Redbooks
ibm.com/redbooks

IBM Db2 Analytics Accelerator V7 High Availability and Disaster Recovery

Ute Baumbach

Frank Neumann

Information Management

Redpaper

IBM Redbooks

IBM Db2 Analytics Accelerator V7 High Availability and Disaster Recovery

May 2019

REDP-5536-00

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

First Edition (May 2019)

This edition applies to IBM Db2 Analytics Accelerator V7.

This document was created or updated on May 14, 2019.

© Copyright International Business Machines Corporation 2019. 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

© Copyright IBM Corp. 2019. All rights reserved.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119, Armonk, NY 10504-1785, US

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 jurisdictions 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 provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

© Copyright IBM Corp. 2019. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

DB2®
Db2®
FICON®
FlashCopy®
GDPS®
HiperSockets™
HyperSwap®
IBM®
IBM FlashSystem®
IBM Z®
InfoSphere®
MicroLatency®
Parallel Sysplex®
Redbooks®
Redpaper™
Redbooks (logo) ®
Variable Stripe RAID™
z/OS®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

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.

Preface

IBM® Db2® Analytics Accelerator is a workload optimized appliance add-on to IBM DB2® for IBM z/OS® that enables the integration of analytic insights into operational processes to drive business critical analytics and exceptional business value. Together, the Db2 Analytics Accelerator and DB2 for z/OS form an integrated hybrid environment that can run transaction processing, complex analytical, and reporting workloads concurrently and efficiently.

With IBM DB2 Analytics Accelerator for z/OS V7, the following flexible deployment options are introduced:

- Accelerator on IBM Integrated Analytics System (IIAS): Deployment on pre-configured hardware and software
- Accelerator on IBM Z®: Deployment within an IBM Secure Service Container LPAR

For using the accelerator for business-critical environments, the need arose to integrate the accelerator into High Availability (HA) architectures and Disaster Recovery (DR) processes. This IBM Redpaper™ publication focuses on different integration aspects of both deployment options of the IBM Db2 Analytics Accelerator into HA and DR environments. It also shares best practices to provide wanted Recovery Time Objectives (RTO) and Recovery Point Objectives (RPO).

HA systems often are a requirement in business-critical environments and can be implemented by redundant, independent components. A failure of one of these components is detected automatically and their tasks are taken over by another component. Depending on business requirements, a system can be implemented in a way that users do not notice outages (continuous availability), or in a major disaster, users notice an outage and systems resume services after a defined period, potentially with loss of data from previous work.

IBM Z was strong for decades regarding HA and DR. By design, storage and operating systems are implemented in a way to support enhanced availability requirements. IBM Parallel Sysplex® and IBM Globally Dispersed Parallel Sysplex (IBM GDPS®) offer a unique architecture to support various degrees of automated failover and availability concepts.

This IBM Redpaper publication shows how IBM Db2 Analytics Accelerator V7 can easily integrate into or complement existing IBM Z topologies for HA and DR.

If you are using IBM Db2 Analytics Accelerator V5.1 or lower, see *IBM Db2 Analytics Accelerator: High Availability and Disaster Recovery*, REDP-5104.

© Copyright IBM Corp. 2019. All rights reserved.

Authors

This paper was produced by a team of specialists from around the world working at IBM Redbooks, Poughkeepsie Center.

Ute Baumbach is a software developer in IBM's Research & Development Lab in Boeblingen, Germany. During her more than 30 years with IBM, Ute worked as a developer and team leader for various IBM software products, most related to DB2 for Linux, UNIX, and Windows and Db2 z/OS topics and tools. Currently, she works as a member of the Analytics on IBM Z Center of Excellence team, which is part of the Db2 Analytics Accelerator development organization and focuses on Db2 Analytics Accelerator proofs-of-concepts, customer deployment support, and education. She has co-authored various IBM Redbooks® publications, focusing on DB2 for Linux, UNIX, and Windows topics and Db2 Analytics Accelerator.

Frank Neumann is a Senior Software Engineer and technical team lead, in IBM's Research & Development Lab in Boeblingen, Germany. As part of the development organization, he currently focuses on Db2 Analytics Accelerator on IBM Z. During his over 20 years with IBM, Frank worked as a developer, team leader, and architect for various IBM software components and solutions. As a member of the Center of Excellence for Analytics on IBM Z, he provided consultancy for customers deploying Db2 Analytics Accelerator.

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 IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either 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

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  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`

1

Business continuity and disaster recovery

Business continuity and disaster recovery (DR) act at different levels in an organization. Business continuity is the strategy at the enterprise level; DR is the solution at the IT level. The concept of DR focuses on recovering IT systems from an unplanned outage to provide business continuity.

When we describe high availability (HA) concepts in this document, we imply that these concepts provide IT services to achieve business continuity.

Business continuity solutions are based on the concept of data duplication between two or more data center sites, which are often called *primary site* and *secondary site*. If a primary site that processes production workload is struck by a disaster and IT systems are unavailable, IT operations are provided by the secondary site. Without a second site (or recovery site) providing the same IT services, business continuity is compromised.

Depending on the chosen implementation of the data duplication between primary and secondary sites, businesses might encounter outages until operations are restored at the secondary site. Also, depending on the chosen implementation, transactions can be preserved or lost in a disaster.

This chapter includes the following topics:

© Copyright IBM Corp. 2019. All rights reserved.

1.1 Recovery time objective

Recovery time objective (RTO) refers to how long your business can afford to wait for IT services to be resumed following a disaster. Regarding the IBM Db2 Analytics Accelerator, we refer to this term to resume operations not in Db2 for z/OS only, but with an accelerator being ready to process query workloads.

1.2 Recovery point objective

Recovery point objective (RPO) refers to how much data your company is willing to re-create following a disaster. Regarding the IBM Db2 Analytics Accelerator, we refer to this term in relation to the data available for queries on the accelerator after recovery. For data loaded from Db2 for z/OS, this goes back to RPO characteristics of Db2 for z/OS. For the incremental update feature, specific recovery considerations will be discussed.

2

High availability and disaster recovery in IBM Z enterprise environments concepts

Most, if not all, enterprise transaction systems must implement high availability (HA) and disaster recovery (DR) concepts to support continuous business operation. In the context of DB2 for z/OS, the IBM Z platform offers a set of typical implementation patterns.

Db2 Analytics Accelerator is a logical extension of DB2 for z/OS. Therefore, it is useful to review topologies to understand how the Accelerator fit into these implementation patterns.

In this chapter, we list typical HA/DR scenarios and refer to GDPS product components and their implementation.

This chapter includes the following topics:

- 2.1, “Storage failures and synchronous and asynchronous data mirroring” on page 4
- 2.2, “Db2 Active-Active configuration with GDPS Metro” on page 5
- 2.3, “Db2 Active-Passive configuration with GDPS Global” on page 6
- 2.4, “Db2 Active-Passive configuration with GDPS Metro” on page 7
- 2.5, “Db2 multi-site configurations and tests” on page 8

For more details about GDPS refer to *IBM GDPS Family: An Introduction to Concepts and Capabilities*, SG24-6374.

© Copyright IBM Corp. 2019. All rights reserved.

2.1 Storage failures and synchronous and asynchronous data mirroring

On IBM Z, storage (disks) is provided by a separate, dedicated storage subsystem, which is connected to the IBM Z CPC through, FICON® adapters (for example). The GDPS Metro solution can maintain a synchronous copy of data on a secondary set of disks.

If a failure occurs on the primary disks, GDPS detects this problem and triggers a “swap” operation such that the connected component (for example, LPAR) uses the secondary disks (with mirrored data) instead. This scenario is shown in Figure 2-1with two LPARs that are connected to primary and secondary disks.

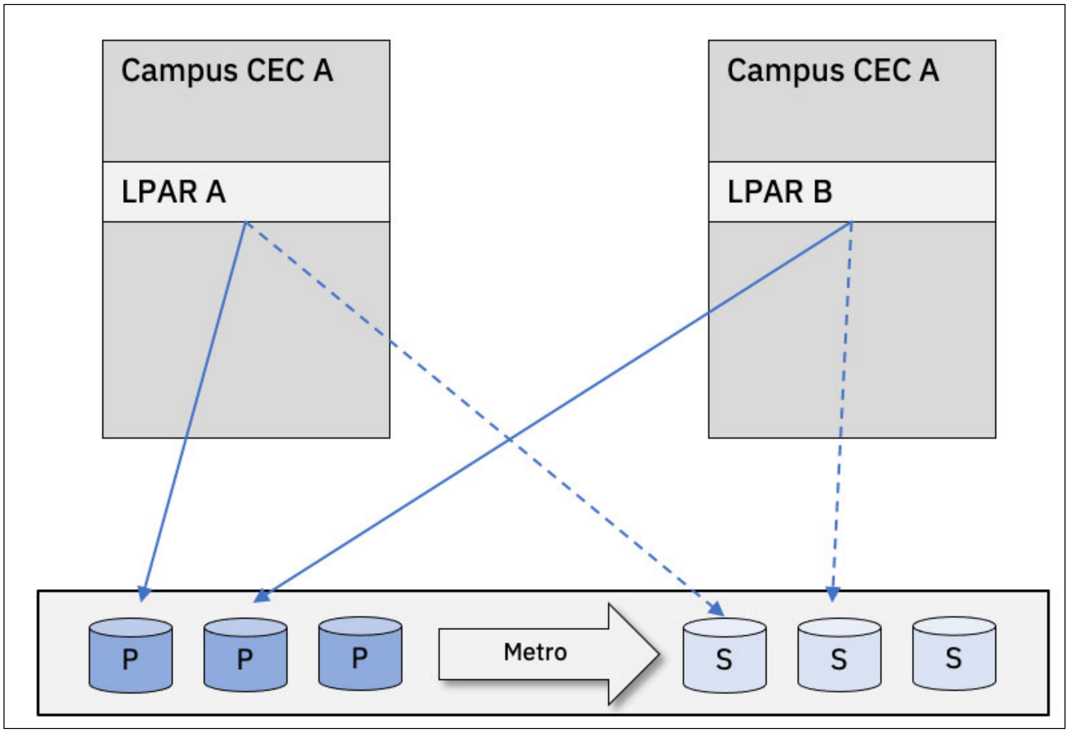


Figure 2-1 Disk mirroring with GDPS Metro

Components, such as z/OS, support the IBM HyperSwap® mode where operation is paused for a short time and then continue by using secondary disks. At the time of this writing, HyperSwap is not supported for operating systems, such as Linux on Z.

Db2 Analytics Accelerator on IIAS comes with its own disk built into the appliance. Disk failures within the system are addressed by fail-over to redundant disks. Db2 Analytics Accelerator on Z can store its data on IBM Z storage (ECDK or FCP/SCSI) and leverage enterprise storage capabilities. Together with GDPS, the Accelerator can participate in a swap operation, but does not support HyperSwap. Details regarding this support are described in 6.4, “Storage failures and HyperSwap” on page 56.

The synchronous point-to-point copy of data is limited to metropolitan distances between primary and secondary disks. The reason is increasing signal latency because I/O operations must be committed on the secondary disks first.

To cover larger distances between primary and secondary sites, an asynchronous data replication/mirror protocol is used and part of the GDPS Global product. Still, bandwidth must be large enough to cope with the change volume, but distance is no longer a problem.

However, asynchronous protocols include the risk that data-in-flight might be lost; for example, if an I/O operation completed on the primary site but a failure (such as power outage) prevents the change from being sent to the secondary site. In this case, data might be lost and recovery procedures must take incomplete or even inconsistent data into account, which can lead to an RPO > 0.

2.2 Db2 Active-Active configuration with GDPS Metro

A typical HA environment deploys a DB2 for z/OS data sharing group with active members in two CPCs (CEC A and CEC B) at two sites, as shown in Figure 2-2. These active data sharing group members access data on a primary storage subsystem. The two sites are in metro distance, which is typically on the same campus or only a few kilometers or miles apart.

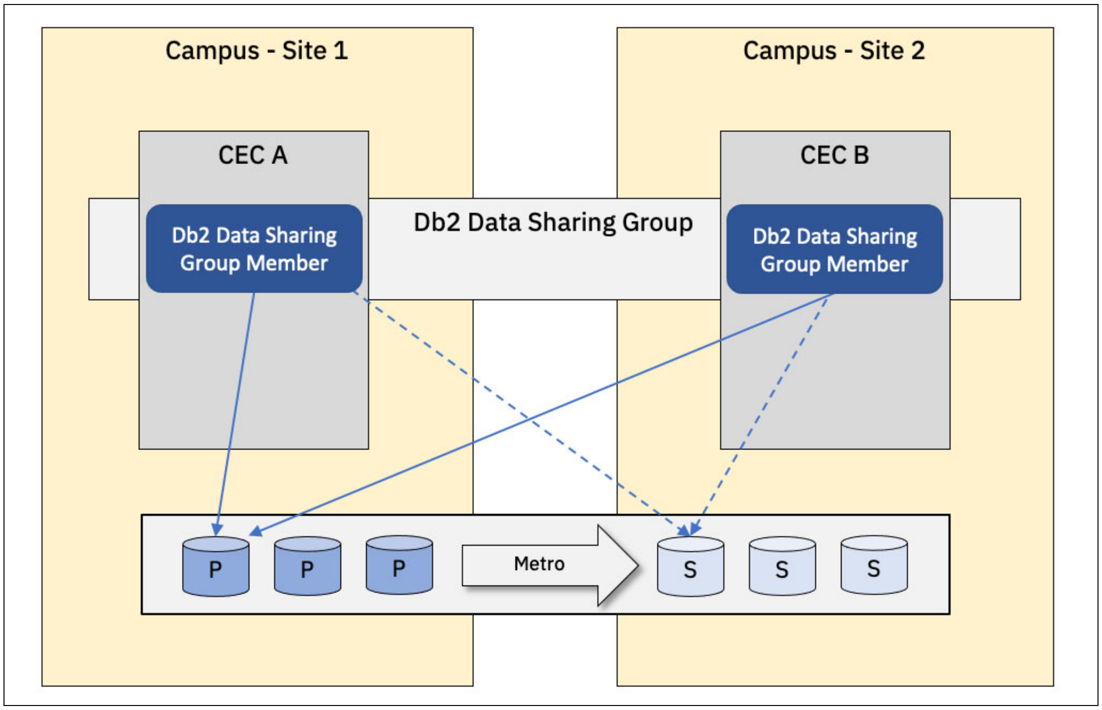


Figure 2-2 Db2 Active-Active Configuration with GDPS Metro

DB2 for z/OS follows a “shared everything” model; therefore, all active members can access all database data. This configuration differs from a “shared nothing” approach with partitioned database management systems where members can access their data partition only.

