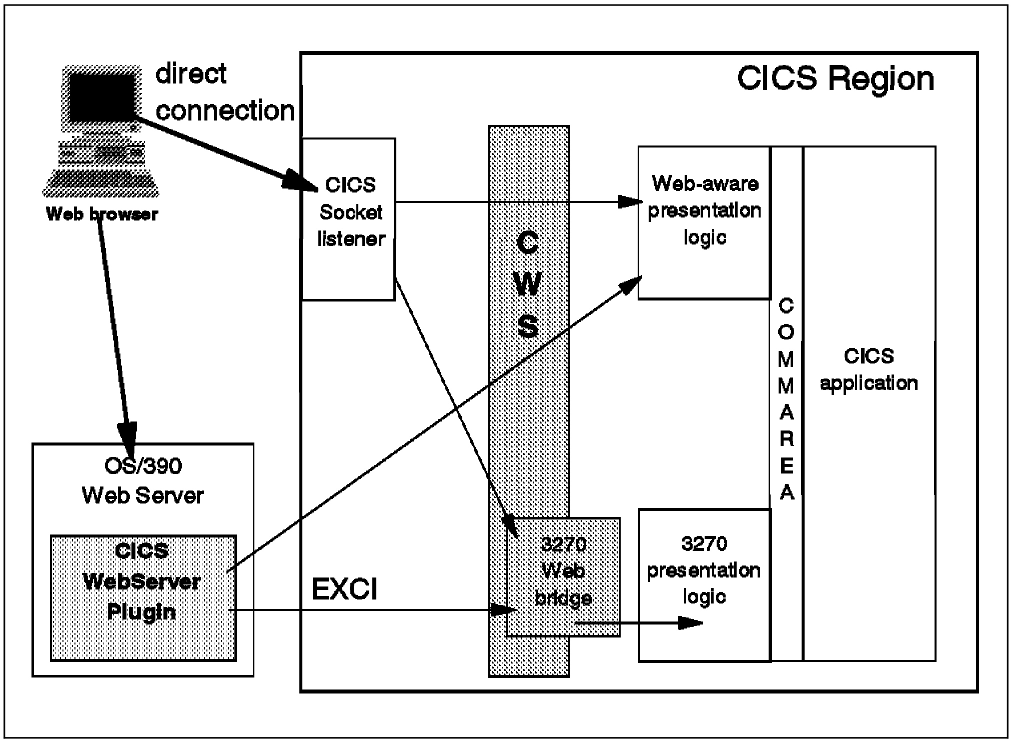


Figure 36. CICS Web Support

Using CWS, there are two ways for the client to access CICS:

1. A direct connection from a Web browser to CICS. This uses the facilities of the CICS Sockets listener to pass the request directly into CICS Web support.
2. Through the OS/390 Web Server using the facilities of the CICS WebServer Plugin. This is a CICS-supplied GWAPI extension to the OS/390 Web Server. It routes requests from the OS/390 Web Server into the CICS region using the EXCI communication mechanism.

In option 1, the client attaches directly to CICS using the sockets interface. Using the workload balancing features in TCP/IP, the client connect request could in fact be routed to any CICS region in the CICSplex that has the Sockets listener enabled, thereby shielding the client from the unavailability of a CICS region.

In option 2, the client communicates with the Web Server using HTTP, and the Web Server uses EXCI to communicate with the target CICS region. The communication with the Web Server can benefit from the TCP/IP workload balancing capabilities, and the communication between the Web Server and

CICS can use XCF and the DFHXCURM exit to potentially communicate with any CICS region in the sysplex.

With CWS, the most significant potential affinity to take note of is the affinity when using a 3270 transaction with CWS and the 3270 Web bridge. In this case, the transaction cannot be transaction-routed to another CICS region. If you are *not* using the bridge, (that is if you are using a Web-aware CICS program), then you can use dynamic DPL to route a LINK from the HTTP presentation logic module to the business logic module. However, any Web API or HTML generation in CICS must be done in the Web Owning Region.

Further information about CICS Web Support may be found in *CICS Internet Guide*, SC34-5445, and the IBM Redbook *CICS TS OS/390 1.3 Web Support and 3270 Bridge*, SG24-5480.

4.1.4.3 CICS CORBA Client Support

The Internet Inter-ORB protocol (IIOP) is an industry standard that defines formats and protocols to provide client/server semantics for distributed object-oriented application programs in a TCP/IP network. It is part of the Common Object Request Broker Architecture (CORBA) specification.

CICS Transaction Server for OS/390 Release 3 provides support for inbound requests to Java application programs, using the IIOP protocol. Execution of Java server programs requires the VisualAge for Java Enterprise ToolKit for OS/390. For information about building Java applications to run in CICS, and the use of the CICS Java classes, see *CICS Application Programming Guide*.

All communications using the CICS CORBA client support use the CICS TCP/IP listener, and therefore can use any of the available TCP/IP workload balancing options.

For more information about this option, see *CICS Internet Guide*, and the IBM Redbook *Java Application Development for CICS*.

4.1.4.4 Other Options

There are a number of other methods of accessing CICS from the Web, including CGI, ICAPI, Sockets, Templates, and others. However, these methods are not strategic and we will not discuss them here.

An *excellent* guide to help you determine the best way of leveraging your existing CICS applications and knowledge in an e-business world is available on the Internet at:
`http://www.ibm.com/software/ts/cics/library/whitepapers/cicsweb`

4.1.5 DB2

Of the three subsystems we discuss in this section, DB2 is the simplest from an e-business access perspective. If you have DB2 and wish to make DB2 data available via the Web, you have four basic options:

- Use CICS
- Use IMS
- Use some form of Java programming
- Use Net.Data
- Use a C/C++ CGI/GWAPI program with WebSphere

If you are going to use CICS or IMS, the e-business considerations for those subsystems are the same regardless of whether the data they are accessing is in DB2, IMS DB, or VSAM. And, because we cover those subsystems in 4.1.4, “CICS” on page 148 and 4.1.6, “IMS” on page 160, we will not discuss them again in this section. Instead, we will concentrate on the Java and Net.Data options.

Both the Net.Data and Java options provide flexibility as to where the “frontend” runs—it can either be in the same OS/390 as the DB2 subsystem resides on, or it can be on another system (S/390 or otherwise). If the frontend runs on the same OS/390 as the DB2 subsystem you wish to access, then local DB2 access is used to communicate with DB2. If the frontend runs on another system, then DRDA must be used to communicate with DB2. Figure 37 shows these two options diagrammatically.

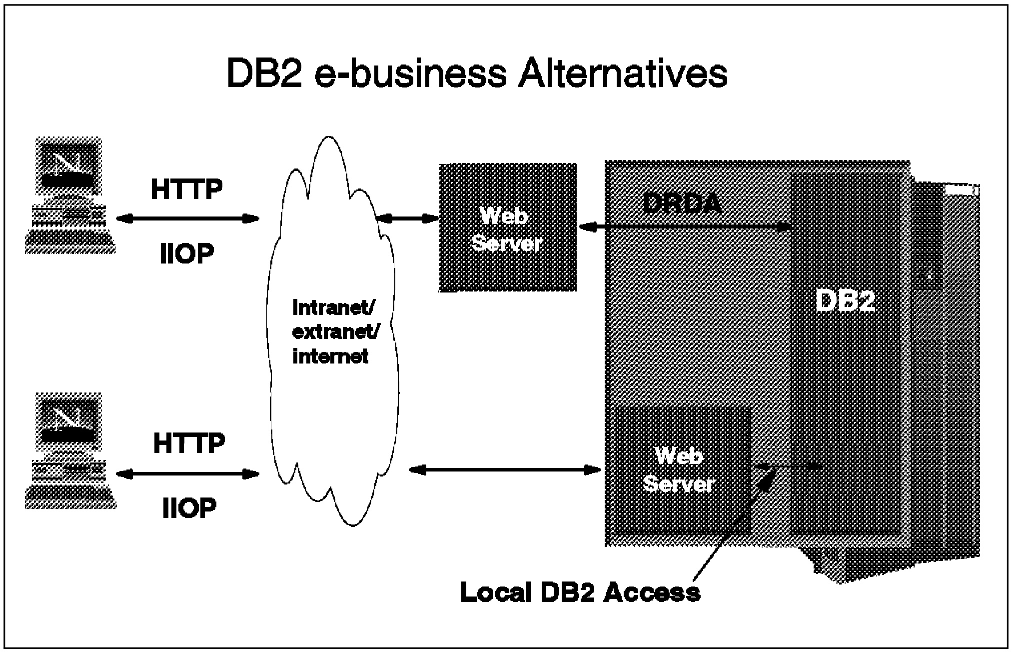


Figure 37. DB2 Access Options

If your frontend is not on the same system as the target DB2, you must use DRDA to communicate with DB2. DRDA can use either LU6.2 or TCP/IP to communicate with DB2. If you are using TCP/IP, you can use one of the TCP/IP workload balancing mechanisms to distribute the requests across the members of the DB2 data sharing group.

If you are using LU6.2, you *can* use the VTAM GR support to achieve the same thing—however, this is not the most efficient option. Web Servers typically have many sessions, and VTAM GR would funnel all of those into the one DB2 subsystem. A better choice is to use the WLM-based workload balancing for SNA sessions which is built into both DB2 for OS/390 (with DDF) and DB2 Connect.

If you are using local access to communicate with DB2, there must be a DB2 subsystem on the same system as the frontend. This DB2 does not necessarily have to be the one that you will get the data from—it is possible to have the local DB2 use a DRDA connection to get your data from a remote DB2 subsystem.

If you are interested in the highest levels of availability, one option to consider is to place two DB2 subsystems that are in the same data sharing group on the OS/390 system(s) that are running your frontend(s). In the normal course of events, whichever subsystem is placed first in IEFSSN will be the one that receives all DB2 calls that specify the Group Attach Name. However, if that DB2 is stopped, perhaps to apply maintenance, all calls using the Group Attach Name will get routed to the remaining DB2 subsystem. When defining the initialization parameters for Net.Data or the Web Server, you have the option of specifying the DB2 subsystem name or the Group Attach Name. Specifying the Group Attach Name allows you to use this two-DB2 setup. This does *not* provide any workload balancing, however, it does provide improved availability if you ensure that only one of the two DB2 subsystems is ever stopped at a time.

4.1.5.1 DB2 Web Components

There are a number of terms and components that will be used in most of the DB2 Web connection options. Before we get into discussing the options, we will briefly describe each of these components:

JDBC

Java Database Connectivity (JDBC) allows the use of dynamic SQL statements in a Java program, giving you the ability to access a relational database (in this case, DB2), from a Java program. JDBC is actually a driver that must be available when the Java program runs.

SQLJ

Structured Query Language for Java (SQLJ) is basically an enhancement to JDBC that permits the use of static SQL from within a Java program. As with JDBC, SQLJ is a driver that must be available when the Java program executes.

Net.Data

Net.Data is a product that provides a set of macros that can be used to access DB2 data. Net.Data programs consist of a set of Net.Data macros, and run under the control of a Web Server. Net.Data is included as part of DB2 for OS/390, and is a replacement for the earlier DB2 WWW Connection.

DB2 Connect

DB2 Connect is IBM's product that provides an interface to DRDA (there are similar products available from other vendors). DB2 Connect would be used as the middleware to connect any distributed server to DB2. There is a DB2 Connect Personal Edition, which is designed to be installed on the client workstation and permit access to DB2 directly from that workstation. There is also a DB2 Connect Enterprise Edition which is designed to run on a distributed server platform. In this case, clients communicate with the server using JDBC or SQLJ, and the server communicates with DB2 using DB2 Connect to translate the SQLJ and JDBC calls into DRDA calls.

DB2 Connect V6 introduced specific sysplex support. If WLM is running in Goal mode, and the SYSPLEX installation option is specified in DB2 Connect, WLM will return information to DB2 Connect about the members of the DB2 data sharing group, and the available capacity on the systems each of the members are running on. DB2 Connect then uses this information when deciding which DB2 member to send a given request to. More information about this aspect of DB2 Connect can be found in *DB2 Connect Quick Beginnings for UNIX*, available on the Internet at: `ftp://ftp.software.ibm.com/ps/products/db2/info/vr6/pdf/letter/db2ixe60.pdf`

DRDA

Distributed Relational Data Architecture (DRDA) is an architecture that permits external access to a relational database. In the context of what we are discussing here, DRDA would normally be used as a way for distributed platforms to communicate directly with DB2. DRDA also supports communication between two DB2 for OS/390 subsystems that are not in the same data sharing group.

4.1.5.2 DB2 Java Options

JDBC and SQLJ are used to include SQL statements in Java programs. The decision regarding where to place the Java program (as an applet on the client platform, as a servlet on a distributed platform, or as a servlet running under the Web Server on OS/390) is determined more by performance considerations than functionality.

It is possible to include the JDBC and/or SQLJ code in an applet; however, the size of the resulting file (which would have to be downloaded to the client for every invocation) would probably not provide very good performance. Also, the client would have to have DB2 Connect or a similar product installed—this might be an acceptable situation for an intranet application, but not one that will be used for customers coming in from the Internet.

The solution that is recommended to most installations is to run the Java program as a servlet on OS/390, using VisualAge for Java to compile the Java programs for improved performance.

4.1.5.3 Net.Data

The other option is to use Net.Data. Net.Data can be run in either a distributed Web Server or on a Web Server on OS/390. To invoke the Net.Data program, the URL specified by the client will include the name of the Net.Data program. When the client call reaches the Web Server, the Net.Data program will be invoked. If the Web Server is running on a distributed platform, DB2 Connect would then be used to communicate with DB2 on OS/390.

There is no difference in functionality, and once again, your decision is likely to be driven by performance and system management considerations.

4.1.5.4 Other Options

There are numerous other options for accessing DB2 data from the Web. Some of the ones that are provided by IBM are:

- CGI
- GWAPI
- QMF for Windows
- Host Publisher
- DB2 Forms

There are also a number of products available from other vendors.

Like Net.Data and Java, CGI and GWAPI programs can run in either a Web Server that is local to DB2, or in a distributed Web Server. However, the use of CGI and GWAPI is no longer considered strategic—new applications should be developed using Java or Net.Data instead.

QMF for Windows, Host Publisher, and DB2 Forms are all aimed at specific application requirements, and would not be as generic as Java or Net.Data.

4.1.5.5 Further Information

For more information about the use of DB2 with the Web, refer to the IBM Redbooks *Accessing DB2 for OS/390 Data From the World Wide Web*, SG24-5273, *WOW! DRDA Supports TCP/IP: DB2 Server for OS/390 and DB2 Universal Database*, SG24-2212.

4.1.6 IMS

Just as for CICS and DB2, there are numerous options available that allow existing IMS applications and data to be included as a viable part of the e-business world. These options are in a state of evolution and include connectivity solutions for both SNA and TCP/IP networks.

Once the request reaches the IMS environment to be processed, IMS additionally provides a variety of Parallel Sysplex capabilities that answer the requirements for performance, capacity and availability.

To better understand this whole environment for IMS, this section provides a brief overview of:

- Communication architectures and solutions for IMS
- IMS workload balancing methods
- Parallel Sysplex solutions

4.1.6.1 IMS Communication Architectures and Solutions

IMS can be accessed across SNA and TCP/IP networks using a variety of communication protocols and architectures. It is worth noting that regardless of which option is chosen, there is no corresponding requirement to modify the IMS transaction. The same transaction can be invoked concurrently from different environments. The IMS solutions were created to allow the transactions to be independent from, and unaware of, the network interfaces. Figure 38 on page 161 shows some of the different options.

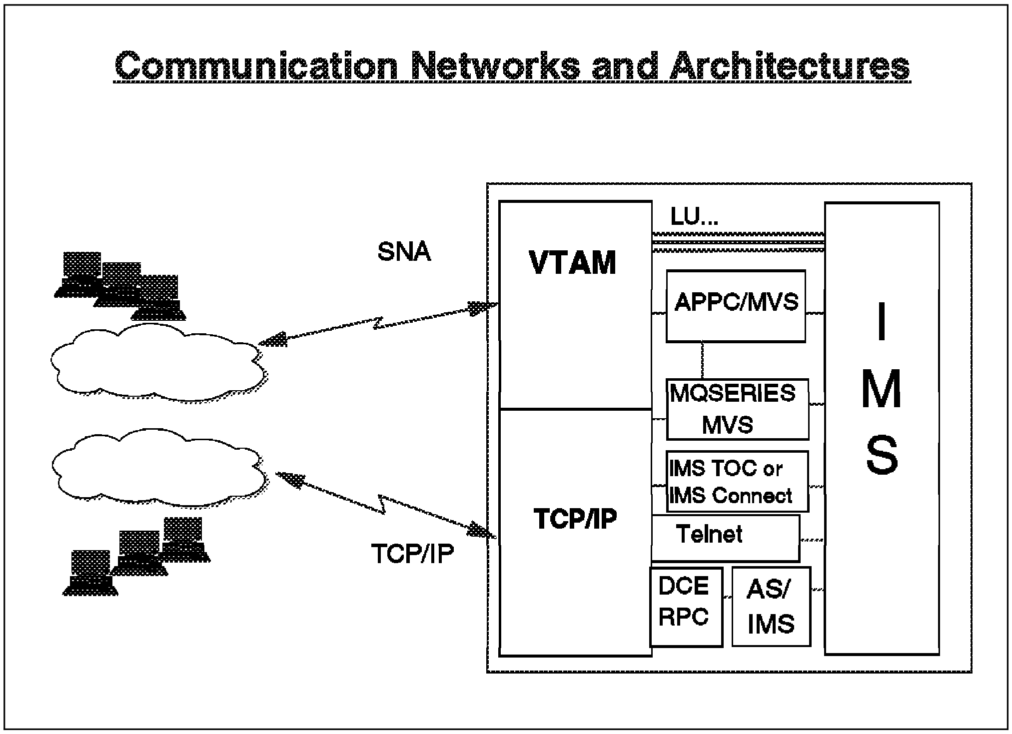


Figure 38. IMS Communication Options

The traditional *SNA* communication protocols (LU1, LU2, etc.) can be used to access IMS applications since IMS is a VTAM application. The most commonly used SNA protocol in the e-business world is LU 6.2 with its associated implementation called Advanced Program to Program Communications (APPC). APPC provides a standard, connection-oriented, Application Programming Interface (API) for one program to directly communicate with another, even when separated by a network.

Access to IMS across a *TCP/IP* network falls into three primary categories:

1. Remote login using Telnet and 3270 emulation
2. The Remote Procedure Call (RPC) interface using Distributed Computing Environment (DCE) services
3. Using the standard TCP/IP connection-oriented API called *sockets*

Note that the sockets API provides a program-to-program communication interface for the TCP/IP environment much as APPC provides for SNA networks. Although IMS is not a TCP/IP application, accessibility across all three categories is fully supported. For remote login, the Telnet server actually creates an LU2 session into IMS through VTAM. For DCE/RPC, a product called Application Server/IMS communicates with remote DCE clients and interfaces with IMS. For sockets support, IMS provides a connector (called IMS TOC in IMS/ESA Version 6, and IMS Connect in IMS/ESA Version 7—we will refer to them generically as IMS TOC/IMS Connect in this document) that provides the conversion between sockets-based messages and IMS messages.

Of note is the queue-based application architecture and the use of the MQSeries set of products. MQSeries on MVS provides a bridge that retrieves the inbound message from the MQ queue, sends it to the IMS message queue, and then receives replies from IMS and sends them to the appropriate MQ outbound queue. MQSeries hides the network specifics from the application programs and supports deployment across both SNA and TCP/IP networks.

Any of these protocols and associated products can be used when developing e-business applications that access IMS. To assist in the development and implementation effort, several options and solutions are currently available. These include:

- IMS WWW templates (IWT)
- Component Broker support
- IMS TOC/IMS Connect
- The IMS Client for Java
- The IMS Connector for Java
- Open Database Access (ODBA)
- Classic Connect
- Host on Demand
- Net.Commerce

A brief overview of each of these options follows.

Accessing IMS Transactions Using APPC: The *IMS WWW templates* are C programs, written and supplied by IBM, that provide a flexible, interactive and portable interface to IMS transactions using APPC. The templates function as a Web Server program that access IMS transactions and map the results into an attractive page for the browser users. They are provided with a choice of Web application interfaces:

- Common Gateway Interface (CGI)
- Internet Connection API (ICAPI), which is a higher-performing interface than CGI and is unique to IBM Web Servers
- The more recent Java servlet wrapper

These programs can be used as "canned" solutions or as models when creating user-written Web Server programs.

Component Broker is IBM's implementation of Common Object Request Broker Architecture (CORBA) and provides the capability for applications to access transactions and data using standard object technology. An application adaptor for IMS is provided to facilitate access to IMS transactions. Remote access to IMS uses APPC. When Component Broker and IMS reside in the same OS/390 image, access is provided using a direct interface to the Open Transaction Manager Access (OTMA) component of IMS. If Component Broker is not on the same OS/390 image as IMS, APPC will be used as the interface.

Open Transaction Manager Access (OTMA): It is worth noting at this point a function called Open Transaction Manager Access (OTMA) which was introduced in IMS V5. OTMA supports a high-performance interface between IMS and any OS/390 application for the exchange of transaction messages and commands. The OS/390 applications are expected to use MVS XCF services and to prepare the messages in a format IMS understands. Messages that are received by the OTMA component in IMS are processed in the same manner as any other transaction, that is, the message queue services, logging, and other components of IMS are used as required by the transaction. Examples of applications that use OTMA include:

- The MQSeries Bridge for IMS
- AS/IMS which supports DCE/RPC access
- IMS TOC/IMS Connect which supports TCP/IP sockets access

Due to the use of XCF as the communications vehicle, OTMA can be used even if IMS is not active in the same OS/390 image as the client.

Accessing IMS Transactions Using TCP/IP Sockets: IMS TOC/IMS Connect provide the architectural framework that supports connectivity to IMS from any TCP/IP sockets application. The IMS TCP/IP OTMA Connection (IMS TOC, also known as ITOC) product code can be obtained from the IMS Web site, rather than on standard product tapes. At the time of writing, IMS TOC is scheduled to run out of service in September, 2001. IMS Connect, which is a separately priced feature of IMS V7 Transaction Manager, replaces IMS TOC and provides enhanced functionality.

IMS TOC/IMS Connect provides services to translate (including ASCII-to-EBCDIC, if needed) data sent from the TCP/IP client into a format understood by the IMS OTMA component. Optionally, the data is also validated (userid and password/passticket checks). Reply messages are also translated into a format understood by the TCP/IP sockets program.

It is important to note that IMS TOC/IMS Connect can interface with multiple IMS systems even if an IMS resides on a different MVS in the sysplex. Figure 39 shows the role of IMS TOC/IMS Connect.

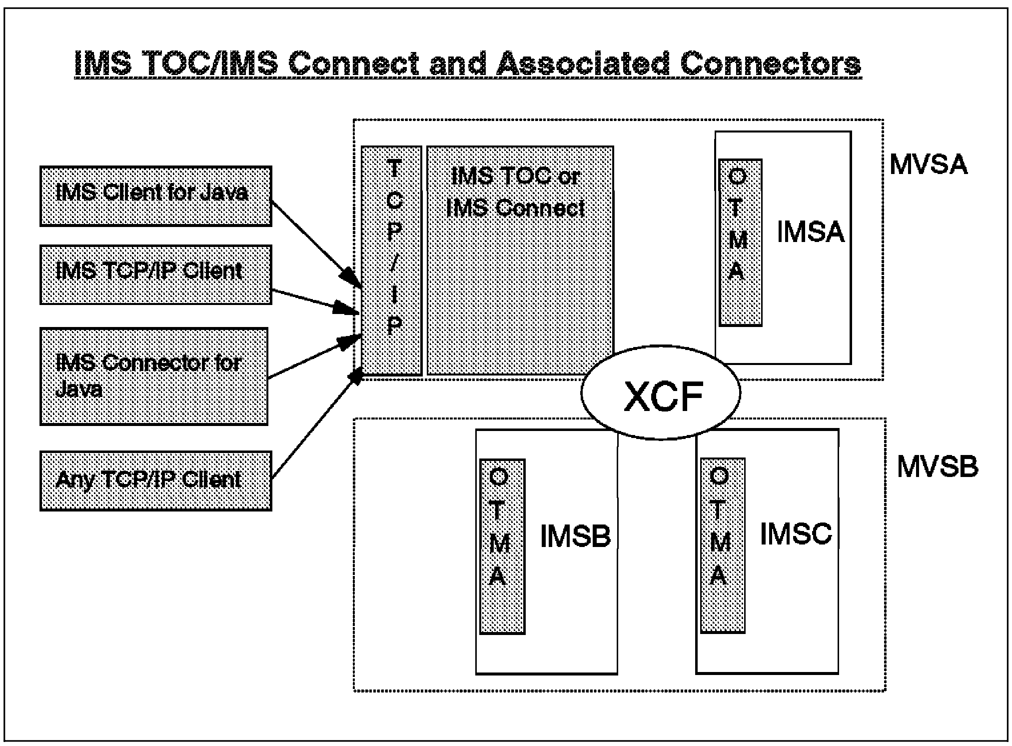


Figure 39. Role of IMS TOC/IMS Connect

The *IMS Client for Java* consists of both a sample Java program that issues TCP/IP socket calls to access IMS transactions and an associated user exit routine for IMS TOC/IMS Connect. The exit translates the message into the format required by the IMS OTMA interface. Both the sample Java program and exit routine are delivered with IMS TOC/IMS Connect. They are provided as a prototype of TCP/IP connectivity into IMS and can be used as is or, since the source is provided, modified as needed.

The *IMS Connector for Java* provides a way to generate and create Java applications using a standard toolkit. It is delivered as a component of the VisualAge for Java (VAJava) development environment, and provides the code to invoke IMS transactions using TCP/IP sockets.

With additional support from IBM WebSphere Studio and the WebSphere Application Server, the capability is also provided to build and run Java servlets.

Accessing IMS Databases: *Open Database Access (ODBA)* is an IMS function, introduced in IMS V6, that provides a callable interface for OS/390 applications to directly access IMS databases. The use of ODBA supports the allocation of one or multiple PSBs and the ability to issue any DL/I call against the associated databases. The OS/390 application can be a stored procedure invoked by a remote SQL client, an OS/390 Web server program, or any MVS application.

Classic Connect is a separately orderable component of DataJoiner that provides read-only access to data stored in IMS databases and Virtual Storage Access Method (VSAM) data sets. Classic Connect, in conjunction with a DataJoiner instance, allows users from a variety of client platforms to submit a standard SQL query that accesses IMS DB and VSAM data. For example, a user on a PC (with DB2 client software) can issue an SQL join across IMS, VSAM, DB2 for MVS, Informix, Oracle, Sybase, and DB2 common server data sources.

IMS DataPropogator maximizes the value of IMS DB assets by enhancing IMS DB and DB2 coexistence. IMS DataPropagator enables the propagation of changed data between IMS DB and DB2 databases. A companion product, DataPropagator Relational, can further distribute this data to distributed DB2 systems across the enterprise.

Other IBM-provided Solutions for Web-Enabling IMS: Host on Demand (HOD) provides a quick and easy Web interface to back-end enterprise systems like IMS. Access is provided across a TCP/IP network and uses 3270 emulation and Telnet. HOD also provides the necessary tools to create an attractive presentation interface for the Web browser.

Net.Commerce provides a solution that quickly, easily, and securely enables electronic commerce on the World Wide Web. It helps companies integrate existing business processes and legacy systems with the Web, as well as grow new Web-based businesses. Net.Commerce is scalable, open to customization and integration, and capable of handling any level of transaction volume growth. It comes complete with catalog templates, setup wizards, and advanced catalog tools to help build effective and attractive electronic commerce sites.

4.1.6.2 Workload Balancing Methods

Once the e-business application is developed and deployed, the next area to consider is the requirement to support growth and availability. Several workload balancing solutions are provided that address access across both SNA and TCP/IP networks. It is assumed that the target IMS subsystems are part of an IMS data sharing group, and thus are capable of accepting and subsequently processing or routing any transaction request.

SNA: VTAM Generic Resources (VTAM GR) is a function provided by VTAM to minimize the knowledge that an end-user needs to logon to one of several like-instances of an application (for example, one of the IMSs) in a Parallel Sysplex. To the end user, there is a common name, called the Generic Resource Name, which is used by VTAM to refer to *any* of the members of a Generic Resource Group. VTAM decides which member to use when actually establishing a session request from an end-user node that logs on using the generic name.

Figure 40 shows how remote applications and devices that use SNA communication protocols, e.g., LU 2 (3270 devices) and LU 6.2 (APPC) programs, can connect to an IMS system in a Parallel Sysplex using a generic name. Any of the options that use SNA to access IMS can therefore get the availability benefits provided by VTAM GR.

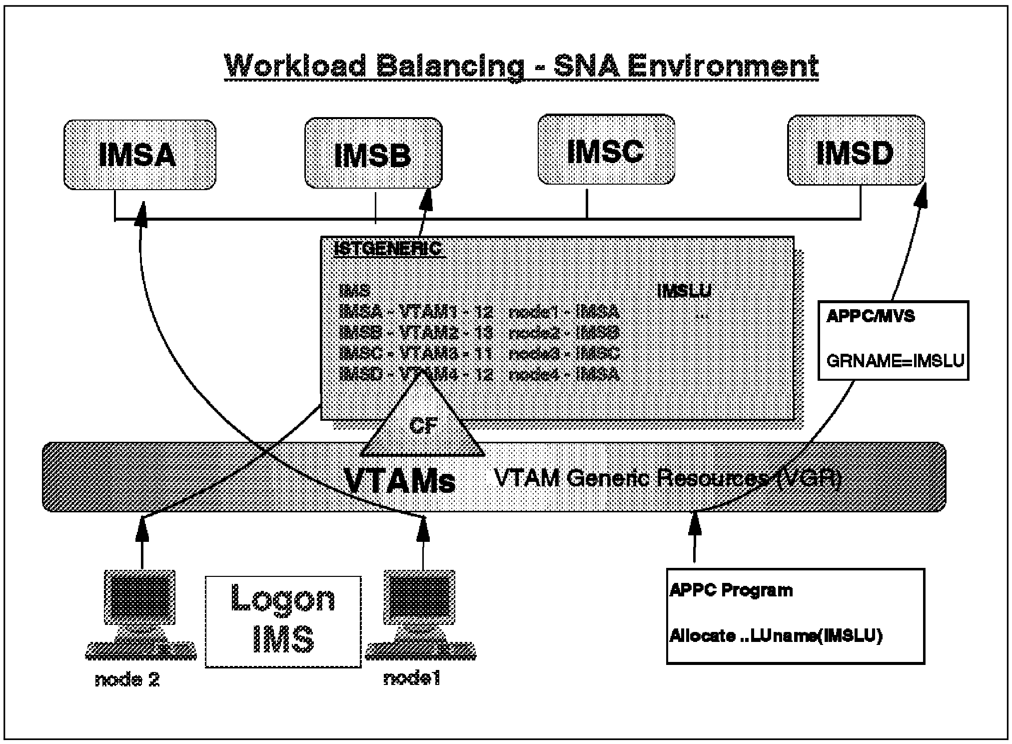


Figure 40. IMS Workload Balancing with VTAM GR

TCP/IP: TCP/IP provides a number of mechanisms for balancing workloads across servers.

For long-running sessions and connections that use Telnet, the Telnet server on OS/390 supports DNS/WLM. Telnet registers with WLM, allowing incoming requests to be routed to the most appropriate Telnet server in the sysplex. Once a specific Telnet server is picked to process the request, the Telnet-to-IMS session (which uses VTAM) can take advantage of VTAM GR to access one of the IMS subsystems in the data sharing group. This is shown in Figure 41 on page 166.

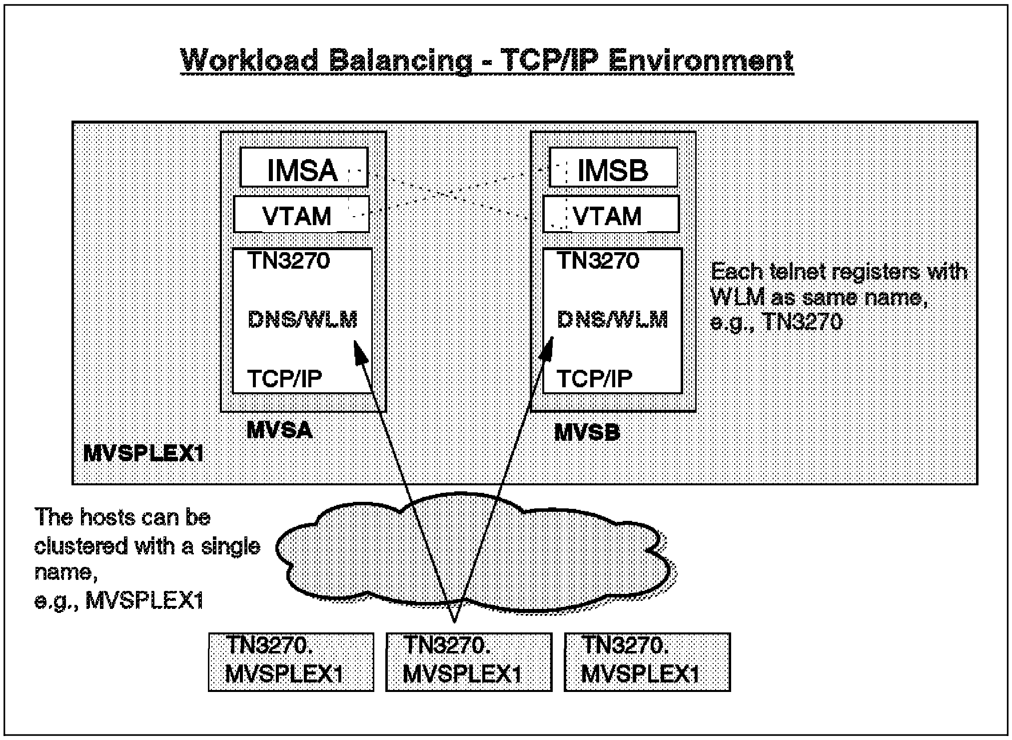


Figure 41. Load Balancing across Telnet Servers

Note that at the time of writing, neither IMS itself nor IMS TOC provide support for DNS/WLM—the use of DNS/WLM in relation to IMS is limited to the Telnet Server's support of DNS/WLM.

For short-lived requests, such as Web requests to IMS TOC/IMS Connect, the use of DNS/WLM is not suitable because of the overhead of the DNS/WLM interaction for each transaction request. Prior to the availability of OS/390 V2R10, the supported methods for such short-lived requests consists of the use of functions provided by the Interactive Network Dispatcher (IND) and external devices such as the IBM 2216 router and CISCO MNLB.

These types of capabilities allow for a generic TCP/IP hostname to resolve to a specific IP address, for example, that of the router. The software on the router chooses and establishes a connection to one of the IP stacks in the sysplex using a mechanism such as a round-robin approach or even by querying the WLM. Each of the IP stacks in the sysplex specifies an IP address loopback alias equal to the IP address of the front-end router. All the inbound messages are sent via the router to the target IP address, but reply messages are sent directly by the back-end IP stack (this can be done because of the loopback alias definition). This is shown in Figure 42 on page 167.

Once a back-end IP stack is chosen, an application such as IMS TOC/IMS Connect can be configured on the same port number on each IP stack. This is important because a socket connection is actually a combination of IP address and port number.

The IMS TOC/IMS Connect address space can then be configured to communicate with all the IMS systems in the sysplex, using OTMA.

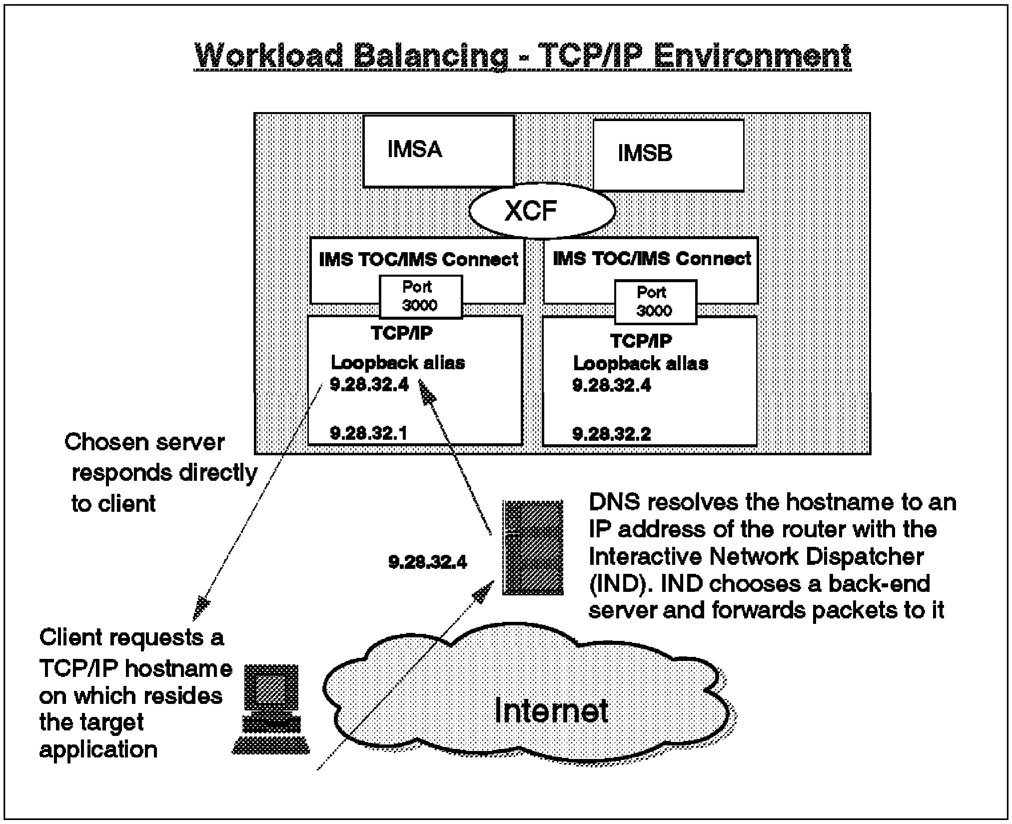


Figure 42. IMS Load Balancing with IND

IMS TOC provides an exit that can be used to decide which IMS subsystem to send a given transaction to. The exit can contain the names of a number of IMS subsystems. However, the exit has no knowledge of whether a given subsystem is available or not. If the subsystem that the exit selects is unavailable, the transaction will fail.

One of the new capabilities provided with IMS Connect (available as a feature in IMS V7) is a datastore table which keeps track of the current status (active or inactive) of all the IMS systems that IMS Connect is configured to reach. An associated exit interface allows user-written code to take action based on this information. IMS Connect, as a member of the XCF group, is informed whenever the status of any other members of the group changes—for example, if an IMS subsystem starts or stops. The message exits can use this information to reroute the request to an alternative IMS system in case the target IMS is not available. Refer to Figure 43 on page 168 as an example of what can be done. Note that this capability is not available with IMS TOC.

