

# IBM z/OS V2R2: Performance

Keith Winnard

Jose Gilberto Biondo Jr

Wilson de Figueiredo

Paul Robert Hering

Alvaro Salla



**z Systems**





International Technical Support Organization

**IBM z/OS V2R2: Performance**

December 2015

**Note:** Before using this information and the product it supports, read the information in “Notices” on page v.

**First Edition (December 2015)**

This edition applies to Version 2, Release 2 of IBM z/OS (5650-ZOS).

**© Copyright International Business Machines Corporation 2015. All rights reserved.**

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

# Contents

<b>Notices</b> .....	v
Trademarks .....	vi
<b>IBM Redbooks promotions</b> .....	vii
<b>Preface</b> .....	ix
Authors .....	ix
Now you can become a published author, too! .....	xi
Comments welcome .....	xi
Stay connected to IBM Redbooks .....	xii
<b>Chapter 1. Resource Measurement Facility</b> .....	1
1.1 Overview .....	2
1.2 RMF monitor III PCIe support .....	2
1.2.1 Monitor III display .....	4
1.3 RMF Support for large page enhancements .....	7
1.3.1 Enhancements .....	8
1.3.2 Postprocessor paging activity report changes .....	8
1.3.3 Postprocessor virtual storage report changes .....	11
1.3.4 Monitor III STORM report changes .....	12
1.3.5 SMF records new fields .....	12
1.4 Global Resource Serialization use statistics in RMF monitor III .....	15
1.4.1 Usage and invocation .....	16
1.5 RMF zFS monitoring enhancements .....	18
<b>Chapter 2. Workload Manager</b> .....	19
2.1 WLM basic concepts .....	20
2.2 WLM large memory enhancements .....	20
2.2.1 IEAOPTxx parameter MCCFXTPR .....	20
2.2.2 IEAOPT parameter RCCFXTT .....	21
2.2.3 IEAOPT parameter RCCEXTT .....	21
2.3 IEAOPT parameters available to monitoring products .....	22
2.4 WLM buffer pool management .....	22
2.4.1 WLM buffer pool management .....	22
2.4.2 Housekeeping function improvement .....	26
2.5 WLM health-based routing services .....	27
2.5.1 WLM routing services overview .....	27
2.5.2 Health factor effecting routing Weight .....	28
2.6 Improvements in z/OS V2.R2 IWM4HLTH service .....	29
2.6.1 New IWM4QHLLT query service .....	30
<b>Chapter 3. Real Storage Management scalability</b> .....	31
3.1 Overview .....	32
3.2 Performance gains because of large memories .....	32
3.3 Improving performance by using large memories .....	33
3.4 Large memory studies .....	33
3.5 RSM modifications at z/OS V2R2 .....	37
3.5.1 Dumping Large Memory by using Standalone Dump utility .....	38
3.5.2 Change in default for LFAREA INCLUDE1MAFC(YES) .....	38

<b>Chapter 4. zFS Monitoring Enhancements</b> . . . . .	41
4.1 Overview of zFS monitoring enhancements . . . . .	42
4.2 Using the zFS monitoring enhancements . . . . .	42
4.2.1 RMF Monitor III Data Gatherer option. . . . .	42
4.2.2 RMF Monitor III zFS Sysplex Reports. . . . .	43
4.2.3 RMF Monitor III ZFSOVW Report . . . . .	44
4.2.4 Monitor III ZFSOVW I/O data . . . . .	45
4.2.5 Monitor III ZFSOVW User Cache data . . . . .	45
4.2.6 Monitor III ZFSOVW vnode Cache data . . . . .	46
4.2.7 Monitor III ZFSOVW metadata cache data . . . . .	46
4.2.8 Monitor III Online Help for zFS panels . . . . .	46
4.2.9 RMF Monitor III ZFSFS report. . . . .	47
4.2.10 RMF Monitor III ZFSKN report . . . . .	48
4.2.11 zFS user and metadata considerations. . . . .	49
<b>Related publications</b> . . . . .	53
IBM Redbooks . . . . .	53
Other publications . . . . .	53
Online resource . . . . .	53
Help from IBM . . . . .	54

# Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*

**The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:** INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include 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.


## COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

# Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

CICS®	OS/390®	z Systems™
DB2®	Parallel Sysplex®	z/OS®
DS8000®	Redbooks®	z10™
FICON®	Redbooks (logo)  ®	z13™
IBM®	Resource Measurement Facility™	zEnterprise®
IBM z13™	RMF™	
IMS™	System z®	

The following terms are trademarks of other companies:

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.



## Find and read thousands of IBM Redbooks publications

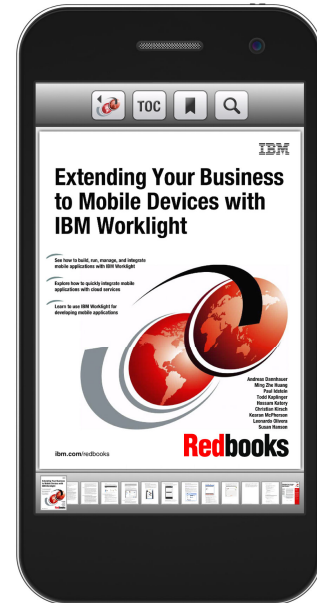
- ▶ Search, bookmark, save and organize favorites
- ▶ Get up-to-the-minute Redbooks news and announcements
- ▶ Link to the latest Redbooks blogs and videos

Get the latest version of the Redbooks Mobile App



Download  
Now

iOS



## Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!



[ibm.com/Redbooks](http://ibm.com/Redbooks)

About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

# Preface

This IBM® Redbooks® publication helps you to become familiar with the technical changes that were introduced into the performance areas with IBM z/OS® V2R2.

This book is one of a series of IBM Redbooks publications that take a modular approach to providing information about the updates that are contained within z/OS V2R2. This approach has the following goals:

- ▶ Provide modular content
- ▶ Group the technical changes into a topic
- ▶ Provide a more streamlined way of finding relevant information based on the topic.

We hope you find this approach is useful. We value your feedback.

## Authors

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

**Keith Winnard** is the z/OS Project Leader at the International Technical Support Organization, Poughkeepsie Center. He writes extensively and is keen to engage with customers to understand what they want from IBM Redbooks publications. Before joining the ITSO in 2014, Keith worked for clients and Business Partners in the UK and Europe in various technical and account management roles. He is experienced with blending and integrating new technologies into the traditional landscape of mainframes.

**Jose Gilberto Biondo Jr** is an IT Specialist in Integrated Technology Delivery, ServerSystems Operations/Storage Management in IBM Brazil. He has seven years of experience in z/OS, working with storage management since 2007. Jose works mainly with IBM storage products (DFSMSdfp, DFSMSdss, DFSMShsm, and DFSMSrmm). He also works with OEM software products. Jose's areas of expertise include installing and maintaining storage products, and process automation.

**Wilson de Figueiredo** is z/OS System Programmer. He manages the operations support team at Banco do Brasil, a government bank in Brazil. He has more than 11 years of experience in mainframe systems. He holds a system analysis, internet consulting, and business administration degrees. His areas of expertise include IBM Parallel Sysplex®, z/OS security, and z/OS availability.

**Robert Hering** is an IT Specialist at the ITS Technical Support Center, Mainz, Germany. He provides support to clients with z/OS and z/OS UNIX related questions and issues. He has participated in several ITSO residencies since 1988, writing about UNIX related topics. Before supporting IBM OS/390® and z/OS, Robert worked for many years with the IBM VM operating system and its variations (VM/370, VM/HPO, VM/XA, and VM/ESA).

**Alvaro Salla** is a Senior IT Consultant for the ITSO. He has more than 40 years teaching and developing educational material that covers the z/OS mainframe platform and consulting services focusing on performance. Alvaro has co-authored Redbooks publications about IBM DS8000®, DFSMS, WLM, SYSPLEX, ABCs, IBM RMF™, and disaster recovery. He started with IBM in 1969.

Thank you to the following people for their valuable contributions to this project:

**Dieter Wellerdiek** (IBM) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Andreas Henicke** (IBM) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Karl-Hans Holder** (IBM) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Matthias Gubitz** (IBM) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Michelle Li** (IBM) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Harald Bender** (IBM) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Juergen Baumann** (Software Engineer, RMF Development team, Boeblingen Center) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Juergen Kuhn** (Software Engineer, RMF development team, Boeblingen Center) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Karin Genther** (Software Engineer, RMF development team, Boeblingen Center) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Steve Partlow** (Senior Software Engineer, z/OS Memory Management, Poughkeepsie Center) for providing valuable advice and guidance throughout the creation of this IBM Redbooks publication.

**Bob Haimowitz** (Development Support Team (DST), Poughkeepsie Center) for setting up and maintaining the systems, and providing valuable advice, guidance, and assistance throughout the creation of this IBM Redbooks publication.

**Rich Conway** (Development Support Team (DST), Poughkeepsie Center) for setting up and maintaining the systems, and providing valuable advice, guidance, and assistance throughout the creation of this IBM Redbooks publication.

**Peter Bertolozzi** (Systems Management specialist, IBM Redbooks residency support, Poughkeepsie Center), Poughkeepsie Center) for setting up and maintaining the environments within syslab in which residents worked.

**John Gierloff** (Operations, Poughkeepsie Center) Residency set up and support.

**Don Brennan** (DST, Poughkeepsie Center) for setting up and maintaining the systems hardware used in the creation of this IBM Redbooks publication.

**Ella Buslovich** (Graphics specialist, location) for providing guidance and specialist graphics for this IBM Redbooks publication.

**Ann Lund** (ITSO Administration, Poughkeepsie Center) for administrative support to enable the residences publication.

**Cheryl Gera** (ITSO Administration, Poughkeepsie Center) for managing the business operations for this IBM Redbooks publication.

## Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience by using leading-edge technologies. Your efforts help to increase product acceptance and customer satisfaction as you expand your network of technical contacts and relationships. Residencies run 2 - 6 weeks in length and you can participate in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

[ibm.com/redbooks/residencies.html](http://ibm.com/redbooks/residencies.html)

## Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

[ibm.com/redbooks](http://ibm.com/redbooks)

- Send your comments in an email to:

[redbooks@us.ibm.com](mailto:redbooks@us.ibm.com)

- Mail your comments to:

IBM Corporation, International Technical Support Organization  
Dept. HYTD Mail Station P099  
2455 South Road  
Poughkeepsie, NY 12601-5400

## Stay connected to IBM Redbooks

- ▶ Find us on Facebook:  
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:  
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:  
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:  
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:  
<http://www.redbooks.ibm.com/rss.html>



# Resource Measurement Facility

This chapter describes the modifications that were implemented in IBM Resource Measurement Facility™ (RMF) in V2R2.

RMF is an IBM product for z/OS performance measurement and management.

This chapter includes the following topics:

- ▶ 1.1, “Overview” on page 2
- ▶ 1.2, “RMF monitor III PCIe support” on page 2
- ▶ 1.3, “RMF Support for large page enhancements” on page 7
- ▶ 1.4, “Global Resource Serialization use statistics in RMF monitor III” on page 15
- ▶ 1.5, “RMF zFS monitoring enhancements” on page 18

## 1.1 Overview

RMF is the base product to collect performance data for z/OS and sysplex environments. It monitors system performance behavior and provides reports and the opportunity to tune and configure your system according to your priorities and needs.

## 1.2 RMF monitor III PCIe support

Peripheral Component Interconnect Express (PCIe) is a high-speed serial computer bus extension that is used internally on a computer for connecting various parts. On the IBM z13™, a PCIe I/O Drawer includes the following optional features that are connected to processor units (at CPC Drawers) through PCIe buses:

- ▶ IBM FICON® Express 16S.
- ▶ OSA Express 5S.
- ▶ 10 GbE Remote Direct Memory Access (RDMA) over Converged Ethernet Express (RoCE), which is a new communication protocol that provides transparent acceleration for a sockets-based TCP/IP applications and middle ware.
- ▶ IBM zEnterprise® Data Compression Express (zEDC), a fast data compression and deflation outbound processor. The CPACF inbound coprocessor is still available. The z/OS component in charge of data compression (CMF) favors zEDC, depending on the algorithms.
- ▶ Flash Memory Express.



Figure 1-1 shows the IBM z13 PCIe I/O Drawer.

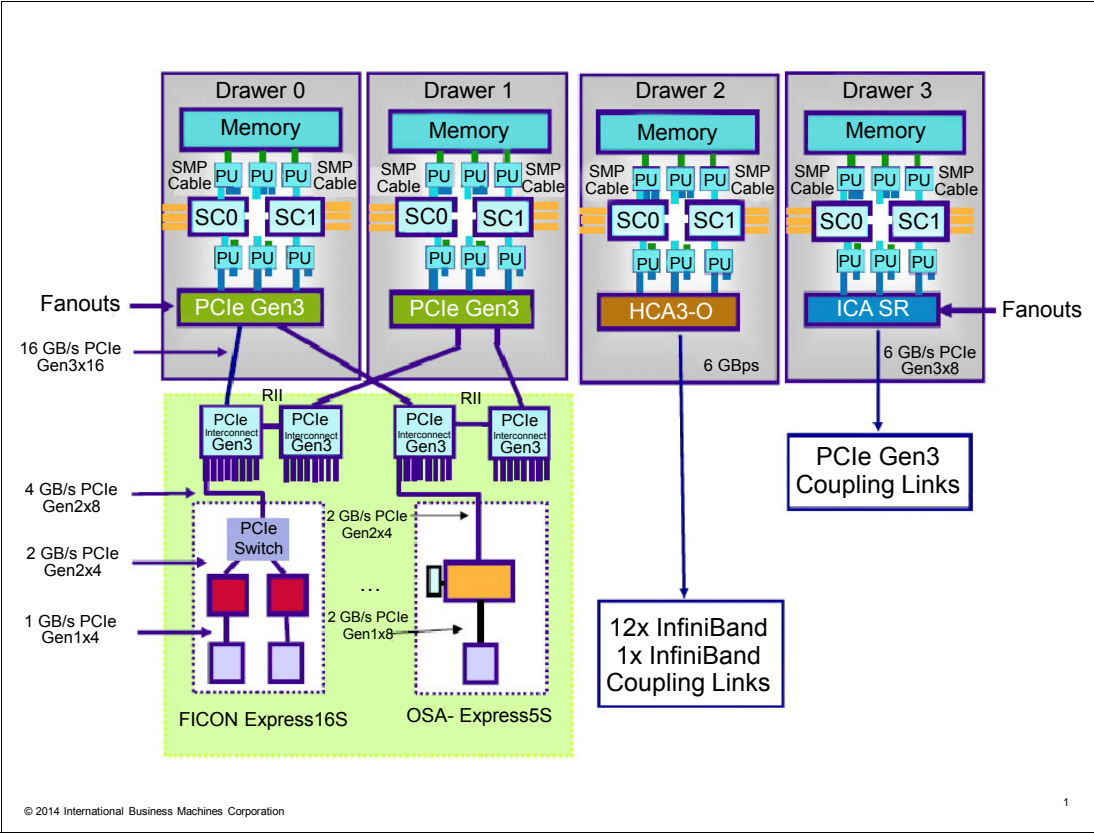


Figure 1-1 z13 PCIe I/O Drawer

In Figure 1-2, the PCIE I/O Drawer is shown with the slots where the optional features are installed. Those features also are included as options in a zEC12 and a zBC12.