If a failure occurs on site 1, the DB2 for z/OS data sharing group member becomes unavailable and the storage on site 1 is no longer accessible. However, the member on site 2 continues to work. After a HyperSwap operation, it uses secondary disks on site 2 that were mirrored and are up-to-date.

Depending on available system resources, the surviving site 2 member can be dynamically assigned more capacity (CPs, RAM) to handle the now increasing workload.

An active-active configuration includes with zero downtime (RTO=0), but covers only local issues (for example, power outage in one building, but not an earthquake that causes damage for an entire region).

Db2 Analytics Accelerator is based on a “shared nothing” data store; therefore, two active accelerator instances cannot share data. In 6.5, “Active-passive configuration with standby Accelerator and GDPS Metro” on page 57, we describe ways how the Accelerator on IBM Z can support an active-active configuration.

2.3 Db2 Active-Passive configuration with GDPS Global

In an active-passive configuration, only one active Db2 subsystem is available. This subsystem can be a member of a data sharing group but also a “dormant” copy of a stand-alone Db2 subsystem in site 2, as shown in Figure 2-3.

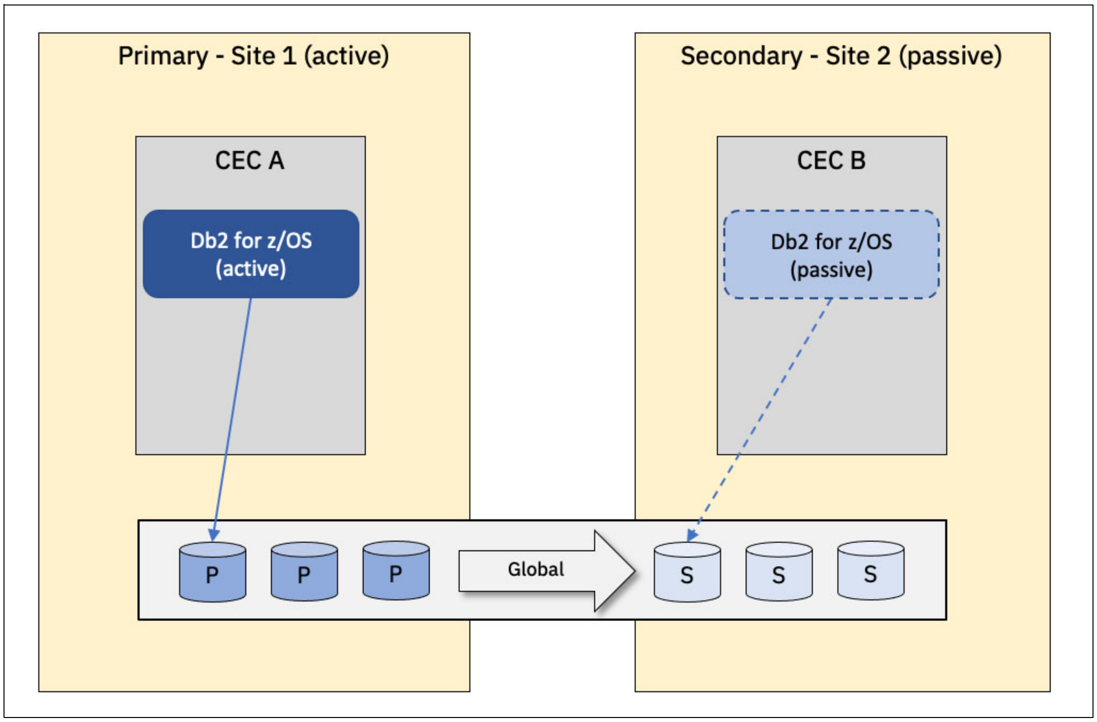


Figure 2-3 Db2 Active-Passive configuration with GDPS Global

Although this configuration is not tightly coupled with GDPS Global and asynchronous replication/mirroring, it is common to have a passive Db2 subsystem in a remote disaster recovery site.

If a failure is detected on site 1, GDPS activates the Db2 subsystem on site 2 and allows it to use the asynchronously copied set of disks on site 2.

This process leads to an RTO > 0. Because of its asynchronous data mirroring technique, a GDPS Global architecture includes RPO > 0 and some data loss might occur.

Startup and recovery take a little longer, but this configuration also can address issues that are occurring in an entire geographical region.

IBM Z pricing models might consider active and passive sites, such that maintaining a passive site costs less than another active site.

With Db2 Analytics Accelerator, several options are available to cover this scenario. The Accelerator on IBM Z currently is not supported by GDPS Global in a 2-site configuration. As an alternative, we recommend one of the following options

- An active Accelerator in the disaster recovery site (described next)
- For Db2 Analytics Accelerator, definition of a storage environment for second site. This requires manual startup of the passive Accelerator and is described in 6.1, “IBM Z integration, network, and storage management” on page 52.

2.4 Db2 Active-Passive configuration with GDPS Metro

Another popular variation is a mix of the two previously explained concepts. It has storage subsystems with synchronously mirrored disks that use GDPS Metro, but only one active site for a DB2 for z/OS subsystem. This configuration is shown in Figure 2-4.

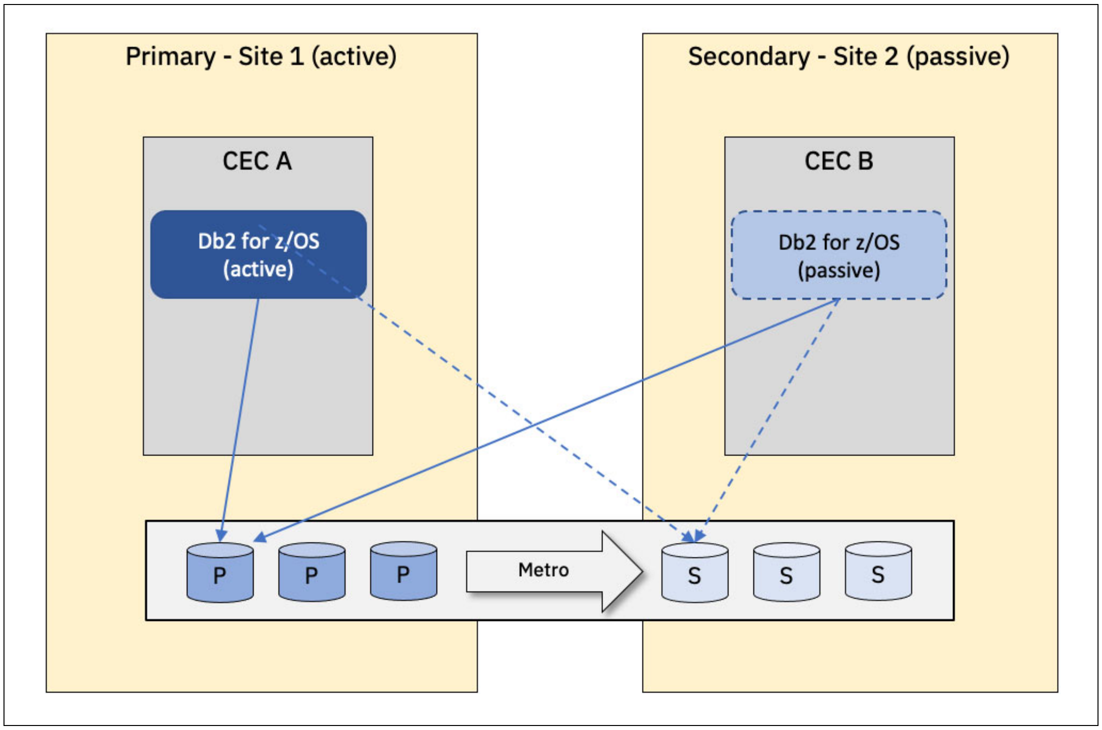


Figure 2-4 Db2 Active-passive configuration with GDPS Metro

Synchronous disk mirroring protects against storage subsystem failures and loss of data (RPO=0). The active/passive concept for the DB2 for z/OS subsystems is also a good option for stand-alone subsystems (that is, non-data-sharing-groups). However, if one site encounters a disaster beyond storage, a service interrupt is perceived until the passive subsystem is serving requests, which leads to an RTO>0.

Db2 Analytics Accelerator on IBM Z can complement such a configuration by following the same architecture with an active and passive installation.

2.5 Db2 multi-site configurations and tests

The active-active and active-passive concepts can further be combined and extended.

A three-site setup combines an active-active combination with a third passive site used for disaster recovery.

GPDS introduced and supports the concepts of a "cascaded" and a "multi-target" topology which refers to the way how data is mirrored/replicated for such multi-site environments.

Figure 2-5 shows an example for a three-site setup which uses GDPS Metro for synchronous disk mirroring between the active Db2 sites 1 and 2 in Region A, and GDPS Global GM for asynchronous mirroring to a DR site 3.

With a GDPS cascaded topology the primary disks in Site 1 (P) are mirrored to the secondary disks in Site 2 (S) using Metro Mirror and then secondary disks are mirrored to DR disks in Site 3 using Global Mirror. This is depicted in Figure 2-5.

After a swap from primary (P) to secondary (S) disks, the setup turns into a GDPS multi-target topology where secondary (S) disks are mirrored to primary (P) disks using Metro Mirror and, at the same time, secondary (S) disks are mirrored to the DR disks using Global Mirror.

Such a configuration also supports failures of site 1 or 2 (the remaining site will then do the asynchronous mirroring to the DR site) or complete region switches.

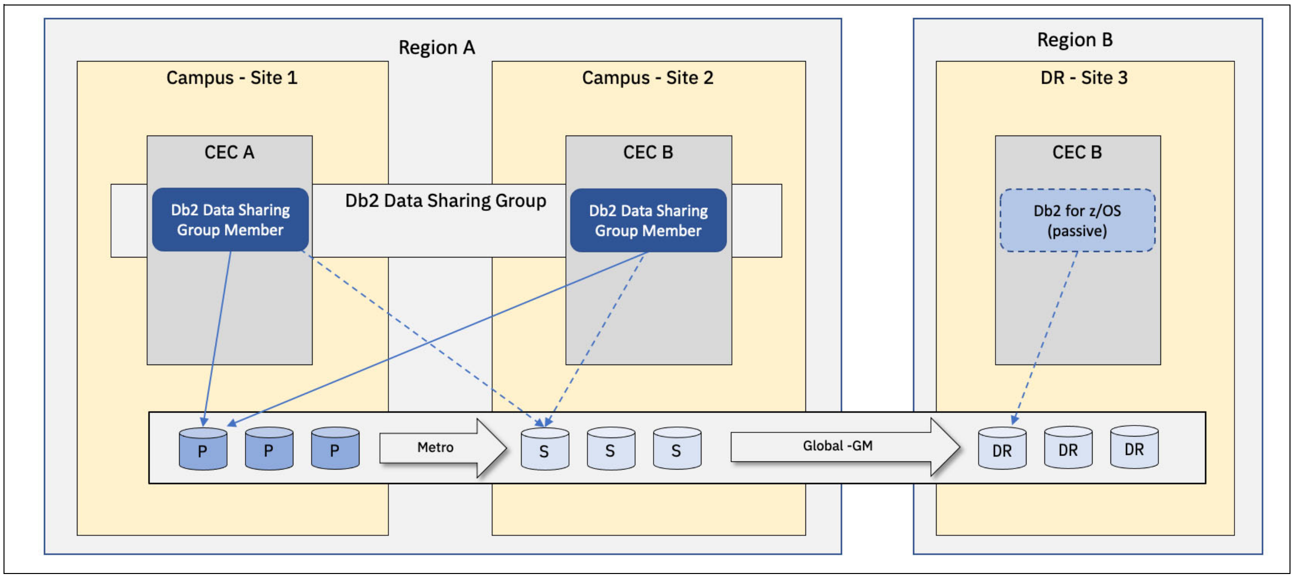


Figure 2-5 Db2 3-site configuration with GDPS Metro and Global GM

A four-site setup duplicates an active-active configuration in a second site and provides the same level of availability in geographically distinct regions.

As with the 3-site solution, the 4-site solution can dynamically switch between a cascaded topology and a multi-target topology. The 4-site configuration combines GDPS Metro with GDPS GM as does the 3-site solution. Figure 2-6 shows an example of a 4-site configuration.

Compared to a 3-site setup, the 4-site setup provides the same level of high availability (and resources) in the both regions, allowing a planned or unplanned switch of production between the regions.

This fourth copy of data is created using asynchronous Global Copy that can be switched to synchronous-mode (that is, Metro Mirror) during a planned or unplanned region switch, which provides the HA copy in that region.

Some industries or organizations enforce regular swaps between environments in "global" distance to ensure and validate operational availability in both regions.

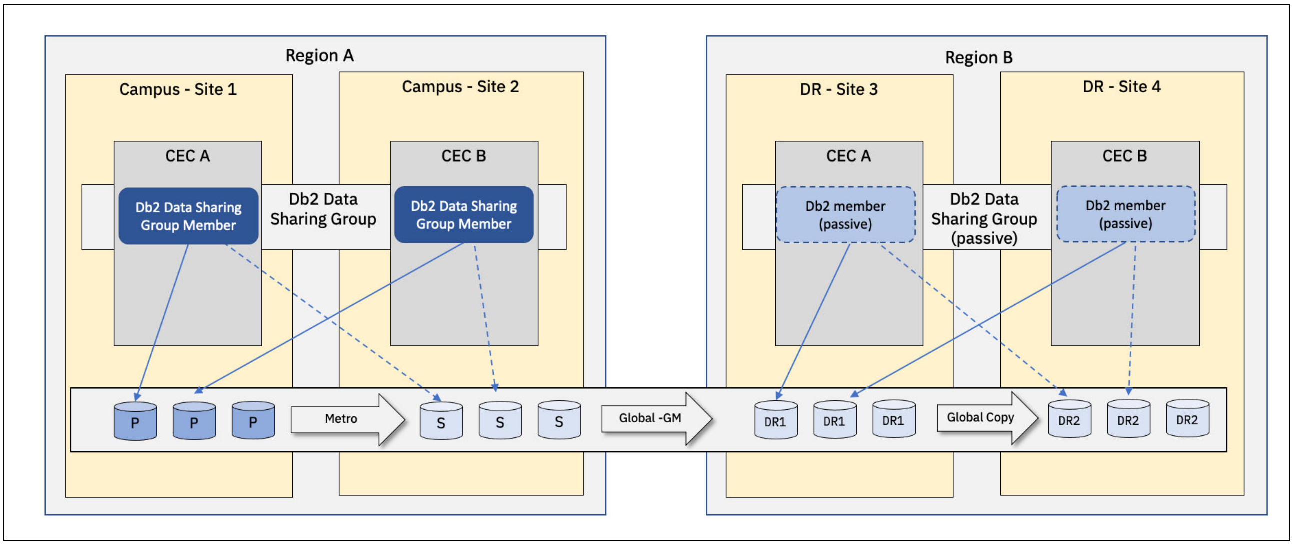


Figure 2-6 Db2 4-site configuration with GDPS Metro and Global GM

Db2 Analytics Accelerator on IBM Z supports GDPS Metro and GDPS Global GM in three-site and four-site setups. Note, that Db2 Analytics Accelerator on IBM Z does only 3-site or 4-site configurations using GDPS Global GM. Configurations using GDPS Global XRC are not supported.