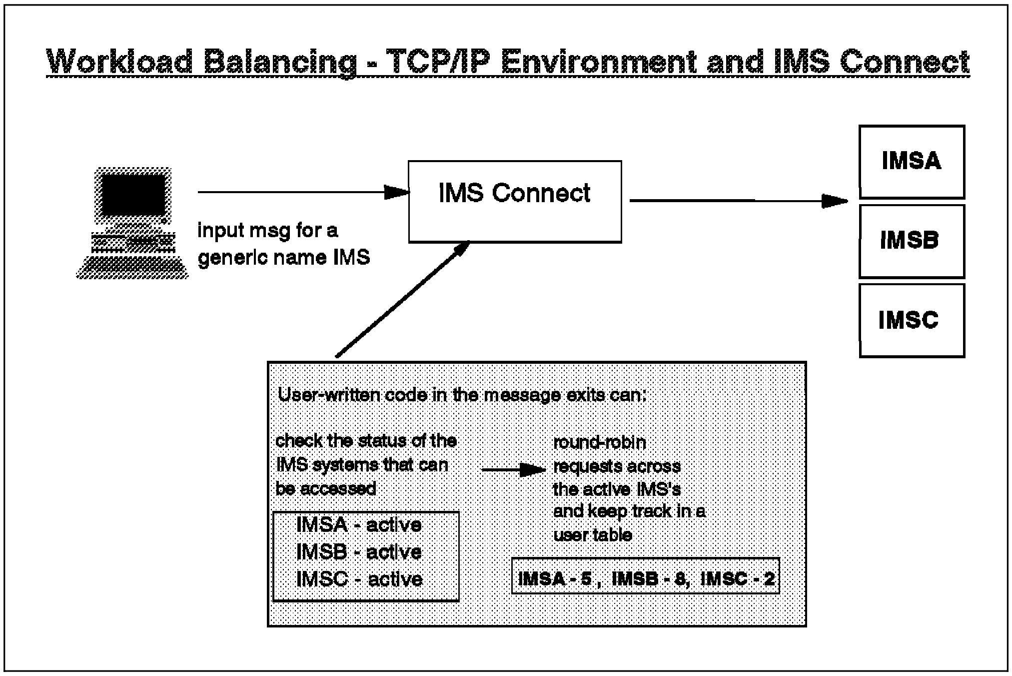


Figure 43. IMS Connect Workload Balancing Capability

4.1.6.3 Parallel Sysplex Support

Once the transaction reaches the IMS message queue, the normal IMS Parallel Sysplex features can be used to ensure the transaction gets processed by one of the members of the data sharing group. The Parallel Sysplex capabilities of IMS are described in detail in 4.2.5, "IMS Transaction Manager in Parallel Sysplex" on page 187 and 4.3.2, "IMS DB Data Sharing" on page 203.

4.1.6.4 Additional Information

Further information about each of these options is available on the IMS Web site, at:
`http://www.ibm.com/software/data/ims/imswwwc.html#9`

In addition, various white papers, manuals, and redbooks are pointed to from the Web site:
`http://www.ibm.com/software/data/ims/library.html`

IBM Redbooks that may prove useful are:

- *E-business Connect Using the IMS Connectors*, SG24-5427
- Chapter 10 of *IBM Web-to-Host Integration*, SG24-5237
- *IMS/ESA V6 Parallel Sysplex Migration Planning Guide for IMS TM and DBCTL*, SG24-5461,
- *Connecting IMS to the World Wide Web: A Practical Guide to IMS Connectivity*, SG24-2220

4.2 Transaction Management in Parallel Sysplex

In a Parallel Sysplex environment, it is possible to achieve data sharing without any transaction manager support, because data sharing is provided by the database manager. However, transaction manager support is required for *dynamic transaction routing*. It is therefore recommended that you upgrade the transaction manager to the correct level to support Parallel Sysplex and the data sharing level of the database manager. This upgrade can be classified as a "business as usual" activity.

4.2.1 Dynamic Transaction Routing

Dynamic transaction routing is the ability to route the execution of an instance of a transaction to any system, based on operational requirements at the time of execution. Operational requirements are such things as required performance or transaction *affinity*. Affinities are further discussed in 4.2.2.5, "Affinities and CICS" on page 177, and in "Affinities and IMS TM" on page 189. Dynamic transaction routing in a Parallel Sysplex delivers dynamic workload balancing. For further discussion, refer to 2.6, "Dynamic Workload Balancing in Parallel Sysplex" on page 83.

To establish a dynamic transaction routing environment, the following steps are required:

1. Understand any transaction affinities
2. Remove or tolerate these affinities
3. Establish multiple application regions
4. Clone the application

For a CICS application to be suitable for cloning, it must be able to run in more than one CICS region. The first application that is chosen to use the Parallel Sysplex data sharing environment is likely to be one that can be cloned without change. Cloning can occur on a single image. Cloning across images provides additional availability. The transaction manager definitions must be modified to route the transactions accordingly.

Now we will take a look at two IBM transaction managers, CICS and IMS TM.

4.2.2 CICS Transactions in a Parallel Sysplex

In this section, we look at the topology for CICS in a Parallel Sysplex, and also at how affinities may affect the Parallel Sysplex configuration.

4.2.2.1 Multiple Region Operation

For many years, CICS has had the ability to route transactions from one CICS address space to another. CICS provides two mechanisms to do this:

- Multiple region operation (MRO)
- Inter-system communication (ISC)

Prior to CICS/ESA V4, CICS MRO could only be used to route transactions between CICS regions within the *one* image. MRO transactions used one of two methods of communication: cross memory, or interregion communication (IRC).

If you wished to route transactions between multiple systems, you would use ISC. ISC uses VTAM connections between the CICS regions in different images.

Starting with CICS V4, however, MRO is able to connect CICS regions in different images within a sysplex using XCF. The communication is achieved using XCF as the service provider, which has a lower overhead than the VTAM connection in an ISC setup. The increase in throughput achieved by the use of XCF instead of ISC can be significant.

More information about the benefits of XCF for MRO can be found in *CICS/ESA V4 Performance Comparison with CICS/ESA V3*. This is available to IBM representatives on MKTTOOLS in the CICSE41P package. Also, detailed information about the different options, including the possible throughput increase when using XCF/MRO compared to VTAM CTCs for the connections between CICS systems, is found in *CICS TS for OS/390 Intercommunication Guide*, SC33-1695.

The amount of CF resource that you will need depends on how you are using the CF. Using the CF for XCF may require *substantial* resources if there is a very high rate of XCF traffic from CICS function shipping and MRO transaction routing, for example.

Today, a CICS system with only a single CICS address space is rare. For capacity and availability reasons, most CICS users have already implemented MRO. Availability is enhanced by separating workloads of different types from each other. It is not unusual to have a CICS region dedicated to a single application suite. Capacity is improved within the CICS application owning regions (AORs) by splitting off the terminal functions into a separate terminal owning region (TOR).

Figure 44 on page 171 shows the evolution of CICS from a single system to multiple regions running on multiple images, all communicating through XCF. CICS installations that are not using MRO, whether they have few or many CICS regions, may want to implement MRO as one of the first plan items for migrating to a Parallel Sysplex.

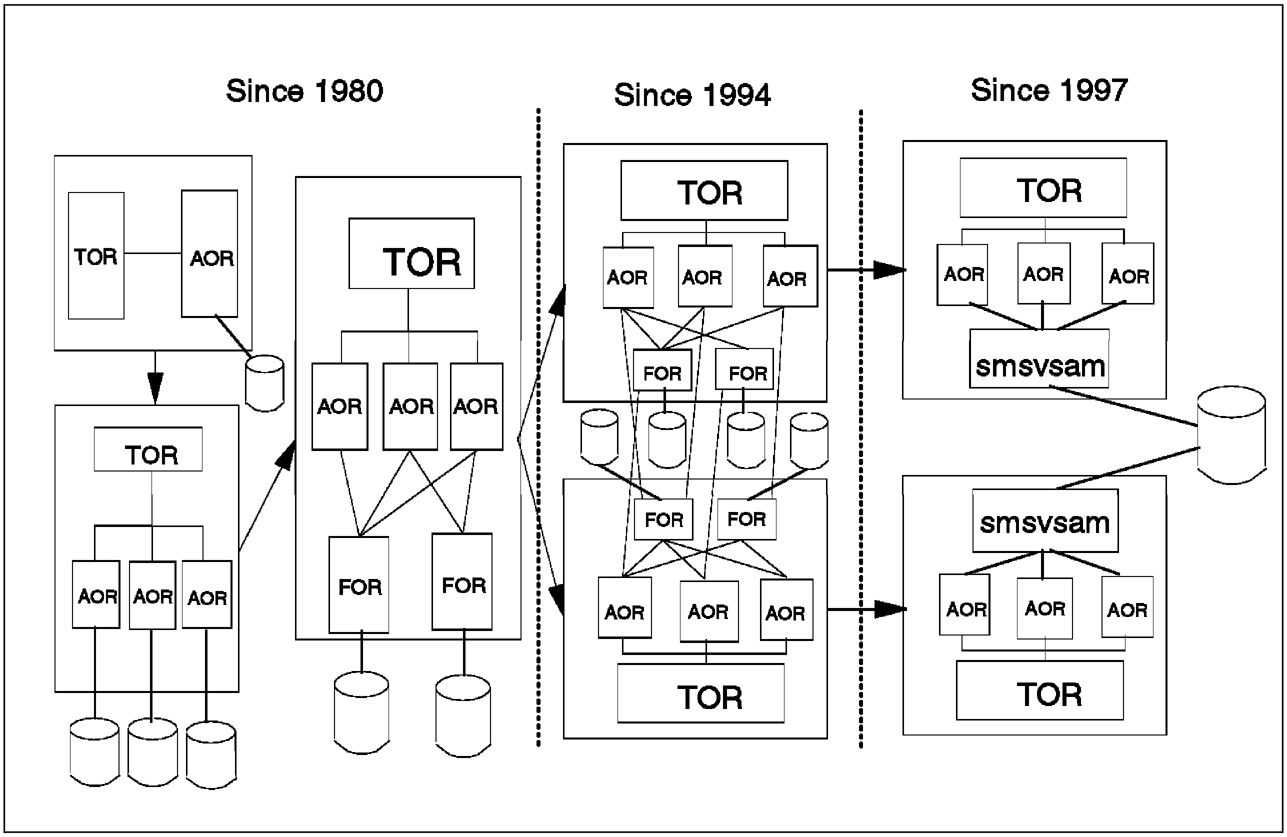


Figure 44. Evolution of Multi-Region Operation in CICS

Another exploitation of MRO is called *function shipping.* This is used to allow data sharing between CICS regions. Function shipping works by passing all requests for a particular resource (usually files or a database) to a specific CICS region. With this capability came a new type of CICS region, the file-owning region (FOR).

Also, you can substitute DB2 or IMS DBCTL for the FOR. For a discussion on database managers, refer to 4.3, "Database Management in Parallel Sysplex" on page 194.

Note: Local DL/1 support in IMS has been withdrawn at IMS/ESA V5.

4.2.2.2 Dynamic Transaction Routing

In CICS/ESA V3, a dynamic transaction routing exit was added. This exit allows routing decisions to be made based on programmable criteria rather than on static table-based definitions. Also, IBM introduced a product, CICSPlex SM, which provides sophisticated routing algorithms that can provide workload balancing and failure avoidance. CICSPlex SM provides a single point of control or management for *all* CICS regions in the Parallel Sysplex environment. CICSPlex SM provides a single system image for CICS operations, and real time analysis, and monitoring. CICSPlex SM supports CICS/ESA V.4.1 and above, and CICS/VSE V2.2 and above. Some examples of the facilities provided by CICSPlex SM are:

- **Single system image:** CICSPlex SM keeps track of the location and status of every CICS resource, allowing CICS regions to be managed as a single system. Actions affecting multiple resources can be achieved with a single command, even though the resources are spread among multiple CICS systems on different CPCs.

- **Single point of control:** CICSPlex SM allows all the CICS systems in an enterprise to be managed from a single point of control. This supports installations who have business reasons for organizing their systems into more than one CICSplex (for example, one for testing and another for production).
- **Run Time Analysis (RTA):** RTA provides management by exception, drawing attention to potential deviations from service level agreements. CICS users will see improved reliability and higher availability of their systems, because potential problems are detected and corrected before they become critical.

You should be aware that there is a cost in transaction routing due to the path length of the routing code, plus the actual time to move the execution from one place to another. To minimize this cost, CICSPlex SM has a check in its balancing algorithm to see if the *local* region is suitable for execution of the transaction. If it is, it will *favor* this system, as MRO can then use cross-memory services to route the transaction, rather than XCF/MRO or ISC.

Starting with CICS TS for OS/390 R3, CICSPlex System Manager is an exclusive element, fully integrated with the base CICS Product. Standalone releases of the CICSPlex SM product (5695-081) do not support CICS TS 1.3.

For more details on CICSPlex SM, refer to the IBM Redbook *CICSPlex Systems Manager Business Application Services: A New Solution to CICS Resource Management*, SG24-5267, *CICS for OS/390 and Parallel Sysplex*, GC33-1180, *CICSPlex Systems Manager V1 Concepts and Planning*, GC33-0786 or section 4.2.3, “CICSPlex Systems Manager (CICSPlex SM)” on page 185 of this book.

CICS and Parallel Sysplex

CICS/ESA and the Parallel Sysplex have a very *complementary structure,* especially when using MRO.

It is recommended that you use XCF MRO services when connecting CICS regions within a Parallel Sysplex.

CICS systems that are part of a Parallel Sysplex can be at *different* release levels. All releases of CICS that are still in support can coexist within a single CICSplex, sharing a single CSD if you wish.

The use of CICSPlex SM is highly recommended when running CICS in the Parallel Sysplex environment.

4.2.2.3 What Will My Data Sharing Configuration Look Like

This section describes an example of a target data sharing configuration that is recommended for a Parallel Sysplex. The main subsystem components of this target configuration are described in 4.2.2.4, “The Target Subsystems in a Parallel Sysplex” on page 173. For a complete discussion on this topic, refer to *OS/390 MVS Sysplex Application Migration*, GC28-1863.

The target configuration is designed with considerations for:

- **Availability:** It provides maximum availability of online applications by cloning as many of the sysplex components as possible. These include:
  - The images
  - The VTAM nodes
  - The transaction processing subsystems (CICS and IMS TM)
  - The DBCTL, DB2, and VSAM record level sharing (RLS) subsystems, providing IMS DB, DB2, and VSAM multisystem data sharing
- **Capacity:** Provide a growth capability for adding additional capacity without disrupting production work. The sysplex provides this kind of nondisruptive growth capability, enabling you to easily upgrade or add a new CPC.
- **Systems management:** It provides better systems management of multiple images, with OS/390 clones offering easier installation and administration of additional images.

You should aim for as much symmetry as possible in a Parallel Sysplex. For example, there is no reason why you should not install a CICS terminal-owning region on each image.

4.2.2.4 The Target Subsystems in a Parallel Sysplex

The target configuration shown in Figure 45 on page 174 includes the following elements:

- One terminal-owning region (TOR1 through TOR4) in each of the OS/390 images.
- Twelve application-owning regions allocated across the four OS/390 images.

  Note: Although the transaction processing workload that runs in this Parallel Sysplex configuration is assumed to be a mixture of DBCTL, DB2-based, and VSAM applications, all the application-owning regions are capable of running all transactions. That is, they are clones of each other, and any workload separation is controlled by workload management policies and the CICS dynamic routing mechanism. Therefore, all the application-owning regions require a connection to DBCTL, DB2, and VSAM RLS subsystems in their respective images.
- Four DBCTL environments allocated across the four OS/390 images to support the CICS-DL/1 workload processed by the application-owning regions. Each DBCTL consists of a database control (DBCTL) address space, a database recovery control (DBRC) address space, and a DL/1 separate address space (DLISAS).
- Four DB2 subsystems allocated across the images to support the CICS-DB2 workload processed by the application-owning regions.
- Four SMSVSAM systems allocated across the images to support the CICS-VSAM workload processed by the application-owning regions.
- Four DFHCFMN server regions allocated across the images to support access to data tables in the CF.
- Eight Integrated Resource Lock Managers (IRLMs), two for each image (a separate IRLM is required for DBCTL and DB2 on each image).

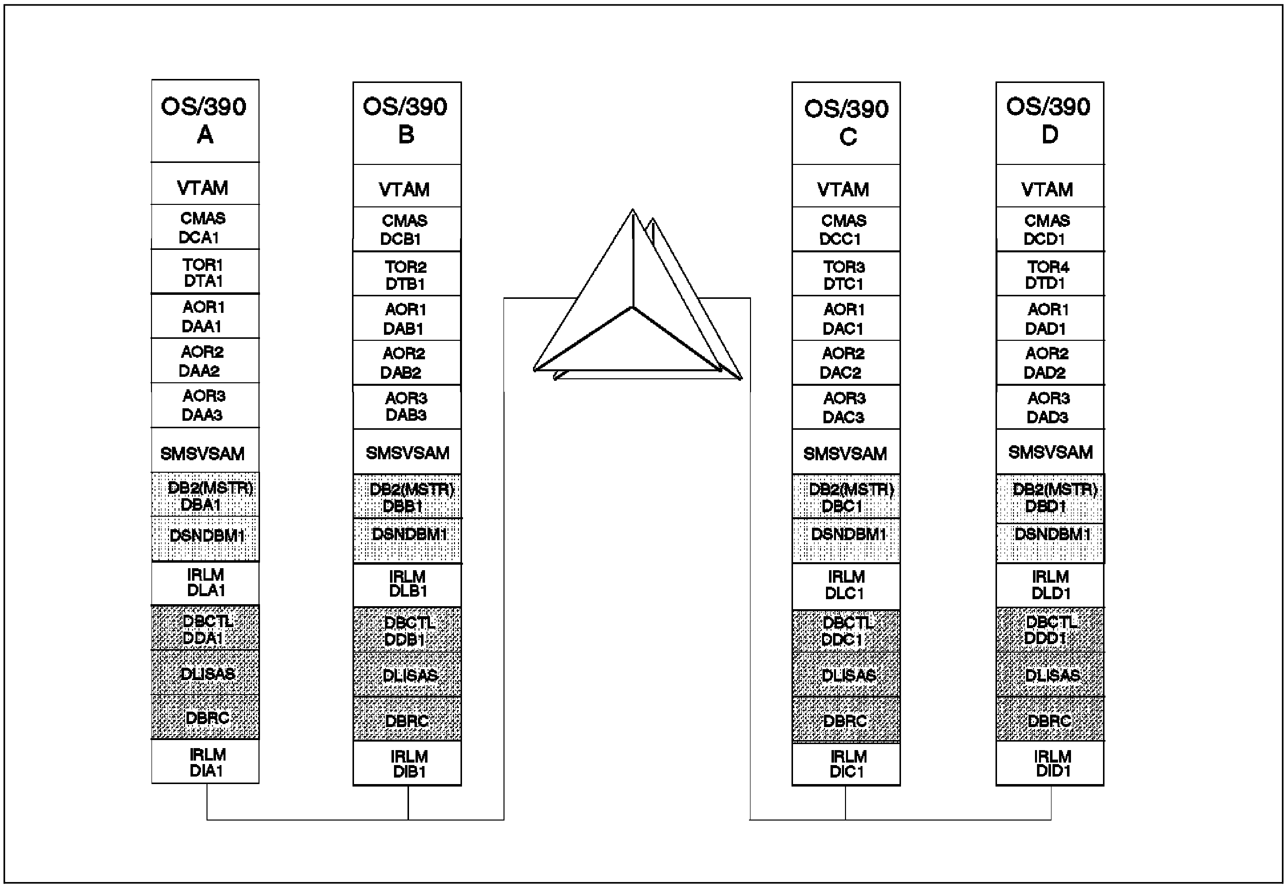


Figure 45. Sample CICS Data Sharing Subsystem Configuration

The CICS Terminal-Owning Regions: With the VTAM generic resources function, all the terminal-owning regions in the CICSplex can be represented by one generic application (APPL) name, and appear as one to terminal users. This means that regardless of which application the users want, they log on to only a single CICS application ID. VTAM generic resources resolves the generic name to the specific APPLID of one of the terminal-owning regions. Thus, the CICSplex appears as a single system to the end user.

Fast TOR restart can be implemented by using *single node persistent sessions* introduced in CICS/ESA V4.1; *Multi-Node Persistent Sessions* are introduced in CICS TS for OS/390 R1.

The terminal-owning regions are identical in every respect except for their external identifiers. This means that:

- They have different specific APPLIDs to identify them to VTAM and their partner MRO regions.
- They each have a unique local name specified on the SYSIDNT system initialization parameter.
- They each have a unique CICS monitoring subsystem identifier for RMF performance reporting, specified on the MNSUBSYS system initialization parameter.

Generally, apart from the identifiers just listed, you should try to make your terminal-owning regions identical clones, defined with identical resources (such as having the same system initialization parameters).

Exact cloning may not be possible if you have some resources that are defined to only one region. For example, if your network needs to support predefined auto-connected CICS terminals, you have to decide to which region such resources should be allocated and specify them accordingly. In this situation you cannot use the same group list GRPLIST system initialization parameter to initialize all your terminal-owning regions. However, starting with CICS/ESA V4.1, the GRPLIST system initialization parameter was extended, allowing you to specify up to four group list names, which makes it easier to handle variations in CICS startup group lists.

The reasons for having multiple terminal-owning regions are as follows:

- **For continuous availability**

  You need to ensure that you have enough terminal-owning regions to provide continuous availability of the CICSplex.

  Fewer users are impacted by the failure of one terminal-owning region. If a terminal-owning region fails, the users connected to other terminal-owning regions are unaffected, while the users of the failed region can logon again immediately, using the VTAM generic resource name, without waiting for the failed terminal-owning region to restart.

  Furthermore, the Multi-Node Persistent Session support introduced with CICS TS for OS/390 allows CICS sessions to remain connected across VTAM, hardware, or software failures. When CICS is restarted in the same system or another system, the session can be re-established with no disruption to the end user, other than the need to logon again.

- **For performance**

  To service several application-owning regions requires many MRO send and receive sessions. It is better to allocate the required sessions across several terminal-owning regions than to try to load them all into just one or two systems.

  In the sample configuration, we balanced the number of subsystems and CICS regions to fully exploit images running on multiprocessor CPCs.

- **For faster restart**

  If a terminal-owning region fails, restarting is faster because of the smaller number of sessions to be recovered.

The CICS Application-Owning Regions: The application-owning regions are defined as sets, with each set containing identical regions (AOR clones). Each set of clones should be capable of handling one or more different applications. The terminal-owning regions achieve workload balancing and availability by dynamically routing the incoming transactions to the best candidate application-owning region within a cloned set.

If you have split your CICS regions into separate regions based on applications, the data sharing, workload balancing environment of the Parallel Sysplex allows you to collapse regions together again. If your reason for splitting applications into separate regions is to provide some form of storage protection between applications, the introduction of transaction isolation in CICS/ESA 4.1 may make splitting no longer necessary.

Note: The AORs can be at different levels of CICS, as long as each AOR provides support for the facilities used by the applications in those AORs.

The CICS File-Owning Regions: Prior to VSAM RLS, a file-owning region was required if you wished to share a VSAM file between multiple CICS regions. However, files that are accessed in RLS mode are now accessed directly from each CICS region, rather than having to be routed through a central owning region.

However, you will still need a file-owning region within the CICSplex to provide access to other data files which currently do not support data sharing through the CF, such as BDAM files or VSAM files that are not accessed in RLS mode. The application-owning regions function ship file requests to the file-owning region. We have not shown any file-owning region in our configuration, but you might need to have one or more.

The CICS Queue-Owning Regions: When using releases prior to CICS TS for OS/390 R1, the inclusion of a queue-owning region in the CICSplex is important. This avoids any intertransaction affinities that might occur with temporary storage or transient data queues. Defining queues to the application-owning regions as remote queues, accessed through a queue-owning region, ensures that they are accessible by any application-owning region through function shipping requests. An alternative to a queue-owning region is to make the file-owning region a combined FOR/QOR.

With CICS TS for OS/390 and later, there is no need for the QOR. CICS TS for OS/390 introduces shared temporary storage queues between different CICS systems using the CF. Shared temporary storage provides multisystem data sharing for non-recoverable temporary storage queues; the queues are stored in CF list structures.

The CICS Web-Owning Region:

A recent addition to the set of CICS regions is a new region that is used exclusively for handling transactions from the Web. There are a number of regions why you may wish to use a separate region for this function:

- You may wish to protect your Web users from problems that may arise with traditional applications.
- The WOR may contain special function (such as the TCP/IP listener) that you have not activated on your other CICS regions.
- You may wish to be able to control the performance of the WORs separately from the traditional regions. This could either be to guarantee them good performance, or to restrict them to a certain amount of resource usage.
- For security reasons, you may wish to place the WOR in a different LPAR to your production systems.
- By passing the Web transactions through a WOR, you effectively get the same benefit as a TOR, and therefore the ability to route the transactions across multiple AORs.

The IMS DBCTL Environment: To exploit IMS data sharing, our target configuration includes one IMS DBCTL environment in each OS/390 that has CICS application-owning regions, with each such region connected to the DBCTL in its image.

The DB2 Subsystems: To exploit DB2 data sharing, our target configuration includes one DB2 subsystem in each OS/390 that has CICS application-owning regions, with each such region connected to the DB2 in its image.

The CICSPlex SM Address Space: Our configuration shows a CICSPlex SM address space (CMAS) in each OS/390 image that runs CICS regions. Together they provide a single system image of the CICSplex, shown in Figure 45 on page 174. CICSPlex SM is described in 4.2.3, "CICSPlex Systems Manager (CICSPlex SM)" on page 185.

A CMAS is the hub of any CICSPlex SM configuration, being responsible for managing and reporting on CICS regions and their resources. A CICSplex is managed by one or more CMASes. In our case, we have installed one CMAS in each image. The CMAS does the monitoring, real time analysis, workload management, and operational functions of CICSPlex SM, and maintains configuration information about the CICS regions for which it is responsible. See *CICS for OS/390 and Parallel Sysplex*, GC33-1180 for more information about a CMAS.

4.2.2.5 Affinities and CICS

CICS transactions use many different techniques to pass data from one transaction to another. Some of these techniques require that the transactions exchanging data must execute in the same CICS AOR, and therefore impose restrictions on the dynamic routing of transactions. If transactions exchange data in ways that impose such restrictions, there is said to be an affinity between them.

There are two categories of affinity, intertransaction and transaction-system affinity.

The restrictions on dynamic transaction routing caused by transaction affinities depend on the duration and scope of the affinities. Clearly, the ideal situation for a dynamic transaction routing program is to have no transaction affinities at all. However, even when transaction affinities do exist, there are limits to the scope of these affinities. The scope of an affinity is determined by:

- **Affinity relation:** This determines how the dynamic transaction routing program is to select a target AOR for a transaction instance associated with the affinity.
- **Affinity lifetime:** This indicates how long the affinity exists.

Intertransaction affinity is an affinity between two or more CICS transactions. It is caused when transactions pass information between one another, or synchronize activity between one another, by using techniques that force them to execute in the same CICS AOR. Intertransaction affinity, which imposes restrictions on the dynamic routing of transactions, can occur in the following circumstances:

- One transaction terminates, leaving state data in a place that a second transaction can access only by running in the same CICS AOR.
- One transaction creates data that a second transaction accesses while the first transaction is still running. For this to work safely, the first transaction usually waits on some event, which the second transaction posts when it has read the data created by the first transaction. This synchronization technique requires that both transactions are routed to the same CICS region.

Transaction-system affinity is an affinity between a transaction and a particular CICS AOR (that is, it is not an affinity between transactions themselves). It is caused by the transaction interrogating or changing the properties of that CICS region.

Transactions with affinity to a particular system, rather than another transaction, are not eligible for dynamic transaction routing. Usually, they are transactions that use INQUIRE and SET commands or have some dependency on global user exit programs.

An affinity lifetime is classified as one of the following types:

System: The affinity lasts while the target AOR exists, and ends whenever the AOR terminates (at a normal, immediate, or abnormal termination). The resource shared by transactions that take part in the affinity is not recoverable across CICS restarts.

Permanent: The affinity extends across all CICS restarts. The resource shared by transactions that take part in the affinity is recoverable across CICS restarts. This is the most restrictive of all the intertransaction affinities.

Pseudo-conversation: The LUname or user ID affinity lasts for the entire pseudo-conversation, and ends when the pseudo-conversation ends at the terminal.

Logon: The (LUname) affinity lasts for as long as the terminal remains logged on to CICS, and ends when the terminal logs off.

Signon: The (user ID) affinity lasts for as long as the user is signed on, and ends when the user signs off.

Note: For user ID affinities, the pseudo-conversation and signon lifetime are only possible in situations where only one user per user ID is permitted. Such lifetimes are meaningless if multiple users are permitted to be signed on at the same time with the same userid (at different terminals).

An affinity relation is classified as one of the following:

Global: A group of transactions in which all instances of transactions initiated from any terminal must execute in the same AOR for the lifetime of the affinity. The affinity lifetime for global relations is system or permanent.

LUname: A group of transactions whose affinity relation is defined as LUname is one in where all instances of transactions initiated from the same terminal must execute in the same AOR for the lifetime of the affinity. The affinity lifetime for LUname relations is pseudo-conversation, logon, system, or permanent.

User ID: A group of transactions whose affinity relation is defined as user ID is one in which all instances of transactions initiated from a terminal and executed on behalf of the same user ID must execute in the same AOR for the lifetime of the affinity. The affinity lifetime for user ID relations is pseudo-conversation, signon, system, or permanent.

In a dynamic transaction routing environment, your dynamic transaction routing program must consider transaction affinities to route transactions effectively. Ideally, you should avoid creating application programs that cause affinities, in which case the problem does not exist. However, where existing applications are concerned, it is important that you determine whether they are affected by transaction affinities before using them in a dynamic transaction routing environment.

This is the task the CICS Transaction Affinities Utility is designed to do. It can build a list of affinities, which you can then review. Once this has been done, you decide whether the affinity can be removed (usually through a small application modification or CICS table definition change), or whether it is to be tolerated. When many affinities are tolerated, the ability to perform workload balancing is reduced. The utility can build a list of the affinities that are to be tolerated, and it is possible to put this information into CICSPlex SM. The CICSPlex SM balancing algorithms will then ensure that the affinity is respected.

Recommendation to Use CAU

It is strongly recommended you use the CICS Transaction Affinities Utility (CAU) to identify possible affinities in your CICS systems.

CAU is a part of CICS Transaction Server for OS/390.

IBM Global Services can provide assistance in analyzing the data provided by this utility, which can result in significant time savings.

More details are found in the IBM Redbook *OS/390 MVS Parallel Sysplex Application Considerations Presentation Guide*, SG24-4743, *CICS/TS for OS/390: CICS Transaction Affinities Utility Guide*, SC33-1777 or 4.2.4, "CICS Transaction Affinity Utility" on page 187 of this document.

4.2.2.6 CICS TS for OS/390 and Parallel Sysplex

In this section we look at CICS TS for OS/390 V1.1, V1.2 and V1.3. The functions included in this releases that are related to Parallel Sysplex are:

- CICS TS for OS/390 V1.1
  - CICS/VSAM record level sharing (RLS)
  - Temporary storage data sharing
  - The CICS log manager
  - Enhancements to VTAM generic resources
- CICS TS for OS/390 V1.2
  - Support for the DASD-only option of the system logger
- CICS TS for OS/390 V1.3
  - Dynamic routing and load balancing of distributed program link (DPL) and EXEC CICS START requests
  - CF data tables support
  - Sysplex Wide Enqueue (ENQ) and Dequeue (DEQ)
  - Named Counter Server
  - CICSPlex (R) System Manager enhancements
  - Resource definition online (RDO) for CICS temporary storage

For a complete overview of the functions in CICS TS for OS/390, refer to the *CICS TS for OS/390 Release Guide*, GC33-1570.

The use of the DASD Logger, introduced in OS/390 V2R4, and CICS TS for OS/390 R2, is *not* covered here as it is only intended for a single system environment rather than a Parallel Sysplex.

CICS/VSAM Record Level Sharing (RLS): RLS is a VSAM function, provided by DFSMS/MVS V1.3, and exploited by CICS TS for OS/390 R1 and follow-on releases. RLS enables VSAM data to be shared, with full update capability, between many applications running in many CICS regions. For a description of CICS/VSAM RLS, refer to 4.3.3, "CICS/VSAM Record Level Sharing Considerations" on page 207.

RLS provides many benefits to CICS applications, improving the way CICS regions can share VSAM data. Using RLS you may:

- **Improve availability:** Availability is improved in a number of ways:
  - The FOR is eliminated as a single point of failure. With an SMSVSAM server in each image in the Parallel Sysplex, work can be dynamically routed to another image in the event of a system failure.
  - Data sets are not taken offline in the event of a backout failure. If a backout failure occurs, only the records affected within the unit of work remain locked; the data set remains online.
- **Improve integrity:** Integrity is improved in RLS mode for both reading and writing of data. RLS uses the shared lock capability to implement new read integrity options. CICS supports these options through extensions to the application programming interface (API). For CICS/VSAM RLS usage and sizing related to the CF, refer to 2.12, "CICS/VSAM RLS Structures" on page 137 in Volume 2: Cookbook, SG24-5638 and the CF Sizer tool, available at:
  `http://www.s390.ibm.com/cfsizer/`
- **Reduce lock contention:** For files opened in RLS mode, VSAM locking is at the record level, not at the control interval level, which can improve throughput and reduce response times.
- **Improve sharing between CICS and batch:** Batch jobs can read and update, concurrently with CICS, *non-recoverable* data sets that are opened by CICS in RLS mode. Conversely, batch jobs can read (but not update), concurrently with CICS, *recoverable* data sets that are opened by CICS in RLS mode.
- **Improve performance:** Multiple CICS application-owning regions can directly access the same VSAM data sets, avoiding the need to function ship to file-owning regions. The constraint imposed by the capacity of an FOR to handle all the accesses for a particular data set, on behalf of many CICS regions, does not apply.

Temporary Storage Data Sharing: The CICS TS for OS/390 temporary storage data sharing facility supports the Parallel Sysplex environment by providing shared access to CICS non-recoverable temporary storage queues.

Temporary storage data sharing enables you to share non-recoverable temporary storage queues between many applications running in different CICS regions across the Parallel Sysplex. CICS uses temporary storage data sharing to store shared queues in a structure in a CF, access to which is provided by a CICS temporary storage server address space.

CICS stores a set of temporary storage queues that you want to share in a temporary storage pool. Each temporary storage pool corresponds to a CF list structure defined in the CFRM policy. You can create single or multiple temporary storage pools within the Parallel Sysplex, to suit your requirements, as the following examples show:

- You could create separate temporary storage pools for specific purposes, such as for production or for test and development.
- You could create more than one production pool, particularly if you have more than one CF and you want to allocate temporary storage pool list structures to each CF.

The benefits of temporary storage data sharing include:

- **Improved performance compared with the use of remote queue-owning regions:** Access to queues stored in the CF is quicker than function shipping to a QOR.
- **Improved availability compared with a QOR:** The availability of a temporary storage server is better than that with a QOR because you can have more than one temporary storage server for each pool (typically one server in each image in the Parallel Sysplex). If one temporary storage server or image fails, transactions can be dynamically routed to another AOR on a different image, and the transaction can then use the temporary storage server in that image.
- **Elimination of intertransaction affinities:** Temporary storage data sharing avoids intertransaction affinity.

  For CICS temporary storage usage and sizing related to the CF, refer to 2.13, "CICS Shared Temporary Storage Structures" on page 143 in Volume 2: Cookbook, SG24-5638 and the CF Sizer tool available at: `http://www.s390.ibm.com/cfsizer/`

The CICS Log Manager: The CICS log manager replaces the journal control management function of earlier releases. Using services provided by the system logger, the CICS log manager supports:

- The CICS system log, which is also used for dynamic transaction backout. (The CICS internal dynamic log of earlier releases does not exist in CICS TS for OS/390.)
- Forward recovery logs, auto journals, and user logs (general logs).

The system logger is a component of OS/390 which provides a programming interface to access records on a log stream. Refer to Appendix C, "System Logger Considerations" on page 215 in Volume 2: Cookbook, SG24-5638, for further information.

The CICS log manager uses the services of the system logger to enhance CICS logging in line with the demands of the Parallel Sysplex environment. In particular, it provides online merging of general log streams from different CICS regions which may be on different images in the Parallel Sysplex. The CICS log manager, with the system logger, improves management of system log and dynamic log data (all of which are written to the system log stream) by:

- Avoiding log wraparound
- Automatically deleting obsolete log data of completed units-of-work

All CICS logs (except for user journals defined as type SMF or DUMMY) are written to system logger log streams. User journals of type SMF are written to the SMF log data set.

There are a number of tasks that you must complete in order to set up the CICS TS for OS/390 log manager environment. Refer to *CICS TS for OS/390 Migration Guide*, GC34-5353 and to 2.10, "System Logger Structures" on page 125 in Volume 2: Cookbook, SG24-5638. Specifically, we recommend that you:

- Carefully plan your CF configuration. The system logger requires at least one CF. However, the ability to rebuild CF structures becomes vital if you want to avoid disruptions to your processing environment in case of a CF failure. Therefore, it is recommended that you have two CFs.
- When planning structure sizes, ensure that you allow enough storage to prevent the log stream associated with the CICS system log from spilling to DASD.
- Use the CICS logger CF sizing utility (DFHLSCU) to calculate the amount of CF space you need and the average buffer size of your log streams.
- Make sure that your specifications for the log streams are such that the system logger copies to staging data sets if the CF is (or becomes) volatile.
- Specify each staging data set to be at least the same size as the log stream share of the CF, but round up the average block size to a multiple of 4096.
- Make sure you define a naming convention that is sysplex-wide consistent for CF structures, DASD log data sets, and DASD staging data sets. This will help you in identifying these resources.

The CICS log manager provides several benefits for all users. Refer to the *CICS TS for OS/390 Release Guide*, GC33-1570 for further explanation.

Enhancements to VTAM Generic Resources: Support for VTAM generic resources is enhanced in CICS TS for OS/390 R1, with the aim of improving the usability of generic resources with LU6.2 (APPC) connections. The main benefit of these changes is to facilitate inter-sysplex communication, in which both sysplexes use generic resources. The changes mean the following:

- There is more flexibility in communicating by APPC links between generic resources in partner sysplexes, because routing of connections between sysplexes is controlled by CICS.
- You can inquire of an APPC connection between generic resource sysplexes to see which member of the remote generic resource is in session with a member of the local generic resource.
- Generic resource members can now use APPC, as well as MRO, connections within the CICSplex.

Dynamic Routing and Load Balancing of Distributed Program Link (DPL) and EXEC CICS START Requests: CICS dynamic routing facility is extended to provide mechanisms for dynamically routing transactions started by distributed program link (DPL) requests, and a subset of START commands. Dynamic balancing for DPL includes:

- DPL requests from an external CICS interface (EXCI) client.
- External Call Interface (ECI) requests from any of the CICS Client workstation products.

The routing mechanisms allow workload balancing to be managed by CICSPlex SM, allowing integrated workload balancing for EXCI clients, CICS clients, and started tasks. The ability to dynamically route all types of program link requests improves the performance and reliability of:

- CICS Transaction Gateway
- CICS Web Support
- EXCI calls
- CICS Client ECI calls
- DCE/RPC
- ONC/RPC
- Any function that issues an EXEC CICS LINK Program request

Similarly, the performance and reliability of applications that use the subset of START commands, which can be dynamically routed, are improved.

More information about the use of Dynamic Routing in CICS TS 1.3 can be found in Chapter 10 of the IBM Redbook *CICS TS 1.3 Implementation Guide*, SG24-5274 and in Chapter 4 of *CICS TS 1.3 Release Guide*, GC34-5352.

CF Data Tables Support: CICS CF data tables support allow user applications running in different CICS regions that reside in one or more OS/390 images within a Parallel Sysplex, to share working data with update integrity.

Data in a CF data table is accessed through the CICS file control API, enabling existing applications to use it, either without any modification, or with minimum changes, depending on the level of function required. CF data tables provide efficient sharing of data with integrity, and behave much like a sysplex-wide equivalent of user-maintained data tables. Key lengths greater than 16 bytes are not supported.