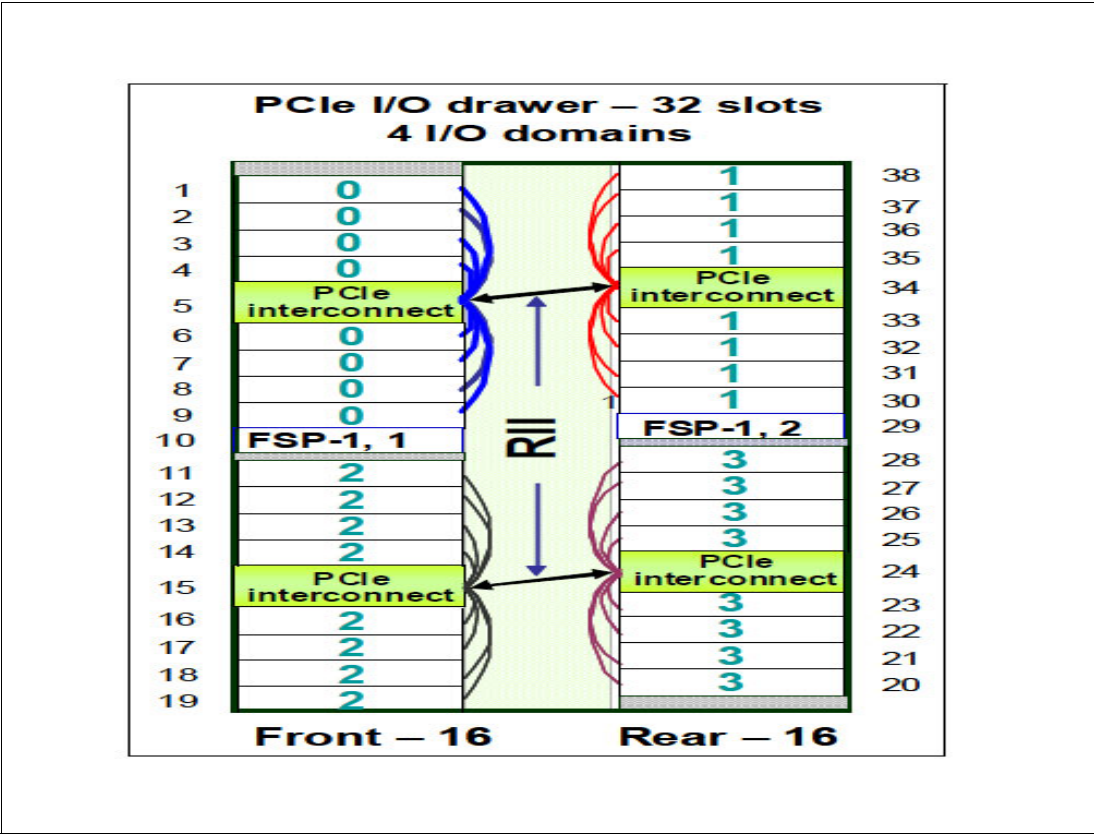


Figure 1-2 A PCIe I/O Drawer

Before RMF V2R2, the performance analysis on RoCE feature and zEDC hardware accelerators was done through the SMF 74.9 records and the RMF Postprocessor only. z/OS V2R2 introduces the PCIE Activity Report for the RMF Monitor III and the RMF Distributed Data Server (DDS) to provide online optional short-term reporting for the RoCE and zEDC.

### 1.2.1 Monitor III display

Users can control whether they want Monitor III to collect PCIE activity data by specifying data gathering option PCIE/NOPCIE in the RMF PARMLIB member. The default value PCIE is set in PARMLIB member ERBRMF04.

To request the Monitor III PCIE Activity Report data, select **option 3** from the Primary Menu, then select **option 14** from the Resource Report Selection Menu, or enter command **PCIE** or **PCI** from the RMF Monitor III primary menu. Figure 1-3 on page 5 shows a sample PCIE Activity display.

```

      . . . . . RMF V2R2 PCIE Activity . . . . . Line 1 of 2
Command ==> Scroll ==> CSR
Samples: 10 System: SC81 Date: 08/28/15 Time: 15.20.40 Range: 10 Sec
----- Function ----- Alloc - PCI Operations Rate - -Xfer Rate -
ID PCID Type Jobname ASID Status Time% Load Store Block Refr Read Write
0025 01BC zEDC FPGHWAM 0013 Alloc 100 0 1895 0 4972
0035 027C zEDC FPGHWAM 0013 Alloc 100 0 1891 0 4972

```

Figure 1-3 PCIE Activity panel

In the main panel, the metrics that are displayed are independent of the type hardware feature and reflect the activity of the z/OS system on which RMF data collection occurred. Table 1-1 lists the fields in the PCIE Activity panel.

Table 1-1 PCIE Activity description

Field	Description
Function ID	Hexadecimal identifier of the PCIe feature for which performance data is reported.
Function PCID	Physical channel identifier (PCHID) for the PCIe function.
Function Type	Device type for the PCIe function: <ul style="list-style-type: none"> <li>▶ HWA: Hardware Accelerator</li> <li>▶ Oth: Unknown device type</li> <li>▶ RoCE: Remote Direct Memory Management</li> <li>▶ zEDC: zEnterprise Data Compression</li> </ul>
Function Jawbone	Name of the job that allocated the PCIe function.
Function ASID	Address space ID of the job that allocated the PCIe function.
Function Status	The PCIe function status at the end of this reporting interval: <ul style="list-style-type: none"> <li>▶ Allow: The function is allocated and in use.</li> <li>▶ Depend: The function is in the process of de-allocation.</li> <li>▶ Error: The function is in permanent error.</li> <li>▶ DeAlloc: The function is de-allocated.</li> <li>▶ Unknown: The function status is unknown.</li> </ul>
Alloc Time%	Percentage of reporting interval for which the PCIe function was allocated or in the process of de-allocation.
PCI Operations Rate Load	Rate of PCIe Load operations per second that were run within this reporting interval.
PCI Operations Rate Store	Rate of PCI Store operations per second that were run within this reporting interval.
PCI Operations Rate Block	Rate of PCI Store Block operations that were run within this reporting interval.
PCI Operations Rate Refresh	Rate of Refresh PCI Translations operations that were run within this reporting interval.
XFER Rate Read	Number of MB per second that were transferred by DMA (RoCE) reads from all defined DMA address spaces to the PCIe function.

Field	Description
XFER Rate Write	Number of MB per second that were transferred by DMA (RoCE) writes from the PCIe function to all defined DMA address spaces.

If the pointer is placed on one of the cursor sensitive fields Function ID, Function PID, or Function Type, more metrics are displayed for zEDC Accelerators in a pop-up window called “Hardware Accelerator And Compression Activity”. (More information about RoCE devices is not available.) Figure 1-4 shows the extra information for zEDC Accelerators.

```

      . . . . . RMF V2R2      PCIE Activity      . . . . . Line 1 of 2
Command ==>                                     Scroll ==> CSR
Samples: 10      System: SC81  Date: 08/28/15  Time: 15.20.40  Range: 10      Se

----
ID      -      RMF Hardware Accelerator And Compression Activity
0025      Press Enter to return to the Report panel.
0035

Function ID : 0035      Alloc Time % : 100      More: +
Allocated   : 14.08.14      on 08/28/15

Hardware Accelerator
Time Busy % : 36.1      Transfer Rate : 765

Request
Execution Time : 139      Std. Deviation: 16.2
Queue Time : 19.1      Std. Deviation: 30.0
Size : 294

Buffer Pool
Memory Size : 16      Utilization % : 0

Compression      Decompression
Request Rate : 2604      0
Throughput : 661      0

```

Figure 1-4 zEDC Accelerator and Compression Activity panel

Table 1-2 lists the fields in the Hardware Accelerator And Compression Activity window.

Table 1-2 HW Accelerator And Compression Activity fields

Field	Description
Allocated	Date and time when the Hardware Accelerator was allocated.
Hardware Accelerator Time Busy%	The percentage of time that this partition kept the Hardware Accelerator busy.
Hardware Accelerator Transfer Rate	The number of megabytes per second that were transferred by zEDC operations.
Request Execution Time	The average time in microseconds the Hardware Accelerator took to process a request.
Request Execution Time Std. Deviation	The standard deviation of the request execution time.
Request Queue Time	The average queue time in microseconds that was spent for a request.
Request Queue Time Std. Deviation	The standard deviation of the request queue time.
Request Size	The average number of kilobytes that were transferred per request.
Buffer Pool Memory Size	The total size of memory in megabytes that is allocated to the buffer pool.

Field	Description
Buffer Pool Utilization%	The average use percentage of the buffer pool that z/OS kept for in-use buffers.
Request Rate	The number of compression or decompression requests per second.
Throughput	The number of megabytes that were compressed or decompressed per second. The number of bytes compressed is greater than the decompressed bytes.
Ratio	The ratio between input and output bytes that were compressed or decompressed within this reporting interval.

Alternatively, you can use RMF (DDS) to view a browser-based version of the Monitor III PCIE Activity Report. To support PCIE reporting in the DDS, the new resource PCIE was added as a child of resource I/O\_SUBSYSTEM to the DDS resource model. Under the PCIE resource, you find all PCIE functions that are allocated in the system with all available relevant performance metrics in RMF Monitor III.

## 1.3 RMF Support for large page enhancements

The use of large 1 MB pages improves the Enhanced Dynamic Address Translation (E-DAT) virtual translation process, which reduces CPU time. Fixed and pageable large pages are available. Consider the following points:

- ▶ Fixed pages are never paged out.
- ▶ Pageable pages can be paged out.
- ▶ Small pages (4 KB) can be paged out to page data sets or SCM (Flash memory).
- ▶ Large pages (1 MB) are paged out to SCM only.

Shared pageable large pages is a virtual storage concept to allow large 1 M pages in high virtual storage (above the bar) to be shared between specific address spaces. A shared page is not a common page. Consider the following differences:

- ▶ A common page is shared among all address spaces.
- ▶ A shared page is shared among a set of specific address spaces.

One example is the shared pages between the two IBM DB2® address spaces, DDF, and DBM1. Shared pages improve performance running the cross memory activity to be faster. The size of the shared area is defined by the keyword HVSHARE at IEASYSxx PARMLIB member.

Virtual storage that is above the bar is acquired by issuing the IARV64 macro with one of the following requests:

- ▶ GETSTOR: For private storage
- ▶ GETCOMMON: For common storage
- ▶ GETSHARED: For shared storage

The storage is allocated objects and these objects are assigned in units of 1 MB. An object consists of one or more large pages.

Do not confuse 1 M page objects with 1 M large pages. Before RMF at V2R2, it is possible to see how many memory objects are allocated, how much central storage is used for high virtual shared memory, and whether the virtual storage is backed by standard real storage frames (4 KB) or large real storage frames (1 MB). RMF V2R2 adds this data about Shared Pageable Large Pages.

### 1.3.1 Enhancements

RMF supports shared pageable large pages reports for Postprocessor Paging Activity Report (PAGING), Virtual Storage Activity Report (VSTOR), and Real Time Monitor III (STORM).

System-wide usage data of shared pageable large pages is provided by the following components:

- ▶ SMF 71 records
- ▶ Postprocessor Paging Activity Report (and Overview Conditions)
- ▶ RMF Monitor III STORM Report, System Summary section
- ▶ RMF Data Portal STORM Report

Address space related usage data of shared pageable large pages is provided by the following components:

- ▶ SMF 78-2 records
- ▶ RMF Postprocessor Virtual Storage Activity Report (VSTOR), under High Virtual Memory Usage section
- ▶ RMF Data Portal STORM Report

**Hardware dependency:** E-DAT architecture must be available to support Large Memory Objects and 1 MB frames. E-DAT was introduced in IBM system z10™ hardware. It uses the STE-format control bit in the segment table entry to determine whether to produce a page-frame real address or a segment-frame absolute address.

### 1.3.2 Postprocessor paging activity report changes

The Postprocessor Paging Activity Report was updated (at MEMORY OBJECTS AND HIGH VIRTUAL STORAGE FRAMES section) to include the extra information regarding 1 M shared large pages. Figure 1-5 on page 9 shows a sample view of the updated report.

P A G I N G   A C T I V I T Y							
z/OS V2R2		SYSTEM ID RMFA		DATE 02/17/2015		INTERVAL 02.00.000	
		RPT VERSION V2R2 RMF		TIME 14.15.00		CYCLE 1.000 SECONDS	
OPT = IEAOPT00   LFAREA SIZE = 134217728   MEMORY OBJECTS AND HIGH VIRTUAL STORAGE FRAMES							
-----							
MEMORY OBJECTS	COMMON	FIXED 1M	SHARED	SHARED 1M			
-----				-----			
MIN	70	0	1	0			
MAX	70	0	17	16			
AVG	70	0	7	6			
1 MB FRAMES	-----	FIXED	-----	-----	PAGEABLE	-----	
	TOTAL	AVAILABLE	IN-USE	TOTAL	AVAILABLE	IN-USE	
MIN	128	0	3	256	0	256	
MAX	128	125	128	256	0	256	
AVG	128	113	15	256	0	256	
HIGH SHARED FRAMES	TOTAL	CENTRAL STORAGE	BACKED 1M			AUX DASD	AUX SCM
		-----	-----				
MIN	136902.1M	225	0			0	0
MAX	136902.1M	45,689	100			0	0
AVG	136902.1M	9,413	8			0	0
HIGH COMMON FRAMES	TOTAL	CENTRAL STORAGE	BACKED 1M	FIXED	FIXED 1M	AUX DASD	AUX SCM
		-----	-----	-----	-----		
MIN	17301504	16,479	0	8,894	0	0	0
MAX	17301504	16,479	0	8,894	0	0	0
AVG	17301504	16,479	0	8,894	0	0	0

Figure 1-5 Sample Postprocessor Paging Activity report

The following changes are featured:

- In MEMORY OBJECTS, column “1 MB” was renamed to “FIXED 1M” to clarify the meaning of large pages. The order of columns “FIXED 1M” and “SHARED” was switched to group column “SHARED” with column “SHARED 1M”. The following columns are available:
  - COMMON: Number of memory objects that are allocated in the high virtual common storage of the system.
  - FIXED 1M: Number of fixed memory objects that are allocated in the system and can be backed by 1 MB frames.
  - SHARED: Number of memory objects that are allocated in the high virtual shared storage of the system.
  - SHARED 1M: Number of shared memory objects that are allocated in the system and can be backed in 1MB frames.
- In HIGH SHARED FRAMES, column “BACKED 1M” was added. The following columns are available:
  - TOTAL: Size of high virtual shared area in units of 4 K pages.
  - CENTRAL STORAGE: Number of pages from high virtual shared storage that is backed in central storage (in units of 4 KB).
  - BACKED 1M: Number of high virtual shared memory 1 MB pages storage that is backed in central storage.
  - AUX DASD: Number of auxiliary storage slots that are used for high virtual shared pages that are backed on DASD.
  - AUX SCM: Number of auxiliary storage slots that are used for high virtual shared pages that are backed on SCM storage.

- In HIGH COMMON FRAMES, column “BACKED 1M” was added. Column “FIXED 4K” was changed to “FIXED”. Column “FIXED 1M” was added. The only common area above the bar is the HCSA. The size of the common area above the bar is defined through the keyword HVCOMMON at IEASYSxx. The following columns are available:
  - TOTAL: Size of high virtual common area in units of 4 K pages.
  - CENTRAL STORAGE: Number of pages from high virtual common storage that is backed in central storage (in units of 4 KB).
  - BACKED 1M: Number of high virtual common memory 1 MB pages that are backed in central storage.
  - FIXED: Number of pages from high virtual common storage that is fixed in central storage (in units of 4 K).
  - FIXED 1M: Number of high virtual common memory 1 MB pages that are fixed in central storage.
  - AUX DASD: Number of auxiliary storage slots that are used for high virtual common pages that are backed on DASD.
  - AUX SCM: Number of auxiliary storage slots that are used for high virtual shared pages that are backed on SCM storage.

The Postprocessor Paging Activity Report also was updated in the Frame And Slot Counts - SHARED FRAMES/SLOTS section to include more information about 1M shared pages.

The new enhancements are shown in Figure 1-6.

P A G I N G   A C T I V I T Y										
z/OS V2R2	SYSTEM ID RMFA		DATE 02/17/2015		INTERVAL 15.00.000					
	RPT VERSION V2R2 RMF		TIME 14.15.00		CYCLE 1.000 SECONDS					
PAGE 2										
OPT = IEAOPT00 LFAREA SIZE = 134217728 CENTRAL STORAGE MOVEMENT AND REQUEST RATES - IN PAGES PER SECOND										
-----										
SYSTEM UIC: MIN = 3322 MAX = 3501 AVG = 3411										
[...]										
FRAME AND SLOT COUNTS										
-----										
(90 SAMPLES)										
CENTRAL STORAGE FRAMES	TOTAL	AVAILABLE	SQA	LPA	CSA	LSQA	REGIONS+SWA	HV SHARED	HV COMMON	
MIN	753,792	462,840	5,561	18,898	5,683	14,928	166,298	225	16,479	
MAX	753,792	536,690	5,561	18,980	5,688	15,070	194,615	45,689	16,479	
AVG	753,792	517,708	5,561	18,908	5,686	14,982	176,192	9,413	16,479	
FIXED FRAMES	TOTAL	NUCLEUS	SQA	LPA	CSA	LSQA	REGIONS+SWA	<16 MB	16MB-2GB	
MIN	37,943	3,156	4,969	67	14,944	8,876	5,931	25	10,519	
MAX	58,057	3,156	4,969	67	14,944	8,939	25,982	25	30,396	
AVG	45,722	3,156	4,969	67	14,944	8,899	13,687	25	18,203	
SHARED FRAMES / SLOTS	TOTAL	CENTRAL STORAGE	FIXED TOT	FIXED BEL	HV 1M		HV 4K	AUX DASD	AUX SCM	
MIN	2,958	2,664	24	0	0		1	0	0	
MAX	8,391,566	38,164	24	0	100		9,901	0	0	
AVG	3,068,030	8,284	24	0	8		3,647	0	0	
LOCAL PAGE DATA SET SLOTS	TOTAL	AVAILABLE	BAD	NON-VIO	VIO					
MIN	1,802,699	1,802,699	0	0	0					
MAX	1,802,699	1,802,699	0	0	0					
AVG	1,802,699	1,802,699	0	0	0					
SCM PAGING BLOCKS	TOTAL	AVAILABLE	BAD	IN-USE						
MIN	131,072	115,090	0	15,982						
MAX	131,072	115,090	0	15,982						
AVG	131,072	115,090	0	15,982						

Figure 1-6 Sample Postprocessor Paging Activity Report

The following columns are available:

- TOTAL: The total number of central storage frames and auxiliary slots that are used by shared pages.
- CENTRAL STORAGE: The total number of central storage frames that are used by shared pages.