In addition to regular and controlled “swaps” between sites, DR tests are also configured to prepare for uncontrolled failures. These tests run on a disconnected copy of the data and changes are made during DR tests often are not replicated or mirrored after the test is complete.

Enterprise storage systems offer IBM FlashCopy® operations so that mirrored, active production data remains unchanged for the exercise.

Db2 Analytics Accelerator on IBM Z can also be deployed by using enterprise storage (CKD); therefore, use FlashCopy and other storage operations. These operations are useful for DR tests, software update scenarios, and initial tests. 6.1, “IBM Z integration, network, and storage management” on page 52 discusses options how to start an Accelerator using disks with copied data.

3

Introducing high availability concepts for IBM Db2 Analytics Accelerator

In general, high availability (HA) and disaster recovery (DR) solutions are implemented by adding redundant components.

In this chapter, we describe how multiple accelerators or multiple Db2 subsystems can be configured and managed. The concepts that are described in this chapter act as the foundation to build advanced configurations and integrate them into HA/DR IBM Z environments.

This chapter includes the following topics:

© Copyright IBM Corp. 2019. All rights reserved.

3.1 Workload balancing

A key feature in IBM Db2 Analytics Accelerator is the capability to automatically balance query routing from a Db2 for z/OS subsystem to multiple attached accelerators. With the option to share one or more accelerators between multiple Db2 for z/OS subsystems, this feature provides options for a flexible deployment pattern.

A setup that features one Db2 for z/OS subsystem (or member of a data sharing group) that contains data in tables T1 and T2 is shown in Figure 3-1.

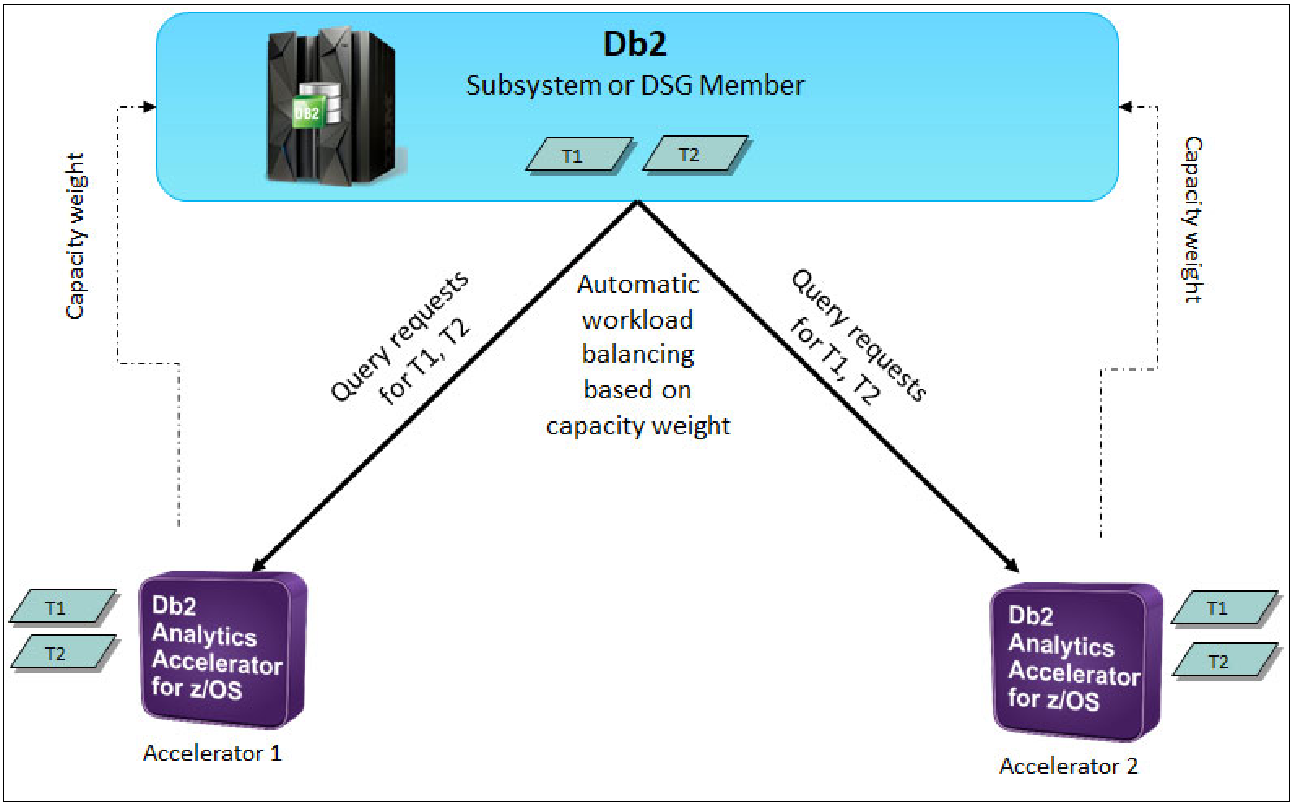

Db2
Subsystem or DSG Member
T1
T2
Capacity weight
Capacity weight
Query requests for T1, T2
Automatic workload balancing based on capacity weight
Query requests for T1, T2
T1
T2
Db2 Analytics Accelerator for z/OS
Accelerator 1
Db2 Analytics Accelerator for z/OS
Accelerator 2
T1
T2

Figure 3-1 Workload balancing with multiple active accelerators

Two accelerators (Accelerator 1 and Accelerator 2) are defined for this Db2 for z/OS subsystem and both have tables T1 and T2 loaded and active. For query processing and query routing, the Db2 for z/OS optimizer now must decide which accelerator to use if a query is eligible for acceleration.

Each accelerator reports its current capacity weight, along with other monitoring parameters to each attached Db2 for z/OS subsystem. This capacity weight value includes information about the current accelerator usage. The Db2 optimizer uses this value to implement automated workload balancing between these two accelerators. The system with a lower capacity weight value receives more queries for execution than the other system.

Because this workload balancing is applied to each query that is run on an accelerator, this feature is useful for adding capacity and for HA scenarios. If one accelerator fails, query requests automatically are routed to the remaining accelerator without any other administration intervention.

3.1.1 Directing queries to specific accelerators

A user of Db2 Analytics Accelerator can influence the Db2 optimizer's decision by directing queries to a specific accelerator. This feature can be useful, for example, in the following situations:

- An accelerator is deployed on a remote location (DR location) in addition to locally deployed accelerators. The capacity weight does not include information about network latency that is caused by a long network distance to a remote location. A user wants to ensure that some high priority queries are sent to local accelerators so that elapsed times are not affected by network latency
- A user wants to direct queries to different accelerators, depending on the priorities of the workload, with high priority queries directed to the fastest, highest capacity accelerator and low priority queries directed to a slower accelerator.

To direct dynamic queries to specific accelerators, use the Db2 special register CURRENT ACCELERATOR. It specifies the name of preferred accelerators to which Db2 routes dynamic queries. If none of the accelerators that are named by CURRENT ACCELERATOR are available or eligible, Db2 considers other available accelerators.

The register uses the following syntax:

```
SET CURRENT ACCELERATOR <accelerator_name>
```

The `<accelerator_name>` is a single accelerator name or an accelerator logical name as stored in the SYSIBM.LOCATIONS table. The logical name represents one or more accelerator names.

To direct static queries to specific accelerators, use the ACCELERATOR BIND option. The ACCELERATOR BIND option specifies the preferred accelerators to which Db2 routes static queries. It can be a single accelerator name or an accelerator logical name as stored in the SYSIBM.LOCATIONS table and representing one or more accelerator names.

The ACCELERATOR BIND option affects only the execution of a static query. If one of the specified accelerators does not exist during BIND time, Db2 issues a warning message, but does not fail the BIND.

The next example shows how to define an accelerator logical name that maps to two physical accelerator names.

- The following syntax is used:

  ```
  INSERT INTO SYSIBM.LOCATIONS (LOCATION, LINKNAME, DBALIAS) VALUES
  ('logical_system_accel_name', 'DSNACCELERATORALIAS',
  'physical_system_accel_name')
  ```

- Example:

  ```
  INSERT INTO SYSIBM.LOCATIONS (LOCATION, LINKNAME, DBALIAS) VALUES (IDAATEST,
  'DSNACCELERATORALIAS', 'ACCLPRO1 ACCLPRO2');
  SET CURRENT ACCELERATOR = IDAATEST;
  ```

3.2 Data maintenance and synchronization with multiple accelerators

Most concepts that are discussed in this IBM Redpaper publication assume that data can be synchronized or loaded consistently from Db2 for z/OS into multiple accelerators. For initial loads of accelerator-shadow tables, selected tables must be unloaded in Db2 for z/OS and loaded into each attached accelerator.

Multiple options exist to keep data current in the accelerators. Those options include a full table reload, reload of one or more identified partitions (manually or by using the automatic change detection feature), or the use of incremental update processing.

The option also is available to use a separate licensed tool (the Db2 Analytics Accelerator Loader) to load data from various data sources (Db2 and non-Db2) into one accelerator only into multiple accelerators in parallel or into one accelerator and Db2 in parallel. The Db2 Analytics Accelerator Loader can load data into accelerator-shadow tables or into accelerator-only tables.

If you use accelerator-only tables, the requirement of loading data consistently into multiple accelerators needs more considerations. An accelerator-only table (AOT) with a certain name can exist only on one accelerator. On another accelerator, an accelerator-only table having the same content must have a different name as on the first accelerator.

The different synchronization options are described next.

3.2.1 Loading into multiple accelerators by using the ACCEL_LOAD_TABLES stored procedure

Data that originates from Db2 for z/OS tables is loaded into IBM Db2 Analytics Accelerator by using the stored procedure SYSPROC.ACCEL_LOAD_TABLES. One stored procedure call can load data into one accelerator only. Multiple calls of this stored procedure are required to load the same data into multiple accelerators.

Each call unloads the data from Db2 (UNLOAD utility) and then loads it into one accelerator into a so-called accelerator-shadow table.

To ensure the same results for all incoming queries (regardless which accelerator processes a query) and to optimally use the workload balancing capabilities of IBM Db2 Analytics Accelerator, it is recommended to load data into all available accelerators that are participating in a high availability setup, as shown in Figure 3-2 on page 15.

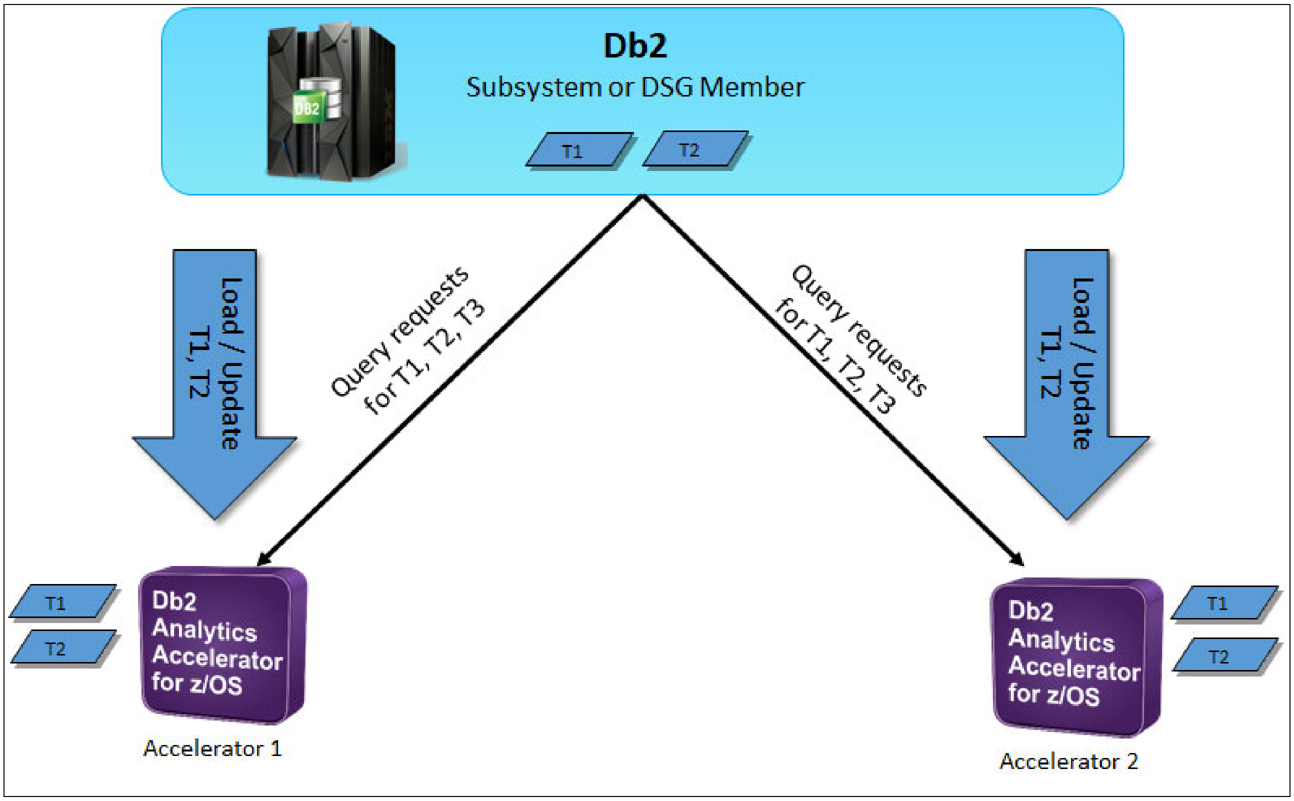

Db2
Subsystem or DSG Member
T1
T2
Load / Update T1, T2
Query requests for T1, T2, T3
Query requests for T1, T2, T3
Load / Update T1, T2
T1
T2
Db2 Analytics Accelerator for z/OS
Accelerator 1
Db2 Analytics Accelerator for z/OS
Accelerator 2
T1
T2

Figure 3-2 Maintaining two active accelerators with the same data

If one accelerator becomes unavailable, the second accelerator (with same data loaded) can continue to serve requests, which provides RTO=0. However, performance might be degraded because of reduced overall query processing capacity (only one accelerator instead of two).