A CF data table (CFDT) pool is a CF list structure, and access to it is provided by a CFDT server. A CFDT server is similar to a shared data tables FOR in terms of the function it performs, and it is operationally similar to a temporary storage data sharing server.

For any given CFDT pool, there must be a CFDT server in each OS/390 that wants to access it. CFDT pools are defined in the CFRM policy, and the pool name is then specified in the startup JCL for the server.

The CFDT server runs as a separate address space within an OS/390 image, as either a started task or a batch job. If all of the CFDTs are in a single pool, only one CFDT server is needed in each OS/390 image in the sysplex. If the CFDTs are divided into two separate pools, two CFDT servers are needed in each OS/390, and so on.

The CFDT server runs in a non-CICS address space, so it requires little management or maintenance. The execution environment for CFDT server regions is provided by a run-time package called the Authorized Cross-Memory (AXM) server environment. The AXM server supports CFDT server regions and cross-memory connection services. This environment is also used by the TS data sharing server.

Sysplex-Wide Enqueue (ENQ) and Dequeue (DEQ): The sysplex-wide (global) enqueue and dequeue function enables CICS transactions running in the same region, or in different regions within a sysplex, to serialize on a named resource using the existing CICS API. By extending the scope of the CICS enqueue mechanism, a major source of inter-transaction affinity is removed, enabling better exploitation of Parallel Sysplex environments, improving price/performance, capacity and availability.

For example, serialization makes it possible for concurrent updates to shared Temporary Storage queues by multiple CICS transaction instances, while locking a shared Temporary Storage queue against concurrent updates. This eliminates the race problem created by relying on serial reuse of a principal facility.

The main points of the changes to the CICS enqueue/dequeue mechanism are as follows:

- Sysplex enqueue and dequeue expands the scope of an EXEC CICS ENQ/DEQ command from region to sysplex, by introducing a new CICS resource definition type, ENQMODEL, to define resource names that are to be sysplex-wide.
- ENQSCOPE, an attribute of the ENQMODEL resource definition, defines the set of regions that share the same enqueue scope.
- When an EXEC CICS ENQ (or DEQ) command is issued for a resource whose name matches that of an installed ENQMODEL resource definition, CICS checks the value of the ENQSCOPE attribute to determine whether the scope is local or sysplex-wide, as follows:
  - If the ENQSCOPE attribute is left blank (the default value), CICS treats the ENQ/DEQ as local to the issuing CICS region.
  - If the ENQSCOPE attribute is non-blank, CICS treats the ENQ/DEQ as sysplex-wide, and passes a queue name and the resource name to GRS to manage the enqueue. The resource name is as specified on the EXEC CICS ENQ/DEQ command, and the queue name is made up by prefixing the 4-character ENQSCOPE with the letters DFHE.
- The CICS regions that need to use sysplex-wide enqueue/dequeue function must all have the required ENQMODELs defined and installed. The recommended way to ensure this is for the CICS regions to share a CSD, and for the initialization GRPLISTs to includes the same ENQMODEL groups.

Named Counter Server: The named counter sequence number server provides a facility for generating sequence numbers for use by application programs, both CICS and batch, in a Sysplex. The named counter server is modeled on the other CF used by CICS and has many features in common with the CFDT server. The unique number generated could typically be used for an order or invoice number, implemented as a new key in a keyed file. Each named counter is held in a pool of named counters, which resides in a CF list structure. Retrieval of the next number in sequence from a named counter is done through a callable programming interface available to CICS and batch programs.

CICSPlex (R) System Manager Enhancements: The CICSPlex System Manager is now fully integrated with the base CICS product. All the CICSPlex SM product functions are available, including:

- Business Application Services (BAS) for defining and installing CICS resources across multiple CICS occurrences on S/390 systems.
- Generating messages and generic alerts for triggering automation actions.
- Runtime analysis (RTA) for detecting potential CICS problems and taking automatic corrective actions.

- As well as managing CICS systems in different OS/390 images and regions, CICS/VSE (R) Version 2.3 (5686-026), CICS Transaction Server for OS/2 Warp Version 4.1 (5622-808), can be managed from a CICSPlex System Manager managing address space (CMAS).
- Extensions to CICSplex system management functions include enable Userkey CICSPlex SM API applications.

Resource Definition Online (RDO) for CICS Temporary Storage: RDO for temporary storage eliminates the need to prepare a temporary storage table (TST) for batch assembly and link-edit. There is now no need to shut down and restart CICS in order to make changes to TS queue definitions. RDO support for TS queues is part of the CICS high availability and continuous operations strategy.

Refer to *CICS TS for OS/390 Release Guide*, GC33-1570 for further explanation.

4.2.3 CICSPlex Systems Manager (CICSPlex SM)

Starting with the V1R3 release, CICSPlex Systems Manager (CICSPlex SM) is a component of CICS TS for OS/390 that monitors CICS regions and manages a CICS multiregion option (MRO) environment (Prior to CICS TS for OS/390 R3, CICSPlex SM was a separate program product). CICSPlex SM provides a single image in a CICS/MRO environment through:

- Single point of control for all CICS regions
- Dynamic workload balancing
- Single system image (operations, monitoring, real time analysis)

Two cases of dynamic workload balancing are to be considered:

- When all candidate AORs are running on images in WLM goal mode
- When at least one of the candidate AORs is running on an image in WLM compatibility mode

An introduction to CICSPlex SM is found in 4.2.2.1, “Multiple Region Operation” on page 169.

4.2.3.1 All Images Operating in Goal Mode

If OS/390 is being operated in goal mode, you still have the choice of running the CICS transactions with the “join shortest queue” algorithm. This is sometimes referred to as “queue mode.” The current recommendation is to use queue mode if all transactions are achieving their goal. As the loads increases and some transactions start missing their goal, this may be a good time to move to goal mode. While the overhead is slightly higher, this is compensated by the improved service to the really critical transactions.

To operate CICSPlex SM in goal mode, the CICS TOR(s) must be at least at CICS/ESA V4.1 level. This is because the TOR must be able to communicate with WLM to get the service classes associated with its transactions. Also, the transactions must be defined with an average response time goal rather than a percentile response time goal. If you use percentile response time goals, CICSPlex SM will revert to the shortest queue algorithm. When CICSPlex SM is operated in goal mode with average response time goals, the following events occur:

1. A transaction arrives at a CICS terminal-owning region (TOR).

2. The TOR passes the transactions external properties, such as LU name, user ID, and so on, to the OS/390 WLM.
3. OS/390 WLM uses this information to assign a service class. The service class name is passed back to the TOR.
4. The TOR calls DFHCRP for transaction routing. Amongst other information, the service class name is passed in a comm_area.
5. DFHCRP in turn calls EYU9XLOP (CICSPlex SM).
6. If CICSPlex SM does not already have information about the goal for that service class, it will request that information from WLM.
7. Having the goal of the transaction, CICSPlex SM selects the "best" AOR. The name of this AOR is passed back to the TOR which then routes the transaction to the selected AOR. The selection process is the following:
   a. Route all transactions belonging to a service class that are failing to meet their goals to a specific AOR,
   b. Those transactions that are meeting their goals are routed to another AOR
   c. Those transactions that are exceeding their goals are routed to another AOR

   These AORs could be in the same OS/390, or in another OS/390 in the Parallel Sysplex, or in a remote OS/390. However, the algorithm will favor local AOR regions, in an effort to minimize the routing overhead. AORs that are prone to abend will not be favored, although they may appear to have a very short transaction response time. Refer to the IBM Redbook *CICS Workload Management Using CICSPlex SM and the MVS/ESA Workload Manager*, GG24-4286, for more information.

Note the following:

- Other than requesting a copy of the service policy, there is no other interaction between CICSPlex SM and the WLM component of OS/390.
- All CICS regions: AORs, TORs, FORs, and so on, will continue to report performance information directly to OS/390 WLM. This behaves in the same manner regardless of the presence or absence of CICSPlex SM.
- CICSPlex SM has agents that run in all of the CICS regions and pass performance information back to the CICSPlex SM CMAS so that it is aware of the performance of all the regions that are part of that CICSplex.

4.2.3.2 At Least One Image Operating in Compatibility Mode

Here, we have the following events:

1. A transaction arrives in a CICS terminal owning region (TOR).
2. TOR passes control to CICSPlex SM for dynamic workload balancing.
3. CICSPlex SM selects the least used AOR, which is the one that has the *smallest queue* (also called queue mode) of waiting transactions. This AOR could be in the same OS/390, or in another OS/390 in the Parallel Sysplex, or in a remote OS/390. However, the algorithm tends to favor local AOR regions.

Note: If you have CICS and you do not install CICSPlex SM, it is possible to have dynamic workload distribution through exit routines in the TOR.

4.2.4 CICS Transaction Affinity Utility

A transaction affinity is a relationship, between transactions of a specified duration, that require them to be processed by the same AOR. Detecting and documenting the affinities in your applications is a daunting task. It is probably almost impossible to perform it manually for most CICS installations. You should use the IBM CICS Transaction Affinities Utility for this task. Refer to 4.2.2.5, "Affinities and CICS" on page 177 for more discussion on CICS and affinities. You can also refer to the IBM Redbook *CICS Workload Management Using CICSPlex SM and the MVS/ESA Workload Manager*, GG24-4286, and *CICS/ESA V4 Dynamic Transaction Routing in a CICSplex*, SC33-1012.

The Transaction Affinities Utility is available as part of the base CICS product starting with CICS TS for OS/390 R1, and can be used to detect transaction affinities in application programs for CICS TS for OS/390 regions only. If you want to detect transaction affinities in releases of CICS earlier than CICS TS for OS/390 R1, the CICS Transaction Server for OS/390 Transaction Affinities Utility (CAU) is available from IBM. The product number is 5696-582. This product includes both data gathering and data analysis functions. You can also get the data gathering part from the Internet at:
`http://www.hursley.ibm.com/cics/txppacs/cs1a.html`.

You would then have to contact the IBM Services organization to have them do the analysis of the data you gathered.

4.2.5 IMS Transaction Manager in Parallel Sysplex

The IMS/ESA V6 Transaction Manager TM provides substantial exploitation of Parallel Sysplex functions. This includes sysplex-wide IMS command processing and the use of system logger and VTAM Generic Resources with their associated CF structures.

Prior to IMS/ESA V6 it was possible for IMS TM to route transactions between one IMS TM control region and another using Multiple Systems Coupling (MSC) facility. However, to exploit this facility you often needed to either code exits or use IMS/ESA Workload Router (5697-B87). True workload balancing is made possible by the introduction of the shared message queues (SMQ), which enable multiple IMS systems to have a single view of the end user workload.

4.2.5.1 IMS Command Processing

In IMS/ESA V6 it is possible for commands entered from a single console to be routed to multiple IMS systems and for the responses to the command to be returned to that console. This capability is discussed in more detail in the section entitled *IMS Command Processing*, in the IBM Redbook *Parallel Sysplex - Managing Software for Availability*, SG24-5451.

4.2.5.2 IMS VTAM Generic Resources

VTAM generic resources (GR) allow a terminal end user to reference a collection of IMS subsystems in a Parallel Sysplex by a single name. This collection is known as a Generic Resource Group (GRG), and IMS systems may join or leave it non-disruptively. Figure 46 on page 188 shows a sample IMS GRG.

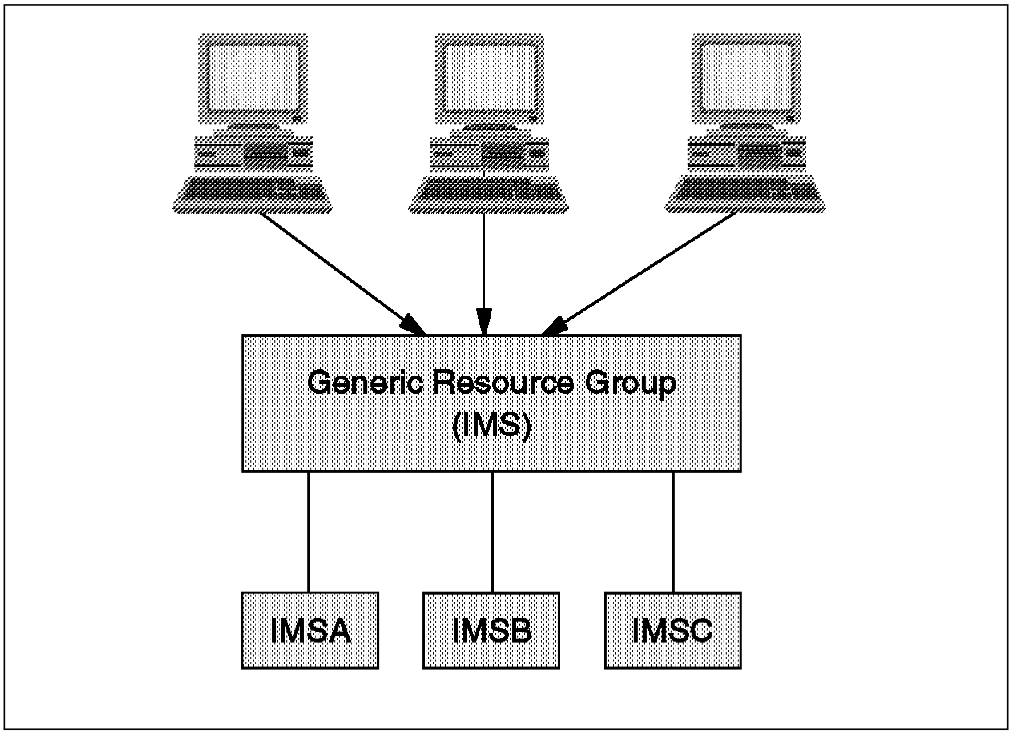


Figure 46. Sample Generic Resource Group in an IMS Environment

If *any* of the OS/390 systems where IMS resides are still in WLM compatibility mode, VTAM attempts to balance the network workload by routing a session request to an IMS in the GRG using the following selection criteria:

- By using existing mapping or affinity
- Session counts
- As directed by the VTAM generic resource Exit

If *all* of the systems are in WLM goal mode, VTAM will choose its recommended instance, based on these criteria. It will then request a recommended application instance from WLM. WLM will make a choice based on available capacity and service consumption by importance on all the systems in the GRG. VTAM will then pass both its own recommended instance, together with WLM's recommended instance, to the generic resource exit. If that exit is missing or inactive, the session will be routed to the instance recommended by WLM.

In the event of an IMS failure, the end users are sometimes able to log back on to the GRG and re-establish their sessions with one of the surviving IMS systems. Depending on the user's status, re-establishing a session may not be possible. Examples of situations where it is not possible to re-establish a session include those users who are in response mode or in an IMS conversation. If there is an OS/390 or CPC failure, the user may not be able to logon again, because IMS may not have been able to delete its affinity to the terminals. For more information about VTAM GR, refer to 4.5.1, “VTAM Generic Resources Function” on page 219. For more information on IMS support for VTAM GR, refer to the IBM Redbook *Using VTAM Generic Resources with IMS*, SG24-5487.

Note: XRF and RSR IMS systems *cannot* be members of a GRG.

IMS/ESA V7, which is not yet generally available at the time of writing, introduces some enhancements to IMS use of VTAM GR. One of these is the ability to have VTAM rather than IMS manage affinities at session termination. This new function is rolled back to IMS/ESA V6 via APAR PQ18590.

IMS support for VTAM Multi Node Persistent Sessions (MNPS) is introduced by IMS/ESA V7, in a new feature called Rapid Network Reconnect (RNR). Using RNR, if IMS fails, user sessions will be automatically reestablished with IMS after it restarts, either on the same system or on another system. RNR is mutually exclusive with XRF. Further information will be available in the IMS manuals when it becomes generally available.

4.2.5.3 IMS Shared Message Queues

Shared message queues provide the mechanism for up to 32 IMS DB/DC and/or DBCTL subsystems to have access to a common set of message queues. There are two types of queues, one for Full Function and one for Fast Path, both of which are stored as list structures in a CF. There are potentially two structures per queue, a primary and an overflow, the overflow structure being optional. The term “structure pair” is used in the following text as a general reference to the message queues whether there is only a primary or both a primary and an overflow. For information on the sizing of these structures, refer to 2.7.2.2, “IMS OSAM and VSAM Structure Sizing” on page 100 in Volume 2: Cookbook, SG24-5638 and the CF Sizer tool available at:
`http://www.s390.ibm.com/cfsizer/`

IMS does not access the structures directly; this is achieved by the use of services provided by an address space, introduced with IMS/ESA V6, the common queue server (CQS).

For a discussion on IMS DB usage of structures, refer to 4.3.2.3, “IMS Data Sharing Groups” on page 204.

Affinities and IMS TM: Implementation of shared message queues is simplified by the basic design of IMS. In general, IMS does not have facilities (such as temporary storage and transient data) that leave data in an address space for later retrieval, because it has always used a multiple region architecture.

Although the problem of “affinities” in IMS is almost nonexistent, you might want to also look at the following packages.

IMS Affinities Analysis

A tool called IMSSPLIT may reveal possible IMS affinities. This tool is discussed in more details in the IMSSPLIT package. This MKTTOOLS package is available to IBM representatives. For a short description of IMSSPLIT, refer to Appendix A, “Tools and Services Catalog” on page 161 in Volume 2: Cookbook, SG24-5638.

Full Function Message Queue: The long and short message queues, which were on DASD in previous releases, are combined into a single structure, referred to as the MSGQ. Message queues are supported either in a CF structure or on DASD but you are *not* allowed to have a combination of the two. The exception to this is XRF, which is discussed briefly in “Common Queue Server (CQS)” on page 191.

To obtain a message from the MSGQ, an IMS registers interest with its local CQS and is notified when a message is placed on the queue. This means that

IMS does not have to request work whenever a resource such as a message region is available.

The flow of work through the shared queues group (SQG) is simple in concept, with each discrete element of work being potentially available to any IMS that has registered an interest. Therefore, terminal input, transaction processing, and terminal output could be processed on one system or three depending on their workloads at the time.

Figure 47 shows a sample of IMS systems sharing message queues.

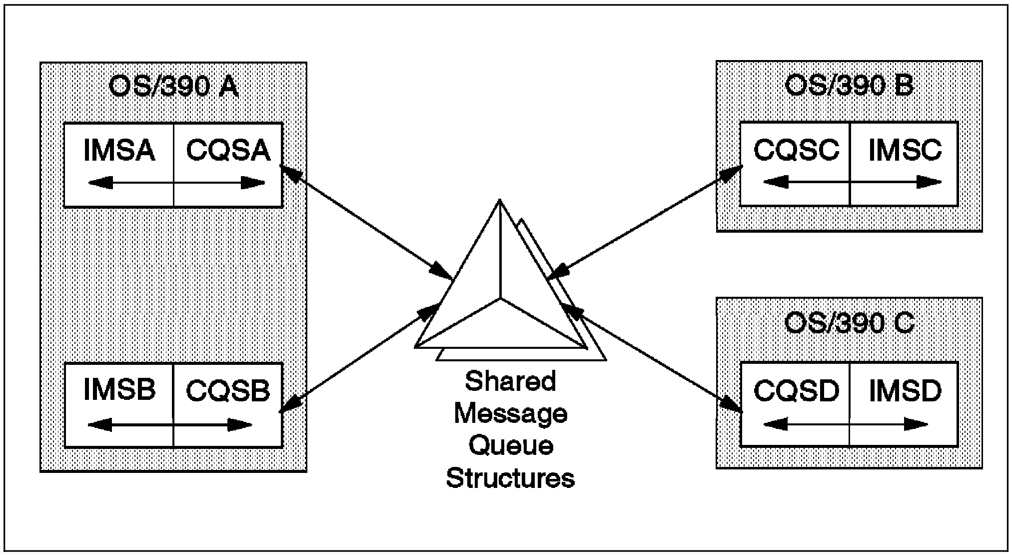


Figure 47. Sample IMS Systems Sharing Message Queues

The package called IMSSMQCP on MKTTOOLS contains a tool to assist you in planning CF resources for the IMS shared message queue structures. For more information about IMSSMQCP, refer to Appendix A, "Tools and Services Catalog" on page 161 in Volume 2: Cookbook, SG24-5638.

Fast Path Queue: In releases of IMS prior to IMS/ESA V6, Fast Path did not use a message queue. IMS V6.1 introduced an optional CF structure, known as the Expedited Message Handler Queue (EMHQ). This is fundamentally the same as the MSGQ, but processing can be directed by exit DBFHAGU0, such that the EMHQ is not used. The exit can select one of three options, *local only, local first,* or *global only.*

1. **Local only:**

   This option ensures that work executes on the receiving system. The EMHQ is not used

2. **Local first:**

   This option will try to have the work executed on the receiving system. If this is not possible, then the EMHQ is used.

3. **Global only:**

   This option always put the message on the EMHQ.

Overflows: The primary structure is divided into queues which represent resources such as transactions and LTERMs. If the primary structure reaches the percentage in use threshold, defined at CQS startup, then the most active queues may, if permitted by a user exit, be moved to the overflow structure.

Common Queue Server (CQS): CQS acts as a server to manage shared queues on behalf of a client, in this case an IMS subsystem. It is also available to non-IMS clients. Up to 32 CQSs may connect to one structure with each CQS only having one client (IMS V7 provides support for up to 32 clients per CQS address space). Each CQS has its own address space and communicates with the other members of the sharing group by using XCF services. Access to the shared queues is via XES services.

CQS is responsible for recovery in the event of a structure or CF failure. This is achieved through a combination of structure recovery data sets (SRDSs), two per structure pair, and a merged log of activity from all sharing CQSs.

The SRDSs are used to periodically checkpoint the queues from their associated structure pairs. Activity between checkpoints is contained in a CQS log stream. If a structure needs to be rebuilt the SRDS is read, and forward recovery is performed using data from the log stream.

An RSR tracking subsystem may support multiple IMS subsystems using shared queues. After an RSR takeover, the remote site may run with either shared queues or message queue data sets.

4.2.5.4 Migration and Compatibility Considerations

To enable IMS support for VTAM generic resources, all VTAMs in the sysplex must be at V4.2 or above.

To exploit shared message queues the minimum requirement is MVS/ESA V5.2 and all IMSs must be V6.1 or above.

For IMS/ESA V7, OS/390 must be at the 2.6 or later level.

4.2.5.5 IMS Workload Routing in an MSC Environment

In IMS TM, there are user exits to effect workload distribution within connected IMS systems. The Input Message Routing exit routine (DFSNPRT0) provides the ability to change the destination system identification (SID) of the input message immediately after it is received by IMS from the input device (or LU6.2 program). By changing the destination name, you can reroute the message to a different destination or IMS system than was originally defined (SYSGEN), or is currently defined (/ASSIGN or /MSASSIGN).

In a Parallel Sysplex environment or in other environments where transaction macros are cloned as local transactions across IMS systems, this exit routine can control the routing of transaction messages to the various IMS systems. This is done by returning a MSNAME or SID to IMS, causing the message to reroute (over the MSC link) to another IMS system for processing. By implementing this exit routine, you are able to implement transaction load balancing within the context of this exit routine. This might be of interest to you as an interim solution prior to your implementation of IMS/ESA V6, which offers dynamic workload balancing via shared queues.

For a more detailed discussion of the IMS TM exit routines refer to the *IMS/ESA V5 Customization Guide*, SC26-8020.

An IBM program product uses these exits to provide workload routing. It is called the IMS/ESA Workload Router (5697-B87). The IMS/ESA Workload Router is designed to enable IMS/ESA V5 (and above) installations to exploit the increased availability, expanded capacity, and cost benefits of the Parallel Sysplex. The product enables IMS transaction routing and workload shifting among CPCs within the Parallel Sysplex, and is offered as an interim solution. The software implementation is generalized and potentially useful to all IMS sites wishing to exploit transaction routing across a Parallel Sysplex.

Figure 48 illustrates how IMS systems (of which one or more may belong to a Parallel Sysplex) can use MSC connections to enable workload routing.

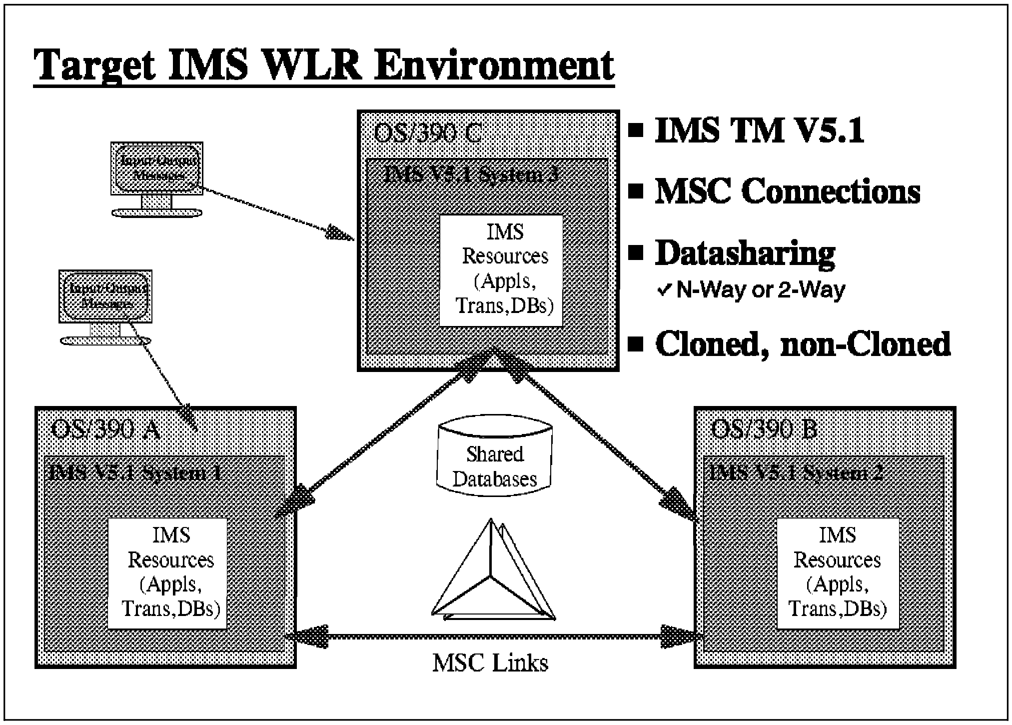


Figure 48. IMS Workload Router

IMS/ESA Workload Router functions as IMS TM user exits to distribute an IMS transaction workload among two or more IMS online systems interconnected through Multiple Systems Coupling (MSC) communications links. Users may view this configuration as a single system image by accessing only one of the IMS systems directly.

The IMS/ESA Workload Router architecture does not preclude direct access to any of the IMS systems by end users. IMS/ESA Workload Router distributes the IMS transaction workload among pre-defined "paths." The paths are equivalent to MSC MSNAMEs defined to the IMS system. Multiple paths may be defined between IMS systems. Each path has an associated goal that represents the percentage of the total number of IMS transactions to be distributed to the path, as the following example shows:

```
Path A GOAL = 20
Path B GOAL = 20
Path C GOAL = 60
```

In this example, 20% of the transactions would be routed to Path A, 20% to Path B, and 60% to Path C. Local processing of the transactions can take place on, for example, Path C. In addition to balancing the workload between IMS systems, the architecture provides for balancing the flow over the MSC links within the configuration.

IMS/ESA Workload Router is sensitive to the availability of MSC link components. If it finds that an MSC link is not available, it will automatically route arriving transactions to another MSC link. When the link becomes available, IMS/ESA Workload Router will automatically reinstate work to it.

IMS/ESA Workload Router provides for the notion of *affinity.* Transactions may be specifically or generically included or excluded from routing. Certain transactions may be directed toward certain paths, processed locally only, or remote-preferred with affinity definitions. The affinity definitions provide for the designation of specific transaction names, generic transaction names, or affinity by IMS transaction scheduler class. IMS/ESA Workload Router includes an online administration interface. The interface executes as an unauthorized IMS MPP in a standard IMS Message Processing Region. It provides for monitoring and dynamically changing the IMS/ESA Workload Router configuration.

IMS/ESA Workload Router routes transactions based on rules established by you. Should a CPC or system fail, you are notified of this. Through automation, you can adjust the flow of transactions away from the failure. IMS/ESA Workload Router is not dynamic. It does not interface with WLM and therefore transactions are truly routed instead of dynamically balanced across the Parallel Sysplex.

For more information on WLR V2, refer to IBM announcement letter ZG96-0664 (296-478), *IMS/ESA Workload Router V2 Offers Greater Routing Flexibility*. Also take a look at the IBM Redbook *IMS/ESA Multiple Systems Coupling in a Parallel Sysplex*, SG24-4750.

4.2.5.6 IMS Performance Analyzer R2

IMS Performance Analyzer R2 provides enhancements to existing functions and support for the IMS/ESA V6 Shared Queues environment. IMS Performance Analyzer R2:

- Supports the processing of multiple IMS subsystem IDs
- Provides log merge support for reporting
- Provides VSCR relief
- Supports VTAM (R) node and APPC reporting
- Removes some current restrictions
- Provides a number of usability enhancements

IMS PA is a functional replacement for both IMSPARS and IMSASAP, and has no GPAR dependency.

The IMS (TM) Performance Analyzer (IMS PA) can help provide you with the information needed to manage IMS resources and system performance. IMS PA provides more than 30 reports to meet a broad range of IMS systems analysis objectives.

Specifically, IMS PA can help:

- Improve transaction response times
- Improve message region and queue use
- Improve buffer pool usage
- Increase availability of resources
- Increase the productivity of system and application programmers

- Reduce the need to run IMS utilities
- Provide ongoing system management and measurement reports

In addition, IMS PA includes:

- An easy-to-use ISPF user interface
- Improved reporting capabilities
- Support for IMS Versions 4, 5, and 6, from a single load library

4.3 Database Management in Parallel Sysplex

In this section we look at database management software. The main focus is on how to set up the software for optimal exploitation of the Parallel Sysplex environment. The following database management software is covered:

- DB2
- IMS DB
- CICS/VSAM RLS

4.3.1 DB2 Data Sharing Considerations

DB2 supports two type of data sharing:

- Shared read-only, which does not exploit the sysplex and allows multiple DB2 systems share data with read access only.
- Full read/write sharing, which requires a Parallel Sysplex and allows multiple DB2 subsystems to have read *and* write access to shared databases.

4.3.1.1 DB2 Read-Only Data Sharing

In a DB2 read-only data sharing environment, one DB2 owns data in a given shared database and has exclusive control over updating the data in that database. A database is a logical construct that contains the physical data in the form of index spaces and table spaces. We use the term owner or owning DB2 to refer to the DB2 subsystem that can update the data in a given database. Other DB2s can read, but not update, data in the owner's database. These other DB2s are read-only DB2s, or readers.

You do not create any physical data objects on the reader, but you do perform data definition on the reader. This is so that the reader's catalog can contain the data definitions that mimic data definitions on the owning DB2. This allows applications running on the reader to access the physical data belonging to the owner. With shared read-only data, the owning DB2 *cannot update* at the same time other DB2 are reading. Any read-only access must be stopped before any updates can be done. Support for shared read-only data is dropped beginning with DB2 UDB for OS/390 V6.

4.3.1.2 DB2 Read-Write Data Sharing

In DB2 V4, IBM introduced function that provides applications with full read and write concurrent access to shared databases. DB2 data sharing allows users on multiple DB2 subsystems to share a single copy of the DB2 catalog, directory, and user data sets. The DB2 subsystems sharing the data belong to a DB2 data sharing group. The DB2 subsystems must reside in a Parallel Sysplex and use of a CF is required. The advantages of DB2 data sharing are as follows:

- Availability
- Flexible configurations

- Improved price performance
- Incremental growth and capacity
- Integration of existing applications

For a more detailed description of the advantages gained by implementing a Parallel Sysplex, refer to 1.4, "Main Reasons for Parallel Sysplex" on page 7.

4.3.1.3 Data Sharing Group

A data sharing group is a collection of one or more DB2 subsystems accessing shared DB2 data. Each DB2 subsystem belonging to a particular data sharing group is a member of that group. All members of the group use the same shared catalog and directory. The maximum number of members in a group is 32. A data sharing environment means that a group has been defined with at least one member. A non-data sharing environment means that no group has been defined. A DB2 subsystem can only be a member of one DB2 data sharing group.

It is possible to have more than one DB2 data sharing group in a Parallel Sysplex. You might, for example, want one group for testing and another group for production data. Each group's shared data is unique to that group. DB2 assumes that all data is capable of being shared across the data sharing group.

Actual sharing is controlled by CPC connectivity and by authorizations. However, DB2 does not incur unnecessary overhead if data is not actively shared. Controls to maintain data integrity go into effect only when DB2 detects inter-subsystem read/write interest on a page set.

The following DASD resources must be shared by a data sharing group:

- Single shared OS/390 catalog; user catalogs for DB2 must be shared to avoid ambiguities.
- Single shared DB2 catalog; any resource dedicated to one system (for example, any DB2 tablespace), must be unique in the DB2 sharing group.
- Single shared DB2 directory.
- Shared databases.
- The LOG data sets are unique to each DB2 member; however, they are read by all members of the data sharing group.
- The boot strap data sets (BSDSs) are unique to each DB2 member; however, they are read by all members of the data sharing group.

It is recommended that DB2 work files are also shared between data sharing members. This enables you to restart DB2 on other systems when needed. Shared work files are required in order to enable the use of Sysplex Query Parallelism.

As we see in Figure 49 on page 196, each DB2 system in the data sharing group has its own set of log data sets and its own bootstrap data set (BSDS). However, these data sets *must* reside on DASD that is shared between all members of the data sharing group. This allows all systems access to all of the available log data in the event of a DB2 subsystem failure.

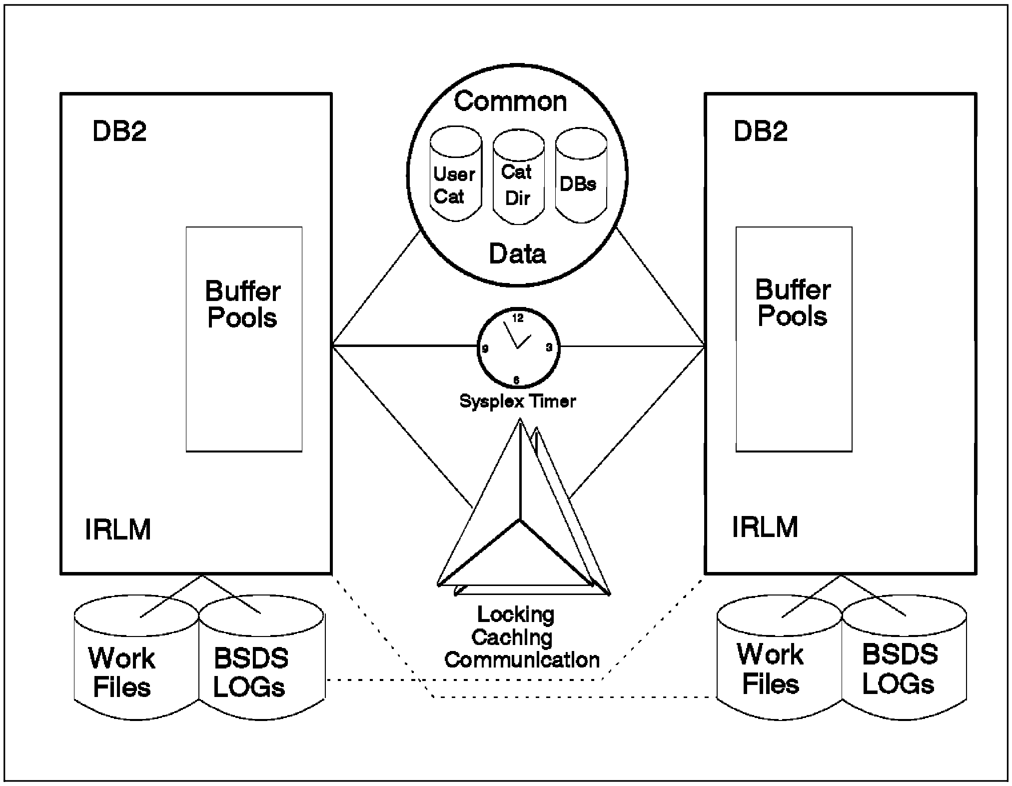


Figure 49. DB2 Data Sharing Group in a Parallel Sysplex

4.3.1.4 Single System Image

A data sharing group presents a single system image using one DB2 catalog and conceptually sharing all user databases. Up to 32 DB2 subsystems can read and write to the same databases. The CF allows tables to be treated like local tables from any DB2 in the data sharing group on different OS/390s.

A transaction executed in one system is dependent on its log files. Therefore, for a non-distributed transaction, only the DB2 that began the transaction keeps all of the information needed to successfully complete it. The same is true for the transaction manager.

All data across a DB2 data sharing group is capable of being shared. Any table or tablespace is assumed to be shared across the DB2 data sharing group, including the DB2 catalog and directory. The physical connections required for data sharing are assumed to be available. DB2 dynamically optimizes data access when only one DB2 is accessing it.

Although authorization changes are effective across the DB2 data sharing group, actual sharing is controlled by physical DASD/CPC connectivity. GRANT and REVOKE need to be issued only once and are valid across the data sharing group.

Data access concurrency is supported at every level, and data sharing is transparent to the DB2 user. For example, row level locking appears to be the same whether done in a data sharing environment or not. Locking is done only by the IRLM - IRLM then uses XES services to communicate with the CF, however, this is transparent to the user or application developer.

4.3.1.5 How DB2 Data Sharing Works

This section provides background information about how shared data is updated and how DB2 protects the consistency of that data. For data sharing, you must have a Parallel Sysplex.

Data is accessed by any DB2 in the group. Potentially, there can be many subsystems reading and writing the same data. DB2 uses special data sharing locking and caching mechanisms to ensure data consistency.

When one or more members of a data sharing group have opened the same table space, index space, or partition, and at least one of them has been opened for writing, then the data is said to be of "inter-DB2 R/W interest" to the members (we shorten this to "inter-DB2 interest"). To control access to data that is of inter-DB2 interest, DB2 uses the locking capability provided by the CF. DB2 also caches the data in a storage area in the CF called a GBP structure, whenever the data is changed.

When there is inter-DB2 interest in a particular table space, index, or partition, it is dependent on the GBP, or GBP-dependent. You define GBP structures using CFRM policies. For more information about these policies, see *OS/390 MVS Setting Up a Sysplex*, GC28-1779.

There is mapping between a GBP and the local buffer pools of the group members. For example, each DB2 has a buffer pool named BP0. For data sharing, you must define a GBP (GBP0) in the CF that maps to buffer pool BP0. GBP0 is used for caching the DB2 catalog, directory table spaces and index along with any other table spaces, indexes, or partitions that use buffer pool 0.

To make a particular database eligible for sharing, you would define a Group Buffer Pool corresponding to the local buffer pool being used by the database. If you have a database that you do not wish to share, you would simply not define a GBP. You can put GBPs in different CFs. GBPs are used for caching data of interest to more than one DB2, to cache pages read in from DASD, and as a cross-invalidation mechanism for buffer coherency across DB2 members.

When a particular page of data is changed by one DB2, DB2 caches that page in the GBP. The CF invalidates any image of the page in the local buffer pools of any other members that currently have a copy of that page. Then, when a request for that same data is subsequently made by another DB2, it looks for the data in the GBP.

Figure 50 on page 198 shows the process of DB2 data sharing in more detail.