- ▶ **FIXED TOT:** The number of central storage frames that are used by shared fixed pages that are allocated below the 2 G bar. This 2 GB bar is in central storage. All the shared pages are above the virtual bar.
- ▶ **FIXED BEL:** The number of central storage frames that are used by shared fixed pages that are allocated below 16 M.
- ▶ **HV 1M:** The number of central storage frames that are used by shared high virtual 1M pages.
- ▶ **HV 4K:** The number of central storage frames that are used by shared high virtual 4K pages.
- ▶ **AUX DASD:** The number of shared pages that are backed on DASD.
- ▶ **AUX SCM:** The number of shared pages that are backed on Flash Memory.

**Note:** AUX DASD and AUX SCM also are called auxiliary storage slots.

### 1.3.3 Postprocessor virtual storage report changes

High virtual memory usage was part of the Private Area Detail section. Now, it is a separate report section.

The high virtual memory usage now has a separate report section; including the introduction of row “SHARED 1 MB”, and the renamed row “1 MB” to “FIXED 1M” to clarify the meaning.

Figure 1-7 shows a sample report of new section. The MEMORY OBJECTS data includes the MIN, MAX, and AVG values for the number of PRIVATE, SHARED, and COMMON memory objects that are allocated with the indicated job. Fixed memory objects and shared memory objects that are backed in 1M frames are also reported.

VIRTUAL STORAGE ACTIVITY						
z/OS V2R2		SYSTEM ID RMFA		DATE 02/17/2015		INTERVAL 02.00.000
		RPT VERSION V2R2 RMF		TIME 14.15.00		CYCLE 1.000 SECONDS
PRIVATE AREA DETAIL						
JOB NAME - TESTJOB1 MEMORY LIMIT - 16384P						
NUMBER OF BYTES OF ALLOCATED BLOCKS BY AREA (BELOW 16 MEG)						
SUBPOOL (AREA)	MIN		MAX		AVG	
230	84K	14.15.00	84K	14.15.00	84K	
[...]						
HIGH VIRTUAL MEMORY USAGE (ABOVE 2GB)						
BYTES	MIN		MAX		AVG	PEAK
PRIVATE	10.00M	14.15.00	10.00M	14.15.00	10.00M	10.00M
SHARED	4.000G	14.15.00	4.000G	14.15.00	4.000G	4.000G
COMMON	0	14.15.00	0		0	0
MEMORY OBJECTS						
PRIVATE	8	14.15.00	8	14.15.00	8	
SHARED	2	14.15.00	2	14.15.00	2	
COMMON	0	14.15.00	0		0	
FIXED 1 MB	0	14.15.00	0		0	
SHARED 1 MB	2	14.15.00	2	14.15.00	2	
1 MB FRAMES						
FIXED	0	14.15.00	0		0	
PAGEABLE	0	14.15.00	0		0	

Figure 1-7 Sample Monitor I VSTOR Report Changes Private Area Detail section

## 1.3.4 Monitor III STORM report changes

On Monitor III STORM report, the 1MB MemObj field under System Summary section was updated, the field “Total” was renamed “Fixed” to clarify the meaning, and “Shared” was included to report the number of shared 1MB memory objects in the system, as shown in Figure 1-8.

RMF V2R2 Storage Memory Objects										Line 1 of 78			
Command ==> _										Scroll ==> CSR			
Samples: 100		System: SC76		Date: 06/30/15		Time: 17.36.40		Range: 100		Sec			
----- System Summary -----													
---MemObj---			---Frames---			-1MB MemObj-		--1MB Fixed--		-1MB Pageable-			
Shared		1	Shared		241	Fixed		0	Total		100	Initial	888
Common		85	Common		18947	Shared		0	Common		0	Dynamic	0
			%Used		10.1	Common		0	%Used		0.0	%Used	2.7
-----													
Jobname		Service		---- Memory Objects ----		--- -1MB Frames-		----- Bytes -----					
C Class		ASID		Total	Comm	Shr	1 MB	Fixed	Pgable	Total	Comm	Shr	
OMVS		S SYSTEM	0015	174	0	0	0	0	14	1176M	0	0	0
TCPIP		S SYSSTC	0070	118	4	0	0	0	0	5594M	2583M	0	0
SMSPDSE		S SYSTEM	0008	29	0	0	0	0	0	48.0M	0	0	0
CSF		S SYSSTC	0075	22	1	0	0	0	0	26.0M	1024K	0	0
JES2		S SYSSTC	0031	13	1	0	0	0	0	836M	1024K	0	0
RMF		S SYSSTC	0083	13	5	0	0	0	0	17.0M	5120K	0	0
HZSPROC		S SYSSTC	0018	10	1	1	0	0	0	8206M	1024K	1024K	0
RESOLVER		S SYSSTC	0019	10	1	0	0	0	0	14.0M	1024K	0	0
TN3270		S SYSSTC	0077	10	2	0	0	0	0	1040M	3072K	0	0
KWRES08		T SYSOTHER	0087	10	0	0	0	0	0	16.0M	0	0	0

Figure 1-8 Sample Monitor III STORM Report Changes ISPF report

## 1.3.5 SMF records new fields

Starting in z/OS V2R2, there are new SMF fields on SMF records 71 and 78 to include specific information that is related to High Virtual Shared Memory. You can use these fields to write programs and collect data from SMF records.

### SMF record 71

Table 1-3 lists the new fields name, length, offset, format, and description.

Table 1-3 New SMF 71 fields

Offset dec   hex	Name	Length	Format	Description
1536 600	SMF71S2M	8	floating	Minimum number of shared memory objects that are allocated in the system and can be backed in 1 MB frames
1544 608	SMF71S2X	8	floating	Maximum number of shared memory objects that are allocated in the system and can be backed in 1 MB frames
1552 610	SMF71S2A	8	floating	Average number of shared memory objects that are allocated in the system and can be backed in 1 MB frames
1560 618	SMF71S3M	8	floating	Minimum number of frames in use for shared high virtual 4K pages

Offset dec   hex	Name	Length	Format	Description
1568 620	SMF71S3X	8	floating	Maximum number of frames in use for shared high virtual 4K pages
1576 628	SMF71S3A	8	floating	Average number of frames in use for shared high virtual 4K pages
1584 630	SMF71S4M	8	floating	Minimum number of high virtual shared memory 1 MB pages that are backed in central storage
1592 638	SMF71S4X	8	floating	Maximum number of high virtual shared memory 1 MB pages that are backed in central storage
1600 640	SMF71S4A	8	floating	Average number of high virtual shared memory 1 MB pages that are backed in central storage
...	...	...	...	Existing, unchanged fields
1680 690	SMF71C2M	8	floating	Minimum number of high virtual common memory 1 MB fixed pages
1688 698	SMF71C2X	8	floating	Maximum number of high virtual common memory 1 MB fixed pages
1696 6A0	SMF71C2A	8	floating	Average number of high virtual common memory 1 MB fixed pages
1704 6A8	SMF71C3M	8	floating	Minimum number of high virtual common memory 1 MB pages that are backed in central storage
1712 6B0	SMF71C3X	8	floating	Maximum number of high virtual common memory 1 MB pages that are backed in central storage
1720 6B8	SMF71C3A	8	floating	Average number of high virtual common memory 1 MB pages that are backed in central storage
...	...	...	...	Existing, unchanged fields
1956 7A4	--	4	--	Reserved
1960 7A8	SMF71CPM	8	floating	Minimum number of high virtual common pages in-use
1968 7B0	SMF71CPX	8	floating	Maximum number of high virtual common pages in-use
1976 7B8	SMF71CPA	8	floating	Average number of high virtual common pages in-use
1984 7C0	SMF714KM	8	floating	Minimum number of 1 MB fixed frames that are used on behalf of 4K page requests
1992 7C8	SMF714KX	8	floating	Maximum number of 1 MB fixed frames that are used on behalf of 4K page requests
2000 7D0	SMF714KA	8	floating	Average number of 1 MB fixed frames that are used on behalf of 4K page requests

Offset dec   hex	Name	Length	Format	Description
2008 7D8	SMF71PLM	8	floating	Minimum number of 1 MB pageable pages that are backed by 1 MB pageable frames
2016 7E0	SMF71PLX	8	floating	Maximum number of 1 MB pageable pages that are backed by 1 MB pageable frames
2024 7E8	SMF71PLA	8	floating	Average number of 1 MB pageable pages that are backed by 1 MB pageable frames

## SMF record 78 subtype 2

The description for existing field name R782LGMO (offset 656 decimal, 290 hexadecimal) was updated to represent the number of *fixed* memory objects that can be backed in 1 MB frames. Also, a new field was created after R782LGMO field, as listed in Table 1-4.

Table 1-4 R782LGMO description change and new R782LSMO field

Offset dec   hex	Name	Length	Format	Description
656 290	R782LGMO	40	Mixed(3)	Number of fixed memory objects that can be backed in 1 MB frames
824 338	R782LSMO	40	Mixed(3)	Number of shared memory objects that can be backed in 1 MB frames

Table 1-5 lists the byte structure of Mixed(3) format.

Table 1-5 Mixed(3) format byte structure

Offset dec   hex	Name	Length	Format	Description
0 0	VSDCMIN	8	floating	Minimum number high virtual memory objects / frames
8 8	VSDCNTME	4	binary	Time stamp for minimum value
12 C	--	4	--	Reserved
16 10	VSDCMAX	8	floating	Maximum number of high virtual memory objects / frames
24 18	VSDCXTME	4	binary	Time stamp for maximum value
28 1C	--	4	--	Reserved
32 20	VSDCTOTL	8	floating	Total for all samples (used to calculate the average)

## 1.4 Global Resource Serialization use statistics in RMF monitor III

Performance monitors products can use resources to report certain types of performance data. A heavy user requests Global Resource Serialization (GRS) via API's GQSCAN and ISGQUERY= QSCAN in a Sysplex about ENQ and Latch contention.

Before z/OS V2R2, problems were reported by heavy GQScan usage, which were difficult to analyze because of missing monitoring support in z/OS. It is suggested that performance monitors do not use such macros but instead be a listener of the Event Notification Facility 51 (ENF 51).

GRS Sysplex replaced GRS Ring topology. The new approach meant that one GRS is not informed about other systems' GRS local contention, which generates much conversation through XCF to derive a total contention picture.

To minimize the overhead, z/OS introduced Contention Notification System (CNS), where the installation selects one GRS to be the CNS. The GRS that is in real time is informed about all of the related activities concerning Sysplex contention, as shown in Figure 1-9.

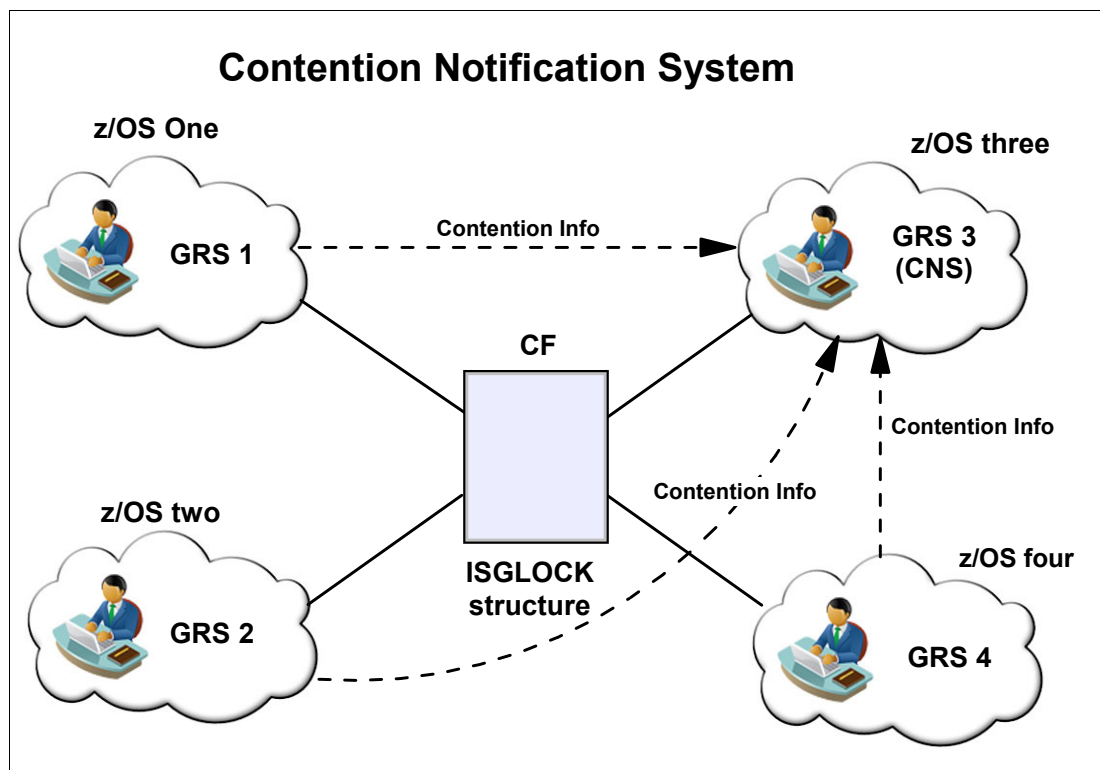


Figure 1-9 GRS Contention Notification System

However, there can be heavy use of resources with this implementation when GRS API's GQSCAN and ISGQUERY= QSCAN are issued. The use can be further increased with the proliferation of performance monitors in one installation.

Performance problems that can occur mainly on large systems. These issues are most problematic to identify because they can be caused by improper usage of those GRS services. The services can be currently started too frequently, have too generic a search, or both.

Snapshot monitoring was available in RMF Monitor II to capture information; however, a solution was needed to improve the reporting and make it easier to understand. The IBM Level 2 and IBM Service Teams designed a new solution to help customers with monitoring GRS activity. They took the information that addresses space-related performance metrics that existed before in Monitor II only and moved it into Monitor III.

In z/OS V2R2, RMF Monitor III introduces a report with job-oriented GRS statistics, the GRS Usage report. It tracks how address spaces are calling the GRS APIs macros to obtain the contention status of resources and requestor of those resources.

## 1.4.1 Usage and invocation

You can view the GQSCAN (and ISGQUERY=QSCAN) information by browsing RMF panels, selecting **option 3** (RMF III), **option 1** (Overview), and **option 4A** (Job Usage).

You can also issue USG from RMF III main menu.

The new RMF Monitor III Job-oriented Usage Report provides detailed GQSCAN (and ISGQUERY=QSCAN) usage statistics on address space level. Also, the report features other job usage statistics regarding CPU, Storage, and I/O resource use. Figure 1-10 shows a sample Job Oriented Usage panel.