Note: To minimize the amount of data that is loaded into multiple accelerators, use the change detection feature of the stored procedure SYSPROC.ACCEL_LOAD_TABLES to load changed data partitions only.

Data consistency across multiple accelerators

Under certain situations, data in multiple accelerators can get out of sync, as shown in the following examples:

- During loading, if some active accelerators are loaded with new data while others are still waiting to be loaded and still have old data
- If Db2 table data changes while accelerator loading is ongoing; in this case, different accelerators can take different snapshots of the data in Db2

In both cases, queries might return different results, depending on which accelerator is chosen for query processing.

Both cases can be addressed by using the HA Load feature of the Db2 Analytics Accelerator Loader product, as described in 3.2.2, “Loading into multiple accelerators using Db2 Analytics Accelerator Loader” on page 16.

Alternatively both cases can be addressing by using the so called HA Load sample program that is delivered with Db2 Analytics Accelerator V7.5, as described in 3.2.3, “Loading into multiple accelerators using the HA Load sample program of Db2 Analytics Accelerator" on page 17.

However, if Db2 Analytics Accelerator Loader is not available or the HA Load sample program cannot be used, the options and techniques that are described next can be considered.

The first case can be addressed with the option to disable and enable a table for acceleration on an accelerator. The following update sequence for multiple accelerators might be used:

1. Disable table for acceleration on all but one accelerator.
2. Load data and enable tables for acceleration on all accelerators sequentially, starting with the accelerator that has the table that is enabled for acceleration.
3. During this time, query requests are processed by the only accelerator that has the table that is enabled for acceleration.

For the second case, table locks can be used to prevent changes during the load process. Under transactional control, the following sequence can be implemented:

1. Lock table(set).
2. Call ACCEL_LOAD_TABLES for first accelerator.
3. Call ACCEL_LOAD_TABLES for second accelerator.
4. Release table(set) lock.

3.2.2 Loading into multiple accelerators using Db2 Analytics Accelerator Loader

The IBM Db2 Analytics Accelerator Loader is a separately licensed tool that you can use to load data from various data sources into IBM Db2 Analytics Accelerator for z/OS or into both IBM Db2 Analytics Accelerator for z/OS and Db2 for z/OS.

One use case of this tool is support for loading Db2 for z/OS or non-Db2 for z/OS data into multiple accelerators at the same time. This process is called *high availability load*.

This process includes the following advantages:

- Data is loaded in parallel to multiple accelerators with reduced elapsed time and CP time.
- Data is unloaded (Db2 UNLOAD utility) only once on Db2 for z/OS and then loaded to multiple accelerators.
- Ensures data consistency across multiple accelerators.

The concept of the use of the HA load feature of Db2 Analytics Accelerator Loader to load data into multiple accelerators in parallel is shown in Figure 3-3 on page 17.

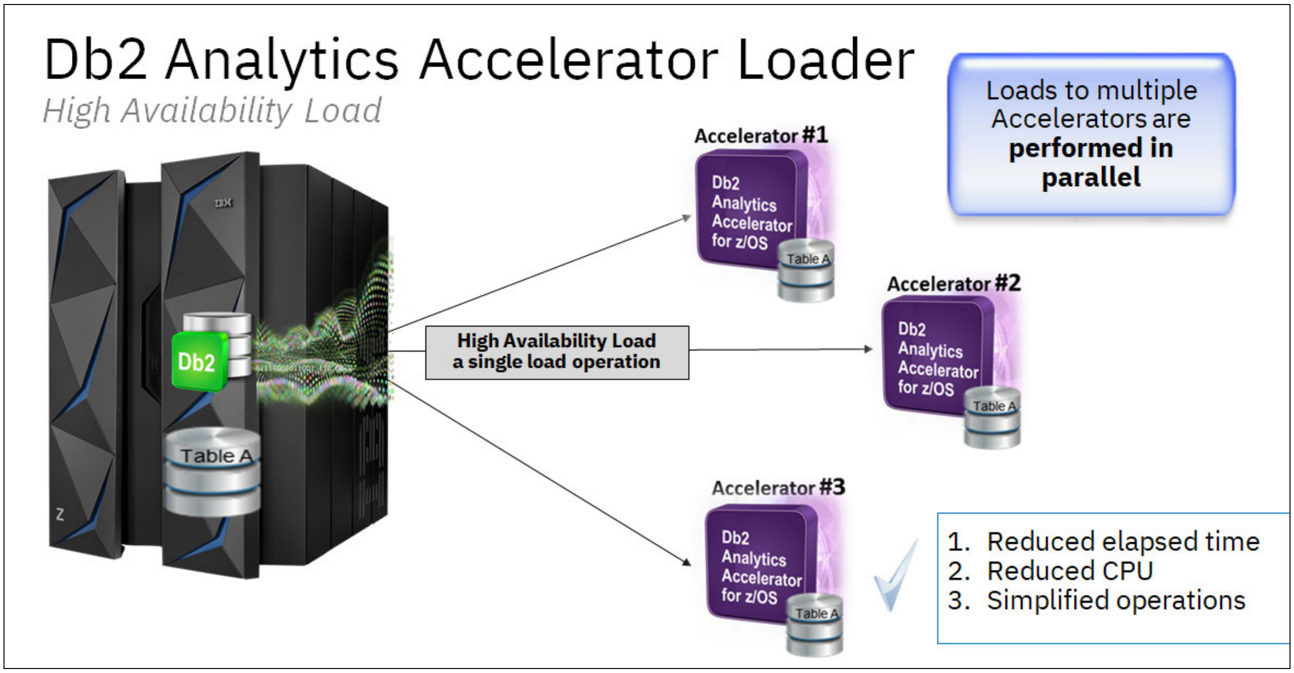


Figure 3-3 High availability load with Db2 Analytics Accelerator Loader

3.2.3 Loading into multiple accelerators using the HA Load sample program of Db2 Analytics Accelerator

Db2 Analytics Accelerator provides a sample program that you can use to load Db2 for z/OS data into multiple accelerators at the same time. The sample progam is delivered as C-code and you have to compile, link and bind it first before you can run it. Jobs to compile, link and bind the program are delivered as well.

The following members of the SAQTSAMP library of your Db2 Analytics Accelerator installation on z/OS belong to the HA Load sample program (source code and JCL jobs):

- AQTLMAIN, AQTLSHMH, AQTLSHMC, AQTLCABA, AQTSJL01, AQTSJL02, AQTSJL03

The following videos describe the setup and usage of the HA Load sample program:

- Overview of HA load sample capability: https://www.youtube.com/watch?v=z7WBk-qLuUQ
- HA load sample hints, tips and best practices: https://www.youtube.com/watch?v=bnWj9RhBypQ
- Installation and setup of the HA Load Sample: https://www.youtube.com/watch?v=EyBtjvh6QOM

3.2.4 Loading a subset of data

Maintenance effort or capacity restrictions might not allow maintaining multiple fully synchronized accelerators.

In this case, a subset of data can be selected and stored on each accelerator. This subset can be defined by application data requirements (to ensure that at least one accelerator has all of the required Db2 data loaded for an application).

A failure or disaster occurs, the remaining data must be loaded into the remaining accelerator first, which causes RTO > 0 for this data. While this load process occurs, Db2 for z/OS can continue to serve query requests, possibly with higher CPU and elapsed times.

3.2.5 Using accelerator-only tables with multiple accelerators

Accelerator-only tables (AOTs) are tables with data that do not originate from Db2 for z/OS base tables. Their data exists only on an accelerator. Typically, AOTs are used to store intermediate or final results of in-database data transformation processes that are run on the Accelerator.

AOTs cannot be loaded by the SYSPROC.ACCEL_LOAD_TABLES stored procedure. AOTs are populated by using INSERT statements or the Db2 Analytics Accelerator Loader tool to load data from external data sources to an accelerator only.

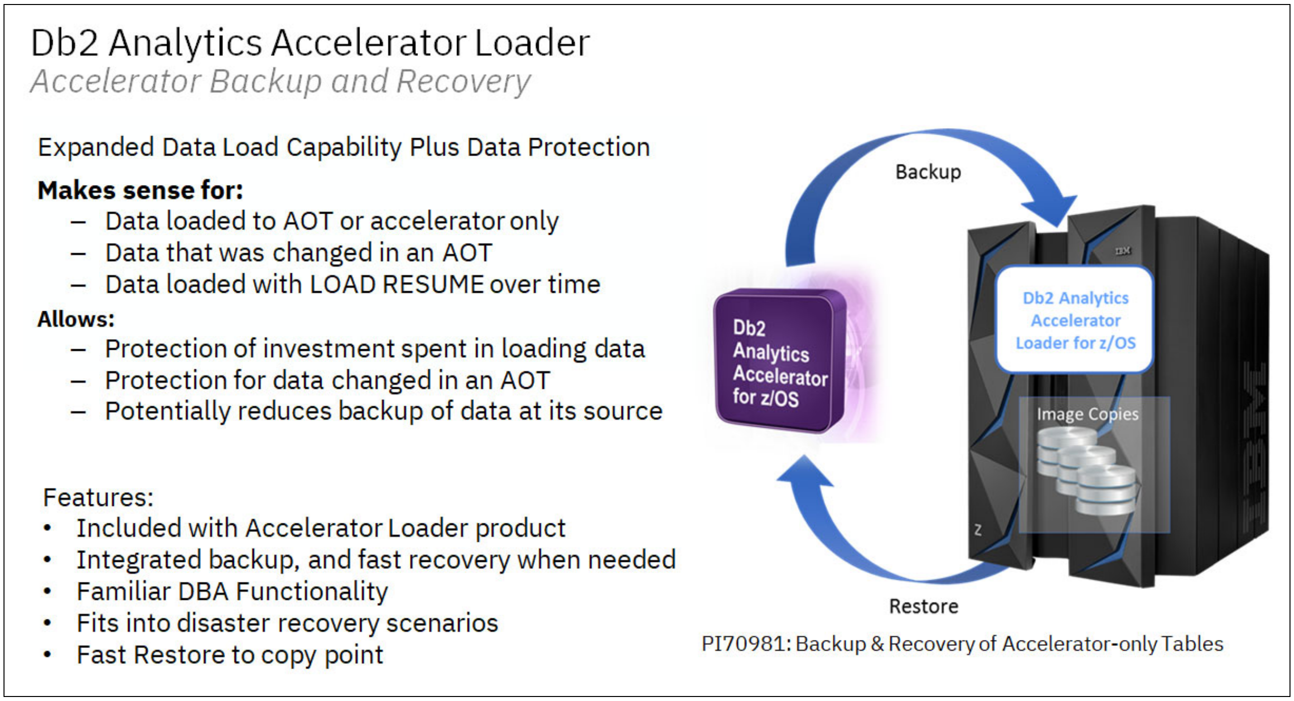
AOTs names across accelerators are unique for a Db2 subsystem. That is, an AOT can be created with the same data on a second accelerator, but it must have a different name if both Accelerators are paired to the same Db2 subsystem.

Therefore, AOTs cannot participate in the workload balancing decisions that Db2 takes to route a query to one or the other accelerator depending on availability and utilization. If the accelerator that hosts an AOT fails, Db2 cannot route a query by using this AOT to another accelerator. Instead, the query fails.

This behavior causes an RTO > 0 for AOTs to make them available on the remaining accelerator. AOTs can be used in high availability environments with two accelerators or more as shown in the following examples:

- Create backups of permanent AOTs; for example, the AOTs that contain final results of in-database transformation processes or the AOTs that were loaded with external data by using Db2 Analytics Accelerator Loader.

  If an accelerator failure occurs, restore the AOTs from the backup to the remaining accelerator. The Db2 Analytics Accelerator Loader contain functionality to back up and recover data that exists on an accelerator only. The backup and recovery functionality of Db2 Analytics Accelerator Loader is shown in Figure 3-4.



Db2 Analytics Accelerator Loader

Accelerator Backup and Recovery

Expanded Data Load Capability Plus Data Protection

Makes sense for:

- Data loaded to AOT or accelerator only
- Data that was changed in an AOT
- Data loaded with LOAD RESUME over time

Allows:

- Protection of investment spent in loading data
- Protection for data changed in an AOT
- Potentially reduces backup of data at its source

Features:

- Included with Accelerator Loader product
- Integrated backup, and fast recovery when needed
- Familiar DBA Functionality
- Fits into disaster recovery scenarios
- Fast Restore to copy point

Backup
Db2 Analytics Accelerator for z/OS
Db2 Analytics Accelerator Loader for z/OS
Image Copies
Restore

PI70981: Backup & Recovery of Accelerator-only Tables

Figure 3-4 Backup and Recovery of Db2 Analytics Accelerator Loader

- If possible, design your AOT creation and population processes in a way that they can rerun on the remaining accelerator to create and populate the same set of AOTs with the same content as on the failed accelerator after a failure.

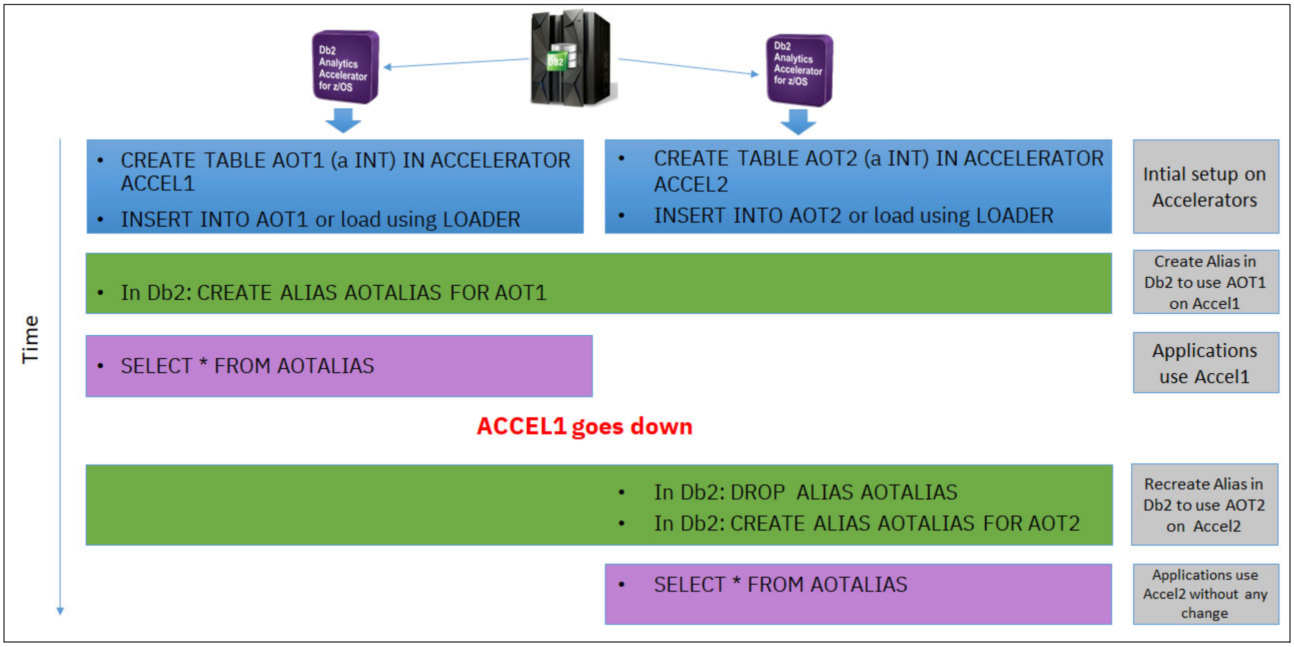
  For example, the use of a logical name for an accelerator within the CREATE TABLE statement ensures that you do not have to adapt CREATE TABLE statements for different accelerators. Only the mapping of a logical name to a physical accelerator name must be changed in the SYSIBM.LOCATION table.