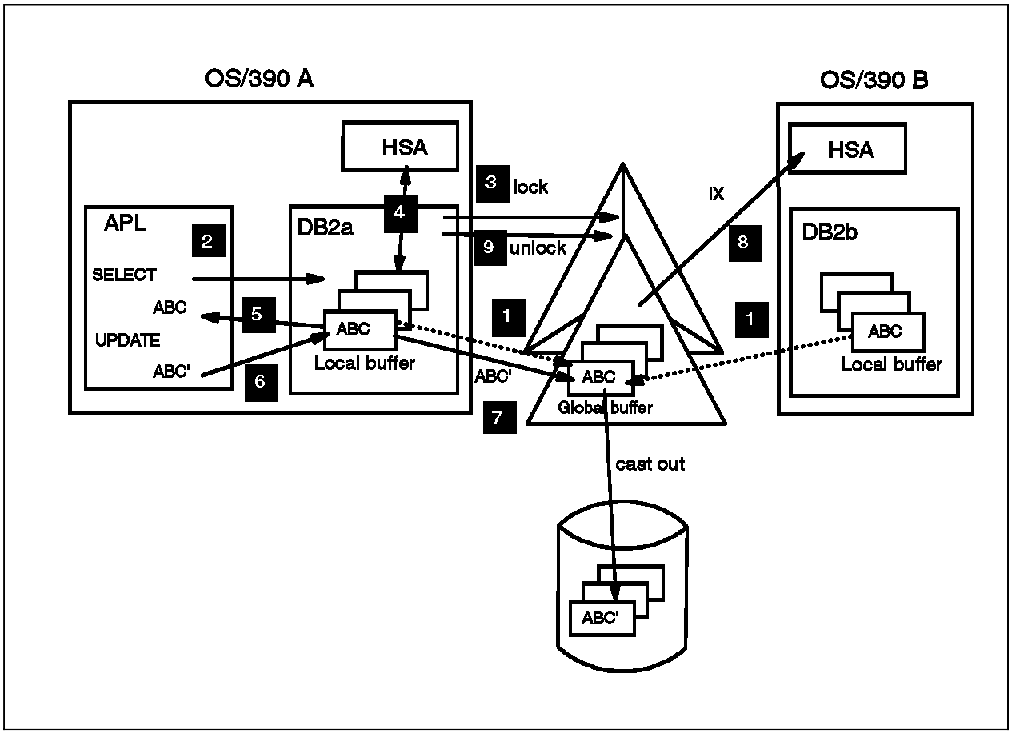


Figure 50. DB2 Data Sharing in a Parallel Sysplex

In Figure 50, both OS/390 A and OS/390 B have registered with the CF (.1.) their interest in page ABC, a current copy of which exists in each systems local buffers.

An application in OS/390 A needs to update the page to ABC' (.2.). DB2 in OS/390 A calls the CF to obtain an exclusive lock for the update (.3.). OS/390 A checks its HSA vector table (.4.), to ensure that the page in its buffer is a valid copy.

OS/390 A changes the page to ABC' (.5.), which is subsequently changed in its local buffer (.6.), and written to the GBP in the CF (.7.). The CF now invalidates OS/390 B's local buffer (.8.) by changing the bit setting in the HSA vector associated with the page ABC. Finally, the lock held by OS/390 A on the page is released (.9.).

For a further description of this process, refer to *DB2 UDB for OS/390 V6 Data Sharing: Planning and Administration*, SC26-9007. For a discussion on buffer pool sizing for DB2, see 2.8, "DB2 Structures" on page 106 in Volume 2: Cookbook, SG24-5638 and use the CF Sizing tool, available at:
`http://www.s390.ibm.com/cfsizer/`

4.3.1.6 Writing Changed Data to DASD

DB2 uses a *castout* process to write changed data to DASD from a GBP. When data is cast out from a GBP to DASD, that data must first pass through a DB2's address space because there is no direct connection from a CF to DASD. This data passes through a private buffer, not DB2's virtual buffer pools. You can control the frequency of castouts with thresholds and checkpoints for each GBP. Any of the DB2 subsystems in the data sharing group may do the castout. The DB2 that is assigned ownership of castout is the DB2 subsystem that had the first update intent (except during restart) on the page set or partition. After the castout ownership is assigned, subsequent updating DB2 subsystems become backup owners. One of the backup owners becomes the castout owner when the original castout owner no longer has read/write interest in the page set. For additional information, please refer to 2.8.2.2, "DB2 GBPs" on page 108 in Volume 2: Cookbook, SG24-5638.

4.3.1.7 DB2 Usage of CF Structures

During startup, DB2 members join one XCF group, and the associated integrated resource lock managers (IRLMs) join another XCF group. To join, they use the names you specify during DB2 installation.

The Sysplex Timer keeps the CPC time-stamps synchronized for all DB2s in the data sharing group. DB2 uses a value derived from the time-stamp to replace the RBA when sharing data.

At least one CF must be installed and defined to OS/390 before you can run DB2 with data sharing capability. Before starting DB2 for data sharing, you must have defined one lock structure and one list structure. One-way data sharing does not require the definition of a cache structure, it is only when you go to more than one-way data sharing that you must define at least one cache structure (GBP 0). DB2 uses the three types of CF structures as follows:

- **Cache structures:** Cache structures are used as GBPs for the DB2 data sharing group. DB2 uses a GBP to cache data that is of interest to more than one DB2 in the data sharing group. GBPs are also used to maintain the consistency of data across the buffer pools of members of the group by using a cross-invalidating mechanism. Cross-invalidation is used when a particular member's buffer pool does not contain the latest version of the data.

  If data is not to be shared (that is, it will be used only by one member), then choose a buffer pool for those non-shared page sets that does not have a corresponding GBP. Assume you choose BP6. Every other member must define its virtual buffer pool 6 with a size of 0 and there should not be a GBP6 defined.

  Depending on the GBPCACHE option specified, GBPs may cache shared data pages, which are registered in the GBP directory. This registration allows XES cache structure support to cross invalidate the data pages when necessary. Changes to a registered resource invalidate all other registered copies of that resource in the local buffer pools.

  Reuse of an invalidated resource results in a reread of that resource. The XES cross invalidation advises DB2 to reread (from either CF or DASD) the page when needed. The CF invalidates a given data buffer in all of the local buffer pools when the data is changed by another subsystem. For more information, refer to 2.8.2.2, "DB2 GBPs" on page 108 in Volume 2: Cookbook, SG24-5638.

  Cache structure services, accessed by DB2, provides the following functions:

  – Automatic notification to affected DB2s when shared data has been changed (cross-system invalidation). The CF keeps track of which DB2s are using a particular piece of data and, the CF updates a bit in the HSA on each CPC that contains a copy of that record.

  – Maintenance of cache structure free space. The CF maintains lists of which entries in a cache structure have been changed and which have not. These lists are kept in the order of most recently used. When the CF needs to reclaim space in a structure, it does so from the list of unchanged entries, using the oldest (or least recently used). Entries on the changed list are not eligible for reclaim.

– Data may be read and written between a GBP and a local buffer pool owned by a single DB2.

– Maintenance of a secondary copy of the GBP. GBP duplexing is available with DB2 V5 (via APAR PQ17797) and later releases.

The GBP in DB2 V4 and later releases is implemented as a CF cache structure. Each GBP contains:

– Directory entry

  The directory entry contains references for each page represented in the GBP. It has slots for all users of that particular page, and is used, for example, to notify all users of cross-invalidation for that page. The entry indicates the position of the data. Shared data can be located in the related local buffer pool of more than one member of a data sharing group at the same time.

– Data entry (GBPCACHE ALL, CHANGED, or SYSTEM)

  The GBP pages are implemented as data entries in the cache structure of the CF. The GBP is maintained by the participating DB2s. It contains the GBP-dependent pages. Data entries are either 4 KB, 8 KB, 16 KB, or 32 KB.

- **List structure:** There is one list structure per data sharing group used as the shared communications area (SCA) for the members of the group. The SCA keeps DB2 data sharing group member information; it contains recovery information for each data sharing group member. The first connector to the structure is responsible for building the structure if it does not exist. For more information, refer to 2.8.2.1, "DB2 Shared Communications Area (SCA)" on page 107 in Volume 2: Cookbook, SG24-5638.

  The SCA is used in DB2 V4 and later releases to track database exception status.

- **Lock structure:** One lock structure per data sharing group is used by IRLM to control locking. The lock structure contains global locking information for resources on behalf of the requesting IRLM and DB2. It protects shared resources and allows concurrency. The system lock manager (SLM), a component of XES, presents the global lock information to the lock structure. For more information, refer to D.3, "IRLM Lock Structure Used for DB2 Data Sharing" on page 229 in Volume 2: Cookbook, SG24-5638.

  A lock structure is used to serialize on resources such as records/rows, pages, or table spaces. It consists of two parts:

  – A CF lock list table, which consists of a series of lock entries that associate the systems with a resource name that has been modified, including a lock status (LS) of that resource.

    A resource is any logical entity such as a record/row, a page, partition, or an entire tablespace. DB2 defines the resources for which serialization is required. A CF lock list table has information about the resource used by a member DB2 and is used for recovery if a DB2 fails. One common set of list elements is used for all members of the data sharing group.

  – CF lock hash table. Each hash entry is made up of one SHARE bit for each member of the data sharing group and one EXCLUSIVE byte.

For more information about DB2 and IRLM structures, refer to 2.8, “DB2 Structures” on page 106 in Volume 2: Cookbook, SG24-5638. For more information about DB2 disaster recovery in Parallel Sysplex, refer to 3.7.4, “DB2 Disaster Recovery Considerations” on page 133 of this book.

4.3.1.8 Data Sharing Enhancements in DB2 V5 and V6

Since the initial introduction of data sharing support in DB2 V4, there have been numerous enhancements to DB2 to further improve performance and availability in a data sharing environment. In this section, we briefly review those enhancements. For more information, refer to *DB2 for OS/390 V5 Release Guide*, SC26-8965, *DB2 UDB for OS/390 V6 Release Guide*, SC26-9013, and *DB2 UDB for OS/390 V7 What's New*, available on the Internet at:
`http://www.ibm.com/software/data/db2/os390/pdf/whatv7.pdf`

The sysplex-related enhancements in DB2 V5 and V6 are as follows:

- **Improvements in query processing:** The full power of a Parallel Sysplex can be used not only to split a read-only query into a number of smaller tasks, but also to run these tasks in parallel across multiple DB2 subsystems on multiple CPCs in a data sharing group. *Sysplex Query Parallelism* is supported by combining the data from all parallel tasks, regardless of the data sharing member on which they were executed. You still get an application view of the thread, just as for *CP Query Parallelism* in DB2 V4.1. TCB times of parallel tasks running on CPCs with different processor speed are adjusted so that they can be analyzed in a meaningful way.
- **Enhanced data sharing support in DB2 V5:**
  - To simplify the monitoring of applications in a data sharing environment, *group-scope* reports can be generated for accounting. This is especially helpful if you use OS/390 Workload Management to dynamically schedule your applications on different data sharing members. Accounting group-scope reports help you get a complete picture of the resources an application has used, regardless of which member it ran on.
  - GBP rebuild makes CF maintenance easier and improves access to the GBP during connectivity losses. Automatic GBP recovery accelerates GBP recovery time, eliminates operator intervention, and makes data available faster when GBPs are lost because of CF failures.
  - Improved restart performance for members of a data sharing group reduces the impact of retained locks by making data available faster when a group member fails.
  - Continuous availability with GBP duplexing in a Parallel Sysplex makes recovery simpler and faster in the event of a CF failure (requires APAR PQ17797).
- **DB2 Usage of Sysplex Routing for TCP/IP:** DB2 provides a service called sysplex routing for TCP/IP DRDA requesters. This allows systems connecting to the DB2 data sharing group using DRDA over TCP/IP connections to have their requests routed to the DB2 server that WLM determines to be the least loaded.

  At the time of writing, the DRDA requesters enabled for this function are another OS/390 for DB2 requester, DB2 Connect Enterprise Edition, and the DRDA client on a PC. This function is similar to that introduced for DRDA requesters using APPC in DB2 V4.1.

- **Query performance enhancements in DB2 V6 include:**
  - Query parallelism extensions for complex queries, such as outer joins and queries that use non-partitioned tables.
  - Faster restart and recovery with the ability to postpone backout work during restart, and a faster log apply process.
  - Increased flexibility with 8 KB and 16 KB page sizes for balancing different workload requirements more efficiently, and for controlling traffic to the CF for some workloads.
  - An increased log output buffer size (from 1000 4 KB to 100000 4 KB buffers) that improves log read and write performance.
- **Data sharing enhancements in DB2 V6:**
  - Continuous availability with GBP duplexing in a Parallel Sysplex makes recovery simpler and faster in the event of a CF failure.
  - More caching options for using the CF improve performance in a data sharing environment for some applications by writing changed pages directly to DASD.
- **Data sharing enhancements in DB2 V7:**
  - *Restart Light*

    A new feature of the START DB2 command allows you to choose Restart Light for a DB2 member. Restart Light allows a DB2 data sharing member to restart with a minimal storage footprint, and then to terminate normally after DB2 frees retained locks. The reduced storage requirement can make a restart for recovery possible on a system that might not have enough resources to start and stop DB2 in normal mode. If you experience a system failure in a Parallel Sysplex, the automated restart in light mode removes retained locks with minimum disruption.

    Consider using DB2 Restart Light with restart automation software, such a Automatic Restart Manager.
  - *Persistent structure size changes*

    In earlier releases of DB2, any changes you make to structure sizes using the SETXCF START,ALTER command might be lost when you rebuild a structure and recycle DB2. Now you can allow changes in structure size to persist when you rebuild or reallocate a structure.
  - *Faster shutdown of DB2 data sharing members*

    You can more easily apply service or change system parameters. A new CASTOUT(NO) option on the -STO DB2 command enables a faster shutdown of DB2 data sharing members.
  - *New global scope for IFI calls and for options of several commands*

    Several recent enhancements to the Instrumentation Facility Interface (IFI) and some commands help you manage your data sharing environment more easily. Now, information from all the data sharing members can be available with a single call from any one of the members.

4.3.2 IMS DB Data Sharing

IMS DB has long supported two levels of *data sharing:*

- Database level data sharing
- Block level data sharing

For the purposes of this book, we shall only consider *block level data sharing,* because that is the level of sharing that exploits the Parallel Sysplex.

IRLM has always been used to provide data integrity, however, there were limitations. On any image IRLM could support multiple IMSs, but cross-image connectivity was limited to two IRLMs. This restriction was primarily due to the performance of the connectivity, which was using VTAM CTCs.

4.3.2.1 IMS DB Data Sharing in a Parallel Sysplex

When multiple IMS systems share data across *multiple* images and use a CF, it is known as *sysplex data sharing*.

Two elements necessary for data sharing in a Parallel Sysplex are:

- The ability to lock data for update across up to 32 images
- The notification of those sharing systems when data has been modified

Since the introduction of IMS/ESA V5 DB and IRLM V2.1, the number of IRLM connections has increased to 32, within a Parallel Sysplex by the use of a lock structure.

For information on the sizing of this structure, refer to 2.7, "IMS Structures" on page 98 in Volume 2: Cookbook, SG24-5638 and the CF Sizer tool, available at: `http://www.s390.ibm.com/cfsizer/`

The notification of change to data is an IMS DB function that was introduced in IMS/ESA V5. The implementation was through the use of two cache structures, one for OSAM and one for VSAM. These structures were "directory only" and therefore did not contain actual database data, only a list of the subsystems with an interest in the validity of that data.

In IMS/ESA V6, the OSAM structure was changed so that it now holds data as well as the directory; OSAM structures use the store-through cache algorithm. This structure may therefore be substantially larger than the corresponding IMS/ESA V5 structure.

For information on the sizing of these structures refer to 2.7, "IMS Structures" on page 98 in Volume 2: Cookbook, SG24-5638, and the CF Sizer tool, available at `http://www.s390.ibm.com/cfsizer/`

4.3.2.2 Fast Path Databases

IMS/ESA V6 introduced support for n-way data sharing of Data Entry Data Bases (DEDBs) with Virtual Storage Option (VSO) areas and Sequential Dependents (SDEPs).

The support for VSO is implemented by the use of one or two store-in cache structures for each area. The data is initially written to the cache and periodically cast out to DASD. This provides the equivalent function to IMS/ESA V5 where the data was held in a data space and cast out from there.

The support for Sequential Dependents (SDEPs) does not use CF structures; it is implemented through changes to the segments and CIs combined with a algorithm for selecting which CIs the SDEPs are stored in.

4.3.2.3 IMS Data Sharing Groups

As with DB2, there is the concept of a data sharing group. Figure 51 shows a sample IMS data sharing group, with the basic components required for n-way data sharing.

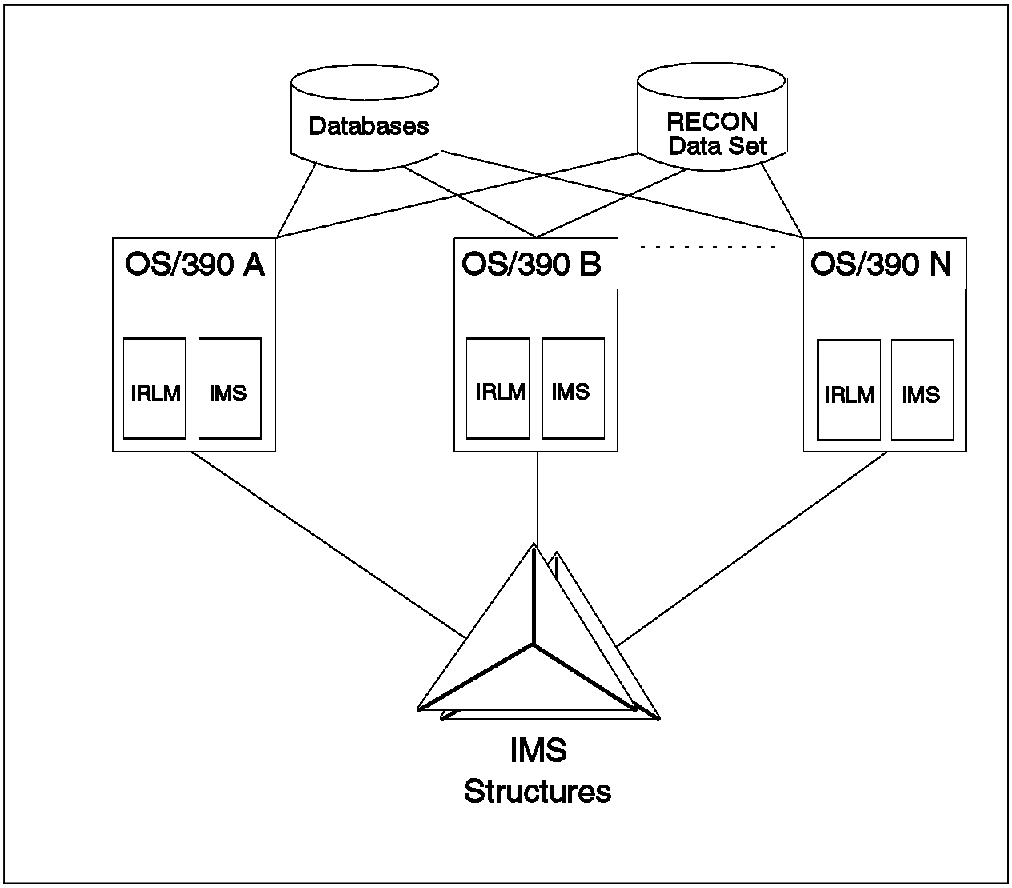


Figure 51. Sample Data Sharing Group in an IMS Environment

The members (IMS subsystems) of the data sharing group share:

- Databases
- A single set of RECON data sets
- One or more CFs
- A single IRLM lock structure in the CF (hereafter called an IRLM structure).
- OSAM, VSAM, and VSO DEDB cache structures in the CF (hereafter called OSAM, VSAM and VSO DEDB structures).

A data sharing group is connected to at least one CF. Without this, IRLM V2.1 is unable to allow inter-processor data sharing. If a CF is not available, IRLM V2.1 only allows intra-processor data sharing. A lock structure is required for sharing lock information between IRLMs.

For a discussion on IMS TM usage of structures, refer to 4.2.5.3, “IMS Shared Message Queues” on page 189.

4.3.2.4 Record Updating and Buffer Invalidation

When a block of data is read into a local buffer by any of the sharing IMSs, it is registered in an entry in the appropriate cache structure. Each entry consists of a field for the buffer ID (known to OS/390 as the resource name) and 255 slots. The slots are for IMS subsystems to register their interest in an entry's buffer. If it is VSAM data, there is no further action; if it is OSAM, then the data may be cached, depending on the definition of the IOBF subpool. There are three options:

- **No caching:** In which case processing is the same as for VSAM
- **Cache all data:** In which case the data is stored in the structure
- **Cache only changed data:** In which case the data is only stored in the cache after it has been changed and written back to DASD

When an IMS wants to modify a block, it must first request a lock from IRLM. The IRLM structure is checked to see if the block is already locked. If the block is not locked, the structure is updated and the lock is granted. When the block has been updated and written back to DASD, the appropriate cache structure is updated to mark the buffer invalid. This causes the CF to notify all IMSs that have registered an interest that their local copy of the block is invalid. Finally, the lock may be released.

Whenever IMS tries to access a buffer, it checks whether the buffer is still valid. If the buffer is invalid, IMS rereads the data from the structure (if it is there) or from DASD. With IMS/ESA V6, OSAM and VSO buffers may be refreshed from the CF rather than from DASD. Thus data integrity is maintained. Buffer invalidation works in all IMS sysplex database environments: DB/DC, DBCTL, and DB batch. In the sysplex environment, IMS supports buffer pools for VSAM, VSAM hiperspace, OSAM, OSAM sequential buffering and VSO DEDB.

4.3.2.5 Connectivity to Structures

For a data sharing group, the first IMS to connect to an IRLM determines the data sharing environment for any other IMS that later connects to the same IRLM. When identifying to IRLM, IMS passes the names of the CF structures specified on the CFNAMES control statement (in the IMS PROCLIB data set member DFSVSMxx), plus the DBRC RECON initialization timestamp (RIT) from the RECON header. The identify operation fails for any IMS not specifying the identical structure names and RIT as the first IMS. The term “connection” can refer to the following subsystems: either an IMS TM/DB subsystem, IMS batch, or a DBCTL subsystem for use by, for example, CICS.

Recommendation to Convert IMS Batch to BMP Batch

It is worthwhile to determine if any of the IMS batch jobs can be converted to batch message processing programs (BMPs). The primary reason for converting an IMS batch job to a BMP is to take advantage of the availability benefits of BMPs. When BMP batch jobs abend, IMS automatically backs them out and does not create any lock reject conditions. This is rarely true for batch jobs. Also, when an IRLM abends or the lock structure fails, batch jobs abend. BMPs do not abend in this situation.

Secondly, BMPs provide a better granularity of locking than IMS batch jobs.

Of relatively minor importance (since the number of connections to a cache structure was raised from 32 to 255) is the fact that a single IMS batch job constitutes one connection to the cache structure. Up to 255 connections are allowed to a cache structure. The BMPs will run using the single connection of the IMS control region.

There may be some additional overhead associated in converting to BMPs, caused by the sharing of the database buffer pools with other dependent regions.

Recommendation to Release Locks Frequently

Ensure that applications issue CHKP calls periodically to release locks held when executed as a BMP.

The same consideration applies to batch.

IMS APARs PQ26416 and PQ26491, and OS/390 APAR OW38840 deliver new function whereby the overhead of connecting and disconnecting to the CF cache structures for batch DL/I jobs is significantly reduced. This is especially beneficial for jobs with very short elapsed times, where the communication with XES can make up a considerable portion of the total elapsed time of the job.

Another enhancement of interest is Long Lock Detection Reporting. This was introduced by IMS APARs PN84685 (IMS V5), PQ07229 (IMS V6), PN79682 (IRLM 2.1), OW20579 (RMF). These APARs provide support for a new RMF report that shows which tasks have been holding a lock for a long time, and which tasks are impacted as a result.

4.3.2.6 IMS Database Types Eligible for Data Sharing

The following database organizations are supported by IMS/ESA V6.1 for data sharing:

- HIDAM
- HDAM
- HISAM
- SHISAM
- Secondary indexes
- DEDBs

Programs using these types of databases usually do not require any changes to function in a data sharing environment.

4.3.2.7 IMS DB V5 Database Types Not Eligible for Data Sharing

The following database organizations are *not* supported by IMS/ESA V5.1 for data sharing:

- DEDBs which use either of the following options:
  - Sequential dependent segments (SDEPs)
  - Virtual storage option (VSO)
    - MSDBs

4.3.2.8 IMS/ESA V5 and V6 Coexistence for OSAM Data Sharing

Data sharing is valid with a mixture of IMS/ESA V5 and IMS/ESA V6 systems, as long as you do not use OSAM data caching.

Data sharing of OSAM data is allowed as long as IMS/ESA V6 does not put data in the structure. At the time of writing, IMS/ESA V6 code is being added to allow IMS/ESA V6.1 to allow the user to specify the ratio of directory entries to data elements. The CFOSAM parameter is used to specify the directory-to-element ratio with few data elements. The ratio may be as high as 999:1. The IOBF statement in the DFSVSMxx member can be used to control the amount of data caching, and should be set to specify *no* data sharing.

For a discussion on directory-to-element ratio, refer to 2.7.2.2, "IMS OSAM and VSAM Structure Sizing" on page 100 in Volume 2: Cookbook, SG24-5638.

4.3.2.9 Fast Database Recovery

IMS/ESA V6 introduced the Fast Database Recovery (FDBR) region which, using XCF monitoring, can track an active IMS while either being on the same or another image.

In the event of the active system failing, the FDBR code will dynamically back out in-flight full function database updates, invoke DEDB redo processing, and purge retained locks from IRLM.

4.3.2.10 XRF and Sysplex Data Sharing

XRF can be used in a sysplex data sharing environment if the CF is available and connected when the alternate system is brought up. For more information about IMS disaster recovery in Parallel Sysplex, refer to 3.7.5, "IMS Disaster Recovery Considerations" on page 134.

4.3.3 CICS/VSAM Record Level Sharing Considerations

CICS/VSAM record level sharing (RLS) is a data set access mode that allows multiple address spaces, CICS application owning regions (AORs) on multiple systems, and batch jobs to access VSAM data at the same time. With CICS/VSAM RLS, multiple CICS systems can directly access a shared VSAM data set, eliminating the need for function shipping between AORs and file owning regions (FORs). CICS provides the logging commit and rollback functions for VSAM recoverable files. VSAM provides record-level serialization and cross-system caching. CICS, not VSAM, provides the recoverable files function. Figure 52 is an illustration of CICS/VSAM RLS implementation in Parallel Sysplex.

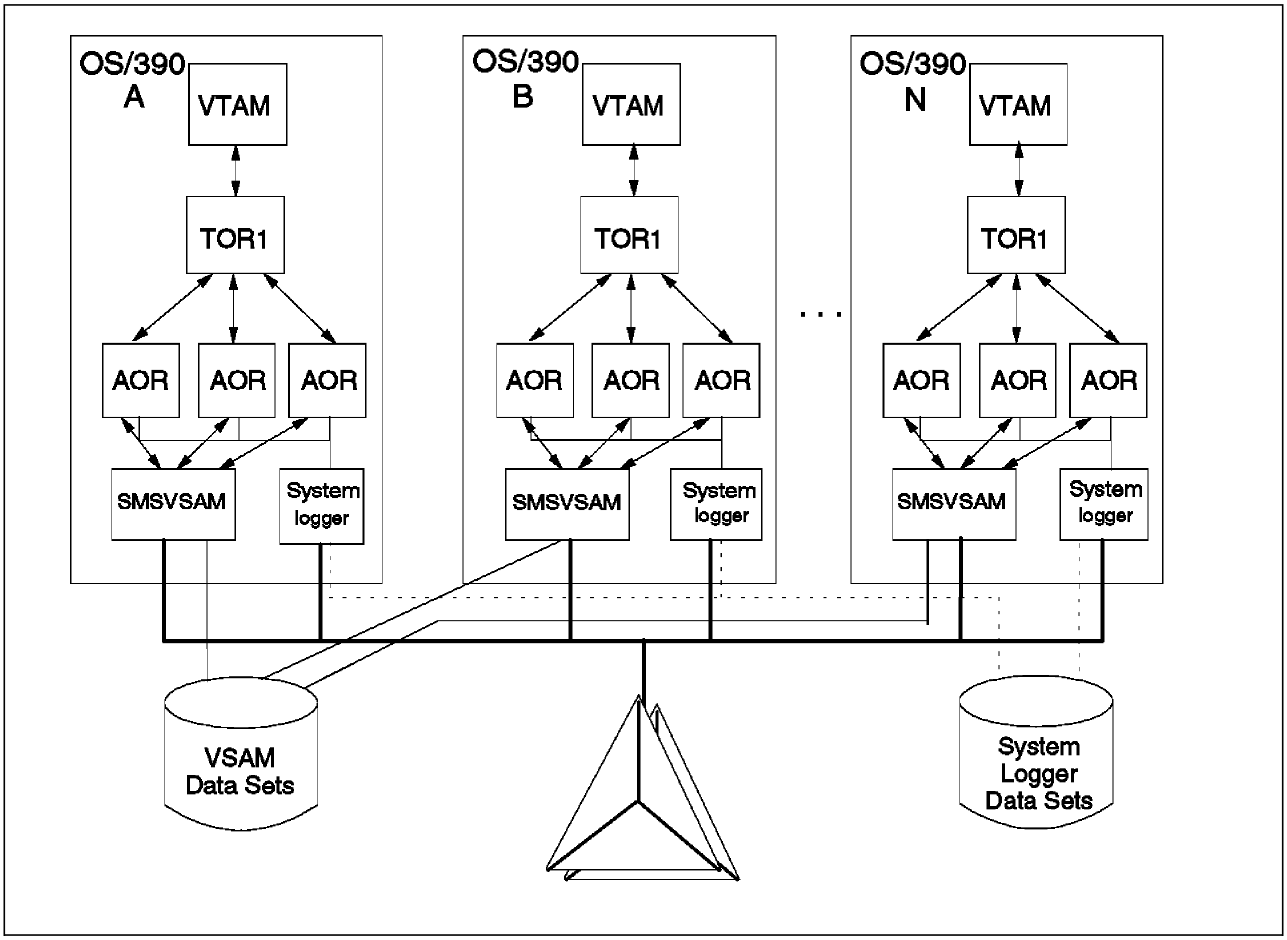


Figure 52. VSAM Record Level Sharing in a Parallel Sysplex

The CICS/VSAM Record Level Sharing (RLS) data access mode allows multisystem access to a VSAM data set while ensuring cross-system locking and buffer invalidation. CICS/VSAM RLS uses XES services to perform data set level locking, record locking, and data caching. CICS/VSAM RLS maintains data coherency at the control interval level. It uses CF caches as store-through caches. When a control interval of data is written, it is written to both the CF cache and to DASD. This ensures that a failure in the CF cache does not result in the loss of VSAM data.

VSAM RLS improves the availability of VSAM data sets during both planned and unplanned outages, because if one of the OS/390 systems or CICS regions is not available, applications can access their data from another system or CICS region. It also provides data integrity and availability through the use of common locking and buffering.

VSAM RLS Improves Availability

VSAM RLS provides improvements for data integrity and availability.

The SMSVSAM server is the system address space used for CICS/VSAM RLS. The data space associated with the server contains most of the VSAM control blocks and the system-wide buffer pool used for data sets opened for record-level sharing. SMSVSAM assumes responsibility for synchronizing this control block structure across the sysplex.

The LOG attribute on the VSAM cluster (contained in the ICF catalog entry) defines a data set as recoverable or non-recoverable. Because CICS maintains the log of changed records for a data set (thus allowing transactional changes to be undone), only VSAM data sets under CICS are recoverable. Whether a data set is recoverable or not determines the level of sharing allowed between applications:

- Both CICS and non-CICS jobs can have concurrent read/write access to non-recoverable data sets.
- Non-CICS jobs can have read-only access to recoverable data sets, concurrent with read/write access by CICS transactions. Full read integrity is ensured.

More information about how VSAM RLS actually works, and comparisons between VSAM RLS processing and traditional VSAM processing, see the IBM Redbooks *CICS and VSAM Record Level Sharing: Planning Guide*, SG24-4765, and *Batch Processing in a Parallel Sysplex*, SG24-5329.

4.3.3.1 Implementing CICS/VSAM RLS

To enable CICS/VSAM RLS, you must define one or more CF cache structures and add these to your SMS base configuration. Cache set names are used to group CF cache structures in the SMS base configuration. In setting up for CICS/VSAM RLS processing, you also need to define the CF lock structure.

A data set is assigned to a CF cache structure based on your SMS policies. If the storage class for a VSAM data set contains a non-blank cache set name, the data set is eligible for record-level sharing. When the data set is opened for RLS-processing, the cache set name is used to derive an eligible CF cache structure to use for data set access.

Restrictions are discussed in the CICS TS, OS/390, DFSMS and sysplex reference libraries. Refer to the CICS TS Program Directory for specific product requirements. For planning and implementation information relation to VSAM RLS, Refer to the IBM Redbook *CICS and VSAM Record Level Sharing: Planning Guide,* SG24-4765.

Refer to CICS APAR II09698, *Installation Planning Information*, for the latest installation recommendations.

For more information on VSAM RLS, refer to *DFSMS/MVS V1.5 Planning for Installation*, SC26-4919.

For more information on the tool *CICS/VSAM RLS Capacity Planning Tool* (CVRCAP), refer to Appendix A, “Tools and Services Catalog” on page 161 in Volume 2: Cookbook, SG24-5638.

4.3.3.2 Software Dependencies

CICS/VSAM RLS support has the following prerequisites:

- MVS/ESA V5.2 or higher.
- DFSMS 1.3 or higher.
- CICS TS 1.1 or higher.

In addition, some mechanism (GRS or an equivalent function) is required for multisystem serialization.

There are additional dependencies on the following program products, or equivalent products, for *full-function* support:

- CICSVR V2.3 or higher
- RACF V2.1 or higher
- EPDM V1.1.1 or higher
- Appropriate levels of COBOL, PL/I, FORTRAN, and Language Environment runtime libraries, for batch applications that will use VSAM RLS data access

All products listed in this section need to be at a specific service level or have function-enabling PTFs applied.

4.3.3.3 Hardware Dependencies

To exploit this support, you must have at least one CF connected to all systems that will be taking part in VSAM record level sharing. If you have multiple CFs, those CFs must be accessible from any system that may be using VSAM RLS.

Maximum Availability Recommendation

For maximum availability, it is recommended that you set up at least two CFs with connectivity to all CPCs in the Parallel Sysplex. If a CF is not operational, this allows the storage management locking services to repopulate its in-storage copy of the locks to the secondary CF. A CF must be large enough to contain either a lock structure or a cache structure (or both), and have enough "white space" to allow the structures to be increased in size.

You define a single CF lock structure, IGWLOCK00, to be used for cross-system record-level locking. The CF lock structure *must* have connectivity to all systems that will be taking part in VSAM RLS. A nonvolatile CF for the lock structure is not required, but it is highly recommended for high availability environments. If you have a volatile CF lock structure, and a power outage causes a Parallel Sysplex failure resulting in loss of information in the CF lock structure, all outstanding recovery (CICS restart and backout) must be completed before new sharing work is allowed.

In order for VSAM RLS to use the structures you have defined in the CFRM policy, you must also define them to SMS. CF cache structures associated with a given storage class must have, at a minimum, connectivity to the same systems as the storage groups mapped to that storage class.

In summary, we recommend you set up multiple CFs for maximum availability and workload balancing, and ensure that these CFs have connectivity to all systems that will use CICS/VSAM RLS.

4.3.3.4 Coexistence Issues

Within a Parallel Sysplex, an SMS configuration can be shared between systems running DFSMS/MVS V1.3 and other DFSMS/MVS systems; however, toleration PTFs must be applied to all pre-DFSMS/MVS V1.3 systems so that they do not conflict with RLS access. An SMS configuration containing VSAM RLS information must be activated by a system that is at least at the DFSMS 1.3 level.

For fallback from DFSMS/MVS V1.3, a DFSMS/MVS V1.3-defined SMS control data set is compatible with down-level DFSMS/MVS systems, if all toleration PTFs are applied to those systems.

GRS is required to ensure cross-system serialization of VSAM resources and other DFSMS/MVS control structures altered by CICS/VSAM RLS. Open requests from a system that is not at the DFSMS 1.3 (or higher) level are not allowed when RLS access to a data set is active, or if RLS transaction recovery for the data set is pending.

In a JES3 environment, you must be careful to define cache set names only in the SMS storage classes that are used by data sets opened for RLS processing. A non-blank cache set name causes a job to be scheduled on an RLS-capable system. If all storage classes have non-blank cache set names, then all jobs accessing SMS-managed data sets are scheduled to DFSMS/MVS V1.3 systems, causing an imbalance in workload between the DFSMS/MVS V1.3 systems and down-level systems.

4.3.3.5 SMS Changes

To implement CICS/VSAM RLS, you need to modify your SMS configuration in the following ways:

- Add the CF cache structures to your SMS base configuration.

  You can have up to 255 CF cache set definitions in your base configuration. Each cache set can have one to eight cache structures (or buffer pools) defined to it, allowing data sets to be assigned to different cache structures in an effort to balance the workload.

  CF cache structures associated with a given storage class must have, at a minimum, the same connectivity as the storage groups mapped to that storage class.

  Having multiple CF cache structures defined in a cache set could provide improved availability. If a CF cache structure fails and a rebuild of the structure is not successful, SMS dynamically switches all data sets using the failed CF structure to other CF structures within the same cache set. For more information, refer to 2.12.4, "CICS/VSAM RLS Structure Rebuild Considerations" on page 142 in Volume 2: Cookbook, SG24-5638.

- Update storage class definitions to associate storage classes with CF cache set names. You also need to define the direct and sequential CF weight values in the storage classes.

- Change the ACS routines to recognize data sets that are eligible for VSAM RLS so that they can be assigned to storage classes that map to CF cache structures.

Storage requirements for SMS control data sets have also changed to accommodate CICS/VSAM RLS information.

4.3.3.6 Batch Considerations

Stop Press

Just as this document was going to the publishers, OS/390 V2R10 was announced. Part of that announcement was a Statement Direction for a function know as Transactional VSAM Services (TVS). TVS adds the ability for batch jobs to update recoverable VSAM files concurrent with updates from CICS regions.

Whereas CICS provides its own logging capability, TVS will provide this capability for batch jobs. VSAM RLS will be used for locking and buffer coherency. It is envisioned that TVS will be used for files that require near-24x7 availability, but also have to be updated from batch.

More information will be available about TVS later in 2000.

There are a number of restrictions relating to the use of VSAM files by batch jobs, where those files are also accessed in RLS mode by CICS. For a full list of the considerations for batch, refer to the chapter entitled *Using VSAM Record-Level Sharing* in *DFSMS/MVS Version 1 Release 4 Using Data Sets*, SC26-4922, and in the IBM Redbook *CICS and VSAM Record Level Sharing: Implementation Guide*, SG24-4766.

4.3.3.7 CICS Considerations

There are also considerations for CICS, in relation to its use of VSAM data sets in RLS mode:

- **Changes in file sharing between CICS regions:** Prior to VSAM RLS, shared data set access across CICS Application-Owning Regions (AORs) was provided by CICS function-shipping file access requests to a CICS File-Owning Region (FOR). CICS/VSAM RLS can now provide direct shared access to a VSAM data set from multiple CICS regions. The highest performance is achieved by a configuration where the CICS AORs access CICS/VSAM RLS directly.

  However, a configuration that places a CICS FOR between CICS AORs and CICS/VSAM RLS is supported as well. FORs can continue to exist to provide distributed access to the files from outside the sysplex.