RMF V2R2 Job Oriented Usage										
Line 1 of 78										
Scroll ==> CSR										
Command ==> _										
Samples: 100 System: SC76 Date: 06/30/15 Time: 16.51.40 Range: 100										
Jobname	CX	Service Class	--- I/O ---	---	CPU ---	---	- Storage -	-----	QScan	----
			Conn	EXCP	Total	TCB	Total Fixed	Total	Resct	Time
KWRES08	T	SYSOTHER	0.642	17.21	0.18	0.18	3737 17387	0	0.0	0
JES2	S	SYSSTC	0.214	3.24	0.01	0.01	8426 1016	0	0.0	0
XCFAS	S	SYSTEM	0.076	1.47	0.00	0.00	3800 1385	0	0.0	0
RMFGAT	S0	SYSSTC	0.076	0.32	0.09	0.08	12339 175	0	0.0	0
*MASTER*	S	SYSTEM	0.046	0.14	0.02	0.00	3567 1120	2	0.5	17
GRS	S	SYSTEM	0.022	0.00	0.00	0.00	2106 134	0	0.0	0
DUMPSRV	S	SYSTEM	0.016	2.71	0.00	0.00	965 697	0	0.0	0
CATALOG	S	SYSTEM	0.016	0.06	0.00	0.00	1775 203	0	0.0	0
OMVS	S	SYSTEM	0.009	0.70	0.00	0.00	92231 963	0	0.0	0
SMS	S	SYSSTC	0.007	0.91	0.01	0.01	495 86	0	0.0	0
NET	S	SYSSTC	0.005	0.00	0.00	0.00	2811 124	0	0.0	0
PFA	S0	SYSSTC	0.000	2.11	0.01	0.01	1105 109	0	0.0	0
PAGENT	S0	SYSSTC	0.000	0.22	0.00	0.00	3530 17387	0	0.0	0
RMF	S	SYSSTC	0.000	0.00	0.02	0.02	20103 186	0	0.0	0
HZSPROC	S0	SYSSTC	0.000	0.00	0.00	0.00	4761 168	0	0.0	0

Figure 1-10 Job Oriented Usage panel

The fields for new Job Oriented Usage panel are listed Table 1-6.

*Table 1-6 Job Oriented Usage panel fields*

Field	Description
Jobname	Name of the job.
CX	Abbreviation for the job class as follows: <ul style="list-style-type: none"> <li>▶ S: Started task</li> <li>▶ T: TSO</li> <li>▶ B: Batch</li> <li>▶ A: ASCH</li> <li>▶ O: OMVS</li> <li>▶ O as second character: Indicates that the address space is using OMVS services</li> </ul>
Service Class	The name of the service class for this address space.
I/O Conn	Device connect time, in seconds, for this address space in the report interval.
I/O EXCP	Number of EXCP operations per second for this address space in the report interval.
CPU Total	Amount of total processor time, in seconds, for this address space in the report interval as sum of TCB time, global and local non-preemptible SRB time, Enclave SRB and Client SRB time.
CPU TCB	Amount of TCB processor time, in seconds, for this address space in the report interval.
Storage Total	The number of active and idle frames for this address space, averaged over the report interval.
Storage Fixed	Number of fixed frames for this address space, averaged over the report interval.
QScan Total	Total number of QScan requests for this address space, including START and RESUME, but not QUIT requests. <p>These requests are SCANACTION=START/RESUME/QUIT:</p> <ul style="list-style-type: none"> <li>▶ START Indicates to start a search of the GRS queues.</li> <li>▶ RESUME indicates to resume a previously started search.</li> <li>▶ QUIT indicates to quit a previously started search. If a started search did not complete, it must be resumed until it completes or is ended by using QUIT.</li> </ul>
QScan Resct	Average number of resources in contention that is returned by QScan requests for this address space.
QScan Time	Average QScan request time, in microseconds, for this address space. This data is important performance data.

Alternatively, the browser-based version of the report can be requested from the RMF Distributed Data Server (DDS). All metrics in the report can be selected as individual metrics for continuous monitoring from the following applications:

- ▶ RMF Data Portal for z/OS
- ▶ z/OSMF Resource Monitoring

The z/OS V2R2 enhancements help system programmers to more easily identify address spaces that are using the GQSCAN services and analyze the performance and improper usage and to potentially resolve any issues. Also, general job resource use is visible at a glance.

## **1.5 RMF zFS monitoring enhancements**

For more information about the new zFS monitoring functions with RMF in z/OS V2R2, see Chapter 4, “zFS Monitoring Enhancements” on page 41.





# Workload Manager

This chapter describes the enhancements that were implemented for Workload Manager (WLM) that is running in z/OS V2R2.

WLM is a z/OS component that performs dynamic workload management for transaction goals as declared at WLM policy.

This chapter includes the following topics:

- ▶ 2.1, “WLM basic concepts” on page 20
- ▶ 2.2, “WLM large memory enhancements” on page 20
- ▶ 2.3, “IEAOPT parameters available to monitoring products” on page 22
- ▶ 2.4, “WLM buffer pool management” on page 22
- ▶ 2.5, “WLM health-based routing services” on page 27
- ▶ 2.6, “Improvements in z/OS V2.R2 IWM4HLTH service” on page 29

## 2.1 WLM basic concepts

WLM is a z/OS component that works with System Resource Manager (SRM), which is another z/OS component, to perform the role of managing business and infrastructure goals to optimize performance within a single z/OS system or across multiple z/OS systems in a Parallel Sysplex environment. WLM performs the following tasks:

- ▶ Manage the transactions goal, as defined in a WLM policy construct named Service Class. WLM determines dynamically the relative priorities of transactions in internal z/OS queues. This WLM function is called *Policy Adjustment*.
- ▶ The WLM/SRM function Resource Adjustment minimize the waste of system resources.
- ▶ Assist Transaction Managers in implementing WLM Dynamic Workload Balancing function between z/OS systems and their server address spaces.

## 2.2 WLM large memory enhancements

A z13 central electrical complex can have up to 10 TB of memory and a z/OS in a logical partition up to 4 TB. This increase is significant and large memory must be traded by CPU cycles, specifically by avoiding I/O operations. In addition, z/OS 2.2 must adjust to this growth. In this section, we describe the WLM modifications to improve the use of memory.

Large memory enhancements were made to meet the following goals:

- ▶ Provide a simpler set of memory targets with the IEAOPT parameter
- ▶ Encourage more of the workload to run in parallel

Some of the IEAOPTxx parameters have percentage values that make it more difficult to define targets for large logical partitions (LPARs). WLM in z/OS V2R2 changes how some IEAOPTxx members in Parmlib are processed towards the level of page fixing. These changes simplify setting storage targets with the following OPT parameters:

- ▶ Pageable storage and real storage constants: MCCFXTPR=xxx, which is a percentage of fixed memory.
- ▶ MPL adjustment constants:
  - RCCFXTT=(xxxx,yyyy): Low and high threshold of total fixed memory
  - RCCFXET=(xxxx,yyyy): Low and high threshold of fixed memory below 16 M

### 2.2.1 IEAOPTxx parameter MCCFXTPR

Fixing a page in memory means that this page cannot be stolen, and might be done to preserve integrity. However, fixing many pages of 1 M large pages can cause a memory availability problem to z/OS. Those pages cannot be stolen if there is a pageable memory shortage situation.

The MCCFXTPR keyword in the IEAOPTxx Parmlib member specifies the percentage of memory that can be page fixed before a pageable storage shortage is detected and message IRA400E is issued.

In this state, SRM performs the following tasks:

- ▶ Rejects new address spaces creation by denying **START**, **LOGON**, and **MOUNT** commands
- ▶ Avoids Job selection by initiators
- ▶ Swaps out the swappable address space with the greatest number of fixed pages

The current MCCFXTTPR default of 80% requires that 20% (100 minus MCCFXTTPR) of memory remain pageable (not fixed), regardless of the absolute amount of memory. In systems with large amounts of memory, the MCCFXTTPR default of 80% can result in a pageable storage shortage situation being detected when there is still plenty of pageable memory.

In z/OS 2.2, the required pageable memory threshold for large memory LPARs (larger than 320 GB) is always equal or less than 64 GB. This threshold allows you to fix much more memory and reduces the risk of experiencing a pageable storage shortage state.

Therefore, if 100% minus MCCFXTTPR is the total amount of online frames and is greater than 64 GB for large memory logical partition, the MCCFXTTPR keyword is not used in determining the threshold at which a shortage of pageable storage exists and the 64 GB figure is used instead.

The OA44668 rollback to z/OS V2R1 is available.

## 2.2.2 IEAOPT parameter RCCFXTT

The RCCFXTT keyword in the IEAOPTxx Parmlib member specifies the low and high threshold of *total* fixed memory. SRM uses these thresholds to determine whether the system multiprogramming level (MPL) must be increased (address spaces can be swapped in) or decreased (address spaces can be swapped out). The default is 66% and 72%.

Consider the following points:

- ▶ On small systems, such percentages are not a problem.
- ▶ On a 1 TB image (for example), these percentages mean that when 660 GB of memory is fixed, the system is no longer under pressure, and SRM stops decreasing the MPL.

In z/OS 2.2, the SRM can calculate the optimal RCCFXTT lower and upper threshold. The RCCFXTT=AUTO definition activates the threshold calculation by using the system resources manager (SRM). However, the default is still RCCFXTT=(66, 72). It is likely that in the future the default is zero.

## 2.2.3 IEAOPT parameter RCCEXTT

The RCCEXTT keyword in the IEAOPTxx Parmlib member specifies the low and high threshold of fixed real memory *below* 16 M. SRM uses these thresholds to determine whether the system MPL must be increased or decreased and whether the shortage memory situation is occurring. The default is 82% and 88%.

This OPT keyword is also enhanced, mainly to stay consistent with the RCCFXTT keyword. The RCCEXTT=AUTO definition activates the threshold calculation by SRM. The default is still RCCEXTT=(82, 88); which is the same as AUTO.

Then, the MCCFXTTPR change allows more memory to be fixed than before without incurring a shortage message. The changed OPT parameter logic for RCCEXTT and RCCFXTT makes it much easier to define the optimal threshold. An optimal memory threshold increases the MPL and more workload can run in parallel.

## 2.3 IEAOPT parameters available to monitoring products

A good understanding of IEAOPT member keywords is encouraged to help you gain better system performance and a workload and resource balance. Performance monitors, such as RMF and Omegamon in releases before z/OS V2R2, use their own software interface to externalize the IEAOPTxx parameters in the user windows and reports.

z/OS V2R2, WLM provides a new Workload Management Service IWM4OPTQ to make the IEAOPTxx parameters available to z/OS monitoring products. The effect of this availability reduces the implementation and maintenance effort whenever WLM adds a new IEAOPTxx Parmlib option or modifies a Parmlib option.

## 2.4 WLM buffer pool management

A buffer pool is a set of same size I/O buffers in virtual storage. Buffer pools are used as a cache to help avoid I/O.

### 2.4.1 WLM buffer pool management

The goal of buffer pool management in DB2 for z/OS is to keep application data available and closer to the processors and reduce (or preferably eliminate) physical I/Os that are issued to retrieve data from peripheral devices. This configuration improves service levels and reduces transaction elapsed time.

WLM determines whether buffer pool I/O is the predominant delay for a service class period. If so, it uses histograms to determine whether increasing the buffer pool size can help to better achieve the performance goal.

Because WLM has a view of relevant resources (including memory), it can help to manage their buffer pools. This management can decide dynamically the amount of buffers per buffer pool, as shown in Figure 2-1 on page 23.

# WLM Buffer Pool Management

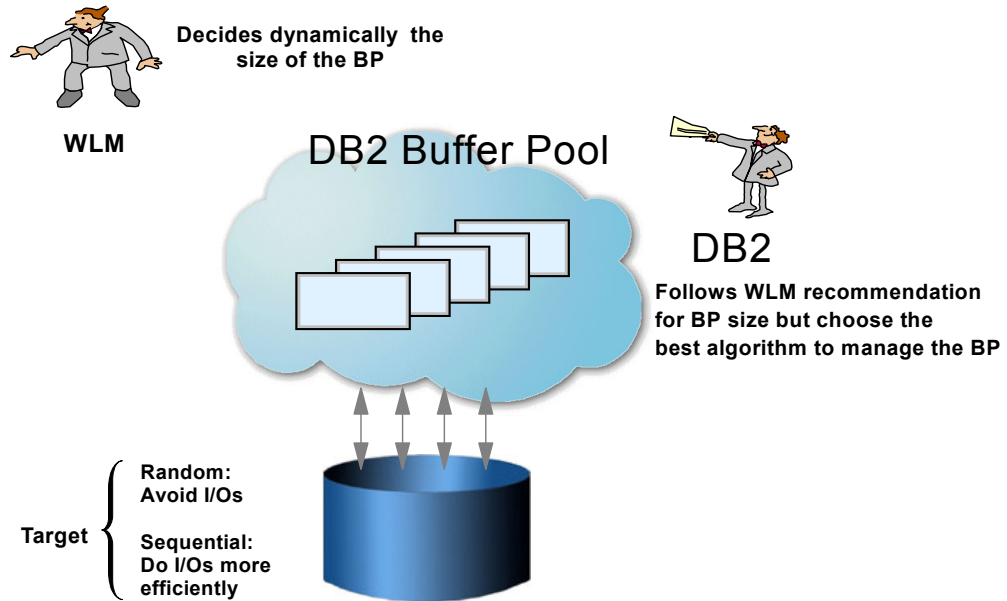


Figure 2-1 WLM buffer pool management for DB2

This capability is activated for each DB2 buffer pool when it is altered by the **ALTER BUFFERPOOL** command with the AUTOSIZE option set to YES. In this case, after the **ALTER** command, DB2 registers the buffer pool with WLM, meaning that DB2 accepts WLM recommendations about the buffer pool size and feeds information back to WLM about this buffer pool's hits and misses, as shown in Figure 2-2 on page 24.

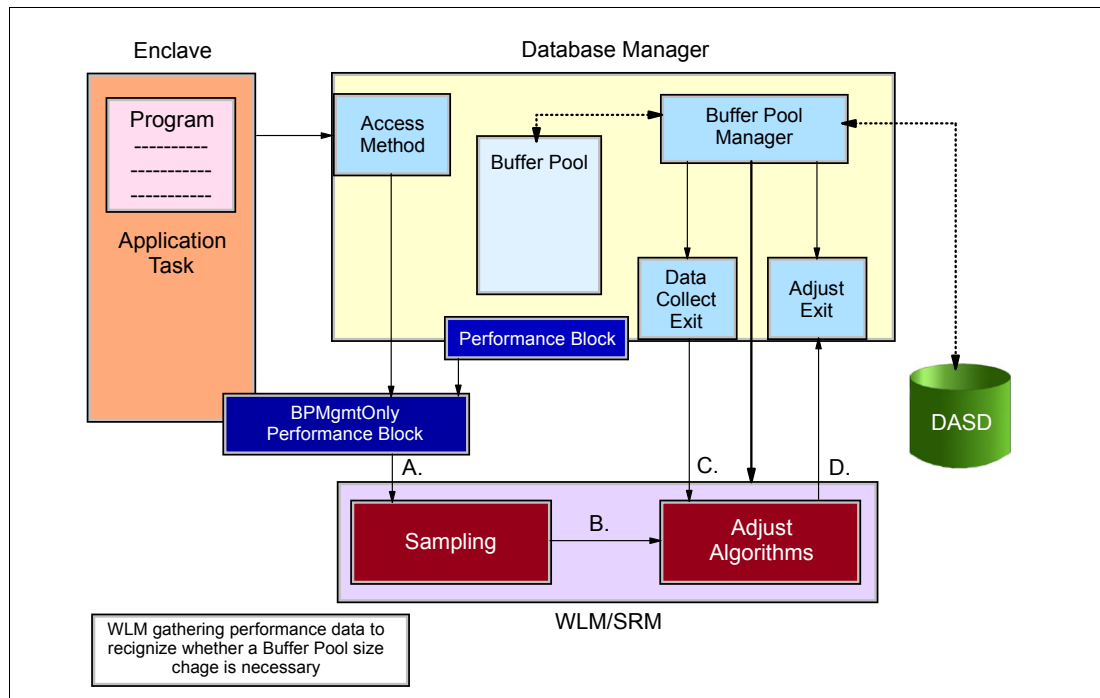


Figure 2-2 A view of WLM BP management

WLM buffer pool management does not completely replace the tools that are used to configure, monitor, and tune buffer pool size. However, when you initially sized your buffer pools, DB2 and WLM can fine-tune the buffer pool size based on long-term trends and steady state growth.

Then, WLM automatically increases or decreases buffer pool size (as appropriate) by as much as 25% of the originally allocated size for DB2 V10 and according to minimum and maximum values that are specified in DB2 V11.

As Transaction Managers (IBM CICS®, IBM IMS™, WebSphere Application Server, DDF) run transactions that demand DB2 data, DB2 reports to WLM that a delay occurred to the transaction task because of a DB2 buffer pool read miss, which forces a synchronous I/O wait.