  The following syntax is used:

  ```
  INSERT INTO SYSIBM.LOCATIONS (LOCATION, LINKNAME, DBALIAS) VALUES
  ('logical_system_accel_name', 'DSNACCELERATORALIAS',
  'physical_system_accel_name')
  ```

- Create and populate your AOTs on all accelerators while they are active. The AOTs have different names on each Accelerator. Design your applications by using AOTs in a way that they use Db2 alias names that refer to AOTs on one active accelerator instead of the use of the physical AOT name.

  After one accelerator fails, re-create the Db2 alias names that refer to the AOTs on the remaining accelerator. From an RTO perspective, this example has a lower RTO than the previous examples. This scenario is shown in Figure 3-5.



Db2 Analytics Accelerator for z/OS
Db2
Db2 Analytics Accelerator for z/OS
Time
CREATE TABLE AOT1 (a INT) IN ACCELERATOR ACCEL1
INSERT INTO AOT1 or load using LOADER
CREATE TABLE AOT2 (a INT) IN ACCELERATOR ACCEL2
INSERT INTO AOT2 or load using LOADER
Intial setup on Accelerators
In Db2: CREATE ALIAS AOTALIAS FOR AOT1
Create Alias in Db2 to use AOT1 on Accel1
SELECT * FROM AOTALIAS
Applications use Accel1
ACCEL1 goes down
In Db2: DROP ALIAS AOTALIAS
In Db2: CREATE ALIAS AOTALIAS FOR AOT2
Recreate Alias in Db2 to use AOT2 on Accel2
SELECT * FROM AOTALIAS
Applications use Accel2 without any change

Figure 3-5 AOTs with multiple accelerators and ALIAS names

3.2.6 Loading accelerator only data with Db2 Analytics Accelerator Loader

The Db2 Analytics Accelerator Loader provides functionality to load data from external data sources into an accelerator only. In addition to loading data into an accelerator-only table (AOT), the Db2 Analytics Accelerator Loader provides the option to load into a special version of accelerator-shadow table instead. By using this option, the Db2 Analytics Accelerator Loader populates the accelerator-shadow table with data on only the accelerator, whereas the corresponding table on Db2 for z/OS is not populated and stays empty.

The Db2 Analytics Accelerator Loader creates the Db2 table and the accelerator-shadow table if they do not yet exist, but populates only the accelerator-shadow table by replacing data or appending to it.

Compared to the use of AOTs, this feature has the advantage that accelerator-shadow tables can exist on multiple accelerators with the same name and can be populated at the same time by using Db2 Analytics Accelerator Loader. Queries that use these tables are routed to one or the other accelerator by using workload balancing. If one accelerator fails, query processing continues on the remaining accelerator (RTO=0).

3.2.7 Incremental Update into multiple accelerators

By using the incremental update capability, logged data changes from INSERT, UPDATE, and DELETE statements on tables can be propagated to multiple accelerators after they are committed. For more information about the architecture and setup of Incremental update, see 3.3, “HA setup for incremental update”.

Db2 Analytics Accelerator provides two techniques for Incremental Update:

- Incremental Update using Integrated Synchronization

  This new advanced technique is available since Db2 Analytics Accelerator V7.5.

- Incremental Update using IBM Change Data Capture (CDC) of InfoSphere® Data Replication for z/OS®

  This techniques is available since Db2 Analytics Accelerator V3.

When using incremental update via CDC and data changes are propagated to multiple accelerators, it is recommended to implement the single log scrape capability (called *log cache*) so that Db2 log data is read only once, which is more efficient than accessing the same log data for each incremental update target accelerator.

Because of the asynchronous data replication protocol between Db2 and the accelerator, data latency might occur and accelerators are slightly out of sync compared to Db2 data.

For SQL queries, relying on latest table data changes the wait-for-data protocol can be used to wait until the latest committed changes are applied before the query runs. This behavior ensures that query processing includes the latest changes, no matter on which accelerator the query is run.

The wait-for-data protocol can be enabled

- For dynamic queries by setting the CURRENT QUERY ACCELERATION WAITFORDATA special register or by setting the ZPARM QUERY_ACCEL_WAITFORDATA
- For static queries by using the ACCELERATIONWAITFORDATA BIND option

  This requires Db2 12 with APAR PH14116.

3.3 HA setup for incremental update

Incremental update is a feature of IBM Db2 Analytics Accelerator where data changes within Db2 for z/OS can be automatically propagated to an accelerator by using Db2 for z/OS logs. This option is useful to track continuous changes in a Db2 for z/OS table without the need to run a full table or partition load into the accelerator.

Db2 Analytics Accelerator provides two techniques for Incremental Update that you can choose from:

- Incremental Update using Integrated Synchronization
- Incremental Update using IBM Change Data Capture (CDC) of InfoSphere® Data Replication for z/OS®

Which technique to use depends on your needs and your environment (for example prerequisites for Integrated Synchronization).

3.3.1 HA setup for incremental update using Integrated Synchronization

Integrated Synchronization is a new advanced data synchronization technique to process incremental updates to the accelerator. This functionality is integrated into Db2 for z/OS. Its purpose is to capture table changes from the Db2 for z/OS log and to apply these changes to the tables on the accelerator. For customers that want to use incremental updates it is no longer necessary to install and configure IBM CDC (Change Data Capture) of InfoSphere® Data Replication for z/OS®.

In addition, Integrated Synchronization provides the following advantages:

- Low latency
- Reduced CPU consumption on z/OS due to a streamlined and optimized design

- On z/OS, the workload to capture the table changes has been massively reduced and the remainder can be handled by IBM Z Integrated Information Processors (zIIPs)
- Simplified administration, packaging, upgrades, and support
- Enterprise-grade HTAP enabler: The integrated low latency protocol is now enabled to support significantly more analytical queries running against the latest committed data
- Supports replication of Db2 archive tables (not supported with CDC)

The underlying technology uses a log data provider tha reads the Db2 log into a memory buffer via a service request block (SRB) that is scheduled in the Db2 address space DBM1. The log data provider is a newly developed, internal Db2 for z/OS component that is provided with Db2 12 APAR PH06628 (PTF UI63356)

On the Accelerator the newly developed log data processor component is responsible for fetching the provided log data regularly to the accelerator into a staging area and applying the data from the staging area to the tables on the accelerator in an optimized high-performance way. For communication between both components the log data processor on the accelerator connects to Db2 for z/OS (DIST address space) via the DDF secure port (SECPORT).

Integrated Synchronization needs to maintain a stable connection to the same log data provider task on the same Db2 subsystem where the session was started.

For data sharing groups ensure it is possible to always connect to the same Db2 member, for example by the following steps :

- Define a dedicated location alias and SECPORT for Integrated Synchronization on all Db2 members. In case you already use a SECPORT for other workloads the SECPORT for Integrated Synchronization would be a different one
- Start the location alias only on the Db2 member on which the Db2 log data should be provided for Integrated Synchronization

This is shown in Figure 3-6 on page 22.

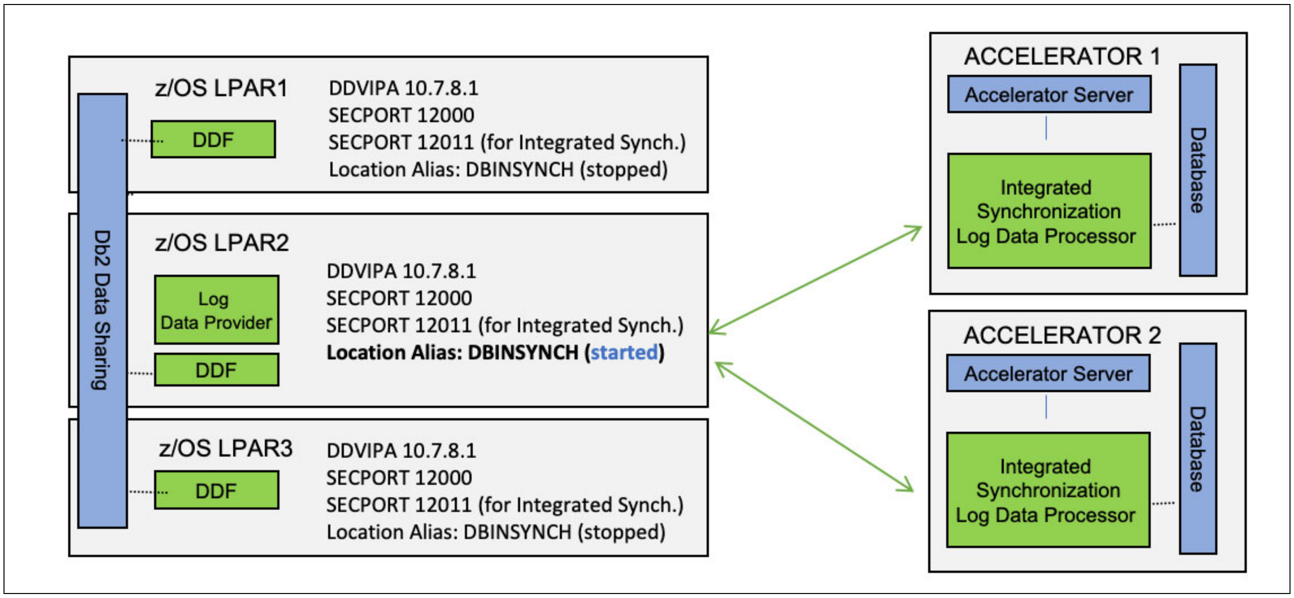

z/OS LPAR1
DDF
DDVIPA 10.7.8.1
SECPORT 12000
SECPORT 12011 (for Integrated Synch.)
Location Alias: DBINSYNCH (stopped)
Db2 Data Sharing
z/OS LPAR2
Log Data Provider
DDF
DDVIPA 10.7.8.1
SECPORT 12000
SECPORT 12011 (for Integrated Synch.)
Location Alias: DBINSYNCH (started)
z/OS LPAR3
DDF
DDVIPA 10.7.8.1
SECPORT 12000
SECPORT 12011 (for Integrated Synch.)
Location Alias: DBINSYNCH (stopped)
ACCELERATOR 1
Accelerator Server
Integrated Synchronization Log Data Processor
Database
ACCELERATOR 2
Accelerator Server
Integrated Synchronization Log Data Processor
Database

Figure 3-6 Data Sharing Group SECPORT and alias definitions for Integrated Synchronization

If the Db2 member that provides the Db2 log data fails, then the location alias must be manually started on a different Db2 member. This way Integrated Synchronization can continue to provide the log data to the Accelerator, At the time of writing an automatic failover of the log data provider from one Db2 member to another is not available.

Because the log data processor on the accelerator(s) connect to the log data provider through a defined TCP/IP address, the log data provider must always be reachable via the same IP address no matter on which Db2 member the log data provider is started. For this process to occur, a distributed dynamic virtual IP address (DDVIPA) is defined for the data sharing group that the accelerator(s) use(s) to establish the connection.

The log data processor component on the accelerator tracks applied changes and notes the last log record sequence number (LRSN) in data sharing or relative byte address (RBA) in non-data sharing, per replicated table. This information is used to continue consistent incremental update processing by requesting past changes from Db2 for z/OS logs if replication was interrupted for some time (either because of manual failover of the log data provider from one Db2 member to another or because of an Accelerator outage).

Information regarding the last applied changes (LRSN/RBA) is used to fully and consistently be compensated by applying the changes since the last committed change on the accelerator.

Built-in consistency checks raise alerts if LRSN/RBA values in the accelerator no longer match with those values in Db2. For more information, see 5.4.2, “Data maintenance by using incremental update” on page 41.

3.3.2 HA setup for incremental update using CDC

Incremental update using CDC is based on IBM InfoSphere® Changed Data Capture (CDC), which is part of the product packaging. The underlying technology uses a capture agent that reads database logs and propagates changes to one or more apply agents. In the context of IBM Db2 Analytics Accelerator, the capture agent is on z/OS and the apply agent is inside the accelerator component.

Figure 3-7 shows a Db2 for z/OS data sharing group with two members. An active log capture agent is present on one Db2 for z/OS data sharing group member. This agent can read and process all logged changes for the entire data sharing group.