- **LOG and LOGSTREAMID parameters for DEFINE CLUSTER:** These parameters replace the corresponding definitions in the CICS File Control Table (FCT). LOG specifies whether the VSAM sphere is recoverable (where CICS ensures backout) or non-recoverable. LOGSTREAMID specifies the name of a forward recovery log stream to use for the VSAM sphere.
- **BWO parameter for DEFINE CLUSTER:** If CICS/VSAM RLS is used, this is the mechanism for specifying backup-while-open in a CICS environment.
- **Sharing control:** Sharing control is a key element of this support. Careful consideration must be given to managing the sharing control data sets that contain information related to transaction recovery.
- **Recovery:** New error conditions affect subsystem recovery. These include loss of the lock structures, loss of a CF cache structure, SMSVSAM server failure, or errors during backout processing.
- **Batch job updates of a recoverable VSAM sphere:** CICS/VSAM RLS does not permit a batch job to update a recoverable sphere in RLS access mode

(however, refer to the “Stop Press” note earlier in this section referring to Transactional VSAM Services). The sphere must be RLS-quiesced first, using a CICS command, and then the batch job can open the VSAM sphere for output using non-RLS protocols. If it is necessary to run critical batch window work while transaction recovery is outstanding, there are protocols that allow non-RLS update access to the VSAM sphere. Backouts done later must be given special handling. If you intend to use this capability, you need to plan for the use of IDCAMS SHCDS PERMITNONRLSUPDATE and DENYNONRLSUPDATE commands and for the special handling of these transaction backouts.

Refer to *CICS TS for OS/390 Release Guide*, GC33-1570, and *CICS TS for OS/390 Migration Guide*, GC34-5353, for more information.

4.3.4 S/390 Partners in Development Databases

Important Notice about S/390 Partners in Development

For the growing list of S/390 partners in development-provided applications available for S/390 Parallel Sysplex, see the Web site:
`http://www.s390.ibm.com/products/s390da/techno/sysplex.html`

There is also information in 2.2.1, “S/390 Partners in Development Software” on page 23 about a number of non-IBM database products that support Parallel Sysplex. Refer to that section for further information.

4.4 Batch Workload Considerations

In this section, we review how the job entry subsystem exploits the Parallel Sysplex environment. Further, we look at techniques available for workload balancing in a batch environment today.

Since MVS/ESA SP V5.1, there has been a component called JESXCF. This component provides the cross-systems communication vehicle (through XCF services) for the use of the job entry subsystem.

All levels of JES2 and JES3 since MVS/ESA SP V5.1 exploit this component. At subsystem initialization time, each JES member automatically joins a JES XCF group. All the members attached to the group then use XCF services to communicate with each other.

4.4.1 JES2 Considerations in Parallel Sysplex

It is recommended that all JES2 members in the sysplex be part of the same JES2 MAS. For discussions and illustrations about how JESplexes map to Parallel Sysplexes, refer to 2.3.3, “What Different 'Plexes Are There?” on page 39.

Note: Be aware that for systems operating at JES2 V5.1 and above in a MAS configuration, a sysplex (*not* necessarily Parallel Sysplex) is mandatory. This is to allow the JESXCF component to provide communication to all members in the MAS through XCF services. However, a Parallel Sysplex does not require JES2 V5.

JES2 can exploit the CF. In addition to its use of XCF, which may be using the CF for communication (see 2.4.1.2, “XCF Signalling Path Considerations” on page 74 for more details), JES2 can also place its checkpoint information into a structure in the CF. For further detail on structures, refer to 2.6, “JES2 Structures” on page 95 in Volume 2: Cookbook, SG24-5638.

N and N+x Considerations for JES2

There are slightly different considerations for coexistence for JES releases in a Parallel Sysplex.

Although IBM generally recommends using the level of JES that is shipped with an OS/390 release, some customers upgrade their JES at a different time to OS/390. So, one thing you must consider is which releases of JES (either JES2 or JES3) are supported with a given level of OS/390.

The other thing to be considered, is which releases of JES can coexist within a JES2 MAS or a JES3 complex.

Rather than describing all the possible combinations here, there is a WSC FLASH (10034) that provides all this information, both for JES2 and JES3. That FLASH also clarifies the coexistence policy for JES for OS/390 V2R10 and later. The FLASH is available on the Internet at: `http://www.ibm.com/support/techdocs/atsmastr.nsf/PubAllNum/Flash10034`

4.4.2 JES3 Considerations in Parallel Sysplex

With JES3 V5.1 and DFSMS/MVS V1.2, it was only possible to have eight members in a JES3 complex. This limitation has been removed with JES3 V5.2 and DFSMS/MVS V1.3 (a corequisite PTF is available for JES3 V5.2.1 to enable more than eight systems).

Starting with OS/390 V2R8 JES3 provides support for WLM-managed batch initiators. With this enhancement the system programmer can specify, by job class group, whether WLM or JES3 will manage batch initiators. When the initiators are managed by WLM, they can be dynamically stopped and started as needed, and placed on systems with sufficient capacity to process the related batch jobs.

There is additional information on JES3 considerations, with emphasis on availability in 3.2.3, “JES3 and Continuous Availability” on page 98.

4.4.2.1 JES3 and Consoles

Be aware that with the introduction of JES3 V5.2, JES3 no longer manages its own consoles, but exploits MCS console support. This may have an impact in your operational procedures.

Read the IBM Redbook *MVS/ESA SP JES3 V5 Implementation Guide*, SG24-4582, for a full discussion on the impact of JES3 and the Parallel Sysplex, and read the IBM Redbook *OS/390 MVS Multisystem Consoles Implementing MVS Sysplex Operations*, SG24-4626 to gain an understanding of console handling in a Parallel Sysplex. See also 4.0, “Consoles and Parallel Sysplex” on page 103 in Volume 3: Connectivity, SG24-5639.

4.4.3 Can I Have JES2 and JES3 in the Same Sysplex

The simple answer to this is yes. You can even have JES3 and JES2 on the same image if you want. With the introduction of JES3 V5.2.1, console management is now provided by MCS. Because MCS is now providing the console support, previous console limitations for mixed JES2 and JES3 sysplexes have been removed.

Therefore if you consider a mixed JES2 and JES3 sysplex, it is recommended that you have both JES3 and JES2 at least at the V5.2.1 level.

Be aware, however, that the JES service organization does not consider this a supported configuration and will not accept APARs caused by interactions between the JES components.

For discussions and illustrations about how JESplexes map to Parallel Sysplexes refer to 2.3.3, "What Different 'Plexes Are There?" on page 39.

4.4.4 Batch Workload Balancing and Parallel Sysplex

Starting with OS/390 V2R4, the workload manager incorporates the capability of managing batch workload across a sysplex. This capability includes multisystem batch workload balancing, dynamic goal-oriented management of the time jobs spend waiting for the initiator, improved reporting of job response time and pre-execution job delays, and reduced operational demands to schedule batch jobs. This support is available starting with OS/390 JES2 V2R4, and OS/390 JES3 V2R8.

There are various techniques available in JES2 to control where a job will run within a JES2 MAS environment. These can provide a form of balancing based on JOBCLASS, SYSAFF, initiator setup, and the use of WLM Scheduling Environments. This can be achieved by operator command, initialization options, JCL/JECL[3] changes, or by coding JES2 exits. See *MVS/ESA SP JES2 V5 Implementation Guide*, SG24-4583, for additional information.

Note: Be aware that even if system affinity is used, JCL conversion can take place on any member of the MAS, so procedure libraries *must* be shared to prevent JCL errors during the conversion process.

JES3 balances the workload among CPCs by considering the resource requirements of the workload. JES3 can route work to specific systems by enabling initiator groups/classes and JECL (*MAIN statement). JES3 is aware of tape and DASD connectivity and thus will schedule work to the CPC that has the correct devices attached. Starting with OS/390 V2R8, JES3 supports WLM-managed initiators.

4.4.5 IBM BatchPipes for OS/390

BatchPipes for OS/390 (5655-D45) is a major enhancement and replacement product for BatchPipes/MVS. It is compatible with the BatchPipes component of SmartBatch for OS/390 1.2.

BatchPipes uses *parallelism* and *I/O optimization* to reduce the batch processing time and balance the workload across the system or Parallel Sysplex. By giving

[3] JECL - JES2 control language. See *OS/390 MVS JCL Reference*, GC28-1757 for more information.

you the ability to run the jobs and job steps in parallel, BatchPipes can reduce the elapsed time of batch jobs and job streams. In addition, elapsed time can be further reduced when BatchPipes is used to minimize I/O resource usage.

BatchPipes provides the following capabilities:

- It allows two or more jobs that formerly ran serially to run in parallel.
- It allows individual job steps in a multi-step job to run in parallel.
- It reduces the number of physical I/O operations where possible by transferring data through CPC storage rather than DASD or tape.

4.4.5.1 Data Piping between Jobs

The traditional batch job stream uses an intermediate data set as a vehicle to pass data from one process to the next process. Job1 writes data to a data set on either tape or DASD. When it completes, the data set is closed and Job2 can start. In essence, Job1 is a *writer* to the data set while Job2 is a *reader* of the data set.

In a BatchPipes environment, the data set is replaced by a *storage buffer*. After Job1 writes the first block of data to the buffer, Job2 can read this block from the buffer. The storage medium that BatchPipes uses to transfer data from writer to reader is known as a *pipe*.

Data in a pipe always flows in one direction, from writer to reader. A writer_pipe_reader set is known as a *pipeline*. The pipe can exist in either a data space or in a CF structure.

4.4.5.2 Data Piping between Systems

The processing of related jobs across systems is achieved by using a common CF structure known as the *pipe*. (See section 2.19, “BatchPipes Structures” on page 158 in Volume 2: Cookbook, SG24-5638 for a discussion of BatchPipes structures.) This set of BatchPipes subsystems is known as a *Pipeplex*.

All the subsystems in the Pipeplex must have the same name and reside on separate systems in the Parallel Sysplex. Cross-system piping can only occur between systems within the same Pipeplex. The Pipeplex can include every system in the Parallel Sysplex, or a subset of the systems. See Figure 9 on page 47 for the recommended Pipeplex configuration.

It is also possible to run multiple Pipeplexes within a system. In Figure 53 on page 217, the Pipeplex BP01 includes all systems in the sysplex. However, Pipeplex BP02 includes only two of the four systems.

Note: BP01 and BP02 cannot be in the same Pipeplex.

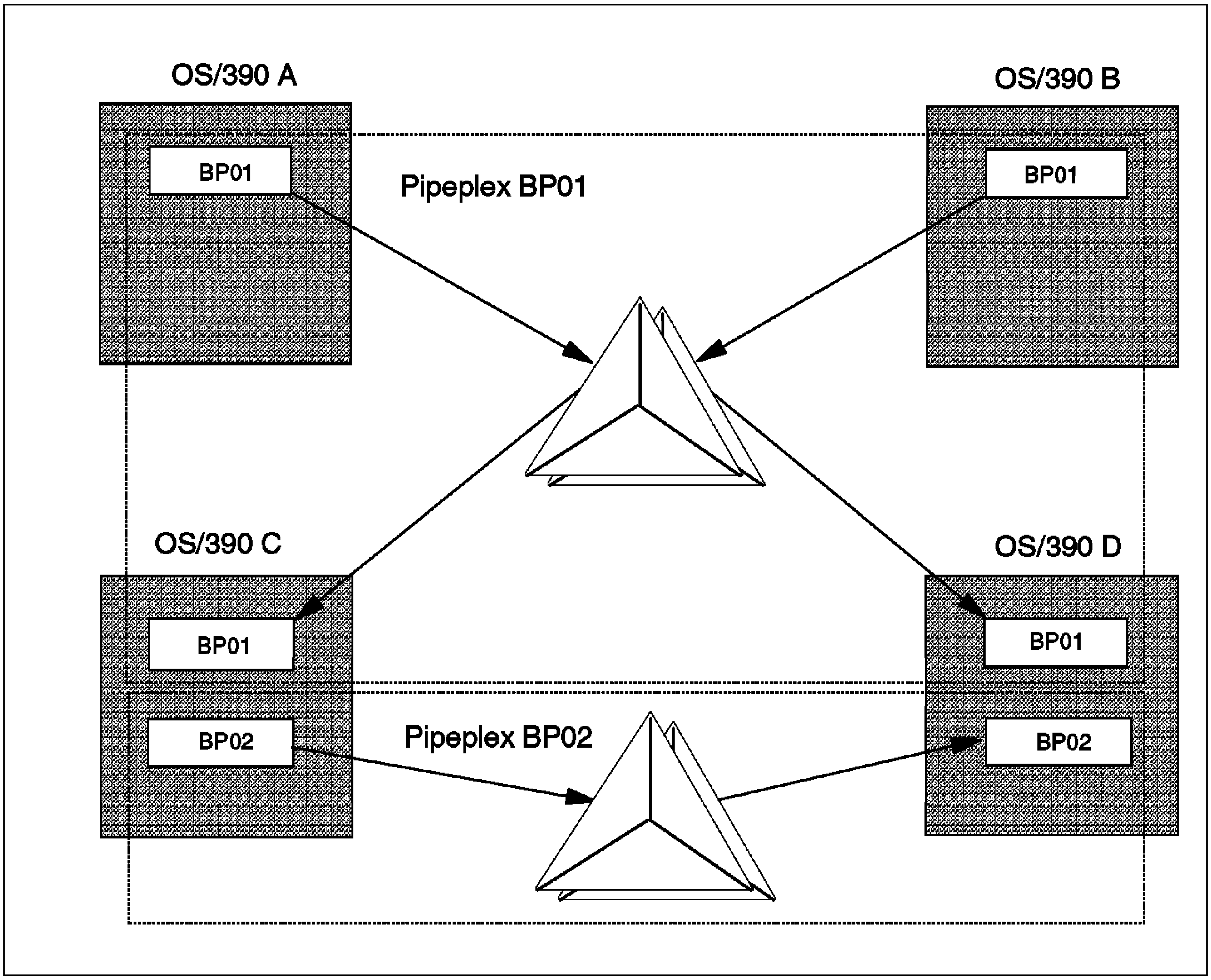


Figure 53. Pipeplex in a Parallel Sysplex

4.4.5.3 Considerations for a Parallel Sysplex

With BatchPipes, jobs that were previously run serially can now run concurrently. Therefore, all resources for those work units must be available concurrently:

- **Initiators:** JES initiators must be available for all jobs in the pipeline. Unless *all* the jobs that will be connected to a pipe have initiators, the jobs will stop when the pipe with no corresponding job fills up. It is recommended that all jobs in a pipeline be assigned one job class and a set of initiators be dedicated to that job class.
- **Tape and DASD:** There should still be sufficient devices to simultaneously hold all job and step data that is outside the control of BatchPipes.
- **Virtual Storage:** Each job has its own need for virtual storage. When multiple jobs run concurrently, the combined storage must be available. Additional storage may be required for a CPC storage pipe.
- **CF storage:** A Pipeplex requires enough CF storage to hold the maximum number of cross-system pipes. See 2.19, “BatchPipes Structures” on page 158 in Volume 2: Cookbook, SG24-5638 for a discussion on the BatchPipes structure.

- **CP:** Each job and job step has its own need for CPs. Since multiple jobs and job steps run concurrently, CPU utilization may increase while the jobs are running.

4.4.5.4 BatchPipes Affect on Charge-Back

BatchPipes may have a major effect on variables that are used in charge-back formulas. For example, TCB time is likely to increase.

Without BatchPipes (that is, when the I/O subsystem moved the data), a significant portion of the system services (that is, TCB time) was not charged to the address space that invoked them.

With BatchPipes, the system resources that the job uses are included in the TCB time. For more details, see *IBM BatchPipes for OS/390 V2R1 User's Guide and Reference*, SA22-7458. Table 14 contains information about the impact that BatchPipes may have on your chargeback mechanisms.

Table 14. Effects of BatchPipes on Charge-Back Variables

Variable	**How BatchPipes Affects the Variable**
TCB Time	TCB time increases. Without BatchPipes (that is, when the I/O subsystem moved the data), significant portion of the system services (that is, TCB time) were not charged to the address space that invoked them. With BatchPipes, the system resources that the job uses are included in the TCB time.
EXCP	EXCP counts decrease. BatchPipes eliminates EXCPs to the pipe data sets.
Tape Mounts	Tape mounts are reduced. Tape and tape mounts for the intermediate data sets that are piped are eliminated, along with tape mount processing and tape storage in a library.
DASD Storage	DASD storage is reduced. DASD storage for the intermediate data sets that are piped is eliminated.
SRB Time	SRB time is reduced. With BatchPipes, fewer SRBs are required to perform data movement.
Service Units (SUs)	• The IOC field increases, where IOC refers to the I/O service of an address space. • The CPU field increases, where CPU refers to the accumulated CPU (that is, TCB) SUs. • The MSO field increases, where MSO refers to accumulated storage SUs.
Uncaptured CPU Time	Uncaptured CPU time is reduced. Compared with accountability of non-BatchPipes jobs, accounting of CPU time is more accurate for BatchPipes jobs.
Working Set Size	BatchPipes does not affect the working set size (WSS). Since processing of the job occurs at a faster rate, the system is likely to keep more pages of the job in the CPC. For example, with fewer physical I/O operations, pages are not paged out while the job waits for those physical I/O operations to occur.
CPU Charges	Because jobs are split and may run on different CPCs, charges for jobs may vary based on the charge rates for the CPC.

4.4.5.5 BatchPipes Services Offering

IBM offers an Integrated Services Offering (ISO) to help you get started using BatchPipes. For a description of the services offered, see Appendix A, "Tools and Services Catalog" on page 161 in Volume 2: Cookbook, SG24-5638.

4.5 Network Workload Balancing Capabilities

There are various ways to connect the network of end users to applications in the sysplex. The two main application types considered in this book are *SNA applications* and *TCP applications*. VTAM provides network access to the sysplex for SNA applications and TCP/IP provides network access to the sysplex for TCP applications. Both VTAM and TCP/IP provide many specific functions to exploit the Parallel Sysplex. Only VTAM, however, currently exploits the Coupling Facility, by its use for Multi Node Persistent Sessions and Generic Resources support.

In this section, we will describe the use of VTAM GR in some detail. Information about how you would connect an SNA network to a Parallel Sysplex is contained in 2.1, "Network Access to Parallel Sysplex Using VTAM" on page 23 in Volume 3: Connectivity, SG24-5639.

For TCP/IP, the role of workload balancing is more closely connected to how your network is configured and connected than is the case for SNA. Therefore, we will only briefly list the workload balancing features of TCP/IP, and leave the detail to be presented in 2.3.2, "Workload Balancing Enhancements" on page 48 in Volume 3: Connectivity, SG24-5639.

4.5.1 VTAM Generic Resources Function

To allow an end user to easily connect to an application that may be available on multiple systems, VTAM provides a function known as *generic resources.* The use of generic resources provides the end user with a single system image of the application no matter where it runs in the Parallel Sysplex.

4.5.1.1 Using Generic Resources

The generic resources function is an extension to the existing USERVAR support. It was introduced in VTAM V4.2 to support applications in a Parallel Sysplex.

The end user accesses the desired application by using the *generic resource name* of the application. VTAM determines the *actual application instance*[4] for this session based on workload and other performance criteria. The generic resources function allows you to add CPCs and applications, and to move applications to different images on the same or different CPCs without impacting the end user.

APPN Requirement: The generic resources function requires the use of Advanced Peer-to-Peer Networking (APPN) in the Parallel Sysplex. This means that the VTAM systems which are running generic resources must be configured either as APPN end nodes or as APPN network nodes. The VTAM end nodes must have a direct connection to at least one VTAM network node inside the Parallel Sysplex. A VTAM end node will select one of the VTAM network nodes to be its network node server, and a direct connection is required for this. For more information on APPN and APPN node types, refer to 2.1.1, "APPN Overview" on page 24 in Volume 3: Connectivity, SG24-5639.

[4] Meaning which application on which image.

Sessions with generic resource applications can be established from either subarea nodes or APPN nodes. The generic resource function can support sessions from all LU types, including LU0, LU1, LU2, LU3, LU6, LU6.1 and LU6.2.

Do You Have to Migrate the Network to APPN?

Generic resources require the use of APPN protocols inside the Parallel Sysplex. This does not mean that the entire network must be running APPN to access these applications.

If the network contains subarea nodes which require access to generic resource applications, one (or more) of the VTAM network nodes in the Parallel Sysplex will provide a boundary between the subarea network and APPN. A VTAM image which provides this function is called an *Interchange Node (IN)* and is discussed in more detail in 2.1.1, "APPN Overview" on page 24 in Volume 3: Connectivity, SG24-5639.

Generic Resource Definition: There are no new VTAM definitions to define generic resources. Subsystems which want to use the VTAM generic resources function will inform VTAM using the SETLOGON macro. The SETLOGON macro is issued after an application opens its ACB with VTAM. The SETLOGON ADD option (with the generic resource name specified in the NIB of the GNAME operand) will indicate to VTAM that this application is to be added into a generic resource. VTAM takes this information, and then updates the CF structure with the new application instance. See "VTAM Use of the CF" on page 221 for more information.

A subsystem can remove itself from a generic resource at any time, without having to close its ACB. It does this with the SETLOGON macro and the DELETE option. If this happens, and the ACB is still open, end users can still logon using the application name, but not using the generic name.

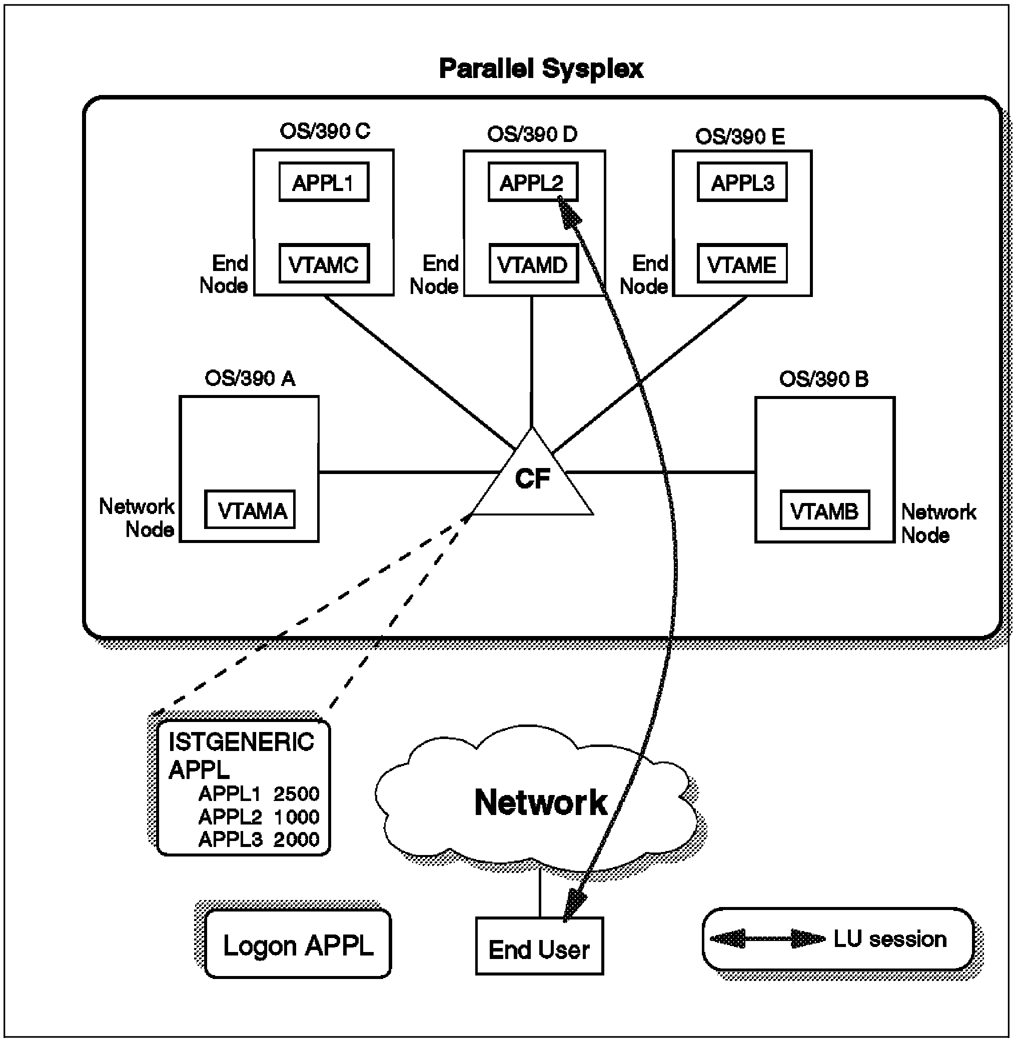


Figure 54. Using VTAM Generic Resources in a Parallel Sysplex

Figure 54 shows a Parallel Sysplex where an end user has logged onto a generic resource name called APPL. There are a number of actual application instances of this generic resource, called APPL1, APPL2, and APPL3. These actual application instances reside on systems VTAMC, VTAMD, and VTAME. One of the network nodes, VTAMA or VTAMB, is responsible for doing the workload balancing, which results in the actual application instance APPL2 being selected for this session request.

4.5.1.2 VTAM Workload Balancing for Generic Resources

VTAM does workload balancing for generic resources using information stored in the CF. For VTAM to use the CF, there must be an active CFRM policy defined for the Parallel Sysplex. All the VTAMs in the Parallel Sysplex that are part of the same generic resource configuration must be connected to the same CF.

VTAM Use of the CF: When VTAM on an OS/390 in a Parallel Sysplex is started, it automatically connects to the CF structure, after first checking that the CFRM policy is active. The first VTAM to become active in the Parallel Sysplex will allocate the storage for the CF structure.

VTAM uses a list structure in the CF to keep the information about the generic resources in the Parallel Sysplex. The default name for this list structure is

ISTGENERIC. Starting with VTAM V4.3, you can override the default name by specifying a different structure name on the VTAM STRGR start option. This allows you to have multiple CF structures for generic resources within your network, which might be useful if you wish to separate test and production structures. For more information on CF usage, sizing, placement, and rebuild configurations, refer to 2.4, "VTAM Structures" on page 85 in Volume 2: Cookbook, SG24-5638 and the CF Sizer tool, available at:
`http://www.s390.ibm.com/cfsizer/`

Inside the VTAM structure, the CF keeps information on the session counts for each application instance in a generic resource. These session counts are updated automatically each time a session is established or terminated.

When a session setup request for a generic resource name is sent into the network from an end user, the request will eventually reach one of the VTAM network nodes in the Parallel Sysplex. The VTAM network node that receives the request will then query the CF structure. The CF will return to the VTAM network node a list of actual application instances, and a session count for each one. The VTAM network node will then choose the appropriate actual application instance. There are a number of ways this is done:

- **VTAM session balancing:** VTAM will use the session counts provided by the CF to balance the number of active sessions on each actual application instance. As a request for a new session is being resolved, the network node VTAM will select the application instance that is running the lowest number of sessions.

  This support was introduced in VTAM V4.2.

- **VTAM call to the OS/390 Workload Manager:** After making the selection based on the session counts, the VTAM network node calls WLM, which can override the instance chosen using VTAM's own logic. WLM will look at the available capacity of the systems running the application instances of the generic resource. The system that has the highest available capacity is selected and the name of the application instance on this system is returned to VTAM.

  This support was introduced in VTAM V4.3.

- **VTAM Generic Resource exit:** Finally, VTAM will call the user-written VTAM generic resource exit (if it exists), which will override the first two choices. IBM provides a default exit called ISTEXCGR, which can be modified to select the application instance based on user criteria. For more information, refer to *VTAM for MVS/ESA V4.4 Customization*, LY43-0068 (available to IBM-licensed customers only).

  This support was introduced in VTAM V4.2.

  A sample VTAM exit is described in the IBM Redbook *Parallel Sysplex - Managing Software for Availability*, SG24-5451.

The reason for providing options to override VTAM's session balancing logic is to take into account the different relative processing capabilities of the systems. In Figure 54 on page 221, system C may have three times the capacity of system D and system E. In this case, using VTAM's session balancing logic alone may result in the underutilization of system C.

Session Affinity: Some sessions have special requirements, such as a dependence on a specific application instance of a generic resource. Examples of sessions with affinities are:

- Parallel LU6.2 sessions
- Parallel LU6.1 sessions

In addition to the application information described earlier, the CF also keeps information about individual LUs in the generic resources structure. If affinities exist, then the VTAM network node uses this information from the CF when it selects the application instance.

In Figure 54 on page 221, if the end user wishes to establish a parallel session after the first session is established with generic resource APPL, the network node doing the instance selection will ensure that APPL2 is selected. In this way, both sessions are between the end user and APPL2 and hence are parallel.

APPN Directory Services: After the actual application instance has been decided, the VTAM network node must then locate the application. It uses APPN's directory services to do this.

When applications on VTAM end nodes first open their ACBs, information is dynamically sent to the VTAM network node server for that end node, using the APPN control point session. There is no longer any need to predefine VTAM applications as cross-domain resources (CDRSCs) on other VTAMs. This process is known as *APPN registration*. If an application closes its ACB, then the VTAM end node will dynamically de-register itself from the network node server.

In the Parallel Sysplex environment, the VTAM network nodes will use the APPN *central directory server* (CDS) database. The CDS is used on every network node in a Parallel Sysplex. Starting with VTAM V4.4, there is no longer a requirement for every network node in the sysplex to be a CDS. In VTAM V4.4, none, some, or all the network nodes may be a CDS.

The CDS will return the name of the VTAM end node where the application instance resides, and also the name of the VTAM network node server for that end node. These two names are used by APPN to verify the application and also to calculate the route to be used for the end user session.

Once the actual application instance has been located and verified, the session request is sent to the VTAM where the instance is located. The session is then established using the selected route.

Local Access to Generic Resources: Special logic is used for VTAM workload balancing. If there are local LUs attached to a VTAM system that runs an actual application instance of a generic resource, the resource selection logic is different. Local LUs can be applications or end-user LUs. If a local LU requests a session with a generic resource name, VTAM will attempt to select the local instance on the same node.

This can result in unbalanced numbers of sessions among the instances of the generic resource. For example, in Figure 54 on page 221, assume there is a large number of local channel-attached devices connected to VTAME. When the end users on these devices request sessions with the generic resource name of APPL, they will *always* be connected to actual application instance APPL3.

4.5.1.3 Generic Resource Subsystem Support

The following IBM subsystems and their follow-on releases support the VTAM generic resources function:

- CICS/ESA V4.1
- DB2 V4.1
- IMS/ESA V6
- APPC/MVS for OS/390 R3
- TSO/E for OS/390 R3
- Netview Access Services V2

In addition, Computer Associates markets a session manager product called "Teleview" that also supports generic resources.

CICS Generic Resources Support: CICS/ESA V4.1 introduces the exploitation of the VTAM generic resources function to *distribute CICS logons* across different TORs in images in the Parallel Sysplex.

It is quite usual to have more than one CICS TOR on the same VTAM system, and the generic resources function can support this.

DB2 Generic Resources Support: DB2 V4.1 and follow-on releases can use the VTAM generic resources function to *balance sessions* from remote DRDA requesters.

DB2 uses the DRDA architecture to communicate with remote requesters, using LU6.2 sessions. The component of DB2 at the DRDA server called the Distributed Data Facility (DDF) is the part that connects to the network. It also opens the VTAM ACB, and can be part of a generic resource. DB2 has another way of doing workload balancing, referred to as *member routing*. With member routing, when a DRDA requester (single user or gateway) connects to a DB2 server in a data sharing group, the server returns a list of available members of the group, together with weighting information supplied by WLM to enable the requester to select the best member of the group to which to direct their request. The DRDA requester must be enabled to support this feature. For SNA DRDA requesters, the requester must support the APPC sysplex transaction program and, at the time of writing, the only DRDA requester that can use this function is DB2 for MVS/ESA V4 and upwards.

DB2 Workload Balancing

- For remote gateways with many parallel LU6.2 sessions from clients, use DB2's member routing.
- For remote clients which are single DRDA users, use VTAM's generic resources.

IMS Generic Resources Support: IMS/ESA V6 uses the VTAM generic resources function to balance sessions across IMS Transaction Managers.

APPC/MVS Generic Resources Support: Starting with OS/390 R3, APPC/MVS can associate a VTAM generic resources name with APPC/MVS LUs, to improve availability of APPC/MVS resources, and to balance sessions among APPC/MVS LUs. Using VTAM generic resources can also reduce the effort, complexity, and cost of managing a distributed processing environment that includes OS/390 systems.

TSO/E Generic Resources Support: Starting with OS/390 R3, TSO/E introduces the use of VTAM generic resources function to balance sessions across all systems in the sysplex.

Netview Access Services Generic Resources Support: Just as this book was going to press, support for Generic Resources was added to Netview Access Services (NVAS). The support is added via APAR PQ35801. With this support, NVAS will register with VTAM as a generic resource, providing the ability to log on to one of a group of NVAS address spaces using a single VTAM name. It also provides support for logging on to applications, using generic resource names, from NVAS.

This support adds a new CF list structure that NVAS uses to record information about user sessions with any of the NVAS address spaces in the generic resource group.

More information about this support can be obtained in INFO APAR II12403.

4.5.1.4 Generic Resource Planning Considerations

The following points summarize the planning considerations for generic resources. For some more general comments on APPN migration, refer to 2.1.2.1, “APPN Planning Considerations” on page 28 in Volume 3: Connectivity, SG24-5639.

- VTAM systems, which support generic resources, must be configured to use APPN (as APPN network nodes or APPN end nodes).
- There must be at least one VTAM network node in the *same* Parallel Sysplex as the VTAM end nodes.
- Each VTAM end node must have a control point session with one of the VTAM network nodes.
- All instances of a generic resource must be in the *same* Parallel Sysplex.
- All VTAM systems that will run generic resources must be in the same network (have the same netid).
- An application can only be known by one generic resource name.
- The generic resource name must be unique in the network. It cannot be the same as another LU or application name.
- A generic resource name cannot be the same as a USERVAR name. This means that any applications that are using generic resources cannot also use XRF.
- It is possible to have more than one actual application instance of a generic resource on a VTAM system.

4.5.2 TCP Workload Balancing

With the increasing emphasis on e-business applications on OS/390, TCP/IP has assumed a critical role in many OS/390 installations. Functions have been added to the TCP/IP implementation on OS/390 to support dynamic routing and workload management within a sysplex.

As we stated previously, which option you use to balance your TCP/IP workload is closely related to how your clients are connected to your systems. Therefore, in this section we will only very briefly describe the workload balancing options, and you should refer to 2.3.2, “Workload Balancing Enhancements” on page 48 in Volume 3: Connectivity, SG24-5639 for more information.

4.5.2.1 Workload Balancing Mechanisms

In TCP/IP, there are two basic mechanisms for balancing connection requests across multiple servers. You can distribute connections based on the *IP host names* of the servers (for example, WTSC63OE), or on an IP address associated with the servers.

If you are going to do the balancing based on host names, you would use Domain Name Servers to do the distribution. TCP/IP V3.2 introduced a function called *connection optimization*, whereby DNS and WLM cooperate to route connection requests to the most appropriate server. The advantage of this mechanism is that each request gets directed to the best server. The disadvantage is that *every* request must go through the overhead of DNS resolution. Also, this mechanism relies on clients not caching and reusing the provided IP address—something that is completely out of your control in a Web environment.

The other mechanism is to distribute the requests based on the IP address of the server. This mechanism would typically be used by an outboard router. The advantage of this mechanism is less overhead, as there is no DSN resolution required. The disadvantage is that server information is typically sampled instead of communicating with WLM for every connection request, so requests can potentially be routed to a less-than-optimal server.

More detailed information about both of these mechanisms, and where each is the most appropriate can be found in 2.3.2, “Workload Balancing Enhancements” on page 48 in Volume 3: Connectivity, SG24-5639.

4.6 APPC/MVS and Parallel Sysplex

Starting with VTAM V4.4, an installation can exploit APPC/MVS support of VTAM generic resources to improve availability of APPC/MVS resources. If one LU in the generic resource group or one system is brought down or fails, APPC/MVS work can continue because other group members are still available to handle requests that specify the generic resource name. Work from remote systems is less affected by the removal of any single APPC/MVS LU or OS/390 system. Additionally, changes in system configuration, capacity, and maintenance have less effect on APPC/MVS work.

APPC/MVS support of VTAM generic resources provides a single-system image for a multisystem APPC/MVS configuration. With generic resource names, transaction programs (TPs) from remote systems can establish conversations with APPC/MVS partner TPs on any system; programmers do not need to know specific partner LU names or update TPs whenever the APPC/MVS configuration changes.

APPC/MVS and Generic Resources Use

APPC/MVS TPs can use generic resource names only for partner LUs, not for local LUs.

With generic resource support, it is easier to expand the APPC/MVS configuration. Additional APPC/MVS LUs associated with the same generic resource name can provide immediate improvement in performance and availability, with few or no required changes for APPC/MVS TPs or side information.

Also, the distribution of work among two or more active APPC/MVS LUs on a single system or in a sysplex is possible, so that each LU is used as efficiently as possible. VTAM and WLM distribute session workload among members of generic resource group, thus reducing contention for specific LUs, and improving performance of systems and TPs.

4.7 TSO/E and Parallel Sysplex

Starting with VTAM V4.4 and OS/390 R3, TSO/E exploits VTAM generic resource support. Functions in the Workload Manager component of OS/390 R3 and the Communications Server allow users to logon to TSO as before, but now using a combination of generic resources and dynamic workload balancing, their sessions are directed to the systems which have the lightest load.

TSO/E and Generic Resources

TSO/E generic resource support provides the capability to balance session distribution across TSO systems in the sysplex.

The ability to reconnect to the original task, if the TSO/E user loses connection, is also included in this support.

If dynamic workload balancing in a TSO environment is of importance to you prior to OS/390 R3, there is a function in VTAM called USERVAR that you may use. Through the USERVAR function, you can provide a pseudonym for multiple TSO systems. The USERVAR exit routine can then be implemented to distribute TSO users to these systems. This allows you to balance your workload, but it is not possible to have functions, such as reconnect, work in a predictable way. However, prior to OS/390 R3, the USERVAR function might satisfy your needs. For further information on the USERVAR exit routine, refer to *VTAM for MVS/ESA V4.4 Network Implementation Guide*, SC31-8370.

For issues relating to the use of session managers in a Parallel Sysplex environment, refer to 2.1.5, “Session Manager Implications in Parallel Sysplex” on page 33 in Volume 3: Connectivity, SG24-5639.

4.7.1 MAS Considerations

The same TSO user ID cannot be logged onto more than one member in the MAS at any one time. It is possible to make a modification in JES2 code to allow the same TSO user ID to be logged onto multiple members in the MAS. This, however, is recommended only for short term migration purposes.

HASPCNVT Modification

```
BZ  XTDUPEND      *** Instruction Deleted ***  @420P190 05990900
B   XTDUPEND    Skip duplicate logon check MODIFICATION 05990901
```

Note: The source for module HASPCNVT is found in SYS1.HASPSRC. With this modification, beware of the following exposures and problems:

- You must be careful not to edit or update the same data set from duplicate TSO user IDs on different systems because there is no SYSTEMS level ENQ to guarantee serialization.
- TSO user notifications are "random." (They go to the first TSO user logged on in the sysplex.)
- Watch out for TSO logon ENQ - if it is SYSTEMS scope in the GRS RNL, then TSO will reject the second logon with the message `already logged on`.
- Ensure the ISPF data set names are unique by qualifying them with the system name.