At WLM Policy Adjustment routine (every 10 seconds), the most poorly performing (Performance Index greater than 1.0) in the highest importance Service Class Period (SCP) is selected for help.

WLM analyzes the various delays (CPU, I/O, server address space service, DB2 buffer pool) that are tracked against the SCP transactions. If the largest portion of delays were the result of particular buffer pool I/O waits, WLM can trigger an alter (increase) to the size of such buffer pool. When this situation occurs, the DB2 DSNB555I message is issued. Consider that DB2 does not necessarily use up the complete size of a buffer pool as recommended by WLM. The size that is recommended by WLM is an upper limit for DB2 to use. However, DB2 decides by its own algorithm how much of this recommended size it needs. DB2 only GETMAINS the space it uses.

A buffer pool can be decreased in size if WLM observes, for example, that other SCPs need memory.

Other than for that reason, WLM-managed DB2 buffer pools were rarely shrunk before z/OS V2R1. In z/OS V2R1, a housekeeping function was added to shrink buffer pools when their usage or misses slowed down.

However, that version of the housekeeping function did not consider the importance of the SCP transactions that were using the buffer pool.

The enhancements that are included in z/OS V2R2 provide the solution that if multiple buffer pools qualify to be shrunk, the choice is made based on the one used by the least important SCP.

When the **ALTER BUFFERPOOL [VPSIZE(s)] AUTOSIZE(YES)** command is issued in the new z/OS V2R2 implementation, the following process occurs:

- ▶ DB2 registers the buffer pool to WLM:
  - MIN size =  $0.75 \times \text{VPSIZE}$
  - MAX size =  $1.25 \times \text{VPSIZE}$
  - Initial USED size between MIN size and MAX size, the size that is used by DB2 when the buffer pool is registered
- ▶ WLM makes the following recommends:
  - Grow the size of the buffer pool when the Performance Index of an SCP is affected and buffer pool delays are a significant contributor. WLM grows the size of the buffer pool up to at most MAX size.
  - Shrink the size of the buffer pool:
    - Because of memory donation to a suffering SCP
    - Because of regular housekeeping cycles. For more information, see 2.4.2, “Housekeeping function improvement” on page 26.

WLM shrinks the size of the buffer pool down to at most:

- Initial USED size during regular housekeeping
- MIN size if memory donation to a suffering SCP is needed, when system is memory constraint.

- ▶ DB2 decides how much of that recommended size it uses, based on its own algorithms:
  - The used (by DB2) size of a buffer pool can be below its WLM recommended size
  - DB2 GETMAINS up to the used size, not up to the recommended size.
- ▶ When the command **ALTER BUFFERPOOL AUTOSIZE(NO)** is issued, DB2 unregisters the buffer pool from WLM management.

The size of buffer pool does not change and DB2 keeps the last recommended buffer pool size. There also is a new VPSIZE for reallocation.

As shown in Figure 2-3 on page 26, the WLM recommendation is not taken immediately by DB2 and the minimum and maximum sizes are respected.

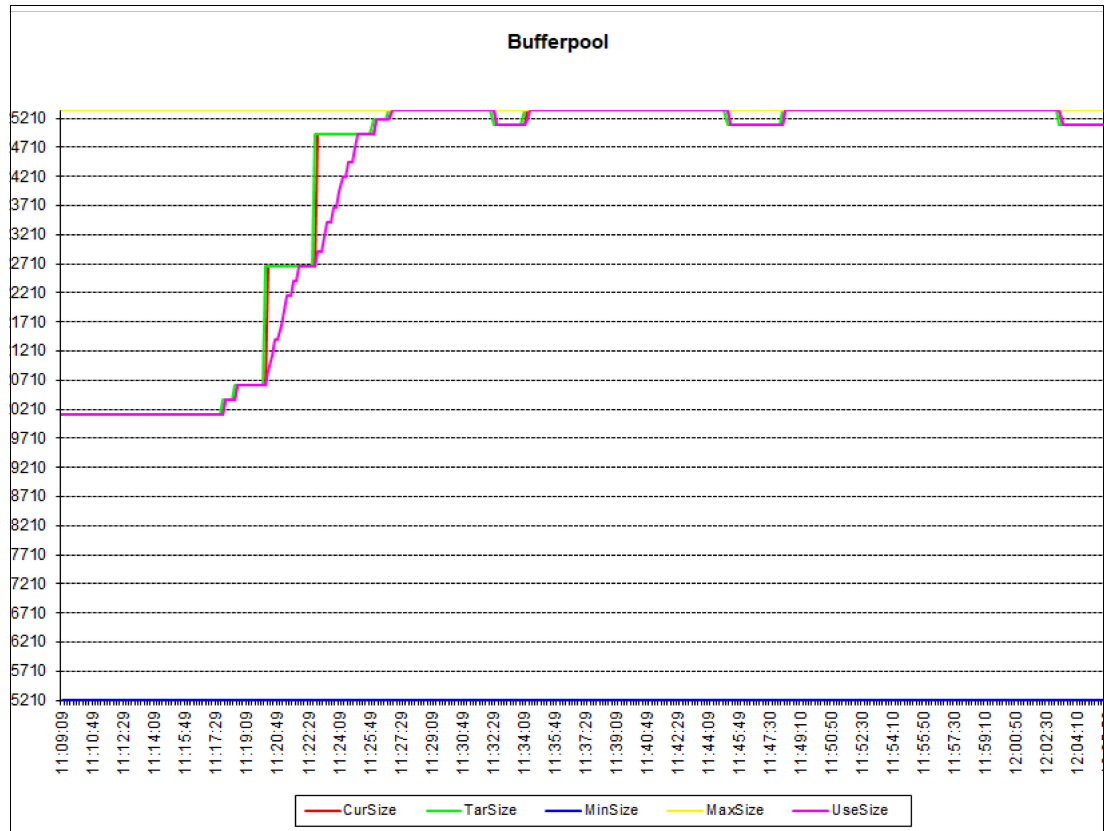


Figure 2-3 Example of WLM/DB2 Bufferpool Management

## 2.4.2 Housekeeping function improvement

This housekeeping function is improved with z/OS V2R2 when shrinking a buffer pool. Consider the following points:

- ▶ A buffer pool can be shrunk at most every 5 minutes. Therefore, when a buffer pool is shrunk by the housekeeping function, it is not considered for housekeeping within the next 5 seconds.
- ▶ At every 10 seconds, WLM selects a buffer pool candidate based on the following criteria:
  - If found, one that has zero references.
  - If found, one that has zero or insignificant number of delays.
  - Some buffer pool such that periods having delays from the buffer pool are not affected in their PI when housekeeping recommends reducing the size of the buffer pool. If multiple such buffer pools are found, the one that is used by the least important periods is shrunk. The latter function aims to reduce buffer pools that are still in use but can become oversized for the usage pattern in the period.
- ▶ If a candidate buffer pool for shrinking is found, it is checked whether not all the space up to the recommended size is used by DB2:
  - If the unused size does not change for 5 minutes, give recommendation to decrease by 50% of unused size.
  - If unused size does not change for 10 minutes, give recommendation to decrease by 100% of unused size.



- Else, if the buffer pool has no unused space, give recommendation to decrease by 1% of maximum size.
- ▶ If the buffer pool must grow within the next 5 minutes after a housekeeping action, WLM increases the check interval (until it is eligible for a housekeeping action again) by an extra 5 minutes.
- ▶ With this housekeeping function, dormant buffer pools shrink over time while AUTOSIZE(YES).

## 2.5 WLM health-based routing services

In this section, we describe some concepts regarding WLM routing services.

### 2.5.1 WLM routing services overview

In Figure 2-4, we see the terminology and the participants of the WLM Routing Services function.

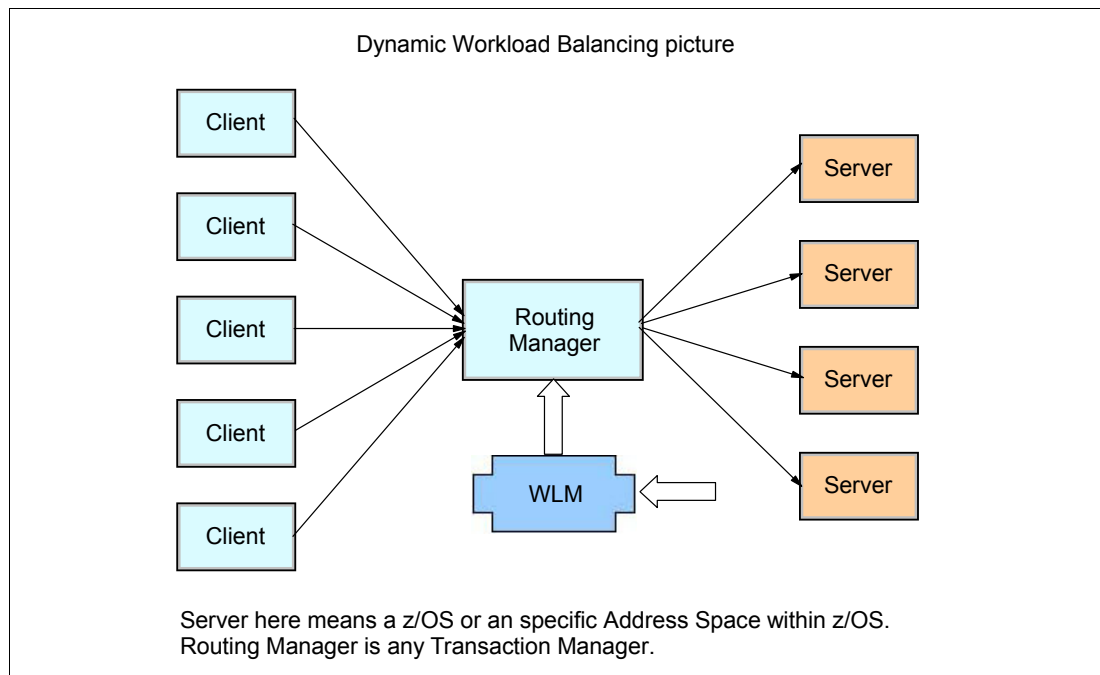


Figure 2-4 Dynamic Load Balancer example

The primary objective of WLM dynamic routing services is to dynamically balance the load among servers. This balancing is aims to improve the following aspects:

- ▶ Capacity: Route work according to capacity available
- ▶ Performance: WLM goal attainment because of the following factors:
  - Less queue time
  - Less PU time because of less transaction concurrency
  - Less Transaction Manager overhead
- ▶ Availability: Avoid shortages
- ▶ Reliability: Avoid not healthy servers to be routed

WLM Routing Services is a WLM function that is used by Transaction Managers (exploiters). These exploiters ask for guidance about where to run (servers) an arrival transaction, online sessions, and network connection ports.

WLM responds with Weights (1 - 64) that are associated to each server by using APIs, such as IWM4SRSC. The higher the Weight number of a server, the better to run transactions in such recipient. These Weights are used by Transaction Managers to distribute transactions proportionally among servers along a time interval. After expiring, new Weights set are requested by WLM, as shown in Figure 2-5.

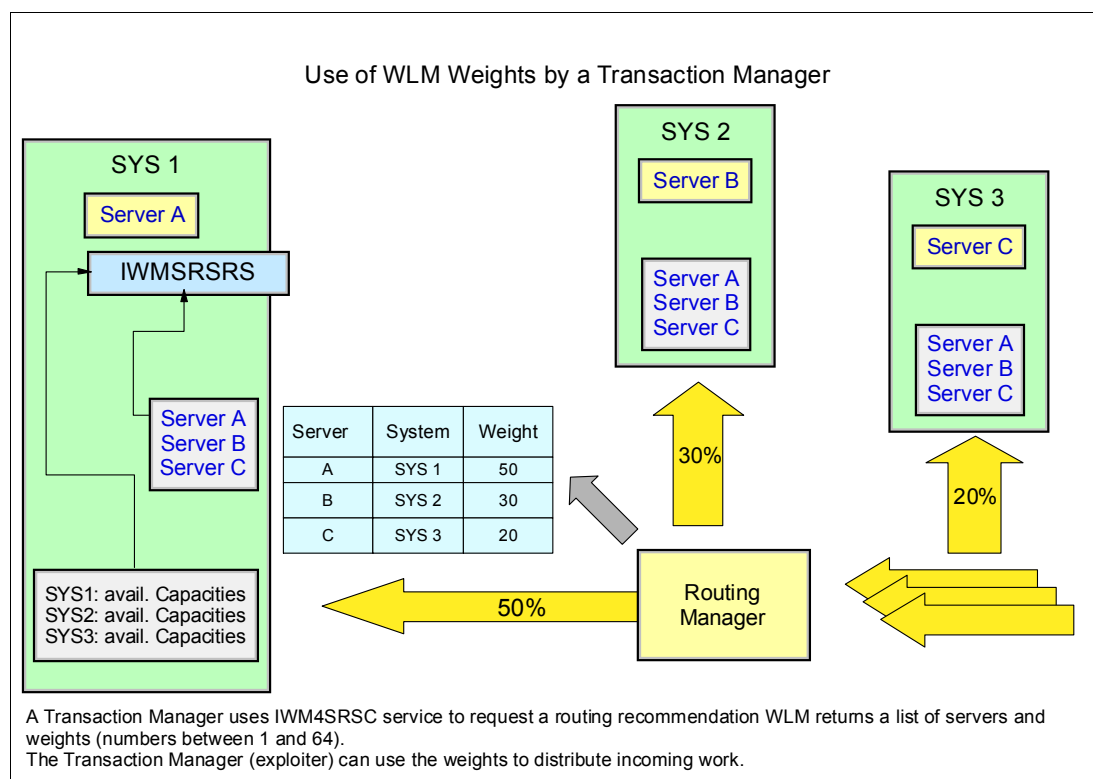


Figure 2-5 Use of WLM Weights by a Transaction Manager

To be an exploiter, the Transaction Manager program must register to WLM Routing Services through IWMSRSRG API macro.

The granulator of the returning Weights is per multiple z/OS systems on Sysplex and one or more address spaces per z/OS system.

The following formula for the IWM4SRSC service is used:

$$\text{Weight} = \text{Available LP Capacity Factor} * \text{PI Factor} * \text{Enclaves\_Queue\_Time\_Ratio} * \text{Health\_Factor}$$

## 2.5.2 Health factor effecting routing Weight

The health indicator can be set per server address space. Consider the following points:

- Health=100 is the default and remains in effect until a different value is set via IWM4HLTH API macro, often issued by the Transaction Manager. The IWM4HLTH service is used to inform WLM about the health state of a server.

Each IWM4HLTH invocation replaces previous health indicator values.

- If the health indicator of a server is less than 100, its capability of running transactions is reduced.
- The server Weight is reduced by applying a factor of health that is divided by 100.

An example of a IWM4HLTH macro is shown in Example 2-1.

*Example 2-1 IWM4HLTH macro example*

---

```
IWM4HLTH STOKEN=STKN,FUNCTION=SET,HEALTH=80, X
SUBSYS=SUBTYPE,SUBSYSNM=SUBSNAME, X
HEALTHRSN=HRSN,RETCODE=RC,RSNCODE=RSN
```

---

The health indicator is a number that shows the percentage of how well the server is performing, as shown in Figure 2-6.