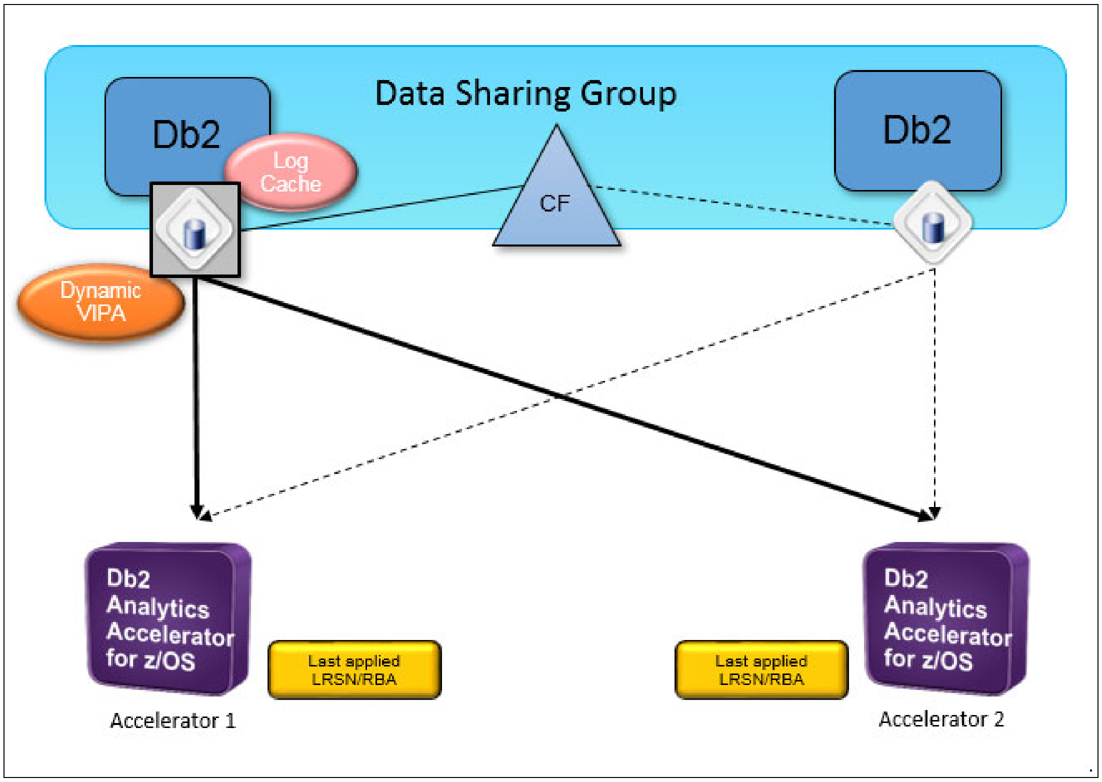


Figure 3-7 Incremental Update with multiple accelerators and active/standby capture agents

Figure 3-7 also shows a standby capture agent that is associated with another member of the Db2 data sharing group. If the active agent fails, this standby capture agent assumes the active role.

Because the components on the accelerator communicate with the active capture agent though a defined TCP/IP address, the active capture agent must always use the same TCP/IP address. This requirement includes the standby capture agent after taking over the active capture agent role. For this process to occur, a dynamic virtual IP address is defined and bound to the active capture agent.

The apply agent component on the accelerator tracks applied changes and notes the last log record sequence number (LRSN) in data sharing or relative byte address (RBA) in non-data sharing, per replicated table. This information is used to continue consistent incremental update processing by requesting past changes from Db2 for z/OS logs if replication was interrupted for some time.

Figure 3-8 shows a failure or outage of the active capture agent log (as seen on the left side). This issue can occur because of a problem with the capture agent software component, Db2 subsystem, or the entire site.

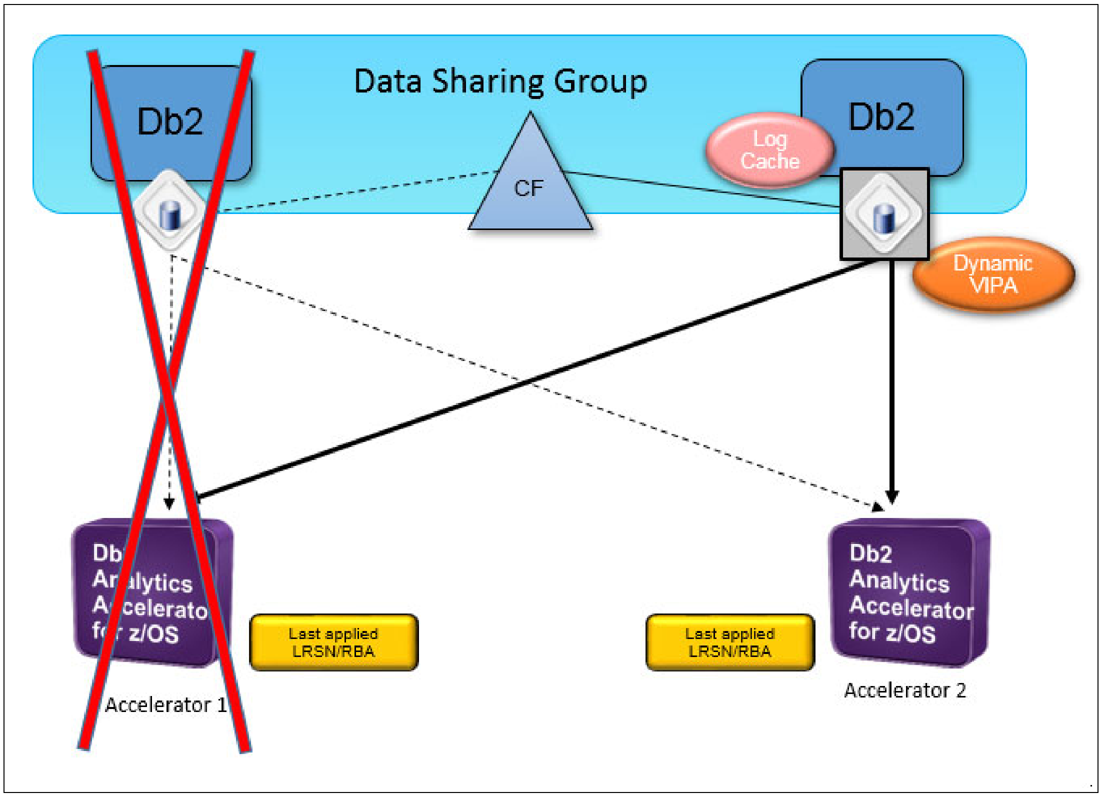


Figure 3-8 Incremental update after the active capture agent failed

The standby capture agent (as seen on the right side of Figure 3-8) is attached to another Db2 for z/OS member and running in another LPAR. This standby capture agent regularly monitors the active capture agent and takes over log capturing automatically if the active capture agent fails.

It also receives the dynamic virtual IP address so that the new active capture agent is accessible from the accelerator by way of the same IP address as before the failure.

Information regarding the last applied changes (LRSN/RBA) is used to fully and consistently be compensated by applying the changes since the last committed change on the accelerator.

Built-in consistency checks raise alerts if LRSN/RBA values in the accelerator no longer match with those values in Db2. For more information, see 5.4.2, “Data maintenance by using incremental update” on page 41.

For administrative purposes, the accelerator server component on the accelerator must establish a connection to the Db2 Data Sharing group by using the Distributed Data Facility (DDF), which is part of Db2.

To open a connection, the DDF port number and an IP address are required. The DVIPA that is configured for high-availability of the CDC Capture agent can be reused for this purpose.

DDF also must be started on all LPARs or Db2 Data Sharing group members that run an active or hot standby CDC Capture Agent. Because the DVIPA is bound to the LPAR that is running the active CDC Capture Agent, the Accelerator Server always connects to the Db2 Data Sharing group by using the DDF started task that runs on the same LPAR as the active CDC Capture Agent.

If the active CDC Capture Agent fails over to another LPAR, the DVIPA is bound to this new LPAR and the DDF started task on this LPAR is used for connecting to Db2 from the accelerator. If it is not started initially, the connection to the Db2 Data Sharing group fails.

For more information about the configuration that is used for this setup, see this IBM Support white paper.

This white paper describes network setup options (private and non-private) for Db2 Analytics Accelerator in general. Also included is a description of how to set up incremental updates and HA for incremental updates within these options.

3.4 High-performance storage saver and multiple accelerators

IBM Db2 Analytics Accelerator can store historical data, which is static data that is not changed. If such data is stored in Db2 for z/OS in range-partitioned tables, it occupies disk storage in your storage system that is attached to z/OS. By archiving this data to IBM Db2 Analytics Accelerator and querying data on an accelerator only, this data (including indexes) is no longer required to be stored on z/OS storage devices. The saved space on z/OS storage can be reused for other purposes while data access to historical (archived) data is still enabled through the accelerator.

In the context of HA, a data strategy must be implemented so that this archived data can still be accessed in a failure or disaster.

IBM Db2 Analytics Accelerator can archive data into multiple accelerators. Because archived data is accessible only for Db2 for z/OS native processing by using image copies that are taken during archiving, it is essential to archive data on multiple accelerators to minimize the risk of data loss.

With multiple accelerators present, after the initial archive process to a first accelerator, a subsequent archiving step to other attached accelerators creates a copy of this archived data in the other accelerators.

Image copies remain the primary backup vehicle and as such, also must be available and accessible on a DR site.

Figure 3-9 shows a table with six partitions. Partitions 1 and 2 include active data, which remains on Db2 for z/OS storage and can be changed. Partitions 3 - 6 include historic (static) data, which does not change and is archived to the accelerators.

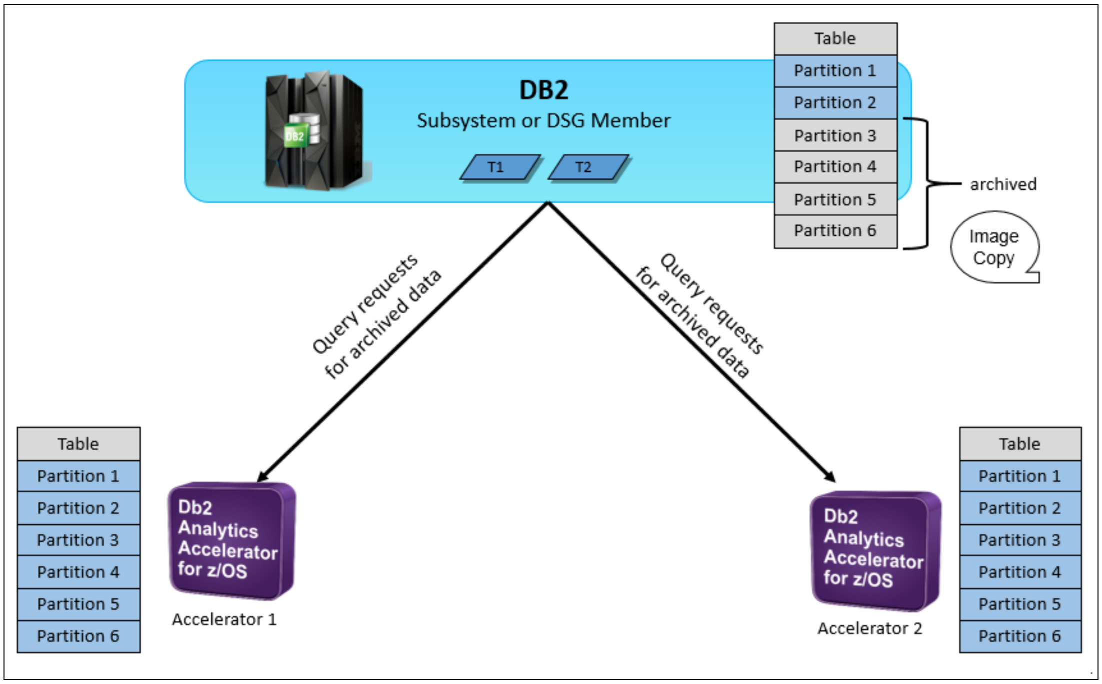


Figure 3-9 High-performance storage server with multiple accelerators

In a first step, an image copy of the archived data (one per partition) is created and the data is stored on Accelerator 1. Functionality to archive data on an accelerator is encapsulated in stored procedure ACCEL_ARCHIVE_TABLES.

In a second step, the same archived data also is stored on Accelerator 2 by calling stored procedure ACCEL_ARCHIVE_TABLES again. Because the data is not available in Db2 for z/OS for unloading, a subsequent invocation of the same stored procedure unloads the data from the image copy that was created during the first invocation of the same stored procedure for the same object.

With this setup, query requests for archived data can be satisfied by both active accelerators. If Accelerator 1 fails, archived data is still accessible through Accelerator 2.

3.5 Federation with multiple accelerators

The term *federation* describes the ability to run accelerated queries against tables that do not belong to or originate from the Db2 subsystem that issues the query.

Usually, an accelerator strictly separates and isolates data for each paired Db2 subsystem. A Db2 subsystem cannot access table data from another Db2 subsystem that happens to be on a paired accelerator.

However, data processing might need data stored from different Db2 subsystems to provide a consolidated view across data. The federation feature provides this access in a controlled way and allows queries to access data originating from other Db2 subsystems than the one submitting the query.

The process to define access to tables that originate from a different Db2 subsystem features the following steps:

1. Using stored procedure ACCEL_GRANT_TABLES_REFERENCES, a user with sufficient rights on the Db2 subsystem that owns the tables (owing Db2 subsystem) grants the right to access a set of tables to a Db2 subsystem that normally cannot access these tables (called the *referencing Db2 subsystem*).
2. Using stored procedure ACCEL_CREATE_REFERENCE_TABLES, a user with sufficient rights on the referencing subsystem creates reference tables. These reference tables contain metadata and pointers to tables on the accelerator that belong to the owning Db2 subsystem with corresponding entries in the catalog of the referencing Db2 subsystem.
3. Queries from the referencing Db2 subsystem can now access these referencing tables.

In an environment with multiple accelerators that are connected to the owning Db2 subsystem and the referencing Db2 subsystem, it is recommended to create the referencing tables on all accelerators (assuming that the owning Db2 subsystem synchronized its tables to all accelerators). This setup is created by calling the ACCEL_GRANT_TABLES_REFERENCES and ACCEL_CREATE_REFERENCE_TABLES stored procedures for each accelerator.

By using this setup, query requests for referencing tables can be satisfied by all active accelerators. If one accelerator fails, the tables are still accessible through another accelerator.

Figure 3-10 on page 29 shows an environment with two Db2 subsystems, two accelerators, and a federation. The owning Db2 subsystem DB2O has two accelerator-shadow tables T1 and T2 synchronized to both accelerators and granted access permission for both accelerators to DB2R as the referencing Db2 subsystem.

The referencing Db2 subsystem DB2R created references to T1 and T2 on both accelerators that are named T1-Ref and T2-Ref. Users of DB2R can run queries on T1-Ref and T2-Ref and join these referencing tables with T3, which is an accelerator-shadow table of DB2R.

The queries are routed to Accelerator 1 or Accelerator 2, depending on usage and availability.