4.7.2 Query Management Facility Workload Considerations

One of the typical workloads for TSO users in a DB2 environment is the use of the query management facility (QMF) for online, often read-only, data analysis. Sometimes QMF uses the *uncommitted read* bind option (ISOLATION(UR)).

Since the access is often read-only, it may be acceptable to use data that is current. If this is true, you may be able to offload the QMF workload to another image and not worry about that DB2 being part of the data sharing group. The QMF queries could then run against the database without having to worry about data sharing overheads. If, however, there is any update of the data required, then the DB2 subsystem *will* have to be part of the data sharing group. Note, however, that TSO attach cannot span across several OS/390s.

4.8 Test Considerations in Parallel Sysplex

In this section we look at implications for testing in a Parallel Sysplex. General testing considerations will not be covered. Depending on the environment implemented for test purposes in Parallel Sysplex, your configuration may vary considerably. For some examples of various test Parallel Sysplex environments, refer to 2.3.1.4, "Pre-Production Environment Cost Options" on page 32.

4.8.1 Testing Implications in Parallel Sysplex

Testing in the Parallel Sysplex environment, which could have up to 32 systems, provides an opportunity for "creativity" because applications should be tested in an environment that mimics the production environment. The challenge is to provide for a thorough, effective test without using excessive resources, and ensuring errors are not introduced when moving or cloning systems. Specific areas for consideration are:

- CF functions
- Dynamic transaction routing
- Failure and recovery scenarios
- Variations or unique functions in systems
- Stress testing
- Shared data validation
- Network
- Maintaining an adequate test environment

This section contains a discussion of each of these areas.

4.8.1.1 CF

Since the CF is vital to the integrity of the system, verifying that it is working correctly with new releases of software is critical.

Considerations include how to test all the various links to each of the systems, verifying that recovery works, validating the data sharing capabilities, and insuring that each user of the CF is working correctly.

The same CF cannot be used for both production and test, because some subsystems have hard-coded the names of the CF structures into their code.

Note: This does not mean that multiple CFs cannot be run in separate LPs on a given CPC. However, multiple CFs in separate LPs on a CPC must all run with the same level of CFCC. For this reason, running your test CFs in the same CPC as the production ones does not provide the capability to test new CF levels in the test sysplex before migrating them to the production sysplex.

4.8.1.2 Dynamic Transaction Routing

The implementation of transaction routing introduces the necessity to verify that routing is occurring as expected. When new applications are introduced, routing may be impacted. Test cases will need to be run to verify the routing. The best test would be to have all the systems participate. Realistically, only a subset of the systems is validated during a test. An availability or risk analysis is needed to determine the optimum test environment.

During the application testing, tests should be developed to validate or detect transaction affinities. It is far better to learn of the affinity during testing than after the application has gone into production and been propagated across several systems. If affinities are detected, appropriate changes can be made to the dynamic routing tables.

4.8.1.3 Failure and Recovery Scenarios

One of the largest causes of problems with CF recovery is a lack of familiarity with the recovery process by those responsible for monitoring and controlling it. Many times, recovery procedures are written, tested, and then never touched again until they are needed, at which point they may be out of date, or the person driving the process has forgotten exactly how the process works. We strongly recommend providing a facility (and an incentive) for system programmers and operators to test their procedures on a regular basis. This allows any problems to be discovered and addressed and helps maintain familiarity with the process and logic behind the procedures.

Test cases will need to be developed to see whether the new environment will recover appropriately when the following failures occur:

- Connectivity failure including:
  - VTAM EN/NN/IN
  - CTCs
  - 3745s
- Couple data sets failure
- CF link failure
- CF failure
- CF structure failure[5]

[5] An event that is unlikely to occur, and difficult to provoke.

- CPC failure
- OS/390 system failure
- Subsystem failure including:
  - Individual CICS AOR failure
  - Individual CICS TOR failure
  - Database managers (for example DB2 and IMS DB)
- Sysplex Timer connectivity failure
- Application failure

Failure tests should be used to get a clear understanding of what all the different structure owners *do* during CF failures (for example, VTAM rebuilds its structure, JES2 goes into CKPT RECONFIG dialog, IRLM rebuilds depending on REBUILDPERCENT, and so on).

Recommended Reference Test Document

OS/390 Parallel Sysplex Test Report, GC28-1963 is an experience-based document. The document is updated quarterly with new experiences and recent product enhancements.

There is also a section on testing in the *Parallel Sysplex Availability Checklist*, available on the Web at: `http://www.s390.ibm.com/ftp/marketing/position/availchk_parsys.pdf`

4.8.1.4 Variations or Unique Functions

To successfully test a new system that is propagated to each of the different hardware platforms in the Parallel Sysplex, all variations must be known. Test cases are needed to validate every variation prior to propagation so that errors do not occur upon implementation because the system has some unique feature or application.

Note: Thus, it is recommended to keep all the images in a Parallel Sysplex as “identical” as possible.

4.8.1.5 Stress or Performance Testing

Many problems, both in application code and system software, only show up under stress. Also, the problems may only surface with a certain workload mix. For these reasons, we strongly recommend that some mechanism to do stress testing should be available, both to the system programmers and also to the application programmers. Further, the stress tests should be set up to mirror the production workload mix as closely as possible. So, if your production workload consists of 50% CICS, 20% DB2, and 10% batch, you should try to match those proportions when running the stress tests.

4.8.1.6 Shared Data Validation

With the introduction of enabling shared data, testing needs to insure that the integrity of the data is being maintained. Testing needs to verify that the expected data is updated correctly and that applications do not break any rules.

4.8.1.7 Network

Note that VTAM has been changed to allow the specification of its CF structure name. The name is specified as a VTAM start parameter. Before this change, VTAM was one of the subsystems that hard-coded the name of its structure in the CF.

The second consideration is to ensure that there is adequate connectivity to the test systems to have a thorough test of the various terminals, connections, and printers within the network. This consideration also applies to non-Parallel Sysplex systems.

4.8.1.8 Maintaining an Adequate Test Environment

Determining what is needed for a valid test environment is a function of your configuration and availability requirements. The trade-off of equipment and resources (costs) versus reduction in errors (availability) needs to be made. The greater the availability requirement, the more robust the test environment and test plan need to be.

For example, to functionally test dynamic transaction routing, the following is needed: at least two systems with the software, CF, CF links and transactions that are to be routed, and shared data. For a general recommendation for a testing environment in Parallel Sysplex, refer to the label box on page 29.

4.9 How to Select Applications to Exploit Parallel Sysplex

Parallel Sysplex provides enhanced application functions that can improve application value, and thus benefit your business. Some application requirements that exist in most installations include:

- Improved application performance
- Continuous application availability
- Application growth

Refer to the IBM Redbook *S/390 MVS Parallel Sysplex Migration Paths*, SG24-2502, for a method to use when actually selecting the workloads to exploit a Parallel Sysplex.

Appendix A. Special Notices

This publication is intended to help large systems customers configure a Parallel Sysplex.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The information about non-IBM ("vendor") products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

Any performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

This document contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples contain the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

ACF/VTAM	AD/Cycle
ADSTAR	Advanced Function Printing
Advanced Peer-to-Peer Networking	AFP
AIX	AnyNet
APPN	BatchPipes
BatchPipeWorks	BookManager
C/370	CBIPO
CBPDO	CICS
CICS/ESA	CICS/MVS
CICS/VSE	CICSPlex
CUA	DATABASE 2
DB2	DB2/2
DB2 Connect	DB2 Universal Database
DFSMS	DFSMS/MVS
DFSMSdfp	DFSMSdss
DFSMShsm	DFSORT
Distributed Relational Database Architecture	DRDA
eNetwork	Enterprise System/9000
Enterprise Systems Architecture/390	Enterprise Systems Connection Architecture
ES/3090	ES/9000
ESA/390	ESCON XDF
ESCON	Extended Services
GDDM	Hardware Configuration Definition
Hiperbatch	Hipersorting
Hiperspace	HPR Channel Connectivity
IBM	IBMLink
IMS	IMS/ESA
Intelligent Miner	Language Environment
LSPR	LSPR/PC
MERVA	MQ
MQSeries	Multiprise
MVS/DFP	MVS/ESA
MVS/SP	MVS/XA
Netfinity	Net.Data
NetView	NTune
Nways	OPC
OpenEdition	Operating System/2
OS/2	OS/390
OS/400	Parallel Sysplex
Personal System/2	Powered by S/390
PowerPC	PowerPC 604
PR/SM	Presentation Manager
Print Services Facility	Processor Resource/Systems Manager
PROFS	PS/2
PSF	QMF
RACF	RAMAC
Resource Measurement Facility	RETAIN

RISC System/6000	RMF
RRDF	RS/6000
S/370	S/390
S/390 Parallel Enterprise Server	SAA
SecureWay	SNAP/SHOT
SQL/DS	Sysplex Timer
System/370	System/390
SystemPac	Systems Application Architecture
Systems Validation Services	SystemView
S/390 Parallel Enterprise Server	VisualAge
VM/ESA	VM/XA
VSE/ESA	VTAM
WebSphere	WIN-OS/2
3090	400

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere., TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

SET, SET Secure Electronic Transaction, and the SET logo are trademarks owned by SET Secure Electronic Transaction LLC.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

Other company, product, and service names may be trademarks or service marks of others.

Appendix B. Related Publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

Note that some of these publications are only available in softcopy—either on one of the CD collections or on the Internet.

B.1 IBM Redbooks

For information on ordering these ITSO publications see "How to get IBM Redbooks" on page 247.

The publications are listed in alphabetical order.

- *A Performance Study of Web Access to CICS*, SG24-5748
- *Accessing CICS Business Applications from the World Wide Web*, SG24-4547
- *Accessing DB2 for OS/390 Data from the World Wide Web*, SG24-5273
- *Automation for S/390 Parallel Sysplex*, SG24-4549
- *Batch Processing in a Parallel Sysplex*, SG24-5329
- *CICS and VSAM RLS: Implementation Guide*, SG24-4766 (available in softcopy only)
- *CICS and VSAM RLS: Planning Guide*, SG24-4765 (available in softcopy only)
- *CICS and VSAM RLS: Recovery Considerations*, SG24-4768 (available in softcopy only)
- *CICS TS for OS/390 V1R2 Implementation Guide*, SG24-2234
- *CICS TS for OS/390: V1.3 Implementation Guide*, SG24-5274
- *CICS TS: Web Interface and 3270 Bridge*, SG24-5243
- *CICS Workload Management Using CICSPlex SM and the MVS/ESA Workload Manager*, GG24-4286
- *Connecting IMS to the World Wide Web: A Practical Guide to IMS Connectivity*, SG24-2220
- *Consolidating UNIX Systems onto OS/390*, SG24-2090
- *Continuous Availability S/390 Technology Guide*, SG24-2086
- *Continuous Availability Systems Design Guide*, SG24-2085
- *Data Warehousing with DB2 for OS/390*, SG24-2249
- *DB2 for MVS/ESA Version 4 Data Sharing Implementation*, SG24-4791
- *DB2 for MVS/ESA V4 Data Sharing Performance Topics*, SG24-4611
- *DB2 for OS/390 Application Design Guidelines for High Performance*, SG24-2233
- *DB2 for OS/390 Capacity Planning*, SG24-2244
- *DB2 for OS/390 V5 Performance Topics*, SG24-2213
- *DB2 on the MVS Platform: Data Sharing Recovery*, SG24-2218
- *DB2 UDB for OS/390 and Continuous Availability*, SG24-5486
- *DB2 UDB for OS/390 Version 6 Performance Topics*, SG24-5351
- *DFSMS Optimizer Usage Guide*, SG24-2235
- *Disaster Recovery Library: Database Recovery*, GG24-3993 (available in softcopy only)
- *Enhanced Catalog Sharing and Management*, SG24-5594
- *Getting Started with DB2 Stored Procedures: Give Them a Call through the Network*, SG24-4693
- *Getting the Most Out of a Parallel Sysplex*, SG24-2073
- *HCD and Dynamic I/O Reconfiguration Primer*, SG24-4037
- *IMS/ESA Data Sharing in a Parallel Sysplex*, SG24-4303

- *IMS e-business Connect Using the IMS Connectors*, SG24-5427
- *IMS/ESA Multiple Systems Coupling in a Parallel Sysplex*, SG24-4750
- *IMS/ESA Shared Queues: A Planning Guide*, SG24-5257
- *IMS/ESA Sysplex Data Sharing: An Implementation Case Study*, SG24-4831
- *IMS/ESA V6 Shared Queues*, SG24-5088
- *IMS/ESA V6 Parallel Sysplex Migration Planning Guide*, SG24-5461
- *IMS/ESA V6 Guide*, SG24-2228
- *IMS Version 7 Release Guide*, SG24-5753
- *Inside APPN: The Essential Guide to the Next-Generation SNA*, SG24-3669
- *JES3 in a Parallel Sysplex*, SG24-4776
- *Modeling Host Environments using SNAP/SHOT*, SG24-5314
- *Net.Commerce for OS/390*, SG24-5154
- *Open Systems Adapter 2 Implementation Guide*, SG24-4770
- *OS/390 eNetwork Communications Server V2R7 TCP/IP Implementation Guide Volume 1: Configuration and Routing*, SG24-5227
- *OS/390 eNetwork Communications Server V2R7 TCP/IP Implementation Guide Volume 2: UNIX Applications*, SG24-5228
- *OS/390 MVS Multisystem Consoles Implementing MVS Sysplex Operations*, SG24-4626
- *OS/390 Parallel Sysplex Application Considerations Presentation Guide*, SG24-4743
- *OS/390 MVS Parallel Sysplex Capacity Planning*, SG24-4680
- *OS/390 MVS Parallel Sysplex Performance Healthcheck Case Study: DMdata/Danske Data*, SG24-5373
- *OS/390 R5 Implementation*, SG24-5151
- *OS/390 Workload Manager Implementation and Exploitation*, SG24-5326
- *Parallel Sysplex CF Online Monitor: Installation and User's Guide*, SG24-5153
- *Parallel Sysplex Automation Guidelines*, SG24-5441
- *Parallel Sysplex Automation: Using System Automation for OS/390*, SG24-5442
- *Parallel Sysplex Continuous Availability Case Studies*, SG24-5346
- *Parallel Sysplex - Managing Software for Availability*, SG24-5451
- *Parallel Sysplex Operational Scenarios*, SG24-2079
- *Planning for IBM Remote Copy*, SG24-2595
- *RACF V2.2 Installation and Implementation Guide*, SG24-4580
- *RAMAC Virtual Array: Implementing Peer-to-Peer Remote Copy*, SG24-5338
- *Revealed! CICS Transaction Gateway and More CICS Clients Unmasked*, SG24-5277
- *Revealed! Architecting Web Access to CICS*, SG24-5466
- *Securing Web Access to CICS*, SG24-5756
- *Selecting a Server - The Value of S/390*, SG24-4812
- *SNA in a Parallel Sysplex Environment*, SG24-2113
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Configuration Guidelines Student Handout*, SG24-4943 (available in softcopy only)
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Configuration Guidelines*, SG24-4927 (available in softcopy only)
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Exercise Guide*, SG24-4913 (available in softcopy only)
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Exercise Installation and Run-Time Procs*, SG24-4912 (available in softcopy only)
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Presentation Guide*, SG24-4911 (available in softcopy only)
- *S/390 G3: Enterprise Server Complex Systems Availability and Recovery Student Handout*, SG24-4942 (available in softcopy only)

- *S/390 MVS Parallel Sysplex Batch Performance*, SG24-2557 (available in softcopy only)
- *S/390 MVS Parallel Sysplex Continuous Availability Presentation Guide*, SG24-4502
- *S/390 MVS Parallel Sysplex Continuous Availability SE Guide*, SG24-4503
- *S/390 MVS Parallel Sysplex Migration Paths*, SG24-2502
- *S/390 MVS Parallel Sysplex Performance*, SG24-4356
- *S/390 MVS/ESA SP V5 WLM Performance Studies*, SG24-4352
- *S/390 OSA-Express Gigabit Ethernet Implementation Guide*, SG245443
- *S/390 OSA-Express Implementation Guide*, SG245948
- *S/390 Parallel Sysplex: Resource Sharing*, SG24-5666
- *TCP/IP in a Sysplex*, SG24-5235
- *Using VTAM Generic Resources with IMS*, SG24-5487
- *World Wide Web Access to DB2*, SG24-4716
- *WOW! DRDA Supports TCP/IP: DB2 Server for OS/390 and DB2 Universal Database*, SG24-2212

B.2 IBM Redbooks collections

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at http://www.redbooks.ibm.com/ for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	Collection Kit Number
System/390 Redbooks Collection	SK2T-2177
Networking and Systems Management Redbooks Collection	SK2T-6022
Transaction Processing and Data Management Redbooks Collection	SK2T-8038
Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
AS/400 Redbooks Collection	SK2T-2849
Netfinity Hardware and Software Redbooks Collection	SK2T-8046
RS/6000 Redbooks Collection (BkMgr Format)	SK2T-8040
RS/6000 Redbooks Collection (PDF Format)	SK2T-8043
Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

B.3 Other Resources

These publications are also relevant as further information sources.

The publications are listed in alphabetical order.

- *A Technical Overview: VTAM V4R2, NCP V6R3, V7R1 and V7R2*, GG66-3256
- *BatchPipes for OS/390 V2 R1 BatchPipeWorks Reference*, SA22-7456
- *BatchPipes for OS/390 V2 R1 BatchPipeWorks User Guide*, SA22-7457
- *BatchPipes for OS/390 V2 R1 Users Guide and Reference*, SA22-7458
- *BatchPipes for OS/390 V2 R1 Introduction*, GA22-7459
- *CICS Transaction Affinities Utility Guide*, SC33-1777
- *CICS IMS Database Control Guide*, SC33-1700
- *CICSPlex Systems Manager V1 Concepts and Planning*, GC33-0786
- *CICS TS for OS/390 Installation Guide*, GC33-1681
- *CICS TS for OS/390 Intercommunication Guide*, SC33-1695
- *CICS TS for OS/390 Migration Guide*, GC34-5353
- *CICS TS for OS/390 Operations and Utilities Guide*, SC33-1685
- *CICS TS for OS/390 Recovery and Restart Guide*, SC33-1698
- *CICS TS for OS/390 Release Guide*, GC33-1570

- *CICS TS for OS/390 System Definition Guide*, SC33-1682
- *DB2 UDB for OS/390 V6 Administration Guide*, SC26-9003
- *DB2 UDB for OS/390 V6 Data Sharing: Planning and Administration*, SC26-9007
- *DB2 UDB for OS/390 V6 Release Guide*, SC26-9013
- *DFSMS/MVS V1.5 Access Method Services for ICF Catalogs*, SC26-4906
- *DFSMS/MVS V1.5 DFSMSdfp Storage Administration Reference*, SC26-4920
- *DFSMS/MVS V1.5 General Information*, GC26-4900
- *DFSMS/MVS V1.5 Managing Catalogs*, SC26-4914
- *DFSMS/MVS V1.5 Planning for Installation*, SC26-4919
- *DFSMS/MVS Remote Copy Guide and Reference*, SC35-0169
- *ES/9000 and ES/3090 PR/SM Planning Guide*, GA22-7123
- *Fiber Optic Link Planning*, GA23-0367
- *Hardware Configuration Manager User's Guide*, SC33-6469
- *IBM RAMAC Scalable Array Storage Introduction*, GC26-7212
- *IBM RAMAC Virtual Array Storage Introduction*, GC26-7168
- *IBM RAMAC Virtual Array Storage Operation and Recovery*, GC26-7171
- *IBM RAMAC Virtual Array Storage Planning, Implementation and Usage*, GC26-7170
- *IBM RAMAC Virtual Array Storage Reference*, GC26-7172
- *IBM Token-Ring Network Introduction and Planning Guide*, GA27-3677
- *IBM 3990/9390 Storage Control Planning, Installation and Storage Administration Guide*, GA32-0100
- *IBM 3990/9390 Reference for Model 6*, GA32-0274
- *IBM 3990/9390 Operations and Recovery Guide*, GA32-0253
- *IBM 9397 RAMAC Electronic Array Storage Introduction*, GC26-7205
- *IMS/ESA V6 Administration Guide: Database Manager*, SC26-8725
- *IMS/ESA V6 Administration Guide: System*, SC26-8730
- *IMS/ESA V6 Administration Guide: Transaction Manager*, SC26-8731
- *IMS/ESA V6 Database Recovery Control (DBRC) Guide and Reference*, SC26-8733
- *IMS/ESA V6 Installation Volume 2: System Definition & Tailoring*, SC26-8737
- *IMS/ESA V6 Operations Guide*, SC26-8741
- *IMS/ESA V6 Release Planning Guide*, GC26-8744
- *IMS/ESA V6 Utilities Reference: Database Manager*, SC26-8769
- *IMS/ESA V6 Utilities Reference: System*, SC26-8770
- *Large Systems Performance Reference*, SC28-1187
- *NCP V7.8 Resource Definition Guide*, SC31-6223
- *Network Products Reference*, GX28-8002
- *Nways Multiprotocol Access Services Configuration Reference Vol 1*, SC30-3884
- *Nways Multiprotocol Access Services Configuration Reference Vol 2*, SC30-3885
- *OS/390 MVS An Introduction to OS/390*, GC28-1725
- *OS/390 MVS Conversion Notebook*, GC28-1747
- *OS/390 MVS Hardware Configuration Definition: User's Guide*, SC28-1848
- *OS/390 MVS Initialization and Tuning Guide*, SC28-1751
- *OS/390 MVS Initialization and Tuning Reference*, SC28-1752
- *OS/390 MVS Parallel Sysplex Application Migration*, GC28-1863
- *OS/390 MVS Parallel Sysplex Hardware and Software Migration*, GC28-1862
- *OS/390 MVS Parallel Sysplex Overview*, GC28-1860
- *OS/390 MVS Parallel Sysplex Test Report*, GC28-1963 (available in softcopy only)
- *OS/390 MVS Planning: APPC/MVS Management*, GC28-1807
- *OS/390 MVS Planning: Global Resource Serialization*, GC28-1759

- *OS/390 MVS Planning: Operations*, GC28-1760
- *OS/390 MVS Planning: Security*, GC28-1920
- *OS/390 MVS Planning: Workload Management*, GC28-1761
- *OS/390 MVS Programming: Assembler Services Guide*, GC28-1762
- *OS/390 MVS Programming: Assembler Services Reference*, GC28-1910
- *OS/390 MVS Programming: Authorized Assembler Services Guide*, GC28-1763
- *OS/390 MVS Programming: Sysplex Services Reference*, GC28-1496
- *OS/390 MVS Programming: Sysplex Services Guide*, GC28-1495
- *OS/390 MVS Recovery and Reconfiguration Guide*, GC28-1777
- *OS/390 MVS System Commands*, GC28-1781
- *OS/390 MVS System Management Facilities (SMF)*, GC28-1783
- *OS/390 MVS RMF Performance Management Guide*, SC28-1951
- *OS/390 MVS RMF Report Analysis*, SC28-1950
- *OS/390 MVS RMF User's Guide*, SC28-1949 (available in softcopy only)
- *OS/390 MVS Security Server System Programmer's Guide*, SC28-1913
- *OS/390 MVS Setting Up a Sysplex*, GC28-1779
- *Planning for S/390 Open Systems Adapter Feature*, GC23-3870
- *S/390 9672/9674/2003 PR/SM Planning Guide*, GA22-7236
- *S/390 9672/9674 Hardware Management Console Guide*, GC38-0453
- *S/390 9672/9674 Managing Your Processors*, GC38-0452
- *S/390 9672/9674 Programming Interfaces: Hardware Management Console Appication*, SC28-8141
- *Sysplex Timer Planning*, GA23-0365
- *System Automation for OS/390 General Information*, GC28-1541
- *Planning for the 9037 Model 2*, SA22-7233
- *Using the IBM 9037 Model 2 Sysplex Timer*, SA22-7230
- *TME 10 OPC Installation Guide*, SH19-4379
- *OS/390 IBM Communications Server: IP Configuration Guide*, SC31-8725
- *OS/390 IBM Communications Server: SNA Migration*, SC31-8622
- *OS/390 IBM Communications Server: SNA Network Implementation Guide*, SC31-8563
- *OS/390 IBM Communications Server: SNA Resource Definition Reference*, SC31-8565
- *2216 Nways Multiaccess Connector Hardware Installation Guide*, GA27-4106
- *2216 Nways Multiaccess Connector Planning Guide*, GA27-4105
- *3745/3746 Overview*, GA33-0180
- *3745 Communications Controller Models A, 3746 Nways Multiprotocol Controller Models 900 and 950 Planning Guide*, GA33-0457
- *3746 Model 900 Migration and Planning Guide*, GA33-0183
- *3746 Model 900/950 Network Node Migration and Planning Guide*, GA33-0349

B.4 Packages

Most of the following packages are available on MKTTOOLS. Some packages come from other tools disks as noted.

MKTTOOLS is accessed by IBM employees in multiple ways:

- OMNIDISK: A VM facility that provides easy access to IBM TOOLS and conference disks
- TOOLCAT: A VM EXEC that provides a catalog of TOOLS disks and easy access to the packages that reside on the disks
- Direct CMS commands (useful if you already know the package name and where it resides)

Many of the packages listed here are available for customer usage; ask your IBM representative for more information.

The publications are listed in alphabetical order.

- ARMPAPER package: *The Automatic Restart Manager*[6]
- ARMWRAP package: *ARM Wrapper*
- BV390 package: *S/390 Parallel Sysplex Business Value*
- BWATOOL package: *Batch Workload Analysis Tool*
- CD13 package: *CICS MVS/ESA CMOS Processor Utilization Tool*[7]
- CHKLIST package: *Hints and Tips Checklist for Parallel Sysplex*
- CICSE41P package: *Performance Comparison of CICS/ESA V4.1 and V3.3*
- CICSTS01 package: *CICS TS for OS/390*
- CICSTS02 package: *CICS TS for OS/390 Presentation Guide*
- CICSTS03 package: *CICS TS for OS/390 Migration Guidance*
- CICSTS05 package: *CICS TS for OS/390 Questions and Answers*
- CICSTS06 package: *CICS TS for OS/390 Support for Parallel Sysplex*
- CICSTS07 package: *CICS TS a Technical Introduction presentation*
- CICSTS09 package: *CICS TS for OS/390 Performance Report*
- CICSTS12 package: *CICS TS Performance Report for CPSM 1.3.0*
- CICSTS13 package: *CICS TS and RLS Performance*
- 6950-22S package: *IBM Performance Management and Capacity Planning Svcs*
- CP2KOVER package: *An Overview of CP2000*[8]
- CP2000 package: *CP2000 OS/2 PC Tool*[9]
- CVRCAP package: *CICS/VSAM RLS Capacity Planning Tool*
- EMEAHELP package: *Parallel Sysplex focal points in EMEA*
- EPSOEMEA package: *EPSO Reference Guide and Reference Card plus EPSO Services Perform Guide and Statement of Work for EMEA*
- FICONPG package: *S/390 FICON Overview (FIPG) -- Connectivity Leadership*
- GBOF7101 package: *S/390 Parallel Sysplex Application Briefs*
- GDPS package: *S/390 Geographically Dispersed Parallel Sysplex Pguide*
- GF225005 package: *A Strategy Report to Customers*
- GF225008 package: *S/390 Technology Leadership White Paper*
- GF225009 package: *Parallel Sysplex Overhead: A Reality Check*
- GF225042 package: *S/390 Coupling Facility Configuration Alternatives*
- GF225114 package: *GDPS S/390 Multi-Site Application Availability Solution*
- GF225115 package: *IBM System/390 Value of Resource Sharing WP*
- G2219017 package: *IMS Workload Router - S/390*
- G3260594 package: *S/390 Software Pricing Reference Guide*
- G3263007 package: *SAP R/3 and S/390 - A Powerful Combination*
- G3263025 package: *S/390 Parallel Sysplex White Paper*
- G3263026 package: *e-business Powered by S/390 White Paper*
- G3263027 package: *S/390 SAP R/3 Whitepaper*
- G3263030 package: *Decision Support White Paper*
- G3263056 package: *S/390 Cost of Scalability Report*
- LPARCE package: *LPAR Capacity Estimator (LPAR/CE)*[10]
- LSPRPC package: *Large Systems Performance Reference (LSPR)*

---

6 TOOLS SENDTO KGNVMC TOOLSMVS MVSTOOLS GET ARMPAPER PACKAGE

7 TOOLS SENDTO WINVMB TOOLS TXPPACS GET CD13 PACKAGE

8 TOOLS SENDTO WSCVM TOOLS CPSTOOLS GET CP2KOVER PACKAGE

9 TOOLS SENDTO WSCVM TOOLS CPSTOOLS GET CP2000 PACKAGE

10 TOOLS SENDTO WSCVM TOOLS CPSTOOLS GET LPARCE PACKAGE

- OPSYSPLX package: *Why Implement an S/390 Operational Single Image (Sysplex)?*
- OSAPERF package: *Performance for S/390 Open Systems Adapter*
- OS390PSR package: *OS/390 Parallel Sysplex Recovery*
- OS390PRF package: *OS/390 Performance Studies*
- OS390TST package: *S/390 Parallel Sysplex Test Report*
- OS390T97 package: *OS/390 Parallel Sysplex Test Report*
- OS390T98 package: *OS/390 Parallel Sysplex Test Report*
- OS390T99 package: *OS/390 Parallel Sysplex Test Report*
- TLLBPAR package: *IBM S/390 Parallel Sysplex Cluster Technology P-Guide*
- TLLBOS79 package: *S/390 Tech Ldrshp Tech Library - OS/390 Version 2 R9*
- PCR package: *CP2000 Processor Capacity Reference (PCR)*[11]
- PSLCU package: *Parallel Sysplex License Charge (US Version)*
- QCBTRACE package: *Queue Control Block Trace*[12]
- QPS390 package: *Quick Pricer for S/390 Support Family*
- RLSLKSZ package: *RLS Lock Structure Sizing Tool*
- RMFTREND package: *RMF Trend Monitor*[13]
- SAFOS390 package: *System Automation for OS/390 Presentations and Infos*
- SA227403 package: *S/390 Open Systems Adapter-Express Customer's Guide*
- SG245176 package: *Introduction to FICON*
- SOFTCAP package: *Software Migration Capacity Planning Aid*
- SPSG390 package: *S/390 services from PSS - package list*
- SPSSZR package: *CP2000 S/390 Parallel Sysplex Quick-Sizer*[14]
- SWPRICER package: *S/390 PSLC Software Pricer (US Version)*
- 6942-08D package: *Operational Support - Parallel Sysplex Exploitation*
- 9037MIGR package: *Migration Planning for the 9037 Model 2 Sysplex Timer*
- 9729 Package: *IBM 9729 Optical Wavelength Division Multiplexer Timer*

B.5 Other References

If you would like to receive a weekly e-mail of IBM announcements tailored to your interest areas, you can subscribe at:
`http://www.ibm.com/isource`

Articles/White Papers

The publications listed are sorted in alphabetical order.

- *Coordination of Time-Of-Day Clocks among Multiple Systems, IBM Journal of Research and Development, Volume 36 No. 4*, G322-0181
- *Coupling Facility Configuration Options*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/marketing/gf225042.html`
- *Five Nines/Five Minutes - Achieving Near Continuous Availability*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/marketing/390avail.html`

11 TOOLS SENDTO WSCVM TOOLS CPSTOOLS GET PCR PACKAGE

12 TOOLS SENDTO KGNVMC TOOLSMVS MVSTOOLS GET QCBTRACE PACKAGE

13 http://www.s390.ibm.com/rmf

14 TOOLS SENDTO WSCVM TOOLS CPSTOOLS GET SPSSZR PACKAGE

- *Geographically Dispersed Parallel Sysplex: The S/390 Multi-Site Application Availability Solution*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/ftp/marketing/position/gf225063.pdf`
- *IBM Systems Journal special issue on Parallel Sysplex Clustering Technology*

  To view or print this S/390 white paper:
  `http://www.research.ibm.com/journal/sj36-2.html`
- *IBM's Parallel Sysplex Overhead: A Reality Check*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/ftp/marketing/position/gf225009a.pdf`
- *Network Implications of the Parallel Sysplex*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/pso/image/syspwp.pdf`
- *OS/390 Maintenance Recommendations for Improving Availability in a Parallel Sysplex Environment*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/marketing/psos390maint.html`
- *Improve Your Availability with Sysplex Failure Management*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/pso/image/sfm.pdf` G321-0127
- *S/390 Parallel Sysplex Cluster Technology: IBM's Advantage*

  To view or print this S/390 white paper:
  `http://www.s390.ibm.com/ftp/marketing/position/gf225015.pdf`

Diskette kit:

- *Hardware Management Console Tutorial*, SK2T-1198 (SK2T-1198 is not orderable on its own, but is part of SK2T-5843).

Washington System Center Flashes (IBMLink/Hone)

The following flashes are available from the Washington Systems Center Web site at:
`http://www.ibm.com/support/techdocs/atsmastr.nsf`

- *Dynamic ICF Expansion*, WSC Flash 98028
- *Integrated Coupling Migration Facility (ICMF) Dispatching Enhancement*, WSC Flash 95028
- *Locating OS/390 Information on the Internet*, WSC Flash 97044
- *LPAR Management Time Performance Update*, WSC Flash 99048
- *MVS Performance Capacity Planning Considerations For 9672-Rxx Processors*, WSC Flash 95005
- *MVS/ESA Parallel Sysplex Performance LPAR Performance Considerations For Parallel Sysplex Environments*, WSC Flash 96009
- *MVS/ESA Parallel Sysplex Performance*, WSC Flash 97031
- *MVS/ESA Parallel Sysplex Performance XCF Performance Considerations*, WSC Flash 10011
- *MVS/ESA V5.1 Performance Information V5 Release To Release Migration Software Performance Impact*, WSC Flash 94041
- *OS/390 Performance Information OS/390 Increased Virtual Storage Requirements*, WSC Flash 96013

- *OS/390 R2 GRS Star APAR/PTF and Lock Structure Sizing*, WSC Flash 97015
- *Parallel Sysplex Configuration Planning for Availability*, WSC Flash 98029
- *Parallel Sysplex Interactive Tools*, WSC Flash 99054
- *Parallel Sysplex Operational Procedures: An Update*, WSC Flash 98022
- *Performance Impacts of Using Shared ICF CPs*, WSC Flash 99037
- *SFM Functions and the Impact of OW30814 for Parallel Sysplex*, WSC Flash 98025B
- *Using a Coupling Facility for the JES2 Checkpoint*, WSC Flash 98048B
- *XCF Service Recommendations to Maximize Sysplex Availability*, WSC Flash 98038

How to get IBM Redbooks

This section explains how both customers and IBM employees can find out about IBM Redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** `http://www.redbooks.ibm.com/`

  Search for, view, download, or order hardcopy/CD-ROM Redbooks from the Redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

  Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders by e-mail including information from the redbook fax order form to:

In United States:	e-mail address: usib6fpl@ibmmail.com
Outside North America:	Contact information is in the "How to Order" section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl/

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the "How to Order" section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl/

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the "How to Order" section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl/

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the Redbooks Web site.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and Redbooks by accessing the IBM Intranet Web site at `http://w3.itso.ibm.com/` and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access `MyNews at http://w3.ibm.com/` for redbook, residency, and workshop announcements.

IBM Redbooks fax order form

Please send me the following:

Title	Order Number	Quantity

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

□ Invoice to customer number

□ Credit card number

Credit card expiration date Card issued to Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Glossary

Explanations of Cross-References:

The following cross-references are used in this glossary:

Contrast with. This refers to a term that has an opposed or substantively different meaning.

See. This refers the reader to multiple-word terms in which this term appears.

See also. This refers the reader to terms that have a related, but not synonymous meaning.

Synonym for. This indicates that the term has the same meaning as a preferred term, which is defined in the glossary.

If you do not find the term you are looking for, see the IBM Software Glossary at the URL: `http://www.networking.ibm.com/nsg/nsgmain.htm`

A

abend. Abnormal end of task.

ACF/VTAM. Advanced Communications Function for the Virtual Telecommunications Access Method. Synonym for *VTAM*.

active IRLM. The IRLM supporting the active IMS subsystem in an XRF complex.

active service policy. The service policy that determines workload management processing if the sysplex is running in goal mode. See *goal mode*.

adapter. Hardware card that allows a device, such as a PC, to communicate with another device, such as a monitor, a printer, or other I/O device. In a LAN, within a communicating device, a circuit card that, with its associated software and/or microcode, enables the device to communicate over the network.

affinity. A connection or association between two objects.

alternate IRLM. The IRLM supporting the alternate IMS subsystem in an XRF complex.

alternate site. Another site or facility, such as a commercial hot site or a customer-owned second site, that will be a recovery site in the event of a disaster.

ambiguous cursor. A database cursor that is not declared with either the clauses FOR FETCH ONLY or FOR UPDATE OF, and is not used as the target of a WHERE CURRENT OF clause on an SQL UPDATE or DELETE statement. The package processes dynamic SQL statements.

architecture. A logical structure that encompasses operating principles including services, functions, and protocols. See *computer architecture, network architecture*, and *Systems Network Architecture (SNA)*.

asynchronous. Without regular time relationship. Unexpected or unpredictable with respect to the program's instructions, or to time. Contrast with synchronous.

authorized program analysis report (APAR). A request for correction of a problem caused by a defect in a current release of a program unaltered the user.

availability. A measure of how much (often specified as a percentage) the data processing services are available to the users in a specified time frame.

B

base or basic sysplex. A base or basic sysplex is the set of one or more OS/390 systems that is given a cross-system coupling facility (XCF) name and in which the authorized programs can then use XCF coupling services. A base sysplex does not include a CF. See also *Parallel Sysplex* and *sysplex*.

basic mode. A central processor mode that does not use logical partitioning. Contrast with *logically partitioned (LPAR) mode*.

batch checkpoint/restart. The facility that enables batch processing programs to synchronize checkpoints and to be restarted at a user-specified checkpoint.

batch environment. The environment in which non-interactive programs are executed. The environment schedules their execution independently of their submitter.

batch message processing (BMP) program. An IMS batch processing program that has access to online databases and message queues. BMPs run online, but like programs in a batch environment, they are started with job control language (JCL).

batch-oriented BMP program. A BMP program that has access to online databases and message queues while performing batch-type processing. A batch-oriented BMP does not access the IMS message queues for input or output. It can access online databases, GSAM databases, and OS/390 files for both input and output.

batch processing program. An application program that has access to databases and OS/390 data management facilities but does not have access to the IMS control region or its message queues. See also *batch message processing program* and *message processing program.*

block level sharing. A kind of data sharing that enables application programs in different IMS systems to update data concurrently.

block multiplexer channel. A channel that transmits blocks of data to and from more than one device by interleaving the record blocks. Contrast with *selector channel*.

BMP program. See *batch message processing program*.

buffer. (1) A portion of storage used to hold input or output data temporarily. (2) A routine or storage used to compensate for a difference in data rate or time of occurrence of events, when transferring data from one device to another.

buffer invalidation. A technique for preventing the use of invalid data in a Parallel Sysplex data sharing environment. The technique involves marking all copies of data in DB2 or IMS buffers invalid once a sharing DBMS subsystem has updated that data.

buffer pool. A set of buffers that contains buffers of the same length. See also *buffer*, *buffer invalidation*, and *group buffer pool*.

byte multiplexer channel. A multiplexer channel that interleaves bytes of data. See also *block multiplexer channel*. Contrast with *selector channel*.