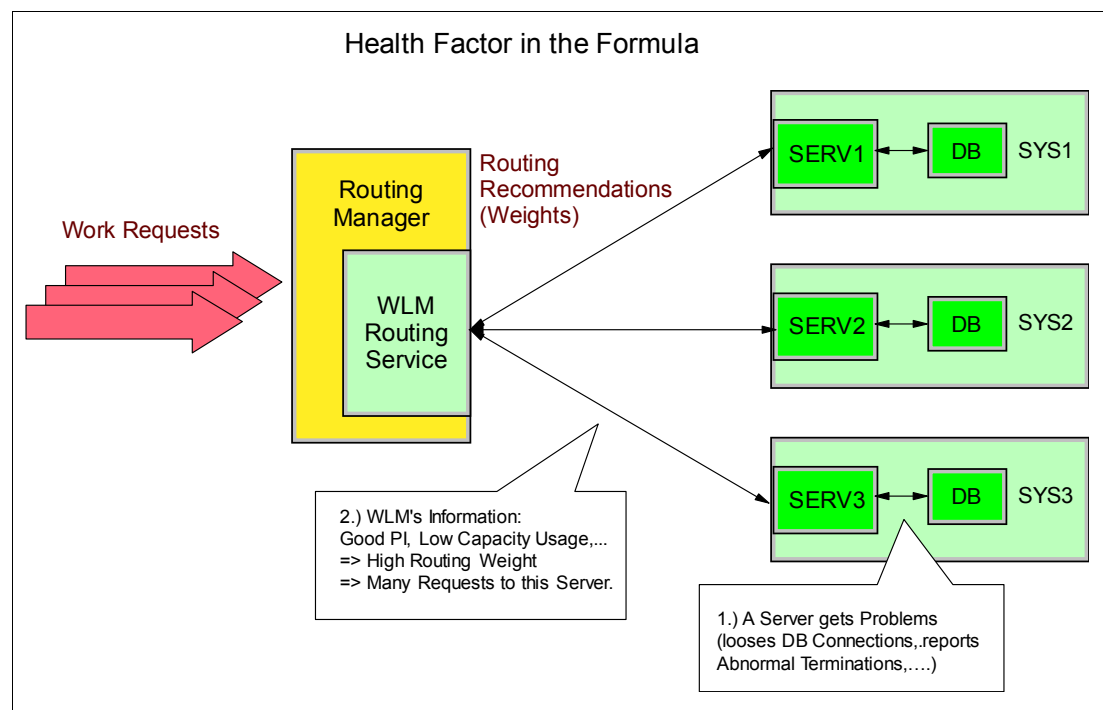


Figure 2-6 Health factor in the Weight formula

## 2.6 Improvements in z/OS V2.R2 IWM4HLTH service

z/OS V2R2 enhancements to dynamic workload routing reduce the business effect of middleware-server health issues by quarantining a severely affected server and rerouting work to a healthy back-end server. Before z/OS V2R2, the programming model for the WLM health services was that only one component (usually the server) provided the health for a specific address space.

In z/OS V2R2, the notion of finger-pointing is introduced, which allows multiple internal and external components to report on the health of a server. The IWM4HLTH service is enhanced to enable multiple callers to report on a server's health, identifying themselves, and providing reasons for their health ratings. Health ratings from different callers are tracked separately, and for workload routing, the minimum of all current health ratings take effect.

For example, with z/OS V2R2, XCF, and XES use the enhanced IWM4HLTH service to reduce the health value of an address space that is associated with an affected XCF group member or XES connector.

IWM4HLTH features the following improvements:

- ▶ Tracks different health values (for example, reported by XCF/XES and the server).
- ▶ The algorithm for determining the health indicator for an address space is changed. The health indicator is the minimum value of all the current settings from the different callers of the service since the last RESET. This change affects the current operation where the server's health indicator is the last value reported.
- ▶ Another function of IWM4HLTH refers to RAS considerations regarding a server's health state. The RESET function restarts setting of a composite health value by specifying an initial value and discarding the values that are reported by other callers.
- ▶ Callers can identify themselves by a subsystem type and subsystem name. WLM uses these parameters to recognize different callers of the service. If no subsystem type is passed, the job name of the caller address space is used instead.
- ▶ Callers can specify the reason or cause of change of the health indicator.
- ▶ Users of this service need to check their programs for sufficient program authorization.

### 2.6.1 New IWM4QHLT query service

WLM also provides a new query service (IWM4QHLT) to obtain reported health indicators for diagnostic purposes. Callers of this service can obtain health values for particular address spaces or for all spaces for which a health value was set.

On z/OS V2R2, Runtime Diagnostics starts the IWM4QHLT service. If any servers have a current health value less than 100, it displays a SERVERHEALTH event in its output along with any other events it finds. As another feature, some of the Predictive Failure Analysis (PFA) checks start Runtime Diagnostics when PFA determines that the checks current value is too low. The new SERVERHEALTH event is returned to PFA by Runtime Diagnostics for any servers with a current health value less than 100 and the event is included in the PFA check's exception report.



## Real Storage Management scalability

In this chapter, we describe how an installation can use the performance advantages of large memories. Also included are the results of a large memory study.

The enhancements in Real Storage Management (RSM) at z/OS V2R2 are also described because of the challenge of supporting and using large memories, such as 4 TB per z/OS system. RSM is a z/OS component that manages such memory.

This chapter includes the following topics:

- ▶ 3.1, “Overview” on page 32
- ▶ 3.2, “Performance gains because of large memories” on page 32
- ▶ 3.3, “Improving performance by using large memories” on page 33
- ▶ 3.4, “Large memory studies” on page 33
- ▶ 3.5, “RSM modifications at z/OS V2R2” on page 37

## 3.1 Overview

Changing business landscapes increase demand for memory usage and improved parallelism in z/OS. In-transaction analytics, subsecond response times, and greater demand because of mobile applications all increase the need for more data with better performance.

Superb Large Memory and n-Way scaling on IBM z Systems™ enable the wide-scale deployment that allows z Systems to host Business Analytics, IT Analytics, Cloud, and big data applications, and serve as the back-end for mobile applications.

z/OS V2R2 recognizes this demand for larger memory and was enhanced to allow logical partitions with up to 4 TB of memory. As a consequence, Real Storage Management (RSM) major algorithms must be modified to keep such components scalable with such large-size memory and allow an increased tasks concurrency. RSM is a z/OS component that is responsible for managing memory (also referred to as *central storage*).

The RSM attempts to use and reuse memory in the most efficient way possible to help your workload perform efficiently. The following major metrics of memory occupancy are important:

- ▶ Highest average UIC close to 64 K seconds
- ▶ Average available frame count of MAXSPACE + 2 x AFQOK can be good for performance, depending on your workload mix
- ▶ Total page fault rate close to zero

Workload efficiency is better achieved by using the available memory and not having to wait for memory to be available.

## 3.2 Performance gains because of large memories

The changes to RSM more efficiently use memory and enable the use of more memory on each z/OS system, which can in turn improve performance. Consider the following points:

- ▶ Substantial latency (delay) reduction for online transaction processing (OLTP) workloads implies significant response time reductions and increased transaction rates.
- ▶ In-memory databases feature dramatic gains in response time by avoiding synchronous I/O waits.
- ▶ Batch window reduction means less queue time and shorter elapsed times for Jobs.
- ▶ Reduced time to insight for analytic workloads:
  - Data is processed more efficiently, which helps organizations keep pace with the influx of data
  - The time it takes to get from raw data to business insight is reduced

Machines, such as the IBM z13, are designed to optimize the use of the caches. It is possible to decrease the workload concurrency due to quicker response times. The quicker response times are helped by the larger memory, which in turn, enables the processor units to run faster. The decreased concurrency lessens the competition for the caches.

### 3.3 Improving performance by using large memories

The following key recommendations can be used to use the benefits of large memory in a z/OS system:

- ▶ Page fix DB2 buffer pools at current number of buffers  
Gain 0 - 6%, clients can see 2 - 3% CPU benefit for buffer pools with high I/O activity. Use Flash or other real memory to mitigate any real memory concerns that are preventing you from page fixing DB2 buffers. If you keep fixed the pages containing I/O buffers fixed you can save the processor consumption for pagefix and pagefree processing for every I/O operation. IBM performance study testing for large memory assumes that all buffer pools are page fixed.
- ▶ Deploy 1M large pages for page fixed DB2 buffers  
Gain another 1 - 2% CPU benefit. Large pages improve the performance of E-DAT along virtual address translations and decrease the processor time.
- ▶ Deploy pageable 1M pages (requires Flash Express)  
Gain up to 1% with 1 MB pages for DB2 11 executable code with z/OS 2.1. Expect to gain another CPU benefit when z/OS 2.2 delivers shared above the bar 1M pageable pages that are used by DB2 (DBM1 and DDF address spaces).
- ▶ Increase size of DB2 local buffer pools to up to 100 GB  
In Sysplex data sharing, also increase the size of Global Buffers Pools in the Coupling Facility (CF) enough to support the local buffer pool size. Gain up to 5%, depending on workload profile and tuning. Be aware that 100 GB per DB2 can mean up to 1 TB or more per z/OS in the CF.

### 3.4 Large memory studies

The information that is shown in Figure 3-1 on page 34, Figure 3-2 on page 35, Figure 3-3 on page 36, and Figure 3-4 on page 37 was obtained as a result of a Large Memory Study that was presented in *IBM zEnterprise System: Performance Report on Exploiting Large Memory for DB2 Buffer Pools with SAP*.

Significant performance improvements were seen when more memory was used for larger DB2 buffer pools while running the SAP Banking Services (SBS) Day Posting workload on an IBM zEnterprise EC12 (zEC12).

The study relates to a customer representative of financial services workload with memory intensive, large number of DB2 tables, and random I/O behavior.

The following test scenarios were used:

- ▶ DB2 V11 Single System
- ▶ 256 GB/512 GB/1024 GB real storage; started with 256 GB of real storage
- ▶ DB2 V11 Single System; minimal number of buffer pools

Improvements were seen when running in single system and data sharing environments. Our measurements showed the following results:

- ▶ Reductions in response time of up to 70%
- ▶ Increases in transaction rates of up to 37%
- ▶ Savings in CPU time per transaction of up to 25%

These performance improvements were a result of up to a 97% reduction in synchronous I/O.

Figure 3-1 shows Posting External Throughput Rate (ETR) versus total buffer pool size. (Posting ETR means in millions the number of DB2 accesses per hour.)

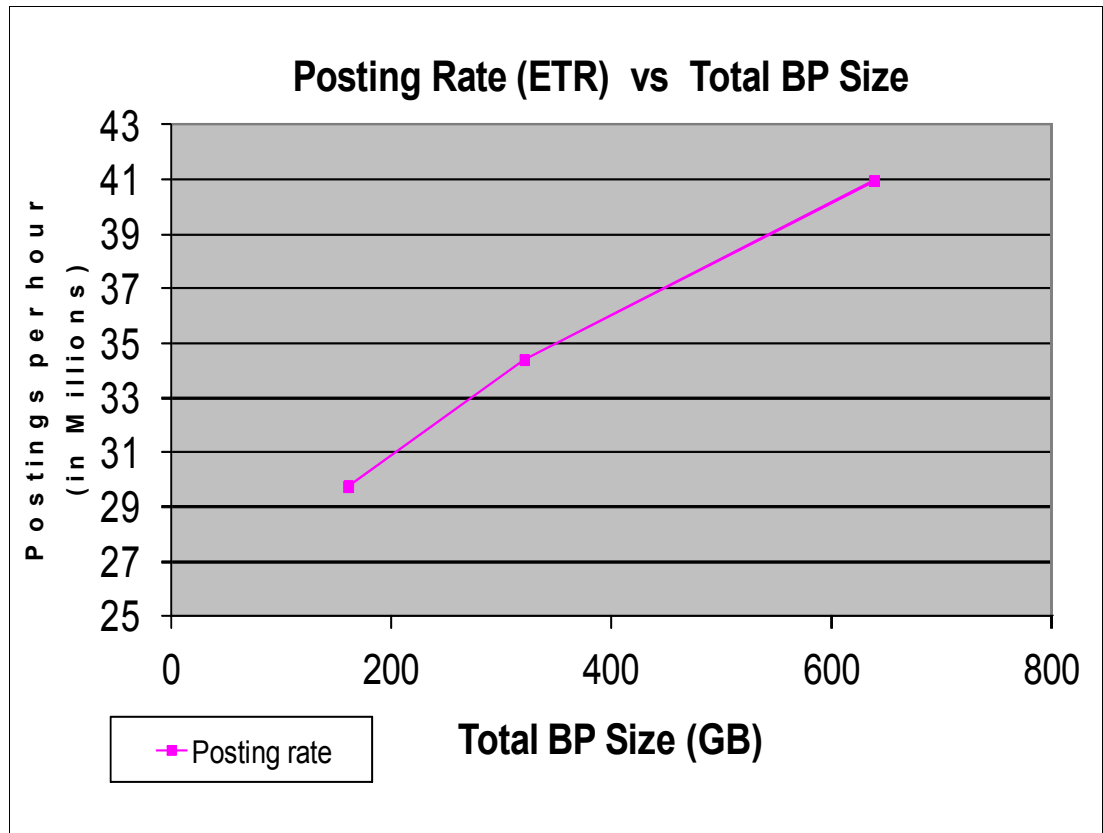


Figure 3-1 Posting ETR versus Total BP size



As shown in Figure 3-2, there is a correlation between transaction response time and the total buffer pool size.

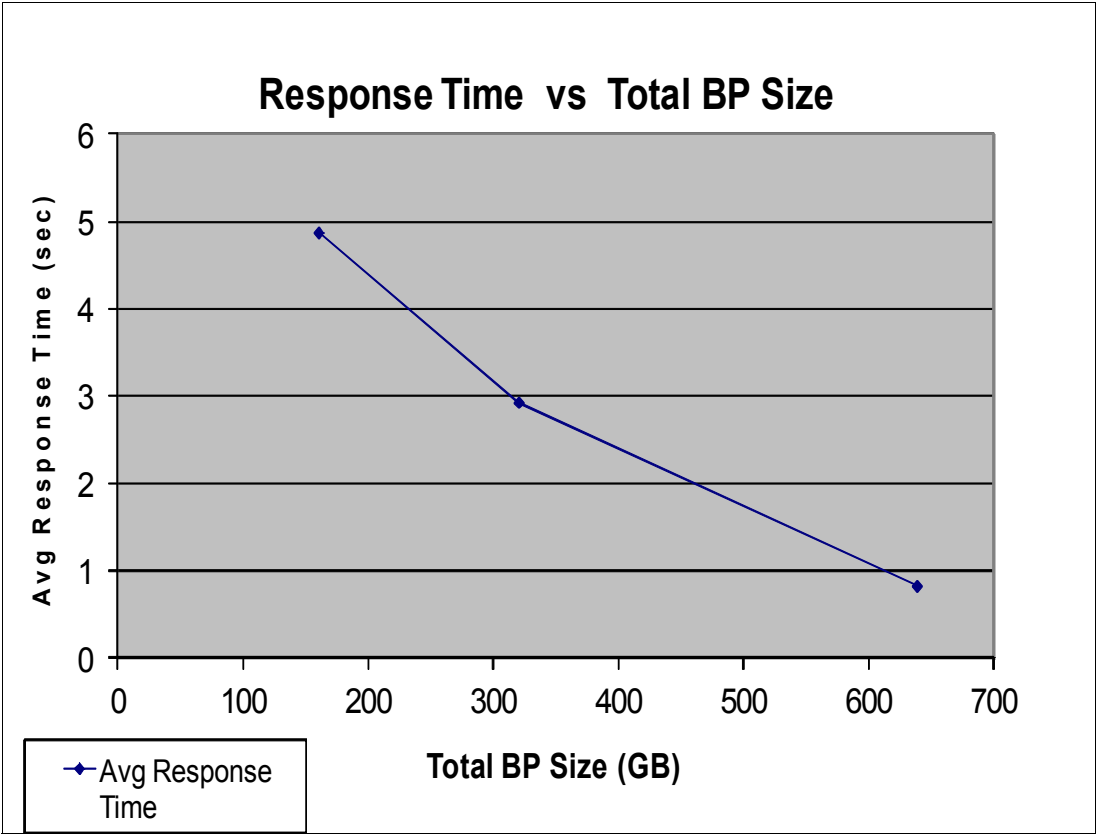


Figure 3-2 Response time versus total buffer pool size

As shown in Figure 3-3, there is a correlation between Internal Throughput Rate (ITR) and the buffer pool size and transaction response time and the buffer pool size.