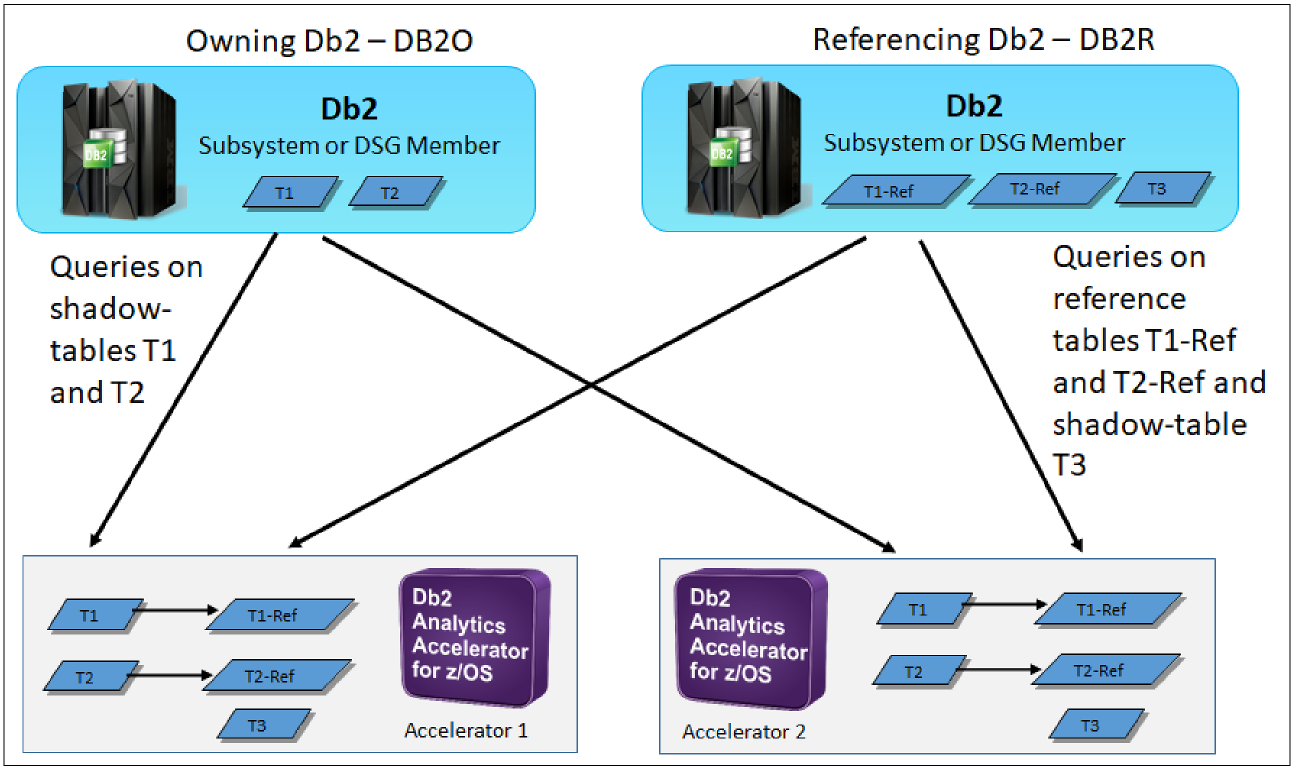


Figure 3-10 Federation with multiple accelerators

3.6 Definition of accelerators in Db2 catalog tables

If a Db2 for z/OS data sharing group was implemented to provide HA capabilities, it is important to understand how and where accelerator definitions are stored.

During the so-called pairing of a Db2 for z/OS subsystem with an accelerator, all configuration parameters are stored in the communication database (CDB) within Db2 for z/OS catalog tables and other pseudo catalog tables. This CDB includes IP addresses of accelerators and an authentication token, which grants access to the data of a paired Db2 for z/OS subsystem or a data sharing group.

Because all members of a data sharing group share these catalog entries, all of them automatically can access a shared accelerator and the data from the respective data sharing group. Assuming that proper installations are completed on all members (for example, stored procedure setup), no other configuration is required if another member (active or standby) wants to access data on an accelerator.

For a DR setup, GDPS can be used to replicate information of an entire Db2 for z/OS subsystem to a remote site. This usually includes Db2 for z/OS setup information and active catalog table content.

If such a dormant Db2 for z/OS subsystem is brought up, it uses definitions in the catalog table to communicate with any connected accelerator, which must be considered.

4

Combining versions and deployment platforms

This chapter provides guidance for what to do if a Db2 subsystem has a V5 Accelerator and a V7 Accelerator, along with the similarities and differences that are offered by these versions.

This chapter includes the following topics:

© Copyright IBM Corp. 2019. All rights reserved.

4.1 Combining V5 and V7 Accelerators

Db2 for z/OS can be connected to multiple Accelerators. The Installation Guide explains how WLM application environments must be configured to support Db2 Analytics Accelerator V5 and V7 at the same time.

If a Db2 subsystem has a V5 Accelerator and a V7 Accelerator that are attached and both include the required table data for a query loaded, it selects one based on configuration precedence settings (for example, ZPARMs or special register).

Running a mix of V5 and V7 accelerators can also be beneficial during an upgrade from V5 to V7 where tables and applications are tested and moved from a V5 to a V7 accelerator.

However, using a V5 Accelerator as a passive backup system or to address disaster recovery cases needs further considerations. Comprehensive application testing is critical.

Because of different capabilities and implementation of the Accelerator between V5 and V7, SQL queries might not be equally supported by both versions. If an application relies on query acceleration and runs well on a primary/active site with a V7 Accelerator, it is crucial to validate that the same queries can also be processed by the backup V5 Accelerator in the secondary site.

4.2 Similarities and differences of IIAS and Z deployments for V7 Accelerators

From an SQL capability and function support level, V7 Accelerators across platforms are identical. Db2 for z/OS does not differentiate between deployment platforms when routing queries to an Accelerator. Therefore, workload balancing is fully supported between all V7 Accelerators, regardless if deployed on IIAS or on IBM Z.

For HA and DR requirements and capabilities, Db2 Analytics Accelerator on IBM Z is more tightly integrated into concepts of the Z platform. Consider the following points:

- IBM Z Enterprise storage (CKD) can be used for operational, runtime, and database data, which enable the use of solutions for backup, mirroring, FlashCopy, and GDPS.
- The Accelerator is deployed as a container in an IBM Z LPAR. An LPAR can be restarted on different hardware by pointing to the same storage. The independence of processors and storage along with IBM Z hardware abstraction makes recovery scenarios easier because “failing servers” do not have to trigger logical and physical data movement and redistribution.

A Db2 Analytics Accelerator on IIAS is not tightly integrated into HA and DR concepts of the Z platform but can complement these concepts by adding Accelerators to the active and passive sites.

However, an Accelerator on IIAS server is designed for HA by containing redundant hardware components. No single point of failure exists. If a hardware component fails (for example, a physical server node), actions are taken automatically to distribute the workload to the remaining nodes.

Because of these architectural differences between IIAS and Z deployments, recommendations for HA and DR configurations with Db2 Analytics Accelerator vary.

5

Planning for HA and DR with IBM Db2 Analytics Accelerator on IIAS

In this chapter, we describe how Db2 Analytics Accelerator on IIAS complements HA and DR topologies for Db2 for z/OS that are managed through GDPS architectures and products.

Typical HA and DR topologies that consist of Db2 active-active or Db2 active-standby configurations or a combination of both were described in Chapter 2, “High availability and disaster recovery in IBM Z enterprise environments concepts” on page 3.

Db2 active-active configurations are managed through GDPS Metro to ensure continuous availability of IT operations, even if a disaster strikes one of the active sites. Db2 active-standby configurations are typically managed by GDPS Global and plan for minimal downtime and a low (but acceptable) data loss to resume normal operations. However, it is also possible to manage a Db2 active-standby through GDPS Metro.

Db2 Analytics Accelerator on IIAS integrates into these topologies and complements them by adding active accelerators to each Db2 active or Db2 standby site and applying established concepts that are described in Chapter 3, “Introducing high availability concepts for IBM Db2 Analytics Accelerator” on page 11.

However, adding a standby accelerator instead of an active one might be considered in certain scenarios. Scenarios of adding active accelerators to Db2 active-active configurations and adding active or standby accelerators to Db2 active-standby configurations are described in this chapter.

© Copyright IBM Corp. 2019. All rights reserved.

This chapter includes the following topics:

5.1 Stand-by accelerators

For most scenarios that are described in this chapter, implementing multiple active accelerators with in-sync data is the preferred option. The benefit is optimal use of accelerator resources and optimal RPO and RTO if an unplanned outage occurs. These benefits include more costs because data must be maintained in multiple accelerators.

The rationale for more stand-by (passive) systems is different for Db2 than for an accelerator. A stand-by Db2 subsystem can be considered because of the following reasons:

- Latency between sites: Longer distances between sysplex members affect transaction times because I/O operations must wait until they are committed on the remote site.
- Licensing: A dormant Db2 for z/OS subsystem causes fewer MLC costs than an active one.

Both of these reasons do not apply to accelerators. Network latency often is not an issue for analytical query processing and the price model for an accelerator is not MLC-based.

However, a stand-by accelerator can be the best choice for the following reasons:

- Lack of network bandwidth between primary and remote site to keep data in-sync between accelerators.
- Non-critical application where a longer recovery time objective (hours to days) is acceptable so that loading the accelerator after a failure is feasible.

5.2 Network setup options

The network setup between IBM Z systems that are hosting the Db2 system (Db2 subsystem or Db2 data sharing group) and the Accelerator on IIAS is independent from the following factors:

- Using GDPS Metro or GDPS Global configurations
- Implementing active or stand-by accelerators

Physical (redundant) network connections must exist between IBM Z systems that are hosting the Db2 systems and all active and stand-by accelerators. Typically, a redundant physical network setup between an IBM Z system and an Accelerator server consists of two OSA cards, two switches, and the appropriate number of redundant cables. TCP/IP settings must be defined for all LPARs that are hosting the Db2 systems to ensure that each Db2 system or member can reach each accelerator by using its IP address.

An example of private data network setup that consists of two LPARs on two IBM Z systems that are connected to an Accelerator on IIAS with three nodes (for example, M4002-003) is shown in Figure 5-1 on page 38.

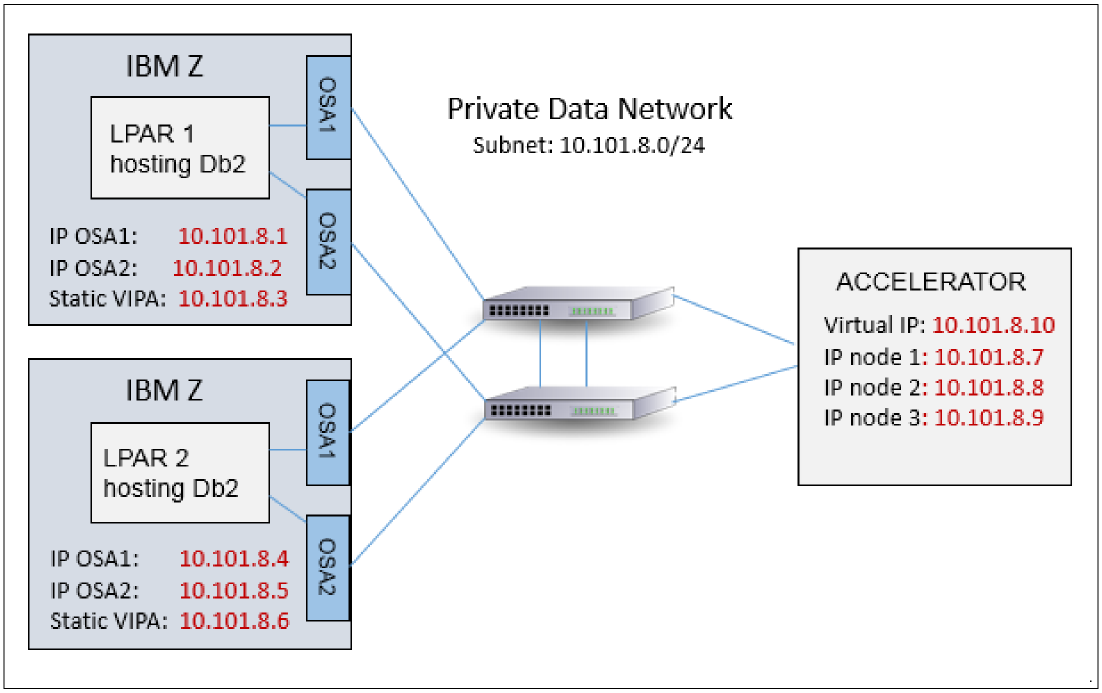


Figure 5-1 High-availability network setup between Db2 and an accelerator

For high-availability (HA) purposes, two OSA cards are used and two switches. This configuration ensures that a route is found to the Accelerator and back, even if one OSA card, one switch, or a cable fails.

Each LPAR has an IP address that is defined per OSA interface. A static virtual IP address (VIPA) is defined over both IP addresses of each LPAR (`10.101.8.3` and `10.101.8.6`). This VIPA ensures that applications on the LPAR can be accessed from the Accelerator, no matter through which OSA card the network traffic flows.

Such a setup is typically used for production systems that have a Db2 Data Sharing Group that is defined across multiple LPARs.

For more information about possible network setup options and high availability aspects for the network setup, see this IBM Support white paper.

5.3 Latency and bandwidth requirements

Network latency and bandwidth requirements for IBM Db2 Analytics Accelerator differ from requirements in a Parallel Sysplex and Db2 data sharing group. Because GDPS Metro implements a guaranteed and synchronous commit protocol for primary and secondary data, network latency affects transaction response times. Practically, this latency limits HA environments to a distance of less than 20 km (12.4 miles).

It is important to understand that an accelerator does not participate in any type of two-phase commit protocol. Therefore, network latency does not have a significant effect on its operation. A query routing request from Db2 for z/OS to an accelerator might take a few milliseconds longer, but with query response times in the range of seconds, minutes, or even hours, this difference usually is not considered to be significant.

However, it is important to provide significant network bandwidth between Db2 for z/OS and the accelerator to allow for data loading and large query result list processing in a reasonable time. Therefore, a 10 Gbps connection is a standard prerequisite to operate an accelerator.

5.4 Db2 active-active configuration with GDPS Metro and active accelerators

In a Db2 active-active configuration with GDPS Metro, one or more active accelerators are connected to the IBM Z servers (CECs) at all active sites. By using the concept of data sharing groups, all accelerators are visible to all Db2 for z/OS members on all active sites.

An example of a data sharing group spanning two sites (Site 1 and Site 2) with Db2 members in each site is shown in Figure 5-2. Primary disks are in Site 1 and secondary disks in Site 2 mirrored with GDPS Metro. The suggested setup includes two accelerators, one for each site. Both sites are connected by way of switches to the Db2 members.