C

cache structure. A CF structure that enables high-performance sharing of cached data by multisystem applications in a sysplex. Applications can use a cache structure to implement several different types of caching systems, including a store-through or a store-in cache. As an example, DB2 uses data sharing group cache structures as GBPs. See also *group buffer pool*, *castout*, and *cache structure services*.

cache structure services. OS/390 services that enable applications in a sysplex to perform operations such as the following on a CF cache structure:

- Manage cache structure resources.
- Store data into and retrieve data from a cache structure.
- Manage accesses to shared data.
- Determine when shared data has been changed.
- Determine whether a local copy of shared data is valid.

card-on-board (COB) logic. The type of technology that uses pluggable, air-cooled cards.

CAS. Coordinating address space.

castout. The DB2 process of writing changed pages from a GBP to DASD.

catalog. A data set that contains extensive information required to locate other data sets, to allocate and deallocate storage space, to verify the access authority of a program or operator, and to accumulate data set usage statistics.

central processing unit (CPU). The part of a computer that includes the circuits that control the interpretation and execution of instructions.

central processor (CP). The part of the computer that contains the sequencing and processing facilities for instruction execution, initial program load, and other machine operations. See also *central processor complex*, *central electronic complex* and *PU*.

central processor complex (CPC). A physical collection of hardware that includes central storage, one or more CPs, timers, and channels.

central storage. Storage that is an integral part of the processor unit. Central storage includes both main storage and the hardware system area.

CF. Coupling Facility. See also *Coupling Facility*.

CFCC. Coupling Facility Control Code. See also *Coupling Facility Control Code*.

CFRM policy. A declaration regarding the allocation rules for a CF structure. See also *structure*.

channel. (1) A functional unit, controlled by a S/390 CPC that handles the transfer of data between processor storage and local peripheral equipment. (2) A path along which signals can be sent. (3) The portion of a storage medium that is accessible to a given reading or writing station. (4) In broadband transmission, a designation of a frequency band 6 MHz wide.

channel subsystem (CSS). A collection of subchannels that directs the flow of information between I/O devices and main storage, relieves the processor of communication tasks, and performs path management functions.

channel-to-channel (CTC). Refers to the communication (transfer of data) between programs on opposite sides of a channel-to-channel adapter (CTCA).

channel-to-channel adapter (CTCA). A hardware device that can be used to connect two channels on the same computing system or on different systems.

CICSPlex. (1) The largest set of CICS systems to be monitored and controlled as a single entity. (2) In a CICSPlex SM environment, the user-defined name, description, and configuration information for a CICSPlex. A CICSPlex can be made up of CICS systems or CICS system groups. See also *CICS system* and *CICS system group*.

CICSPlex SM. CICSPlex System Manager.

CICSPlex SM address space (CMAS). A CICSPlex SM component that is responsible for managing a CICSPlex. A CMAS provides the single system image for a CICSPlex by serving as the interface to other CICSPlexes and external programs. There must be at least one CMAS for each OS/390 image on which you are running CICSPlex SM. A single CMAS can manage CICS systems within one or more CICSPlexes. See also *coordinating address space (CAS)* and *managed address space (MAS)*.

CICSPlex System Manager (CICSPlex SM). An IBM CICS systems management product that provides single system image and single point of control for one or more CICSPlexes, including CICSPlexes on heterogeneous operating systems.

classification. The process of assigning a service class and, optionally, a report class to a work request. Subsystems, together with workload management services, use classification rules to assign work to a service class when it enters a sysplex.

classification rules. The rules workload management and subsystems use to assign a service class and, optionally, a report class to a work request. A classification rule consists of one or more of the following work qualifiers: subsystem type, subsystem instance, user ID, accounting information, transaction name, transaction class, source LU, NETID, and LU name.

CMAS. CICSPlex SM address space. See also *CICSPlex SM address space (CMAS)*.

CMAS link. A communications link between one CICSPlex SM address space (CMAS) and another CMAS or a remote managed address space (MAS). CMAS links are defined when CICSPlex SM is configured.

CNC. Mnemonic for an ESCON channel-attached to an ESCON-capable device.

command. An instruction that directs a control unit or device to perform an operation or a set of operations.

commit. In data processing the point at which the data updates are written to the database in a way which is irrevocable.

compatibility mode. A mode of processing in which the SRM parmlib members IEAIPSxx and IEAICSxx determine system resource management. See also *goal mode*.

complementary metal-oxide semiconductor (CMOS). A technology that combines the electrical properties of positive and negative voltage requirements to use considerably less power than other types of semiconductors.

component. (1) Hardware or software that is part of a functional unit. (2) A functional part of an operating system; for example, the scheduler or supervisor.

computer architecture. The organizational structure of a computer system, including hardware and software.

configuration. The arrangement of a computer system or network as defined by the nature, number, and chief characteristics of its functional units. More specifically, the term *configuration* may refer to a hardware configuration or a software configuration. See also *system configuration.*

connectivity. A term used to describe the physical interconnections of multiple devices/computers/networks employing similar or different technology and/or architecture together to accomplish effective communication between and among connected members. It involves data exchange and/or resource sharing.

console. A logical device that is used for communication between the user and the system. See also *service console*.

construct. A collective name for data class, storage class, management class, and storage group.

continuous availability. The elimination or masking of both planned and unplanned outages, so that no system outages are apparent to the end user.

Continuous availability can also be stated as the ability to operate 24 hours/day, 7 days/week, with no outages apparent to the end user.

continuous operations. The elimination or masking of planned outages. A system that delivers continuous operations is a system that has no scheduled outages.

control interval (CI). A fixed-length area of direct access storage in which VSAM creates distributed free space and stores records. Also, in a key-sequenced data set or file, the set of records pointed to by an entry in the sequence-set index record. The control interval is the unit of information that VSAM transmits to or from direct access storage. A control interval always comprises an integral number of physical records.

control region. The OS/390 main storage region that contains the IMS control program.

control unit. A general term for any device that provides common functions for other devices or mechanisms. Synonym for *controller*.

coordinating address space (CAS). An OS/390 subsystem that provides ISPF end-user access to the CICSPlex. There must be at least one CAS for each OS/390 image on which you are running CICSPlex SM. See also *CICSPlex SM address space (CMAS)* and *managed address space (MAS)*.

couple data set. A data set that is created through the XCF couple data set format utility and, depending on its designated type, is shared by some or all of the OS/390 systems in a sysplex. See also *Sysplex couple data set* and *XCF couple data set*.

Coupling Facility (CF). A special LP that provides high-speed caching, list processing, and locking functions in Parallel Sysplex. See also *Coupling Facility channel*, *Coupling Facility white space*, and *coupling services*.

Coupling Facility channel (CF link). A high bandwidth fiber optic channel that provides the high-speed connectivity required for data sharing between a CF and the CPCs directly attached to it.

Coupling Facility Control Code (CFCC). The Licensed Internal Code (LIC) that runs in a CF LP to provide shared storage management functions for a sysplex.

Coupling Facility Data Tables (CFDT). CFDT enables user applications, running in different CICS regions that reside in one or more OS/390 images, within a Parallel Sysplex, to share working data with update integrity.

Coupling Facility white space. CF storage set aside for rebuilding of structures from other CFs, in case of failure.

coupling services. In a sysplex, the functions of XCF that transfer data and status between members of a group residing on one or more OS/390 systems in the sysplex.

CPU service units. A measure of the task control block (TCB) execution time multiplied by an SRM constant that is CPC-model-dependent. See also *service unit*.

CP TOD. In a CPC with more than one CP, each CP can have a separate TOD clock, or more than one CP might share a clock, depending on the model. In all cases each CP has access to a single clock also called a CPC TOD clock

common queue server (CQS). A server that receives, maintains, and distributes data objects from a shared queue on a CF list structure for its clients.

cross-system coupling facility (XCF). XCF is a component of OS/390 that provides functions to support cooperation between authorized programs running within a sysplex.

cross-system extended services (XES). Provides services for OS/390 systems in a sysplex to share data on a CF.

cryptographic. Pertaining to the transformation of data to conceal its meaning.

Customer Information Control System (CICS). An IBM-licensed program that enables transactions entered at remote terminals to be processed concurrently by user-written application programs. It includes facilities for building, using, and maintaining databases.

CVC. Mnemonic for an ESCON channel-attached to a IBM 9034 (ESCON Converter).

D

daemon. A task, process, or thread that intermittently awakens to perform some chores and then goes back to sleep (software).

data entry database (DEDB). A direct-access database that consists of one or more areas, with each area containing both root segments and dependent segments. The database is accessed using VSAM media manager.

Data Facility Hierarchical Storage Manager (DFHSM). An IBM-licensed program used to back up, recover, and manage space on volumes.

Data Language/I (DL/I). The IMS data manipulation language, a common high-level interface between a user application and IMS. DL/I calls are invoked from application programs written in languages such as PL/I, COBOL, VS Pascal, C, and Ada. It can also be invoked from assembler language application programs by subroutine calls. IMS lets the user define data structures, relate structures to the application, load structures, and reorganize structures.

data link. (1) Any physical link, such as a wire or a telephone circuit, that connects one or more remote terminals to a communication control unit, or connects one communication control unit with another. (2) The assembly of parts of two data terminal equipment (DTE) devices that are controlled by a link protocol and the interconnecting data circuit, and that enable data to be transferred from a data source to a data link. (3) In SNA, see also *link*.

Note: A telecommunication line is only the physical medium of transmission. A data link includes the physical medium of transmission, the protocol, and associated devices and programs; it is both physical and logical.

data set. The major unit of data storage and retrieval, consisting of a collection of data in one of several prescribed arrangements and described by control information to which the system has access.

data sharing. In Parallel Sysplex, the ability of concurrent subsystems (such as DB2 or IMS database managers) or application programs to directly access and change the same data while maintaining data integrity. See also *Sysplex data sharing* and *data sharing group*.

data sharing group. A collection of one or more subsystems that directly access and change the same data while maintaining data integrity. See also *DB2 data sharing group* and *IMS DB data sharing group*.

data sharing Parallel Sysplex. A Parallel Sysplex where data is shared at the record level across more than one system, using a CF structure to guarentee cross-system integrity.

database. (1) A set of data, or a part or the whole of another set of data, that consists of at least one file and is sufficient for a given purpose or for a given data-processing system. (2) A collection of data fundamental to a system. See also *Database Control (DBCTL)*, *data entry database (DEDB)*, *data sharing*, and *data sharing group*.

Database Control (DBCTL). An environment allowing full-function databases and DEDBs to be accessed from one or more transaction management subsystems.

DB2 data sharing group. A collection of one or more concurrent DB2 subsystems that directly access and change the same data while maintaining data integrity.

DDF. Distributed Data Facility (DB2). DB2 subsystem running in an address space that supports VTAM communications with other DB2 subsystems and supports the execution of distributed database access requests on behalf of remote users. This provides isolation of remote function execution from local function execution.

DEDB. See *data entry database*.

delay monitoring services. The workload management services that monitor the delays encountered by a work request.

device. (1) A mechanical, electrical, or electronic contrivance with a specific purpose. (2) An input/output unit such as a terminal, display, or printer.

direct access storage device (DASD). A physical device like IBM's 3390 in which data can be permanently stored and subsequently retrieved using licensed products like IMS and DB2, or using IBM-supported access methods like VSAM in operating system environments like OS/390.

directory. A list of files that are stored on a disk or diskette. A directory also contains information about the file, such as size and date of last change.

disaster. An event that renders IT services unavailable for an extended period. Often the IT facilities must be moved to another site in the event of a disaster.

DNS. See *Domain Name System*.

domain name. In the Internet suite of protocols, a name of a host system. A domain name consists of a sequence of subnames separated by a delimiter character. For example, if the fully qualified domain name (FQDN) of host system is ralvm7.vnet.ibm.com, each of the following is a domain name: ralvm7.vnet.ibm.com, vnet.ibm.com, ibm.com.

domain name server. In the Internet suite of protocols, a server program that supplies name-to-address translation by mapping domain names to IP addresses. Synonymous with *name server*.

Domain Name System (DNS). In the Internet suite of protocols, the distributed database system used to map domain names to IP addresses.

dynamic. Pertaining to an operation that occurs at the time it is needed rather than at a predetermined or fixed time. See also *dynamic connection*, *dynamic connectivity*, *dynamic reconfiguration*, *dynamic reconfiguration management*, and *dynamic storage connectivity*.

dynamic CF dispatching. With dynamic CF dispatching, the CF will monitor the request rate that is driving it and adjust its usage of CP resource accordingly. If the request rate becomes high enough, the CF will revert back to its original dispatching algorithm, constantly looking for new work. When the request rate lowers, the CF again becomes more judicious with its use of CP resource. See also *dynamic ICF expansion*.

dynamic connection. In an ESCON director, a connection between two ports, established or removed by the ESCD and that, when active, appears as one continuous link. The duration of the connection depends on the protocol defined for the frames transmitted through the ports and on the state of the ports.

dynamic connectivity. In an ESCON director, the capability that allows connections to be established and removed at any time.

dynamic ICF expansion. Dynamic ICF expansion provides the ability for a CF LP that is using a dedicated ICF to expand into the pool of shared ICFs or shared CPs. At low request rates, the resource consumption of the shared PU should be 1% to 2%. As the request rate increases, the resource consumption will increase, up to the point where the LP will consume its full share of the shared PU as defined by the LPAR weights. See also *dynamic CF dispatching*.

dynamic reconfiguration. Pertaining to a processor reconfiguration between a single-image (SI) configuration and a physically partitioned (PP) configuration when the system control program is active.

dynamic reconfiguration management. In OS/390, the ability to modify the I/O configuration definition without needing to perform a power-on reset (POR) of the hardware or an initial program load (IPL).

dynamic storage reconfiguration. A PR/SM LPAR function that allows central or expanded storage to be added or removed from a logical partition without disrupting the system control program operating in the logical partition.

E

ECS. Enhanced Catalog Sharing (DFSMS/MVS V1.5)

EMIF. Enhanced multiple image facility (formerly ESCON multiple image facility). A facility which allows the sharing of FICON or ESCON channels between LPs.

emitter. In fiber optics, the source of optical power.

end node. A type 2.1 node that does not provide any intermediate routing or session services to any other node. For example, APPC/PC is an end node.

enhanced catalog sharing. By using a CF cache structure instead of DASD to store catalog sharing control information, shared catalog performance in sysplex environment is improved. This sharing method, called enhanced catalog sharing (ECS), eliminates a reserve, dequeue, and I/O request to the VVDS on most catalog calls.

Enhanced Multiple Image Facility (EMIF). See EMIF.

enhanced sysplex. An enhanced sysplex is a sysplex with one or more CFs. See also *base sysplex* and *sysplex*.

enterprise. A business or organization that consists of two or more sites separated by a public right-of-way or a geographical distance.

Enterprise Systems Connection (ESCON). A set of products and services that provides a dynamically connected environment using optical cables as a transmission medium. See also *ESCD*, *ESCM*, and *ESCON channel*.

Environmental Services Subsystem (ESSS). A component of CICSPlex SM that owns all the data spaces used by the product in an OS/390 image. The ESSS executes at initialization and remains in the OS/390 image for the life of the IPL to ensure that the data spaces can survive the loss of a CICSPlex SM address space (CMAS).

ESA/390. Enterprise Systems Architecture/390.

ESCD. Enterprise Systems Connection (ESCON) Director. See also *ESCD console*, *ESCD console adapter*, and *ESCM*.

ESCD console. The ESCON director input/output device used to perform operator and service tasks at the ESCD.

ESCD console adapter. Hardware in the ESCON director console that provides the attachment capability between the ESCD and the ESCD console.

ESCM. See *ESCON Manager*.

ESCON channel. A channel having an Enterprise Systems Connection channel-to-control-unit I/O interface that uses optical cables as a transmission medium. Contrast with *parallel channel*.

ESCON director (ESCD). A device that provides connectivity capability and control for attaching any two links to each other.

ESCON Extended Distance Feature (ESCON XDF). An ESCON feature that uses laser/single-mode fiber optic technology to extend unrepeated link distances up to 20 km. LPs in an ESCON environment.

ESCON Manager (ESCM). A licensed program that provides S/390 CPC control and intersystem communication capability for ESCON director connectivity operations.

ESCON multiple image facility (EMIF). A facility that allows channels to be shared among PR/SM logical partitions in an ESCON environment.

ESCON XDF. ESCON extended distance feature.

Ethernet. A local area network that was originally marketed by Xero Corp. The name is a trademark of Xerox Corp.

ETR. See *External Time Reference*.

ETR offset. The time zone offset identifies your system location within a network of other systems. This offset is the difference between your local time and Universal Time Coordinate (UTC). See also *Universal Time Coordinate*.

ETS. See *External Time Source*.

exclusive lock. A lock that prevents concurrently executing application processes from reading or changing data. Contrast with *shared lock.*

expanded storage. (1) Optional integrated high-speed storage that transfers 4KB pages to and from central storage. (2) Additional (optional) storage that is addressable by the system control program.

extended Parallel Sysplex. This name is sometimes used to refer to Parallel Sysplexes that exploit data sharing and future enhancements for ultra high availability and disaster recovery.

Extended Recovery Facility (XRF). Software designed to minimize the effect of failures in OS/390, VTAM, the S/390 CPC CP, or IMS/VS on sessions between IMS/VS and designated terminals. It provides an alternate subsystem to take over failing sessions.

External Time Reference (ETR). This is how OS/390 documentation refers to the 9037 Sysplex Timer. An ETR consists of one or two 9037s and their associated consoles.

External Time Source (ETS). An accurate time source used to set the time in the Sysplex Timer. The accurate time can be obtained by dialing time services or attaching to radio receivers or time code generators.

F

false lock contention. A contention indication from the CF when multiple lock names are hashed to the same indicator and when there is no real contention.

Fast Path. IMS functions for applications that require good response characteristics and that may have large transaction volumes. Programs have rapid access to main-storage databases (to the field level), and to direct-access data entry databases. Message processing is grouped for load balancing and synchronized for database integrity and recovery. See also *MSDB* and *DEDB.*

Fast Path databases. Two types of databases designed to provide high data availability and fast processing for IMS applications. They can be processed by the following types of programs: MPPs, BMPs, and IFPs. See also *main storage database* and *data entry database.*

feature. A part of an IBM product that can be ordered separately by the customer.

FICON channel. FIbre CONnection. A S/390 channel which uses industry standard Fibre Channel Standard (FCS) as a base.

file system. The collection of files and file management structures on a physical or logical mass storage device such as a disk.

format. (1) A specified arrangement of things, such as characters, fields, and lines, usually used for displays, printouts, or files. (2) To arrange things such as characters, fields, and lines.

forward recovery. Reconstructing a file or database by applying changes to an older version (backup or image copy) with data recorded in a log data set. The sequence of changes to the restored copy is in the same order in which they were originally made.

frame. For an S/390 microprocessor cluster, a frame may contain one or more CPCs, support elements, and AC power distribution.

frequency. The rate of signal oscillation, expressed in hertz (cycles per second).

full function databases. Hierarchic databases that are accessed through Data Language I (DL/1) call language and can be processed by all four types of application programs: IFP, MPPs, BMPs, and batch. Full function databases include HDAM, HIDAM, HSAM, HISAM, SHSAM, and SHISAM.

G

generic resource name. A name used by VTAM that represents several application programs that provide the same function in order to handle session distribution and balancing in a sysplex.

gigabytes. One billion (10^9) bytes.

global locking (DB2). For data consistency in a data sharing environment, locks must be known and respected between all members. DB2 data sharing uses global locks ensure that each member is aware of all members' locks.

Two locking mechanisms are used by DB2 data sharing to ensure data consistency, logical locks and physical locks.

The two types can be briefly compared as follows:

1. Logical locks

   Logical locks are used to control concurrent access from application processes, such as transactions or batch programs.

2. Physical locks

   Physical locks are used by DB2 members to control physical resourses
   - Page set physical locks are used to track the level of interest in a particular page set or partition and thus determine the nee GBP coherency controls.
   - Page physical locks are used to preserve the physical consistency of pages.

See also *P-lock*.

global resource serialization (GRS). A component of OS/390 used for sharing system resources and for converting DASD reserve volumes to data set ENQueues.

global resource serialization complex (GRSplex). One or more OS/390 systems that use global resource serialization to serialize access to shared resources (such as data sets on shared DASD volumes).

GMT. See *Greenwich Mean Time*.

goal mode. A mode of processing where the active service policy determines system resource management. See also *compatibility mode*.

Greenwich Mean Time (GMT). Time at the time zone centered around Greenwich, England.

group buffer pool. A CF cache structure used by a DB2 data sharing group to cache data and to ensure that the data is consistent for all members. See also *buffer pool*.

group services. Services for establishing connectivity among the multiple instances of a program, application, or subsystem (members of a group running on OS/390) in a sysplex. Group services allow members of the group to coordinate and monitor their status across the systems of a sysplex.

H

Hardware Management Console. A console used to monitor and control hardware such as the 9672 CPCs.

hardware system area (HSA). A logical area of central storage, not addressable by application programs, used to store Licensed Internal Code and control information.

highly parallel. Refers to multiple systems operating in parallel, each of which can have multiple processors. See also *n-way*.

high-speed buffer. A cache or a set of logically partitioned blocks that provides significantly faster access to instructions and data than that provided by central storage.

HiPerLink. A HiPerLink provides improved CF link efficiency and response times in processing CF requests, compared to previous CF link configurations. With HiPerLinks, current data sharing overheads are reduced and CF link capacity is improved.

host (computer). (1) In a computer network, a computer that provides end users with services such as computation and databases and that usually performs network control functions. (2) The primary or controlling computer in a multiple-computer installation.

HSA. See *hardware system area*.

I

IBF. See *Internal Battery Feature*.

IC. See *Internal Coupling Link*.

ICB. See *Integrated Cluster Bus*.

ICF. See *Internal Coupling Facility*.

ICF. Integrated Catalog Facility.

importance level. An attribute of a service class goal that indicates the importance of meeting the goal relative to other service class goals, in five levels: lowest, low, medium, high, and highest.

IMS DB data sharing group. A collection of one or more concurrent IMS DB subsystems that directly access and change the same data while maintaining data integrity. The components in an IMS DB data sharing group include the sharing IMS subsystems, the IRLMs they use, the IRLM, OSAM, and VSAM structures in the CF, and a single set of DBRC RECONs.

IMS system log. A single log made up of online data sets (OLDSs) and write-ahead data sets (WADSs).

in-doubt period. The period during which a unit of work is pending during commit processing that involves two or more subsystems. See also *in-doubt work unit*.

in-doubt work unit. In CICS/ESA and IMS/ESA, a piece of work that is pending during commit processing; if commit processing fails between the polling of subsystems and the decision to execute the commit, recovery processing must resolve the status of any work unit that is in doubt.

indirect CMAS. A CICSPlex SM address space (CMAS) that the local CMAS can communicate with through an adjacent CMAS. There is no direct CMAS-to-CMAS link between the local CMAS and an indirect CMAS. Contrast with *adjacent CMAS*. See also *local CMAS*.

initial microcode load (IML). The action of loading the operational microcode.

initial program load (IPL). The initialization procedure that causes an operating system to start operation.

input/output support processor (IOSP). The hardware unit that provides I/O support functions for the primary support processor (PSP). It also provides maintenance support function for the processor controller element (PCE).

installed service definition. The service definition residing in the couple data set for WLM. The installed service definition contains the active service policy information.

interactive. Pertaining to a program or system that alternately accepts input and then responds. An interactive system is conversational; that is, a continuous dialog exists between user and system. Contrast with *batch*.

interface. A shared boundary. An interface might be a hardware component to link two devices or it might be a portion of storage or registers accessed by two or more computer programs.

Integrated Cluster Bus channel (ICB). The Integrated Cluster Bus channel uses the Self Timed Interface to perform the S/390 coupling communication. The cost of coupling is reduced by using a higher performing (Approximately 280 MB/sec) but less complex transport link suitable for the relatively short distances (The cable is 10 meters; the distance between CPCs is approximately 7 meters).

Integrated Offload Processor (IOP). The processor in the interconnect communication element that detects, initializes, and ends all channel subsystem operations.

Integrated Coupling Migration Facility (ICMF). A PR/SM LPAR facility that emulates CF links for LPs (CF LPs and OS/390 LPs) running on the same CPC to assist in the test and development of data sharing applications.

internal battery feature (IBF). The internal battery feature (IBF) provides the function of a local uninterruptible power source (UPS). This feature may increase power line disturbance immunity for S/390 CPCs.

Internal Coupling channel (IC). The Internal Coupling channel emulates the coupling facility functions in microcode between images within a single CPC. It is a high performance channel transferring data at up to 6Gb/sec. Internal Coupling implementation is a totally logical channel requiring no channel or even cable hardware. However, a CHPID number must be defined in the IOCDS. A replacement for ICMF.

Internal Coupling Facility (ICF). The Internal Coupling Facility (ICF) uses up to two spare PUs on selected S/390 CPCs. The ICF may use CF links or emulated links (ICMF). It can be used initially as an entry configuration into Parallel Sysplex and then maintained as a backup configuration in the future.

interrupt. (1) A suspension of a process, such as execution of a computer program caused by an external event, and performed in such a way that the process can be resumed. (2) To stop a process in such a way that it can be resumed. (3) In data communication, to take an action at a receiving station that causes the sending station to end a transmission. (4) To temporarily stop a process.

invalidation. The process of removing records from cache because of a change in status of a subsystem facility or function, or because of an error while processing the cache image of the set of records. When such a cache image is invalidated, the corresponding records cannot be accessed in cache and the assigned cache space is available for allocation.

IOCDS. I/O configuration data set.

IOCP. I/O configuration program.

I/O service units. A measure of individual data set I/O activity and JES spool reads and writes for all data sets associated with an address space.

J

JES (Job Entry Subsystem). A system facility for spooling, job queuing, and managing I/O.

jumper cable. In an ESCON environment, an optical cable, having two conductors, that provides physical attachment between two devices or between a device and a distribution panel. Contrast with *trunk cable*.

L

latency. The time interval between the instant at which an instruction control unit initiates a call for data and the instant at which the actual transfer of data starts.

leap second. Corrections of exactly one second inserted into the UTC time scale since January 1, 1972. This adjustment occurs at the end of a UTC month, normally on June 30 or December 31. Seconds are occasionally added to or subtracted from the UTC to compensate for the wandering of the earth's polar axis and maintain agreement with the length of the solar day. See also *Universal Time Coordinate(UTC)*.

LIC. See *Licensed Internal Code*.

Licensed Internal Code (LIC). Software provided for use on specific IBM machines and licensed to customers under the terms of IBM's Customer Agreement. Microcode can be Licensed Internal Code and licensed as such.

link. The combination of physical media, protocols, and programming that connects devices.

list structure. A CF structure that enables multisystem applications in a sysplex to share information organized as a set of lists or queues. A list structure consists of a set of lists and an optional lock table, which can be used for serializing resources in the list structure. Each list consists of a queue of list entries.

list structure services. OS/390 services that enable multisystem applications in a sysplex to perform operations such as the following on a CF list structure:

- Read, update, create, delete, and move list entries in a list structure.
- Perform serialized updates on multiple list entries in a list structure.
- Monitor lists in a list structure for transitions from empty to non-empty.

local cache. A buffer in local system storage that might contain copies of data entries in a CF cache structure.

local CMAS. The CICSPlex SM address space (CMAS) that a user identifies as the current context when performing CMAS configuration and management tasks.

local MAS. A managed address space (MAS) that resides in the same OS/390 image as the CICSPlex SM address space (CMAS) that controls it and that uses the Environmental Services Subsystem (ESSS) to communicate with the CMAS.

lock resource. Data accessed through a CF structure.

lock structure. A CF structure that enables applications in a sysplex to implement customized locking protocols for serialization of application-defined resources. The lock structure supports shared, exclusive, and application-defined lock states, as well as generalized contention management and recovery protocols. See also *exclusive lock, shared lock,* and *false lock contention*.

lock structure services. OS/390 services that enable applications in a sysplex to perform operations such as the following on a CF lock structure:

- Request ownership of a lock.
- Change the type of ownership for a lock.
- Release ownership of a lock.
- Manage contention for a lock.
- Recover a lock held by a failed application.

logical connection. In a network, devices that can communicate or work with one another because they share the same protocol.

logical control unit. A group of contiguous words in the HSA that provides all of the information necessary to control I/O operations through a group of paths that are defined in the IOCDS. Logical control units represent to the channel subsystem a set of control units that attach common I/O devices.

logical partition (LP). In LPAR mode, a subset of the processor unit resources that is defined to support the operation of a system control program (SCP). See also *logically partitioned (LPAR) mode*.

logical unit (LU). In *VTAM*, the source and recipient of data transmissions. Data is transmitted from one logical unit (LU) to another LU. For example, a terminal can be an LU, or a CICS or IMS system can be an LU.

logically partitioned (LPAR) mode. A CPC power-on reset mode that enables use of the PR/SM feature and allows an operator to allocate CPC hardware resources (including CPs, central storage, expanded storage, and channel paths) among logical partitions. Contrast with *basic mode*.

loosely coupled. A multisystem structure that requires a low degree of interaction and cooperation between multiple OS/390 images to process a workload. See also *tightly coupled*.

LP. See *logical partition*.

LPAR. See *logically partitioned (LPAR) mode*.

LU. See *logical unit.*

M

m-image. The number (m) of OS/390 images in a sysplex. See also *n-way*.

main storage. A logical entity that represents the program addressable portion of central storage. All user programs are executed in main storage. See also *central storage.*

main storage database (MSDB). A root-segment database, residing in main storage, which can be accessed to a field level.

mainframe (S/390 CPC). A large computer, in particular one to which other computers can be connected so that they can share facilities the S/390 CPC provides; for example, an S/390 computing system to which personal computers are attached so that they can upload and download programs and data.

maintenance point. A CICSPlex SM address space (CMAS) that is responsible for maintaining CICSPlex SM definitions in its data repository and distributing them to other CMASs involved in the management of a CICSPlex.

managed address space (MAS). A CICS system that is being managed by CICSPlex SM. See also *local MAS* and *remote MAS*.

MAS. Managed address space.

MAS agent. A CICSPlex SM component that acts within a CICS system to provide monitoring and data collection for the CICSPlex SM address space (CMAS). The level of service provided by a MAS agent depends on the level of CICS the system is running under and whether it is a local or remote MAS. See also *CICSPlex SM address space (CMAS), local MAS,* and *remote MAS*.

massively parallel. Refers to thousands of processors in a parallel arrangement.

mega-microsecond. A carry out of bit 32 of the TOD clock occurs every 2^{20} microseconds (1.048576 seconds). This interval is sometimes called a "mega-microsecond" (Mµs). This carry signal is used to start one clock in synchronism with another, as part of the process of setting the clocks. See also *time-of-day clock*.

member. A specific function (one or more modules or routines) of a multisystem application that is defined to XCF and assigned to a group by the multisystem application. A member resides on one system in the sysplex and can use XCF services to communicate (send and receive data) with other members of the same group. See *XCF group*, and *multisystem application*.

memory. Program-addressable storage from which instructions and other data can be loaded directly into registers for subsequent execution or processing. Synonymous with *main storage*.

microcode. (1) One or more microinstructions. (2) A code, representing the instructions of an instruction set, that is implemented in a part of storage that is not program-addressable. (3) To design, write, and test one or more microinstructions.

microprocessor. A processor implemented on one or a small number of chips.

migration. Installing a new version or release of a program when an earlier version or release is already in place.

mixed complex. A global resource serialization complex in which one or more of the systems in the global resource serialization complex are not part of a multisystem sysplex.

monitoring environment. A record of execution delay information about work requests kept by the workload management services. A monitoring environment is made up of one or more performance blocks. See also *performance block*.

monoplex. A one system sysplex with sysplex couple data sets that XCF prevents any other system from joining. See also *multisystem sysplex*.

MP. Multiprocessor.

MSDB. See *main storage database.*

MSU. Millions of Service Units. The unit used in IBM PSLC pricing as an estimate of CPC capacity. It is used in this book as an estimate of CPC capacity for CPC and 9674 capacity planning purposes.

multifiber cable. An optical cable that contains two or more fibers. See also *jumper cable, optical cable assembly,* and *trunk cable*.

multimode optical fiber. A graded-index or step-index optical fiber that allows more than one bound mode to propagate. Contrast with *single-mode optical fiber*.

Multi-Node Persistent Session (MNPS). MNPS extends persistent sessions capability across multiple CPCs connected through the CF. MNPS provides for the recovery of VTAM, OS/390, hardware or application failures by restarting the application on another host in the Parallel Sysplex without requiring users to re-logon.

Multiple Systems Coupling (MSC). An IMS facility that permits multiple IMS subsystems to communicate with each other.

multiprocessing. The simultaneous execution of two or more computer programs or sequences of instructions. See also *parallel processing*.

multiprocessor (MP). A CPC that can be physically partitioned to form two operating processor complexes.

multisystem application. An application program that has various functions distributed across OS/390 images in a multisystem environment.

Examples of multisystem applications are:

- CICS
- Global resource serialization (GRS)
- Resource Measurement Facility (RMF)
- OS/390 Security Server (RACF)
- Workload manager (WLM)

See *XCF group*.

multisystem environment. An environment in which two or more OS/390 images reside in one or more processors, and programs on one image can communicate with programs on the other images.

multisystem sysplex. A sysplex in which two or more OS/390 images are allowed to be initialized as part of the sysplex. See also *single-system sysplex*.

Mµs. See *mega-microsecond*.

N

named counter server (CICS). CICS provides a facility for generating unique sequence numbers for use by applications in a Parallel Sysplex environment (for example, to allocate a unique number for orders or invoices). This facility is provided by a named counter server, which maintains each sequence of numbers as a named counter. Each time a sequence number is assigned, the corresponding named counter is incremented automatically so that the next request gets the next number in sequence. This facility uses a CF list structure to hold the information.

NCP. (1) Network Control Program (IBM-licensed program). Its full name is Advanced Communications Function for the Network Control Program. Synonymous with *ACF/NCP*. (2) Network control program (general term).

network. A configuration of data processing devices and software connected for information interchange. See also *network architecture* and *network control program (NCP)*.

network architecture. The logical structure and operating principles of a computer network.

node. (1) In SNA, an endpoint of a link or junction common to two or more links in a network. Nodes can be distributed to S/390 CPC CPs, communication controllers, cluster controllers, or terminals. Nodes can vary in routing and other functional capabilities.

n-way. The number (n) of CPs in a CPC. For example, a 6-way CPC contains six CPs.

O

OLDS. *See* online log data set.

online log data set (OLDS). A data set on direct access storage that contains the log records written by an online IMS system.

open system. (1) A system with specified standards and that therefore can be readily connected to other systems that comply with the same standards. (2) A data communications system that conforms to the standards and protocols defined by open systems interconnection (OSI). Synonym for *node*.

Operational Single Image. Multiple operating system images being managed as a single entity. This may be a basic sysplex, standard Parallel Sysplex or extended Parallel Sysplex.

optical cable. A fiber, multiple fibers, or a fiber bundle in a structure built to meet optical, mechanical, and environmental specifications. See also *jumper cable* and *trunk cable*.

optical receiver. Hardware that converts an optical signal to an electrical logic signal. Contrast with *optical transmitter*.

optical repeater. In an optical fiber communication system, an opto-electronic device or module that receives a signal, amplifies it (or, for a digital signal, reshapes, retimes, or otherwise reconstructs it) and retransmits it.

optical transmitter. Hardware that converts an electrical logic signal to an optical signal. Contrast with *optical receiver*.

OS/390 image. A single occurrence of the OS/390 operating system that has the ability to process work.

OS/390 system. An OS/390 image together with its associated hardware, which collectively are often referred to simply as a system, or OS/390 system.

P

P-lock. There are times when a P-lock must be obtained on a page to preserve physical consistency of the data between members. These locks are known as page P-locks. Page P-locks are used, for example, when two subsystems attempt to update the same page of data and row locking is in effect. They are also used for GBP-dependent space map pages and GBP-dependent leaf pages for type 2 indexes, regardless of locking level. IRLM P-locks apply to both DB2 and IMS DB data sharing.

Page set P-locks are used to track inter-DB2 read-write interest, thereby determining when a page set has to become GBP-dependent. When access is required to a page set or partition through a member in a data sharing group, a page set P-lock is taken. This lock is always propagated to the lock table on the CF and is owned by the member. No matter how many times the resource is accessed through the member, there will always be only one page set P-lock for that resource for a particular member. This lock will have different modes depending on the level (read or write) of interest the member has in the resource. See also *global locking*.

parallel. (1) Pertaining to a process in which all events occur within the same interval of time, each handled by a separate but similar functional unit; for example, the parallel transmission of the bits of a computer word along the lines of an internal bus. (2) Pertaining to the concurrent or simultaneous operation of two or more devices or to concurrent performance of two or more activities in a single device. (3) Pertaining to the concurrent or simultaneous occurrence of two or more related activities in multiple devices or channels. (4) Pertaining to the simultaneity of two or more processes. (5) Pertaining to the simultaneous processing of the individual parts of a whole, such as the bits of a character and the characters of a word, using separate facilities for the various parts. (6) Contrast with *serial*.

parallel processing. The simultaneous processing of units of work by many servers. The units of work can be either transactions or subdivisions of large units of work (batch).

Parallel Sysplex. A Parallel Sysplex is a sysplex with one or more CFs. See also *base sysplex*, *sysplex*, *extended Parallel Sysplex*, and *standard Parallel Sysplex*.

partition. An area of storage on a fixed disk that contains a particular operating system or logical drives where data and programs can be stored.

partitionable CPC. A CPC can be divided into two independent CPCs. See also *physical partition, single-image mode, MP*, and *side*.

partitioned data set (PDS). A data set in DASD storage that is divided into partitions, called *members*, each of which can contain a program, part of a program, or data.

performance. For a storage subsystem, a measurement of effective data processing speed against the amount of resource that is consumed by a complex. Performance is largely determined by throughput, response time, and system availability. See also *performance administration*, *performance block*, *performance management*, and *performance period*.

performance administration. The process of defining and adjusting workload management goals and resource groups based on installation business objectives.

performance block. A piece of storage containing workload management's record of execution delay information about work requests.

performance management. The process workload management uses to decide how to match resources to work according to performance goals and processing capacity.

performance period. A service goal and importance level assigned to a service class for a specific duration. You define performance periods for work that has variable resource requirements.

persistent connection. A connection to a CF structure with a connection disposition of KEEP. OS/390 maintains information about the connection so that when the connection terminates abnormally from a CF structure, OS/390 places the connection in a failed-persistent state, and the connection can attempt to reconnect to the structure.

persistent session. (1) In the NetView program, a network management session that remains active even though there is no activity on the session for a specified period of time. (2) An LU-LU session that VTAM retains after the failure of a VTAM application program. Following the application program's recovery, the application program either restores or terminates the session.

persistent structure. A structure allocated in the CF with a structure disposition of KEEP. A persistent structure keeps its data intact across system or sysplex outages, regardless of whether any users are connected to the structure.

physical partition. Part of a CPC that operates as a CPC in its own right, with its own copy of the operating system.

physically partitioned (PP) configuration. A system configuration that allows the processor controller to use both CPC sides as individual CPCs. The A-side of the processor controls side 0, and the B-side controls side 1. Contrast with *single-image (SI) mode*.

policy. A set of installation-defined rules for managing sysplex resources. The XCF PR/SM policy and sysplex failure management policy are examples of policies.

power-on reset. The state of the machine after a logical power-on before the control program is IPLed.

preference list. An installation list of CFs, in priority order, that indicates where OS/390 is to allocate a structure.

processing unit (PU). The part of the system that does the processing, and contains processor storage. On a 9672 CPC the PU may be assigned as either a CP, SAP, ICF or act as a spare PU.

processor. A processing unit, capable of executing instructions when combined with main storage and channels. See also *processor complex*, *processor controller*, and *processor controller element (PCE* and *CPC)*.

processor complex. A physical collection of hardware that includes main storage, one or more processors, and channels.

processor controller. Hardware that provides support and diagnostic functions for the CPs.

processor controller element (PCE). Hardware that provides support and diagnostic functions for the processor unit. The processor controller communicates with the processor unit through the logic service adapter and the logic support stations, and with the power supplies through the power thermal controller. It includes the primary support processor (PSP), the initial power controller (IPC), the input/output support processor (IOSP), and the control panel assembly.