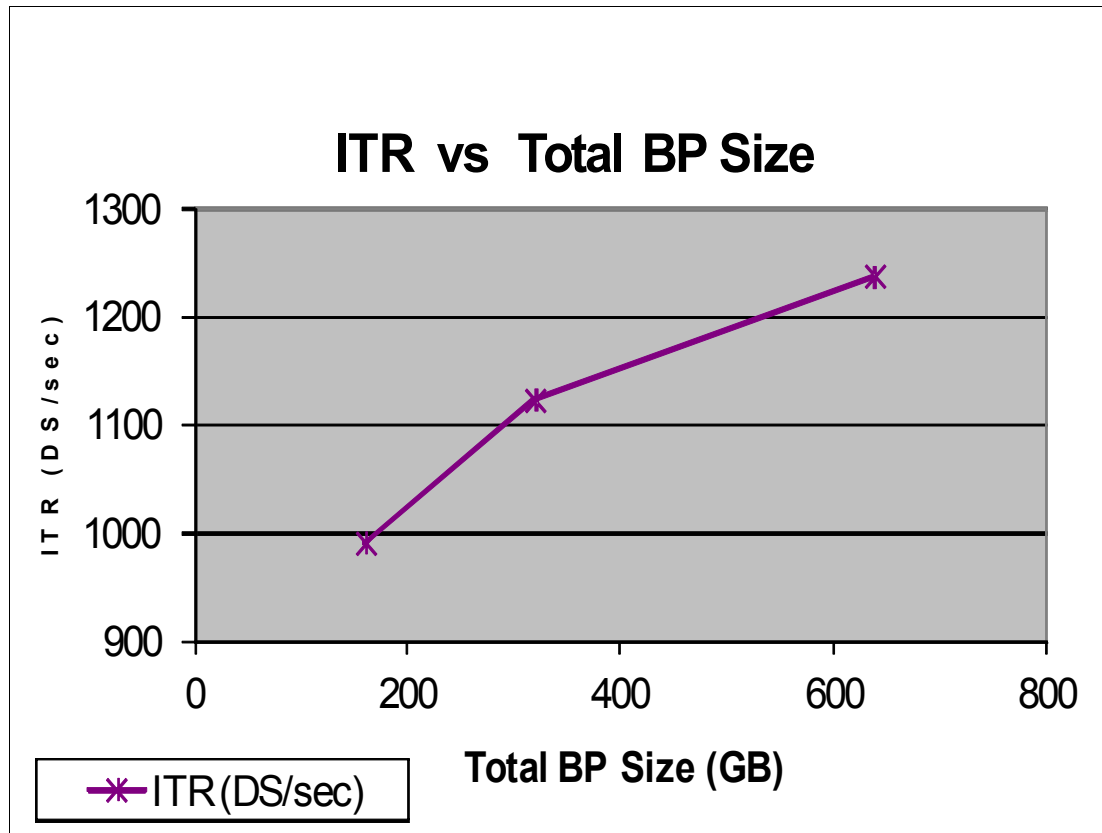


Figure 3-3 ITR versus total buffer pool size

ITR is a processor unit metric that is defined at the large System Performance Reference IBM project by using the following formula:

ITR = Number of processed transactions / total PU time to process such transactions

Often in a Save MSU project, you might want to decrease the ITR and not the average PU usage (because of the hidden demand).

In Figure 3-4, there is a correlation between DB2 request time and the total buffer pool size.

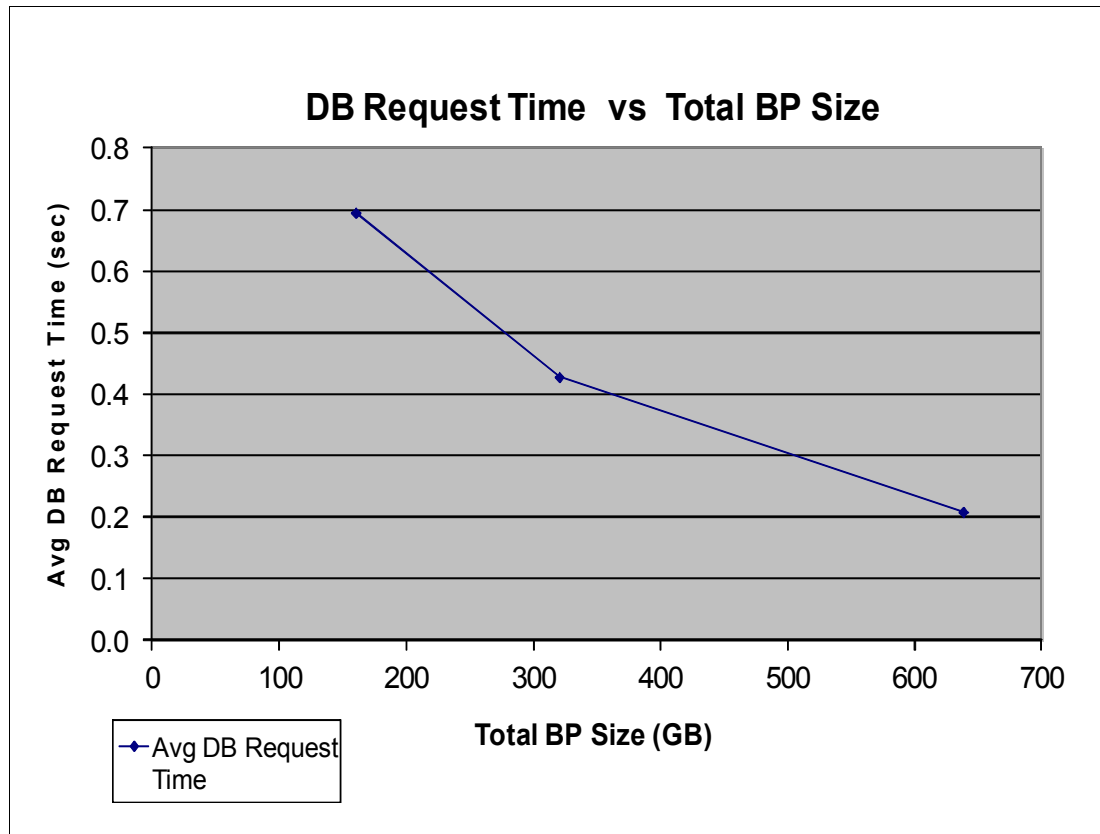


Figure 3-4 Database request time versus total buffer pool size

DB2 request time is the time from running the Program Call (PC) instruction to synchronously starting the DBM1 address space (activating cross memory mode) until the Program Return (PR) instruction is issued. The PR instruction ends the cross memory mode and returns control to the DBM1 caller (client).

## 3.5 RSM modifications at z/OS V2R2

In z/OS V2R2, RSM introduces the following functions to support scalability the large amounts of memory:

- ▶ Dumping Large Memory via stand-alone utility program
- ▶ RSM initialization of memory is more efficient; the processing time while running disabled is reduced by starting frames more efficiently. For more information, see APAR OA44207.
- ▶ Management of 1 MB frames is improved. The area of real storage that is used to contain the Page Frame Table Entry (PFTE)s for the Large Frame Area (LFAREA) is moved to a new area to help prevent the depletion of preferred storage that is below the bar.
- ▶ Processing for the **CONFIG STOR,OFFLINE** and **DISLAY MATRIX** commands was enhanced to be more efficient for large amounts of memory. For more information, see APAR OA44207
- ▶ Frame steal processing is enhanced to be more efficient in scanning PFTEs by skipping over certain areas that are not eligible for steal processing.

RSM was internally modified to increase concurrency at the system and application tasks levels. Consider the following points:

- ▶ Multiple page faults can be dealt with concurrently within an address space or with common storage.
- ▶ Page fixing and unfixing can occur concurrently within an address space or with common storage.
- ▶ GETMAIN or FREEMAIN of storage can occur concurrently with page fixing or freeing and page faults.
- ▶ There is less contention on available frame queues.

### 3.5.1 Dumping Large Memory by using Standalone Dump utility

As applications use more virtual memory, memory dumps become larger and require more real memory and page data sets. If this issue is an issue for you, consider the following suggestions:

- ▶ Avoid dumping the common area.
- ▶ Leave more real memory available for taking memory dumps.

In z/OS V2R2, there is a new REAL parameter option for stand-alone memory dump. Consider the following points:

- ▶ Default of REAL=ALL dumps all real storage.
- ▶ REAL=USED dumps only used real storage and skips any unused (available) real frames, which is recommended.
- ▶ Can be set with REAL=USED when ADSADMP is generated or at run time by using SET REAL(USED).

Additionally, some performance improvements were made to stand-alone memory dump; however, this improvement varies from system to system.

### 3.5.2 Change in default for LFAREA INCLUDE1MAFC(YES)

Available frame queue (AFQ) is a logical queue of frames that are available for hosting pages that are coming from page data sets (page-ins), swap-ins, or are referenced the first time after a GETMAIN.

AFQ is in memory and each frame is represented by an RSM control block that is named PFTE. AFC is a counter that indicates the number of 4 KB frames in the available frame queue. When the AFC is below a threshold that is named AVQLOW, the page stealing process is run. Each frame has 4 KB; however, the frames might be in the following groups:

- ▶ 256 contiguous 4 KB unity to keep 1M large pages
- ▶ 1024 x 2 x 256 contiguous unity 4 KB to keep 2 G large pages

The LFAREA parameter at IEASYSxx Parmlib member specifies the amount of memory that is available at initial program load (IPL) to reserve for backing 1M pages and 2G pages. Before z/OS V2R2, these groups of 4 KB frames were not included in the AFC by default.

#### **APAR OA41968 (before z/OS V2R2)**

This APAR solves the problem that is described in this section. We suggest that you always check for superseding APARs to gain the current technical status and perspective.

RSM does not break up 1M frames in the LFAREA until there are no 4 KB frames available. However, RSM begins page stealing when the AFC drops below AVQLow. This issue can in significant paging stealing (because of the lack of 4 KB available frames), even when there is an abundance of 4 KB frames that are grouped in chunks of 1M available on the system.

The rationale behind waiting until AVQLow is to avoid the use of a limited resource that is reserved for 1M pages until it was necessary. Although RSM attempts to coalesce broken up 1 MB frames, there is no guarantee that coalescing is successful, especially if one of the 4KB frames making up the fixed 1M page is fixed long term when there is demand for such 1 M page.

The solution in this APAR involves changing the content of the available frame count (AFC), by specifying that the 1M set of frames are to be included in the AFC, which increases its value.

Enablement is done through a new Parmlib value INCLUDE1MAFC (YES) that is specified in the IEASYSxx LFAREA parmliB parameter.

By installing this PTF and specifying this new parameter value, system behavior changes in the following manner:

- ▶ RSM performs less paging when there is an abundance of available 1M frames in the memory.
- ▶ RSM is more likely to break up fixed 1M frames to satisfy 4K page demand.

For RMF users, the PTF for RMF APAR OA42510 must be applied before specifying INCLUDE1MAFC. RMF uses the RCEAFC to generate some of their reports. Not applying OA42510 might lead to incorrect RMF reports.

The STGTEST SYSEVENT returns information about the amount of storage that is available in the system. If INCLUDE1MAFC is specified, available fixed 1M pages are included in this amount. SYSEVENT is an API use in the communication with SRM. In this case, some code is asking about the amount of available memory.

### **z/OS V2R2 modification on INCLUDE1MAFC option**

The modifications is to make such option a default. For compatibility with applicable releases before z/OS V2R2, you can also specify INCLUDE1MAFC(YES) as INCLUDE1MAFC.

IBM recommends taking the default, or specifying INCLUDE1MAFC or INCLUDE1MAFC(YES).





## zFS Monitoring Enhancements

The z/OS Distributed File Service File System (zFS) is a z/OS UNIX file system that is used in parallel with the hierarchical file system (HFS). This chapter describes the purpose and usage of the new zFS monitoring functions with Resource Measurement Facility (RMF) in z/OS V2R2.

This chapter includes the following topics:

- ▶ 4.1, “Overview of zFS monitoring enhancements” on page 42
- ▶ 4.2, “Using the zFS monitoring enhancements” on page 42

## 4.1 Overview of zFS monitoring enhancements

Before z/OS V2R2, RMF Monitor III reports included areas in which enhancements can improve to assist users to gain more value. Improvement included the following goals:

- ▶ The single system reports ZFSSUM and ZFSACT offered no possibility to monitor details of zFS related to sysplex awareness of zFS file systems.
- ▶ Performance problems were sometimes experienced when zFS performance data was gathered.

In z/OS V2R2, new RMF Monitor III Sysplex reports are provided with detailed and aggregated zFS data for various aspects of zFS performance analysis.

The following new reports are available:

- ▶ ZFSOVW: Provides wait and cache statistics.
- ▶ ZFSFS: Shows file system statistics.
- ▶ ZFSKN: Features with zFS kernel statistics.

RMF uses new zFS APIs to gather zFS performance data.

**Note:** RMF online monitoring can be used to identify performance problems that are related to zFS distributed file systems.

## 4.2 Using the zFS monitoring enhancements

In this section, we describe the setup and use of the enhancements.

### 4.2.1 RMF Monitor III Data Gatherer option

Users can control whether they want Monitor III to collect zFS activity data by specifying the following data gathering option:

ZFS | NOZFS

Consider the following points:

- ▶ If no option is coded, the default setting ZFS is used.
- ▶ This default value ZFS is set in the sample parmlib member that is named ERBRMF04, as shown in Figure 4-1 on page 43.



```

/*****
/* NAME:          ERBRMF04
/*
/*
/* DESCRIPTION: PARMLIB MEMBER WITH RMF MONITOR III GATHERER OPTIONS
/*              (ALL OPTIONS ARE SET TO DEFAULTS)
/*
...
/*****
CYCLE(1000)          /* SAMPLE EVERY SECOND (1000 MSEC)
DATASET(STOP)        /* NO DATASET SUPPORT
DATASET(NOSWITCH)    /* APPEND TO LAST NON-FULL DATASET
DATASET(WHOLD(7))    /* CONTROLS BUFFER PAGES IN STORAGE
...
VSAMRLS             /* ACTIVATE VSAM RLS GATHERING
OPD                 /* ACTIVATE OMVS PROCESS DATA GATHERING
PCIE                /* ACTIVATE PCIE DATA GATHERING
SCM                 /* ACTIVATE SCM DATA GATHERING
ZFS                /* ACTIVATE ZFS DATA GATHERING
NOSGSPACE           /* NO STORAGE GROUP SPACE GATHERING
NOLOCK              /* NO LOCK DATA GATHERING

```

Figure 4-1 Sample parmlib member ERBRMF04

## 4.2.2 RMF Monitor III zFS Sysplex Reports

To displays the RMF Sysplex Report Selection menu (as shown in Figure 4-2) select **S** from the RMF Monitor III Primary menu. Consider the following points:

- ▶ To request the Monitor III ZFSOVW, ZFSFS, and ZFSKN reports, select option **13**, **14**, or **15**.
- ▶ As an alternative, enter one of the following commands:
  - ZFSOVW
  - ZF0, ZFSFS
  - ZFF, ZFSKN
  - ZFK

RMF Sysplex Report Selection Menu			
Selection ==>			
Enter selection number or command for desired report.			
Sysplex Reports			
1	SYSSUM	Sysplex performance summary	(SUM)
2	SYSRTD	Response time distribution	(RTD)
3	SYSWKM	Work Manager delays	(WKM)
4	SYSENQ	Sysplex-wide Enqueue delays	(ES)
5	CFOVER	Coupling Facility overview	(CO)
6	CFSYS	Coupling Facility systems	(CS)
7	CFACT	Coupling Facility activity	(CA)
8	CACHSUM	Cache summary	(CAS)
9	CACHDET	Cache detail	(CAD)
10	RLSSC	VSAM RLS activity by storage class	(RLS)
11	RLSDS	VSAM RLS activity by data set	(RLD)
12	RLSLRU	VSAM LRU overview	(RLL)
13	ZFSOVW	zFS Overview	(ZF0)
14	ZFSFS	zFS File System	(ZFF)
15	ZFSKN	zFS Kernel	(ZFK)

Figure 4-2 RMF Sysplex Report Selection menu

### 4.2.3 RMF Monitor III ZFSOVW Report

The Monitor III ZFSOVW report is shown in Figure 4-3.

Command ===> █		RMF V2R2	zFS Overview	- PLEX75	Line 1 of 2	
					Scroll ===> CSR	
Samples: 100		Systems: 2	Date: 08/15/15	Time: 21.01.40	Range: 100	Sec
System	-----Wait%-----			-----Cache Activity-----		
	I/O	Lock	Sleep	---User---	--Vnode--	-Metadata-
				Rate Hit%	Rate Hit%	Rate Hit%
SC74	100	0.0	1.2	2.730 100	6.300 100	1.360 94.1
SC75	0.0	0.0	87.9	0.100 100	3.830 100	0.420 100

Figure 4-3 Monitor III ZFSOVW report

Consider the following points regarding this report:

- ▶ It provides a summary of zFS activity, wait percentages, and cache statistics about the current sysplex.
- ▶ This information helps to discover bottlenecks and tune cache behavior.

#### RMF Monitor III ZFSOVW fields

From this panel, you can obtain four other reports that are displayed in pop-up windows with more detailed statistics. You get these reports via the following cursor-sensitive fields for the four sections that are shown in Figure 4-3:

- ▶ Wait% I/O for displaying an I/O Data statistic report.
- ▶ User cache Rate and Hit% for displaying a user data cache statistic.
- ▶ Vnode cache Rate and Hit% for showing a vnode cache statistic.
- ▶ Metadata cache Rate and Hit% for presenting metadata cache statistics.

**Note:** The sensitive fields are marked with red curves for system SC74, as shown in Figure 4-3.

## 4.2.4 Monitor III ZFSOVW I/O data

The Monitor III ZFSOVW I/O data pop-up panel displays a breakdown of I/O requests into three major request types: file system metadata, log file, and user file data, as shown in Figure 4-4.

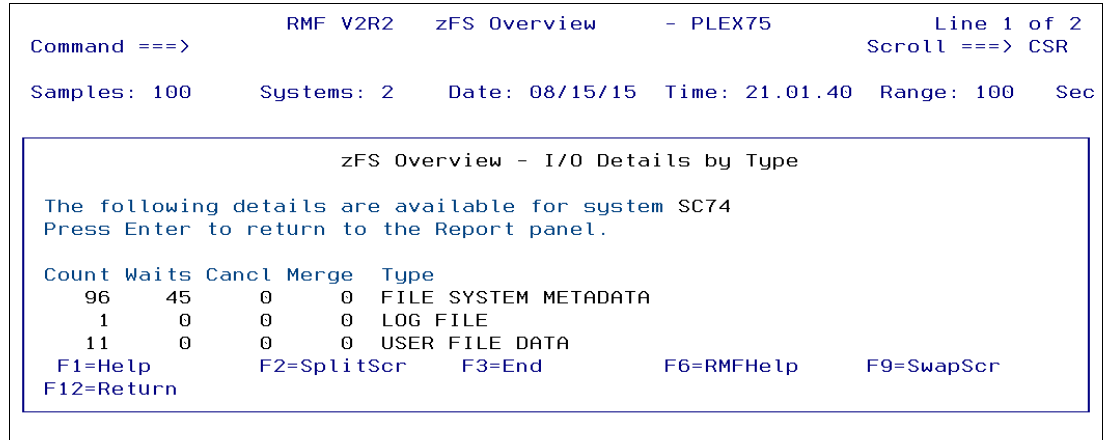


Figure 4-4 Monitor III ZFSOVW I/O data

## 4.2.5 Monitor III ZFSOVW User Cache data

The Monitor III ZFSOVW User cache pop-up panel displays details of the user file cache activity, as shown in Figure 4-5. In zFS, the user file cache is used to cache all regular files.

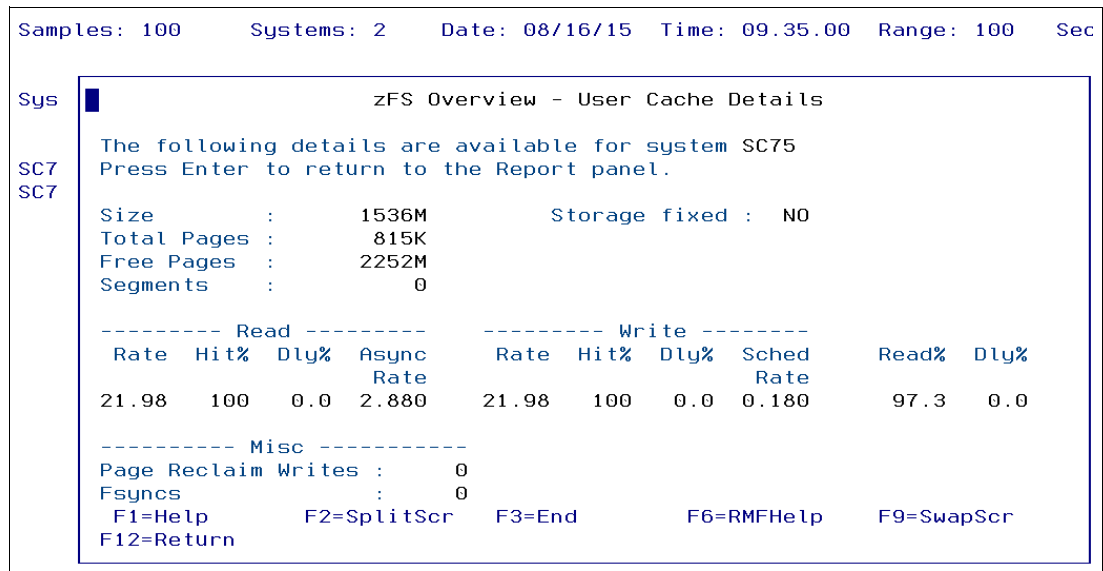


Figure 4-5 Monitor III ZFSOVW User Cache data

**Note:** With heavy workloads, the read hit ratio might be shown much lower. However, this issue does not necessarily mean that there is a problem with the setup.

## 4.2.6 Monitor III ZFSOVW vnode Cache data

The Monitor III ZFSOVW vnode cache pop-up panel displays details of the vnode cache activity. This zFS cache operates on vnodes, which is the data structure to represent file system objects. Sample data is shown in Figure 4-6.

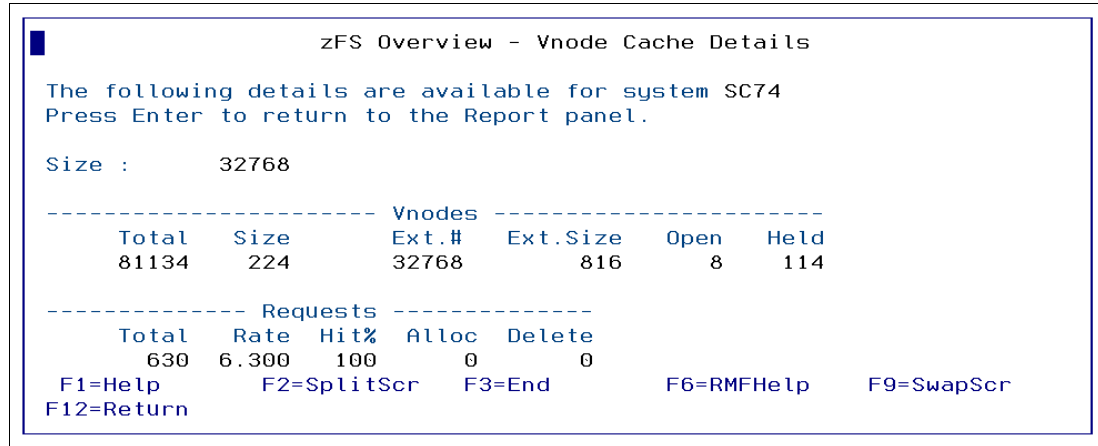


Figure 4-6 Monitor III ZFSOVW vnode Cache data

## 4.2.7 Monitor III ZFSOVW metadata cache data

The Monitor III ZFSOVW metadata cache pop-up panel displays details of the metadata cache activity. This zFS cache contains all file system metadata; for example, directory contents and file status information, as shown in Figure 4-7.

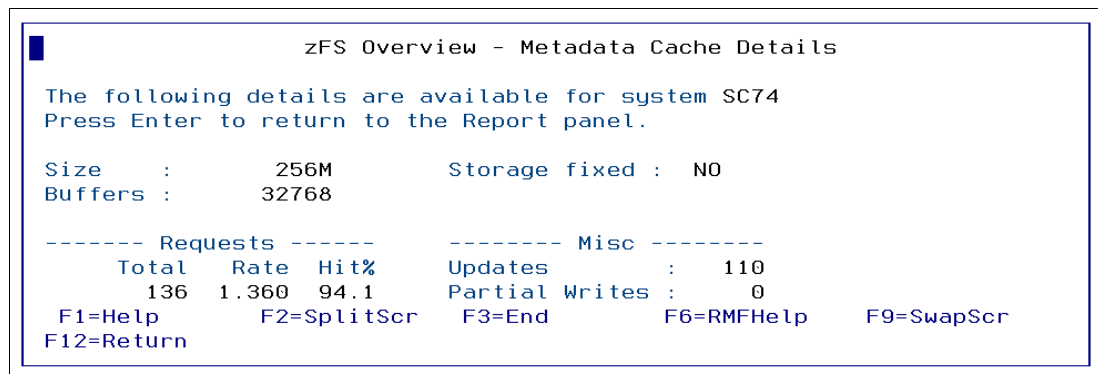


Figure 4-7 Monitor III ZFSOVW metadata Cache data

## 4.2.8 Monitor III Online Help for zFS panels

For more information about all of the fields and the data that is available on the RMF III zFS monitoring panels, see the Online Help.

**Tip:** On all of the RMF III zFS monitoring panels, pressing the Help key (PF1) provides more information about the fields and the information that is shown in these reports.

## 4.2.9 RMF Monitor III ZFSFS report

The Monitor III ZFSFS report provides detailed measurements of zFS activity based on single file systems. This information helps to monitor performance and capacity limits of file systems.

You can access the ZFSFS display by using the following methods:

- ▶ Enter ZFSFS or ZFF in the RMF Monitor III Primary menu.
- ▶ Show the Sysplex panel by selecting option **S** and then option **14**.

The data display also can be customized in zFS file system panels by using one of the following methods:

- ▶ Enter the **ROPTIONS** or **R0** command.
- ▶ Select the name mask that you want to display; for example, HERING.\*.
- ▶ Point to a specific file system to view information for that specific file system.

Figure 4-8 shows an example display of this panel.

RMF V2R2 zFS File System - PLEX75									
Line 1 of 66									
Command ==> █ Scroll ==> CSR									
Samples: 100 Systems: 2 Date: 08/16/15 Time: 09.35.00 Range: 100 Sec									
----- File System Name -----									
System Owner Mode Size Usg% I/O Rate Resp Time Read % XCF Rate									
BBN.V7R0.CONFIG1.ZFS									
*ALL SC74 RW S 416M 95.4 0.000 0.000 0.0 0.000									
CEA.HFS									
*ALL SC74 RW S 1440K 12.2 0.000 0.000 0.0 0.000									
CICST42A.HFS									
*ALL SC74 R0 155M 18.7 0.000 0.000 0.0 0.000									
HERING.TEST.PRELE.ZFS									
*ALL SC74 RW S 26M 3.4 0.000 0.000 0.0 0.000									
HERING.TEST.RW.ZFS									
*ALL SC74 RW S 18M 85.8 0.000 0.000 0.0 0.000									
HERING.TEST.ZFS									
*ALL SC74 RW S 35M 98.0 0.000 0.000 0.0 0.000									
HERING.ZFS									
*ALL SC74 RW S 128M 77.4 0.000 0.000 0.0 0.000									
JES2.ZFS									
*ALL SC74 RW S 7200K 2.4 0.000 0.000 0.0 0.000									

Figure 4-8 RMF Monitor III ZFSFS report

## Displaying information for file systems according to a mask

Figure 4-9 shows the resulting display if the mask setting is changed from ALL to HERING.\* and requesting a detailed view by setting the value for Detail to YES.

RMF V2R2 zFS File System - PLEX75										Line 1 of 16	
Command ==> █										Scroll ==> CSR	
Samples: 100		Systems: 2		Date: 08/16/15		Time: 11.26.40		Range: 100		Sec	
----- File System Name -----											
	System	Owner	Mode	Size	Usg%	I/O Rate	Resp Time	Read %	XCF Rate		
HERING.TEST.PRELE.ZFS											
	*ALL	SC74	RW S	26M	3.4	0.000	0.000	0.0	0.000		
	SC74	SC74	RW S	26M	3.4	0.000	0.000	0.0	0.000		
	SC75	SC74	RW S		0.0	0.000	0.000	0.0	0.000		
HERING.TEST.RW.ZFS											
	*ALL	SC74	RW S	18M	85.8	0.000	0.000	0.0	0.000		
	SC74	SC74	RW S	18M	85.8	0.000	0.000	0.0	0.000		
	SC75	SC74	RW S		0.0	0.000	0.000	0.0	0.000		
HERING.TEST.ZFS											
	*ALL	SC74	RW S	35M	98.0	0.000	0.000	0.0	0.000		
	SC74	SC74	RW S	35M	98.0	0.000	0.000	0.0	0.000		
	SC75	SC74	RW S		0.0	0.000	0.000	0.0	0.000		
HERING.ZFS											
	*ALL	SC74	RW S	128M	77.4	0.000	0.000	0.0	0.000		
	SC74	SC74	RW S	128M	77.4	0.000	0.000	0.0	0.000		
	SC75	SC74	RW S		0.0	0.000	0.000	0.0	0.000		

Figure 4-9 Display with mask setting HERING.\* and Detail YES

## 4.2.10 RMF Monitor III ZFSKN report

The Monitor III ZFSKN report provides measurements about counting the calls that are made to zFS from z/OS UNIX and the average response time of zFS requests.

The reports track sysplex-wide zFS usage with detailed and aggregated zFS data for various aspects of zFS performance analysis.

You can access the ZFSKN display by using one of the following methods:

- ▶ Enter ZFSKN or ZFK in the RMF Monitor III Primary menu.
- ▶ Show the Sysplex panel by selecting option **S** and then option **15**.

Figure 4-10 shows an example of this kernel calls report.

Command ==> <span style="color: green;">█</span>						
RMF V2R2 zFS Kernel - PLEX75				Line 1 of 2		
				Scroll ==> CSR		
Samples: 100		Systems: 2	Date: 08/16/15	Time: 11.40.00	Range: 100	Sec
System Name	- Request Rate -		--- XCF Rate ---		- Response Time -	
	Local	Remote	Local	Remote	Local	Remote
SC74	1.010	0.000	0.000	0.000	22.00	0.000
SC75	0.130	4.020	0.000	0.300	24.00	26.00

Figure 4-10 RMF Monitor III ZFSKN report

#### 4.2.11 zFS user and metadata considerations

The two most important zFS caches are the user cache and the metadata cache. Consider the following points:

- ▶ The user file cache is used to cache all of the regular files.
- ▶ The metadata cache is used to contain all file system metadata; this metadata includes all directory contents, file status information (such as, atime, mtime, size, and permission bits), and file system structures.

**Managing zFS file systems:** Based on Direct I/O (DIO), by continuously enhancing the processes of managing data and the new 64-bit addressing, there is no need to look for the zFS ownership of a file system if it is mounted sysplex-aware, also known as rwshare.

Even with the z/OS V2R2 enhancements it is suggested to look for ownership for metadata because many functions must be processed by the owning system.

For example, in cases of complex **find** processing in a sysplex sharing environment, it is still suggested to always attempt performing the search processing for a specific file system on the zFS owning system to avoid XCF overhead.









# Related publications

The publications that are listed in this section are considered particularly suitable for a more detailed discussion of the topics that are covered in this book.

## IBM Redbooks

The following IBM Redbooks publications provide more information about the z/OS V2R2 updates. Some publications that are referenced in this list might be available in softcopy only:

- ▶ *z/OS V2R2: JES2, JES3, and SDSF*, SG24-8287-00
- ▶ *z/OS V2R2: Security*, SG24-8287-00
- ▶ *z/OS V2R2: Storage Management and Utilities*, SG24-8289-00
- ▶ *z/OS V2R2: Availability Management*, SG24-8290-00
- ▶ *z/OS V2R2: Performance*, SG24-8292-00
- ▶ *z/OS V2R2: Operations*, SG24-8305-00
- ▶ *z/OS V2R2: Diagnostics*, SG24-8306-00
- ▶ *z/OS V2R2: Sysplex*, SG24-8307-00
- ▶ *z/OS V2R2: UNIX System Services* SG24-8310-00
- ▶ *z/OS V2R2: User Interfaces*, SG24-8311-00
- ▶ *z/OS V2R2: ServerPac*, SG24-8500

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft, and other materials at the following website:

[ibm.com/redbooks](http://ibm.com/redbooks)

## Other publications

The following publications are also relevant as further information sources:

- ▶ *z/OS V2R2 RMF User's Guide*, SC34-2666-01
- ▶ *z/OS V2R2 RMF Report Analysis*, SC34-2665-02
- ▶ *z/OS V2R2 RMF User's Guide*, SC34-2664-02
- ▶ *z/OS V2R2 RMF Messages and Codes*, SC34-2666-01
- ▶ *z/OS V2R2 MVS Initialization and Tuning Reference*, SA23-1380-05

## Online resource

For more information and downloads, see the following RMF website:

<http://www.ibm.com/systems/z/os/zos/features/rmf/>

## Help from IBM

IBM Support and downloads:

[ibm.com/support](https://ibm.com/support)

IBM Global Services:

[ibm.com/services](https://ibm.com/services)





SG24-8292-00

ISBN 0738441325

Printed in U.S.A.

Get connected