Processor Resource/Systems Manager (PR/SM). A function that allows the processor unit to operate several system control programs simultaneously in LPAR mode. It provides for logical partitioning of the real machine and support of multiple preferred guests. See also *LPAR.*

program specification block (PSB). The control block in IMS that describes databases and logical message destinations used by an application program.

PR/SM. See *Processor Resource/Systems Manager*.

PSB. See *program specification block.*

public network. A communication common carrier network that provides data communication services over switched, non-switched, or packet-switching lines.

R

RAS. Reliability, availability, and serviceability.

receiver. In fiber optics, see *optical receiver.*

reconfiguration. (1) A change made to a given configuration in a computer system; for example, isolating and bypassing a defective functional unit or connecting two functional units by an alternative path. Reconfiguration is effected automatically or manually and can be used to maintain system integrity.
(2) The process of placing a processor unit, main storage, and channels offline for maintenance, and adding or removing components.

Record Level Sharing (RLS). RLS is an access mode for VSAM data sets supported by DFSMS 1.3 and later releases. RLS enables VSAM data to be shared, with full update capability, between many applications running in many CICS regions across the Parallel Sysplex.

recovery. To maintain or regain system operation after a failure occurs. Generally, to recover from a failure is to identify the failed hardware, to de-configure the failed hardware, and to continue or restart processing.

recovery control (RECON) data sets. Data sets in which Database Recovery Control stores information about logging activity and events that might affect the recovery of databases.

relative processor power (RPP). A unit used to express processor capacity. RPP is a measured average of well defined workload profiles ITR-ratios. ITR (Internal Throughput Rate) is measured in transactions/CPU second. LSPR (Large Systems Performance Reference) measurements predict RPP values for processors running certain releases of operating systems.

remote MAS. A managed address space (MAS) that uses MRO or LU6.2 to communicate with the CICSPlex SM address space (CMAS) that controls it. A remote MAS may or may not reside in the same OS/390 image as the CMAS that controls it.

remote operations. The ability to perform operations tasks from a remote location.

remote site recovery. The ability to continue or resume processing of the critical workload from a remote site.

report class. A group of work for which reporting information is collected separately. For example, you can have a WLM report class for information combining two different service classes, or a report class for information on a single transaction.

request. A service primitive issued by a service user to call a function supported by the service provider.

request for price quotation (RPQ). A custom feature for a product.

resource group. An amount of processing capacity across one or more OS/390 images, assigned to one or more WLM service classes.

Resource Sharing. S/390 Resource Sharing provides the following functionality:

- XCF Signalling - providing multisystem signaling with reduced cost/management
- GRS Star - multisystem resource serialization for increased performance, recoverability and scalability
- JES Checkpointing - multisystem checkpointing for increased simplicity and reduced cost
- Shared Tape - multisystem tape sharing for reduced duplication cost
- Merged Operations Log - multisystem log for single system image/management
- Merged LOGREC - multisystem log for single system image/management

- Shared Catalog - multisystem shared master catalogs/user catalogs for increased performance/simplicity and reduced cost

response time. The amount of time it takes after a user presses the enter key at the terminal until the reply appears at the terminal.

routing. The assignment of the path by which a message will reach its destination.

RPP. See *relative processor power*.

RPQ. See *request for price quotation*.

S

secondary host promotion. Secondary host promotion allows one DFSMShsm system to automatically assume the unique functions of another DFSMShsm system that has failed.

selector channel. An I/O channel that operates with only one I/O device at a time. Once the I/O device is selected, a complete record is transferred one byte at a time. Contrast with *block multiplexer channel.*

serial. (1) Pertaining to a process in which all events occur one after the other; for example, serial transmission of the bits of a character according to V24 CCITT protocol. (2) Pertaining to the sequential or consecutive occurrence of two or more related activities in a single device or channel. (3) Pertaining to the sequential processing of the individual parts of a whole, such as the bits of a character or the characters of a word, using the same facilities for successive parts. (4) Contrast with *parallel*.

serialized list structure. A CF list structure with a lock table containing an array of exclusive locks whose purpose and scope are application-defined. Applications can use the lock table to serialize on parts of the list structure, or resources outside the list structure.

server. A device, program, or code module on for example a network dedicated to a specific function.

server address space. Any address space that helps process work requests.

service administration application. The online ISPF application used by the service administrator to specify the workload management service definition.

service class. A subset of a workload having the same service goals or performance objectives, resource requirements, or availability requirements. For workload management, you assign a service goal and optionally a resource group to a service class.

service console. A logical device used by service representatives to maintain the processor unit and to isolate failing field replaceable units. The service console can be assigned to any of the physical displays attached to the input/output support processor.

service definition. An explicit definition of the workloads and processing capacity in an installation. A service definition includes workloads, service classes, systems, resource groups, service policies, and classification rules.

service definition coefficient. A value that specifies which type of resource consumption should be emphasized in the calculation of service rate. The types of resource consumption are CPU, IOC, MSO, and SRB.

service policy. A named set of performance goals and, optionally, processing capacity boundaries that workload management uses as a guideline to match resources to work. See also *active service policy*.

service request block (SRB) service units. A measure of the SRB execution time for both local and global SRBs, multiplied by an SRM constant that is CPU model dependent.

service unit. The amount of service consumed by a work request as calculated by service definition coefficients and CPU, SRB, I/O, and storage service units.

session. (1) A connection between two application programs that allows them to communicate. (2) In SNA, a logical connection between two network addressable units that can be activated, tailored to provide various protocols, and deactivated as requested. (3) The data transport connection resulting from a call or link between two devices. (4) The period of time during which a user of a node can communicate with an interactive system; usually it is the elapsed time between logon and logoff. (5) In network architecture, an association of facilities necessary for establishing, maintaining, and releasing connections for communication between stations.

shared. Pertaining to the availability of a resource to more than one use at the same time.

shared lock. A lock that prevents concurrently executing application processes from changing, but not from reading, data. Contrast with *exclusive lock.*

side. A part of a partitionable PC that can run as a physical partition and is typically referred to as the A-side or the B-side.

single-image (SI) mode. A mode of operation for a multiprocessor (MP) system that allows it to function as one CPC. By definition, a uniprocessor (UP) operates in single-image mode. Contrast with *physically partitioned (PP) configuration*.

single-mode optical fiber. An optical fiber in which only the lowest-order bound mode (which can consist of a pair of orthogonally polarized fields) can propagate at the wavelength of interest. Contrast with *multimode optical fiber*.

single-OS/390 environment. An environment that supports one OS/390-image. See also *OS/390 image*.

single point of control. The characteristic a sysplex displays when you can accomplish a given set of tasks from a single workstation, even if you need multiple IBM and vendor products to accomplish that particular set of tasks.

single point of failure. An essential resource for which there is no backup.

single GRSplex serialization. Single GRSplex serialization allows several HSMplexes, within a single GRSplex, to operate without interfering with any other HSMplex.

single-system image. The characteristic a product displays when multiple images of the product can be viewed and managed as one image.

single-system sysplex. A sysplex in which only one OS/390 system is allowed to be initialized as part of the sysplex. In a single-system sysplex, XCF provides XCF services on the system but does not provide signalling services between OS/390 systems. See also *multisystem complex* and *XCF-local mode.*

SMS communication data set (COMMDS). The primary means of communication among systems governed by a single SMS configuration. The SMS communication data set (COMMDS) is a VSAM linear data set that contains the current utilization statistics for each system-managed volume. SMS uses these statistics to help balance space usage among systems.

SMS configuration. The SMS definitions and routines that the SMS subsystem uses to manage storage.

SMS system group. All systems in a sysplex that share the same SMS configuration and communications data sets, minus any systems in the sysplex that are defined individually in the SMS configuration.

SSP. See *system support programs*.

standard. Something established by authority, custom, or general consent as a model or example.

standard Parallel Sysplex. A non-data sharing Parallel Sysplex.

STI. Self-Timed Interconnect.

storage. A unit into which recorded data can be entered, in which it can be retained and processed, and from which it can be retrieved.

storage management subsystem (SMS). An operating environment that helps automate and centralize the management of storage. To manage storage, SMS provides the storage administrator with control over data class, storage class, management class, storage group, and ACS routine definitions.

structure. A construct used to map and manage storage in a CF. See *cache structure, list structure*, and *lock structure.*

subarea. A portion of the SNA network consisting of a subarea node, any attached peripheral nodes, and their associated resources. Within a subarea node, all network addressable units, links, and adjacent link stations (in attached peripheral or subarea nodes) that are addressable within the subarea share a common subarea address and have distinct element addresses.

subarea node. In SNA, a node that uses network addresses for routing and whose routing tables are therefore affected by changes in the configuration of the network. Subarea nodes can provide gateway function, and boundary function support for peripheral nodes. Type 4 and type 5 nodes are subarea nodes.

subsystem. A secondary or subordinate system, or programming support, that is usually capable of operating independently of or asynchronously with a controlling system.

support element. A hardware unit that provides communications, monitoring, and diagnostic functions to a central processor complex (CPC).

symmetry. The characteristic of a sysplex where all systems, or certain subsets of the systems, have the same hardware and software configurations and share the same resources.

synchronous. (1) Pertaining to two or more processes that depend on the occurrences of a specific event such as common timing signal.
(2) Occurring with a regular or predictable timing relationship.

sysplex. A set of OS/390 systems communicating and cooperating with each other through certain multisystem hardware components and software services to process customer workloads. There is a distinction between a base sysplex and a Parallel Sysplex. See also *OS/390 system, base sysplex, enhanced sysplex*, and *Parallel Sysplex*.

sysplex couple data set. A couple data set that contains sysplex-wide data about systems, groups, and members that use XCF services. All OS/390 systems in a sysplex must have connectivity to the sysplex couple data set. See also *couple data set*.

sysplex data sharing. The ability of multiple IMS subsystems to share data across multiple OS/390 images. Sysplex data sharing differs from two-way data sharing in that the latter allows sharing across only two OS/390 images.

sysplex failure management. The OS/390 function that minimizes operator intervention after a failure occurs in the sysplex. The function uses installation-defined policies to ensure continued operations of work defined as most important to the installation.

sysplex management. The functions of XCF that control the initialization, customization, operation, and tuning of OS/390 systems in a sysplex.

sysplex partitioning. The act of removing one or more systems from a sysplex.

sysplex query parallelism. Businesses have an increasing need to analyze large quantities of data, whether to validate a hypothesis or to discover new relationships between data. This information is often critical to business success, and it can be difficult to get the information in a timely manner. DB2 V4 and later releases lets you split and run a single query within a DB2 subsystem. With sysplex query parallelism, DB2 V5 and later releases extends parallel processing to allow a single query to use all the CPC capacity of a data sharing group.

Sysplex query parallelism is when members of a data sharing group process a single query. DB2 determines an optimal degree of parallelism based on estimated I/O and processing costs. Different DB2 members processes different ranges of the data. Applications that are primarily read or and are processor-intensive or I/O-intensive can benefit from sysplex query parallelism. A query can split into multiple parallel tasks that can run in parallel across all images (up to 32) in a Sysplex. It can run in parallel on up to 344 CPs within a Parallel Sysplex of 32 systems with 12 CPs each.

sysplex sockets. Socket applications are written generally to communicate with a partner on any platform. This means that the improved performance and scalability Parallel Sysplex is not exploited, unless some application-specific protocol is used; this is not always possible.

The sysplex sockets function provides a standard way to discover information about the connected partner which can then be used to make decisions that can exploit the value of the Parallel Sysplex where applicable.

Sysplex Timer. An IBM unit that synchronizes the time-of-day (TOD) clocks in multiple processors or processor sides. External Time Reference (ETR) is the OS/390 generic name for the IBM Sysplex Timer (9037).

system. In data processing, a collection of people, machines, and methods organized to accomplish a set of specific functions.

system configuration. A process that specifies the devices and programs that form a particular data processing system.

system control element (SCE). The hardware that handles the transfer of data and control information associated with storage requests between the elements of the processor unit.

System Support Programs (SSP). An IBM-licensed program, made up of a collection of utilities and small programs, that supports the operation of the NCP.

systems management. The process of monitoring, coordinating, and controlling resources within systems.

S/390 microprocessor cluster. A configuration that consists of CPCs and may have one or more CFs.

S/390 partners in development. Membership in S/390 Partners in Development is open to companies and organizations developing or planning to develop commercially marketed software executing in an IBM S/390 environment under OS/390, VM or VSE operating systems.

Offerings include low-cost application development and porting platforms, answer to technical questions, and information on IBM's trends, directions and latest technology.

- Special offers on development machines and software
- Bulletin-board Q&A support - Free
- Fee-based technical support
- Choice of S/390 platform library (CD-ROM) - Free
- S/390 Rainbow collection CD-ROM - Free
- Porting and technology workshops
- Remote access to IBM systems for application porting and development
- Solution Partnership Centers for test driving IBM Technologies
- Access to early IBM code
- OS/390, VM and VSE technical disclosure meetings
- Information Delivery - Members get funneled pertinent S/390 information

T

takeover. The process by which the failing active subsystem is released from its extended recovery facility (XRF) sessions with terminal users and replaced by an alternate subsystem.

TCP/IP. Transmission control protocol/Internet protocol. A public domain networking protocol with standards maintained by the U.S. Department of Defense to allow unlike vendor systems to communicate.

Telnet. U.S. Department of Defense's virtual terminal protocol, based on TCP/IP.

throughput. (1) A measure of the amount of work performed by a computer system over a given period of time, for example, number of jobs per day. (2) A measure of the amount of information transmitted over a network in a given period of time.

tightly coupled. Multiple CPs that share storage and are controlled by a single copy of OS/390. See also *loosely coupled* and *tightly coupled multiprocessor*.

tightly coupled multiprocessor. Any CPC with multiple CPs.

time-of-day (TOD) clock. A 64-bit unsigned binary counter with a period of approximately 143 years. It is incremented so that 1 is added into bit position 51 every microsecond. The TOD clock runs regardless of whether the processing unit is in a running, wait, or stopped state.

time-to-live (TTL). In the context of a TCP/IP DNS nameserver, the time-to-live is the time that a DNS nameserver will retain resource records in its cache for resources for which it is not the authoritative name server.

TOD. See time-of-day (TOD) clock.

Token-Ring. A network with a ring topology that passes tokens from one attaching device (node) to another. A node that is ready to send can capture a token and insert data for transmission.

transaction. In an SNA network, an exchange between two programs that usually involves a specific set of initial input data that causes the execution of a specific task or job. Examples of transactions include the entry of a customer's deposit that results in the updating of the customer's balance, and the transfer of a message to one or more destination points.

transmission control protocol/Internet protocol (TCP/IP). A public domain networking protocol with standards maintained by the U.S. Department of Defense to allow unlike vendor systems to communicate.

transmitter. In fiber optics, see *optical transmitter*.

trunk cable. In an ESCON environment, a cable consisting of multiple fiber pairs that do not directly attach to an active device. This cable usually exists between distribution panels and can be located within, or external to, a building. Contrast with *jumper cable*.

TSO. See *TSO/E*.

TSO/E. In OS/390, a time-sharing system accessed from a terminal that allows user access to OS/390 system services and interactive facilities.

TTL. See *time-to-live*

tutorial. Online information presented in a teaching format.

type 2.1 node (T2.1 node). A node that can attach to an SNA network as a peripheral node using the same protocols as type 2.0 nodes. Type 2.1 nodes can be directly attached to one another using peer-to-peer protocols. See *end node*, *node*, and *subarea node*.

U

uniprocessor (UP). A CPC that contains one CP and is not partitionable.

universal time coordinate (UTC). UTC is the official replacement for (and is generally equivalent to) the better known "Greenwich Mean Time."

UP. See *uniprocessor (UP)*.

UTC. See *Universal Time Coordinate*.

V

validity vector. On a CPC, a bit string that is manipulated by cross-invalidate to present a user connected to a structure with the validity state of pages in its local cache.

velocity. A service goal naming the rate at which you expect work to be processed for a given service class or a measure of the acceptable processor and storage delays while work is running.

W

VTS. See *Virtual Tape Server*.

warm start. Synonymous with normal restart.

white space. CF storage set aside for rebuilding of structures from other CFs, in case of planned reconfiguration or failure.

workload. A group of work to be tracked, managed and reported as a unit. Also, a group of service classes.

workload management mode. The mode in which workload management manages system resources on an OS/390 image. Mode can be either *compatibility mode* or *goal mode*.

work qualifier. An attribute of incoming work. Work qualifiers include: subsystem type, subsystem instance, user ID, accounting information, transaction name, transaction class, source LU, NETID, and LU name.

write-ahead data set (WADS). A data set containing log records that reflect completed operations and are not yet written to an online log data set.

X

XCF. See *cross-system coupling facility*.

XCF couple data set. The name for the sysplex couple data set prior to MVS SP V5.1. See *sysplex couple data set*.

XCF dynamics. XCF Dynamics uses the Sysplex Sockets support that is introduced in OS/390 V2R7 IP. Sysplex Sockets allows the stacks to communicate with each other and exchange information like VTAM CPNames, MVS SYSCLONE value and IP addresses. Dynamic XCF definition is activated by coding the IPCONGFIG DYNAMICXCF parameter in TCPIP.PROFILE.

XCF group. A group is the set of related members defined to XCF by a multisystem application in which members of the group can communicate (send and receive data) between OS/390 systems with other members of the same group. A group can span one or more of the systems in a sysplex and represents a complete logical entity to XCF. See *Multisystem application*.

XCF-local mode. The state of a system in which XCF provides limited services on one system and does not provide signalling services between OS/390 systems. See also *single-system sysplex*.

XCF PR/SM policy. In a multisystem sysplex on PR/SM, the actions that XCF takes when one OS/390 system in the sysplex fails. This policy provides high availability for multisystem applications in the sysplex.

XES. See *cross-system extended services*.

XRF. See *Extended recovery facility*.

Y

Year 2000 test datesource facility. The Year 2000 test datesource facility allows you to define several LPs on a single CPC that can enter a sysple with a time and date other than that of the production system. This eliminates the need for dedicating an entire CPC to do year 2000 testing in a multi-member sysplex.

List of Abbreviations

ABARS	aggregate backup and recovery support
ac	alternating current
ACB	access control block
ACDS	alternate control data set
ACDS	active control data set (SMS)
ACS	automatic class selection (SMS)
ACTS	automated computer time service
ADM	asynchronous data mover
ADMF	asynchronous data mover function
ADMOD	auxiliary DC power module
ADSM	ADSTAR distributed storage manager
ALA	link address (ESCON)
AMD	air moving device
AMODE	addressing mode
AMRF	action message retention facility
ANR	automatic networking routing
AOR	application owning region (CICS)
APA	all points addressable
API	application program interface
APPC	advanced program-to-program communication
APPLID	application identifier
APPN	advanced peer-to-peer networking
ARF	automatic reconfiguration facility
ARM	Automatic Restart Manager
AS	address space
AS	auto switchable (Tape Sharing)
ASKQ	ask questions (HONE)
ASYNC	asynchronous
ATL	automated tape library
AVG	average
AWM	alternate wait management (OS/390)
AXM	authorized cross-memory
BAN	boundary access node
BBU	battery backup
BCDS	backup control data set (DFSMShsm)
BCP	basic control program
BCU	basic configurable unit
BCS	basic catalog structure
BDAM	basic direct access method
BGP	border gateway protocol
BIX	build index
BL	block (block-multiplexer channel)
BMP	batch message process (IMS)
BNN	boundary network node (SNA)
BOC	battery operated clock
BOF	bill of forms
BP	buffer pool
BRS	bandwidth reservation
BSC	binary synchronous communication
BSDS	bootstrap data set
BSN	bus switch network
BUF	buffer
BUFND	buffer number data (VSAM)
BUFNI	buffer number index (VSAM)
BUFNO	buffer number
BUFSP	buffer space (VSAM)
BWAT	batch workload analysis tool
BWO	backup while open
BY	byte (byte-multiplexor channel)
CA	Computer Associates
CA	continuous availability
CAA	cache analysis aid
CADS	channel adapter data streaming
CAS	catalog address space
CAS	coordinating address space
CAU	CICS affinity utility
CB	Component Broker
CBIC	control blocks in common
CBIPO	custom-built installation process offering
CBPDO	custom-built product delivery offering
CBU	capacity backup
CBY	ESCON byte multiplexer channel
CCU	communications control unit
CCU	central control unit
CDRM	cross-domain resource manager
CDSC	catalog data space cache
CDS	central directory server
CDS	couple data set
CEC	central electronics complex
CF	Coupling Facility
CFCC	Coupling Facility Control Code
CFCC LP	Coupling Facility Logical Partition
CFDT	Coupling Facility data table
CFFP	Coupling Facility failure policy (OS/390)
CFIA	component failure impact analysis
CFR	Coupling Facility receiver
CFRM	Coupling Facility resources management
CFS	Coupling Facility sender
CFVM	Coupling Facility Virtual Machine (VM)
CHG	changed
CHKP	check point
CHKPT	check point
CHPID	channel path identifier
CI	control interval (VSAM)
CICS	Customer Information Control System
CIU	CICS Interdependencies Utility
CKPT	checkpoint
CLI	call level interface
CLIST	command list
CLO	control link oscillator
CMAS	CICSPlex SM address space
CMC	communications management configuration
CMF	CICS monitoring facility
CMOS	complementary metal oxide semiconductor
CNC	ESCON channel attached to an ESCON-capable device
COB	card-on-board
COMDS	communications data set
CP	control program
CP	central processor
CP/SM	CICSPlex systems manager (CICSPlex SM)
CPC	central processing complex

CPF	command prefix facility
CPU	central processing unit
CQS	common queue server
CR	control region (CICS)
CR	capture ratio
CRA	Configuration Reporting Architecture (S/390)
CDRSC	cross-domain resources
CS	cursor stability (DB2)
CS	central storage
CSA	common systems area
CSAR	complex systems availability and restart
CSECT	control section
CSI	consolidated software inventory (SMP/E)
CSS	channel subsystem
CSTOR	control storage (central storage for CF)
CSU	customer setup
CTC	channel-to-channel
CTCA	channel-to-channel adapter
CU	control unit
CUA	common user access (SAA)
CUoD	Capacity Upgrade on Demand
CV	Central Version (IDMS/CA)
CVC	conversion channel (ESCON)
CVOL	control volume
CWA	common work area
DADSM	direct access device space management
DAE	dump analysis elimination
DASD	direct access storage device
DB	database
DBD	database definition (IMS)
DBRC	database recovery control (IMS)
DCAF	distributed control access facility
DCB	data control block
DCCF	disabled console communication facility
DCE	distributed computing environment
DCFD	Dynamic CF Dispatching
DD	data definition
DDSR	Dynamic Database Session Routing (IDMS/CA)
DDF	distributed data facilities (DB2)
DEDB	data entry database (IMS)
DEQ	dequeue
DFHMS	Data Facility Hierarchical Storage Manager
DFR	deferred
DFSMS	Data Facility Storage Management Subsystem
DFW	DASD fast write
DIB	data in block
DIM	data in memory
DL/1	Data Language 1
DLC	data link control
DLISAS	DL/1, separate address space
DLM	Distributed Lock Manager (Oracle)
DLS	data link switch
DLSw	data link switching
DLUR	dependent LU requester/server
DMA	direct memory access
DMA	dynamic memory array
DNS	Domain Name Server
DP	dispatching priority
DPL	distributed program link (CICS)
DRDA	distributed remote database access
DRDA	distributed relational database architecture
DSG	data sharing group
DSM	distributed security manager
DSNT	data set name table
DSP	data space
DSR	dynamic storage reconfiguration
DTE	data terminal equipment
EC	engineering change
ECB	event control block
ECC	error correction code
ECI	external call interface (CICS)
ECL	emitter coupled logic
ECS	Enhanced Catalog Sharing
EDR	enqueue residency value
EDT	eligible device table
EMC	event monitoring control
EMCS	extended multiple console support
EMH	expedited message handler (IMS)
EMHQ	expedited message handler queue (IMS)
EMIF	ESCON multiple image facility
EMIF	enhanced multiple image facility
EN	end node
ENF	event notification facility
ENQ	enqueue
ENTR	Ethernet/Token-Ring
EP	Emulation Program (3745)
EPDM	Enterprise Data Manager
EPSO	Enhanced S/390 Parallel Sysplex Offering
ERS	enqueue residency value (GRS)
ES	expanded storage
ES	Enterprise Systems
ESCA	ESCON adapter
ESCD	ESCON director
ESCON	Enterprise Systems Connection
ESO	expanded storage only (hiperspace)
ESTOR	non-control storage (expanded storage for CF)
ETOD	Extended time-of-day
ETR	external time reference
ETS	external time source
EXCI	external CICS interface (CICS)
EV	execution velocity (WLM)
FAX	facsimile
FDBR	fast database recovery (IMS)
FF	fast forward
FF	fox fox (hexadecimal)
FF	full function
FICON	FIbre CONnection
FOR	file-owning region (CICS)
FRAD	frame relay access device
FRU	field replaceable unit
FSS	functional subsystem
FTP	file transfer program
FTP	file transfer protocol
FTS	fiber transport services
GA	general availability
GAC	global access checking
GB	gigabyte
GbE	Gigabit Ethernet

GBP	group buffer pool (DB2)
GDG	generation data group
GEM	global enterprise manager
GENCB	generate control block
GMLC	graduated monthly license charge
GMMA	Goal Mode Migration Aid
GMT	Greenwich mean time
GR	generic resource
GRG	generic resource group (IMS)
GRECP	group buffer pool recovery pending (DB2)
GRG	generic resource group
GRS	global resource serialization
GSR	global shared resources
GTF	Generalized Trace Facility
GUI	graphical user interface
GW	gateway
HCD	hardware configuration definition
HDAM	hierarchic direct access method
HFS	hierarchical file system (UNIX)
HIDAM	hierarchic indexed direct access method
HISAM	hierarchic indexed sequential access method
HLQ	high level qualifier
HMC	hardware management console
HOD	host-on-demand
HONE	hands on network environment
HPDT	high performance data transfer (APPC)
HPR	high performance routing (APPN)
HRDW	hardware
HSA	hardware service area
HSB	high speed buffer
HSC	hardware system console
HSSI	high speed serial interface
HW	hardware
HWMCA	hardware management console application
Hz	hertz
I/O	input/output
IBB	internal bus buffer
IBF	internal battery facility
IBM	International Business Machines Corporation
IC	internal coupling
ICB	integrated cluster bus
ICE	I/O controller element
ICF	Integrated Catalog Facility
ICF	Internal Coupling Facility
ICMF	Integrated Coupling Migration Facility
ICP	Interconnect Controller Program (IBM program product)
ICRF	Integrated Cryptographic Feature
ICS	installation control specifications
IFP	IMS fast path
IIOP	Internet Inter-Object Request Block Protocol
IML	initial microcode load
IMS	Information Management System
IMS DB	Information Management System Database Manager
IMS TM	Information Management System Transaction Manager
IOBF	I/O buffer (IMS)
IOC	I/O count (SRM)
IOCP	input output configuration program
IODF	I/O definition file
IOP	I/O processor
IOQ	I/O queue
IOSP	input/output support processor
IPA	IP assist
IP	Internet protocol
IPC	initial power controller
IPCS	interactive problem control system
IPL	initial program load
IPS	installation performance specification
IPsec	IP security protocol
IPX	Internet packet exchange
IRC	interregion communications
IRLM	integrated resource lock manager
ISC	inter-system communications (in CICS and IMS)
ISC	inter-system coupling (CF link type)
ISD	internal disk (channel path)
ISMF	Integrated Storage Management Facility
ISPF	Interactive System Productivity Facility
ISO	integrated services offering
ISR	intermediate session routing
ISV	independent software vendor
ITR	internal throughput rate
ITRR	internal throughput rate ratio
ITSC	International Technical Support Center
ITSO	International Technical Support Organization
ITU	International Telecommunication Union
JCL	job control language
JECL	JES control language
JES	Job entry subsystem
JMF	JES3 monitoring facility
JOE	job output element (JES)
KB	kilobyte
KGTV	Korhonen George Thorsen Vaupel
km	kilometer
LAN	local area network
LASER	light amplification by stimulated emission of radiation
LCU	logical control unit
LE	language environment
LED	light emitting diode
LFS	LAN file server
LIC	licensed internal code
LOB	large object (DB2)
LOC	locate
LOC	location
LP	logical partition
LPAR	logically partitioned mode
LPCTL	logical partition controls (frame)
LPL	logical page list
LSCD	large scale computing division
LSPR	large systems performance reference
LSR	local shared resources
LST	load module temporary store (SMP/E)
LTERM	logical terminal
LU	logical unit (SNA)
LUPS	local UPS
LX	long wavelenght

MAC	medium access control
MAS	multiaccess spool
MAU	multistation access unit
MAX	maximum
MB	megabyte
Mbps	megabits per second
MB/Sec	megabytes per second
MCCU	multisystem channel communications unit
MCDS	migration control data set (DFSMShsm)
MCL	maintenance change level
MCM	multichip module
MCS	multiple console support
MDB	message data block
MDH	migration data host
MES	miscellaneous equipment specification
MIB	management information base (OSI)
MICR	magnetic ink character recognition/reader
MIH	missing interrupt handler
MIM	Multi-Image Manager
MIPS	millions of instructions per second
MLSLEC	maximum list-set-element count
ML1	migration level 1 (DFSMShsm)
ML2	migration level 2 (DFSMShsm)
MODCB	modify control block
MM	multimode
MNPS	multinode persistent session
MP	multiprocessor
MPC+	multipath channel+
MPF	message suppressing facility
MPG	multiple preferred guests
MPL	multiprogramming level
MPNP	multiprotocol network program
MPR	message processing region (IMS)
MPTN	multiprotocol transport networking
MRNS	multiprotocol routing network services
MRO	multiregion operation
MSC	multisystems communication (IMS)
MSDB	main storage database (IMS)
MSGQ	shared message queue (IMS)
MSO	main storage occupancy (SRM)
MSS	multiprotocol switched services
MSU	millions of service units
MTS	macro temporary store (SMP/E)
MTU	maximum transmission unit (Ethernet)
MTW	mean time to wait
MUF	multi-user facility (Datacom/CA)
MULC	measured usage license charge
MVS	Multiple Virtual Storage
N/A	not applicable
NAU	network addressable unit
NDF	non-deferred
NFS	network file system
NIB	node initialization block
NIP	nucleus initialization process
NLP	network layer packets
NN	network node (SNA)
NNP	network node processor
NNS	named counter server (CICS)
NO	number
NOOVLY	no overlay
NSS	national system support
NSSS	networking and system services and support
NVS	nonvolatile storage
OAS	OSA address table
OCDS	offline control data set (DFSMShsm)
OCF	operations command facility
OCR	optical character recognition/reader
ODBC	open database connectivity
OEMI	original equipment manufacturer information/interface
OLDS	online log data set (IMS)
OLS	offline sequence
OMF	operations management facility
ONC	open network computing
OO	object-oriented
OPC	operations planning and control
OS	operating system
OSA	open systems architecture
OSAM	overflow sequential access method
OSI	open systems interconnection
OSPF	open shortest path first
OVLY	overlay
PAF	Processor Availability Facility
PCE	processor controller element
PCM	plug-compatible manufacturers
P/DAS	peer-to-peer dynamic address switching
PD	program directory
PDS	partitioned data set
PDSE	partitioned data set enhanced
PEL	picture element
PEP	Partitioned Emulation Program (3745)
PES	Parallel Enterprise Server
PET	platform evaluation test
PI	performance index
PI	path in
PI	program isolation
PIF	program interface
PLET	product announcement letter
PLO	perform locked operation
PM	Presentation Manager (OS/2)
PMIO	performance management input/output
PMOS	performance management offerings and services
PO	path out
POR	power-on reset
PP	physical partitioned (mode of operation)
PPP	point-to-point protocol
PPRC	peer-to-peer remote copy
PPS	pulse-per-second
PR/SM	processor resource/system manager
PROFS	professional office system
PSB	program specification block
PSLC	Parallel Sysplex license charge
PSMF	Parallel Sysplex Management Facility
PSO	Parallel Server Option (Oracle)
PSP	preventive service planning
PSP	primary support processor
PTF	program temporary fix
PTH	path
PTS	parallel transaction server
PU	physical unit (SNA)
PU	processing unit (9672)

QDIO Queued Direct Input/Output
QOR queue-owning region (CICS)
RACF Resource Access Control Facility
RAMAC Random Access Method of Accounting and Control
RAS reliability availability serviceability
RBA relative block address
RBA relative byte address
RCD read configuration data
RDS restructured database (RACF)
REA RAMAC electronic array
REQ required
RETAIN Remote Technical Assistance and Information Network
RG resource group
RIOC relative I/O content
RIP routing information protocol
RIT RECON initialization timestamp (IMS)
RJP remote job processor
RLL row-level locking
RLS record-level sharing (VSAM)
RLSWAIT RLS wait
RMF Resource Measurement Facility
RMODE residency mode
RNL resource name list (GRS)
ROT rules of thumb
RPC remote procedure call
RPL request parameter list
RPP relative processor performance
RPQ request for price quotation
RR repeatable read (DB2)
RRDF Remote Recovery Data Facility
RRMS Recoverable Resource Management Services (OS/390)
RRS resource recovery services
RRSF RACF Remote Sharing Facility
RSA RAMAC scalable array
RSA ring system authority (GRS)
RSM real storage manager
RSR remote site recovery
RSU recommended service upgrade
RTA real time analysis
RTM recovery termination manager
RTP rapid transport protocol
RVA RAMAC virtual array
SA OS/390 System Automation for OS/390
SAA systems application architecture
SAE single application environment
SAF System Authorization Facility
SAP system assist processor
SAPR systems assurance product review
SCA shared communication area (DB2)
SCDS save control data set (SMP/E)
SCDS source control data set (SMS)
SCE system control element
SCH subchannel
SCKPF Store Clock Programmable Field
SCP system control program
SCS source control data set
SCSI small computer system interface
SCTC ESCON CTC control unit
SCV software-compatible vendor
SDEP sequential dependent (IMS)
SDT shared data tables (CICS)
SE support element
SEC system engineering change
SECS seconds
SETI SE technical information (HONE)
SFM sysplex failure management
SHISAM simple hierarchic indexed sequential access method
SI single image (mode of operation)
SID system identifier
SIE start interpretive execution (instruction)
SIGP signal processor (instruction)
SIMETR simulated external time reference (OS/390)
SLA service level agreement
SLIP serviceability level indicator processing
SLM system lock manager (XCF)
SLMH single-link multihost (IBM 3174 with ESCON interface)
SLSS system library subscription service
SM single-mode
SMF Systems Management Facility
SMOL sales manual online (HONE)
SMP/E system modification program/extended
SMPLOG SMP/E log
SMPLTS SMP/E load module temporary store
SMPMTS SMP/E macro temporary store
SMPSCDS SMP/E save control data set
SMPSTS SMP/E source temporary store
SMQ shared message queue (IMS)
SMS system managed storage
SMSVSAM system managed storage VSAM
SNA systems network architecture
SNAP/SHOT system network analysis program/ simulation host overview technique
SNI SNA network interconnect
SNMP simple network management protocol
SP system product
SPE small programming enhancement
SPOC single point of control
SPOF single point of failure
SPOOL simultaneous peripheral operation online
SPUFI SQL processor using file input (DB2)
SQG shared queues group (IMS)
SQL structured query language (DB2)
SQL/DS structured query language/data system (VM)
SRB service request block
SRDS structure recovery data set (CQS)
SROD shared read-only database (IMS)
SRM System Resources Manager
S/S start/stop
SSCH start subchannel
SSI single system image
SSI subsystem interface
SSL secure socket layer
SSM secondary space management (DFSMShsm)
SSP subsystem storage protect
STC started task control
STCK store clock (instruction)

STCKE	store clock extended (instruction)
STI	self-timed interconnect (S/390)
STS	source temporary store
STSI	Store System Information Instruction (S/390)
SUBSYS	subsystem
SC	Shared Cache (IDMS/CA)
SVC	supervisor call (instruction)
SVS	solutions validation services
SW	software
SX	short wavelength
SYNC	synchronous
SYSAFF	system affinity
SYSCONS	system consoles
SYSID	system identifier
SYSPLEX	systems complex
SYSRES	system residence volume (or IPL volume)
TAI	French for International Atomic Time
TCB	task control block (OS/390)
TCM	thermal conduction Module
TCP/IP	Transmission Control Protocol/Internet Protocol
TCU	terminal control unit
TESTCB	test control block
TG	transmission group
TKE	trusted key entry
TM	transaction manager
TME	Tivoli Management Environment
TMM	tape mount management
TOD	time of day
TOR	terminal-owning region
TP	transaction program (APPC)
TPF	Transaction Processing Facility
TPNS	Teleprocessing Network Simulator
TS	temporary storage (CICS)
TS	transaction server (CICS)
TSC	TOD synchronization compatibility
TSO	Time Sharing Option
TSOR	temporary storage-owning region (CICS)
TSQ	temporary storage queue (CICS)
TTL	time-to-live (TCP/IP)
UBF	user buffering
UCW	unit control word
UD	undeliverable (message)
UDB	universal database (DB2)
UDF	update-time-order (DB2)
UDP	user diagram protocol
UOW	unit of work
UP	uniprocessor
UPS	uninterruptible power supply/system
UR	uncommitted read (DB2)
URL	universal resource locator
UTC	universal time coordinate
VF	vector facility
VGA	video graphics array/adapter
VIO	virtual I/O
VIPA	virtual IP addressing (TCP/IP)
VLF	virtual lookaside facility
VM	Virtual Machine
VPD	vital product data
VR	virtual route (SNA)
VRTG	virtual-route-based transmission group (SNA)
VSAM	Virtual Storage Access Method
VSO	virtual storage option (IMS)
VTAM	Virtual Telecommunications Access Method
VTOC	volume table of content
VTS	virtual tape server
VVDS	VSAM volume data set
WAITRBLD	wait for rebuild (IMS)
WAN	wide area networks
WLM	Workload Manager
WLR	IMS/ESA workload router
WQE	write-to-operator-queue-element
WSC	Washington System Center
WSS	working set size
WTAS	world trade account system
WTO	write-to-operator
WTOR	write-to-operator-with-reply
WWQA	world wide question & answer
WWW	World Wide Web
XCA	external communications adapter
XCF	Cross-System Coupling Facility
XDF	Extended Distance Feature
XES	Cross-System Extended Services
XI	cross-invalidate
XJS	extended job entry subsystem
XRC	extended remote copy
XRF	Extended Recovery Facility

Index

Special Characters

Numerics

A

C

E

H

I

J

L

N

O

P

Q

R

X

IBM Redbooks evaluation

OS/390 Parallel Sysplex Configuration Overview
SG24-5637-00

Your feedback is very important to help us maintain the quality of IBM Redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at http://www.redbooks.ibm.com/
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?
__Customer __Business Partner __Solution Developer __IBM employee
__None of the above

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction ____________

Please answer the following questions:

Was this redbook published in time for your needs? Yes____ No____

If no, please explain:

____

____

____

____

What other redbooks would you like to see published?

____

____

____

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

____

____

____

____

____

SG24-5637-00
Printed in the U.S.A.

OS/390 Parallel Sysplex Configuration Overview

SG24-5637-00