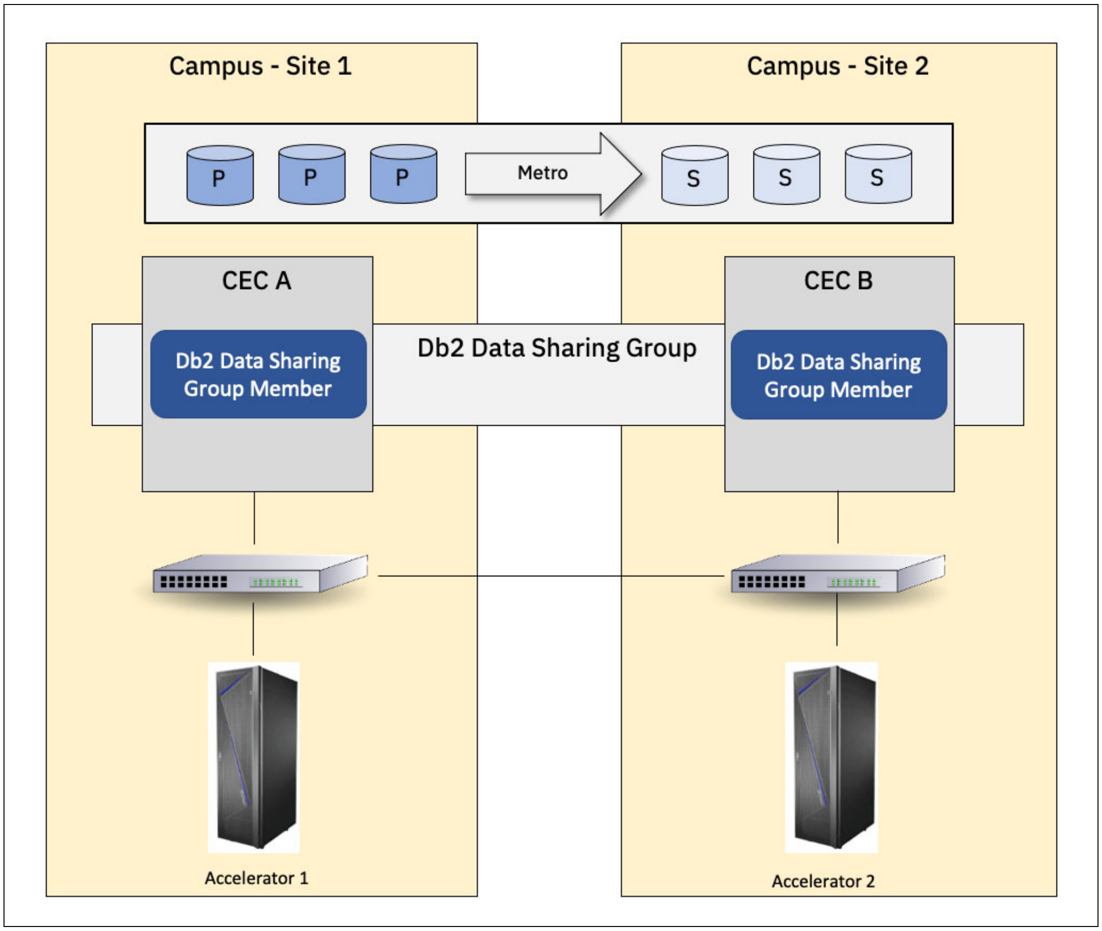


Figure 5-2 Db2 active-active configuration with active IIAS accelerators

As described in 3.2, “Data maintenance and synchronization with multiple accelerators” on page 14, the preferred option is to actively use all available accelerators for query workloads in both sites if resources are available to maintain data in both accelerators.

Figure 5-3 shows how query workload flows with such a setup. The active-active configuration allows the acceptance of requests in both sides. The respective Db2 Data Sharing Group Member then chose dynamically which accelerator is best-suited to process the query request. The workload balancing feature that is described in 3.1, “Workload balancing” on page 12 ensures that workload is distributed among the connected accelerators.

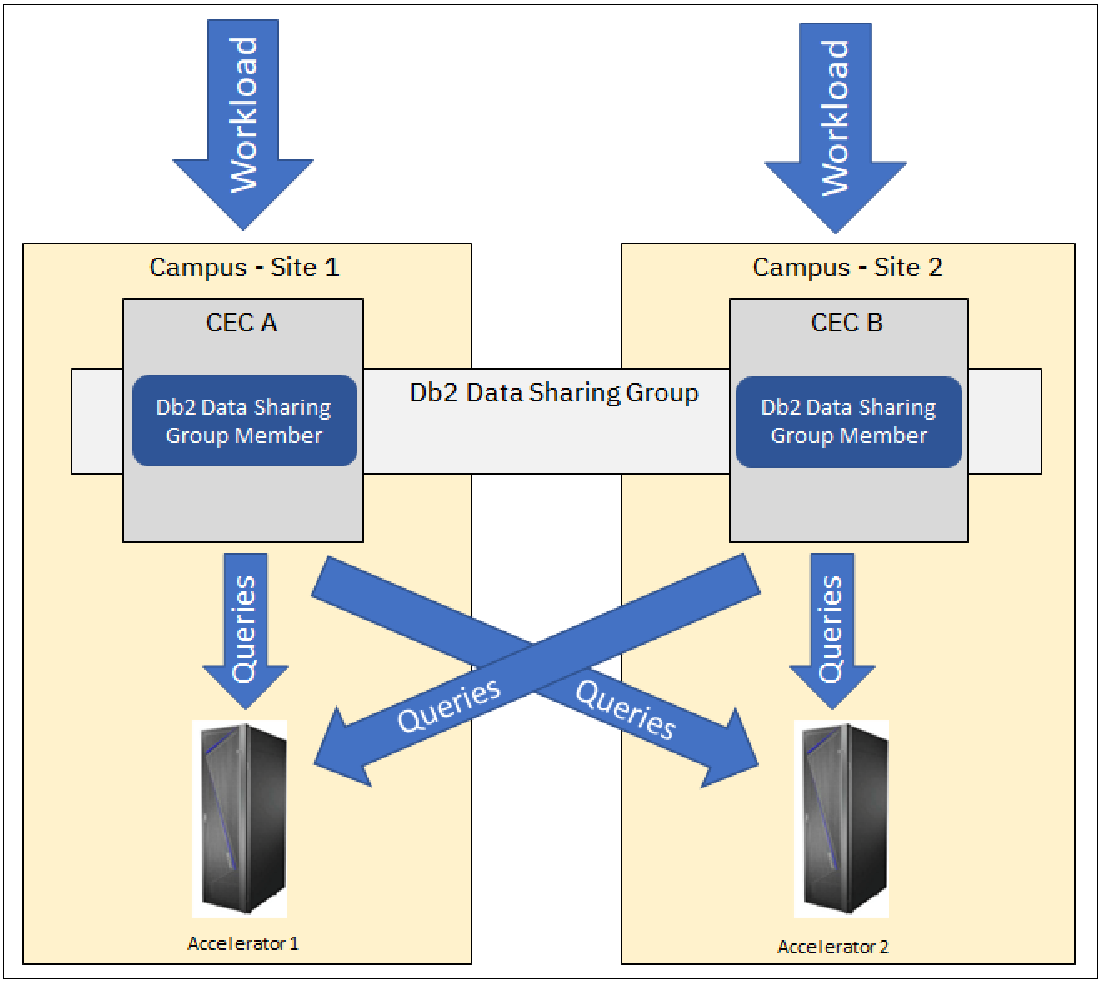


Figure 5-3 Workload flow with active-active configuration and IIAS accelerators

The distance between the sites often does not impose a problem for query processing. In practice, users do not notice a difference if a query was processed on Accelerator 1 or Accelerator 2.

A failure at Site 1 can affect the entire site or certain components. If the entire Site 1 is unavailable, components in Site 2 take over, the secondary disks become active (GDPS HyperSwap), and the Db2 member (or members) in Site 2 no longer use Accelerator 1.

The query workload moves entirely to Site 2. If Accelerator 1 is also affected by the outage, the Db2 subsystem in Site 2 automatically detects that Accelerator 1 is unavailable and continues with the remaining query workload only on Accelerator 2.

If Accelerator 1 is available and only the Db2 subsystem on Site 1 is down, both accelerators are used.

5.4.1 Data maintenance by using regular load cycles

By applying the same data maintenance processes to all connected accelerators, the same data is stored on all connected accelerators. This configuration allows Db2 for z/OS to balance incoming query requests between accelerators, which results in a balanced usage of all accelerators and best possible response times by even workload distribution to appliances.

Synchronous data maintenance to all connected accelerator can easily be achieved by using the HA Load feature of the Db2 Analytics Accelerator Loader product, as described in 3.2.2, “Loading into multiple accelerators using Db2 Analytics Accelerator Loader” on page 16 or by leveraging the HA Load sample program as described in 3.2.3, “Loading into multiple accelerators using the HA Load sample program of Db2 Analytics Accelerator” on page 17".If both is not possible you use the stored procedure ACCEL_LOAD_TABLES.

However, if the stored procedure is called multiple times to load data into more than one accelerator, it is likely that all invocations complete at a different time. Therefore, the newly loaded data is available first on one accelerator, then on the other. The time difference between completing both stored procedure calls shows queries to obtain different result sets. No automation is in place to prevent this issue from occurring.

If different results cannot be tolerated while data is loaded into one but not into another accelerator, a potential mechanism to work around this issue is described in “Data consistency across multiple accelerators” on page 15.

Because all connected accelerators can be reached from each site and all accelerators contain the same data for regularly loaded tables, query acceleration remains active, even if one site is not available.

This concept allows for RPO = 0 (including query acceleration) and maintains the same RTO as it is achieved without an accelerator in the same configuration.

5.4.2 Data maintenance by using incremental update

The suggested option to keep both accelerators in sync also applies to the concept of incremental update. As described in 3.3, “HA setup for incremental update” on page 21, one site has the active capture agent (for Incremental Update with CDC) or log data provider (for Incremental Update with Integrated Synchronization) started and propagating changes to multiple active Accelerators. If this site goes down then the same components on the other site can take over this role of replicating data to the remaining Accelerators.

For Incremental Update with CDC this failover is done automatically if a standby capture agent is setup on the remaining site. This standby capture agent becomes active and continues with incremental update processing. For Incremental Update with Integrated Synchronization this failover currently must be done manually by starting the log data provider on the other site.

This concept is not limited to a single data sharing group. Multiple Db2 data sharing groups or stand-alone Db2 for z/OS subsystems (with their own capture agent instance) can propagate their changes to multiple accelerators.

5.4.3 High-performance storage saver

For an active-active configuration, all archived data should be stored in all active accelerators, as described in 3.4, “High-performance storage saver and multiple accelerators” on page 26.

This configuration allows for continuous processing and for requests that access archived data.

5.5 Db2 active-standby configuration with GDPS Metro and active accelerators

In a Db2 active-standby configuration with GDPS Metro, one or more accelerators are connected to an IBM Z server at the primary and secondary sites. All accelerators are visible to all Db2 for z/OS members on the primary and secondary sites, as shown in Figure 5-4.

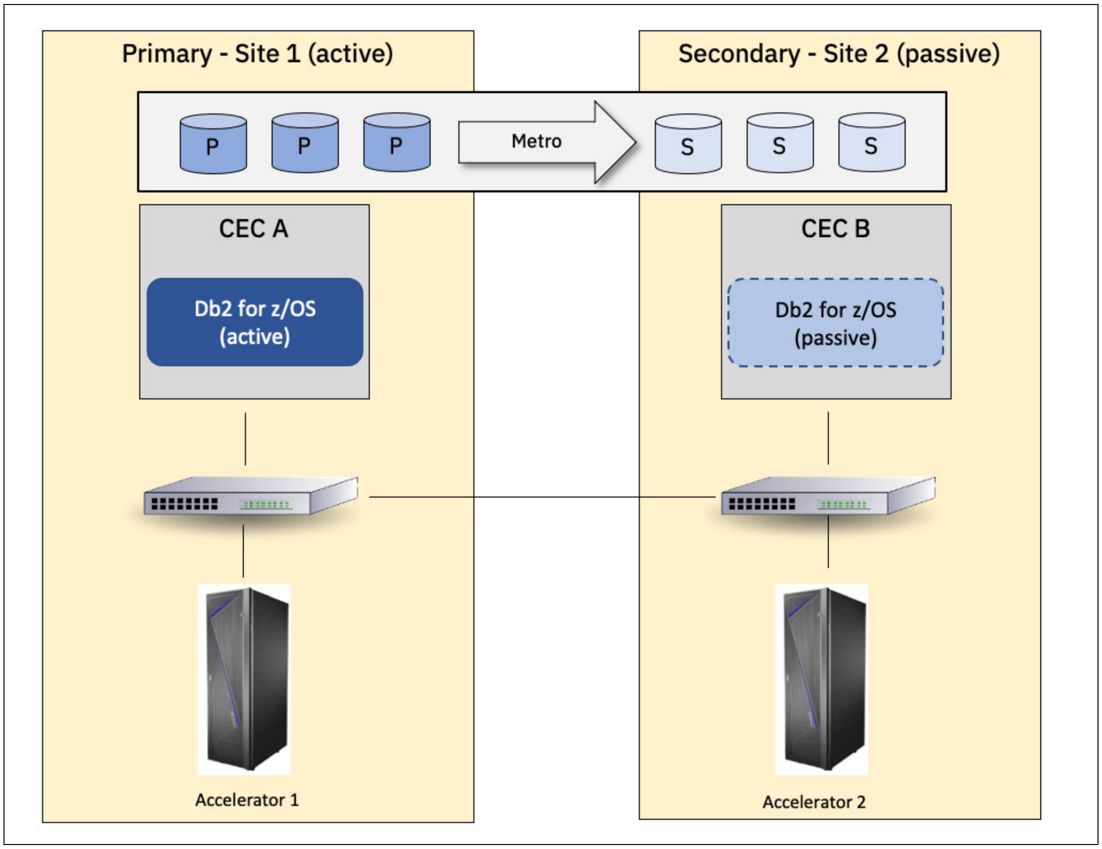


Figure 5-4 Db2 active-standby configuration with GDPS Metro and IIAS accelerators

The suggested configuration for this setup has two active accelerators. Although the Db2 for z/OS subsystem in Site 2 is in standby mode, Accelerator 2 in Site 2 is active.

To support a quick failover if a failure occurs, data on the second accelerator must be kept updated. It is advantageous to use the second accelerator for regular processing as well, except for other considerations, such as power consumption.

Figure 5-5 shows how queries are processed with this setup. Only the active Db2 for z/OS subsystem in Site 1 accepts requests. Because Accelerator 1 and Accelerator 2 are updated and active, query routing considers both accelerators.

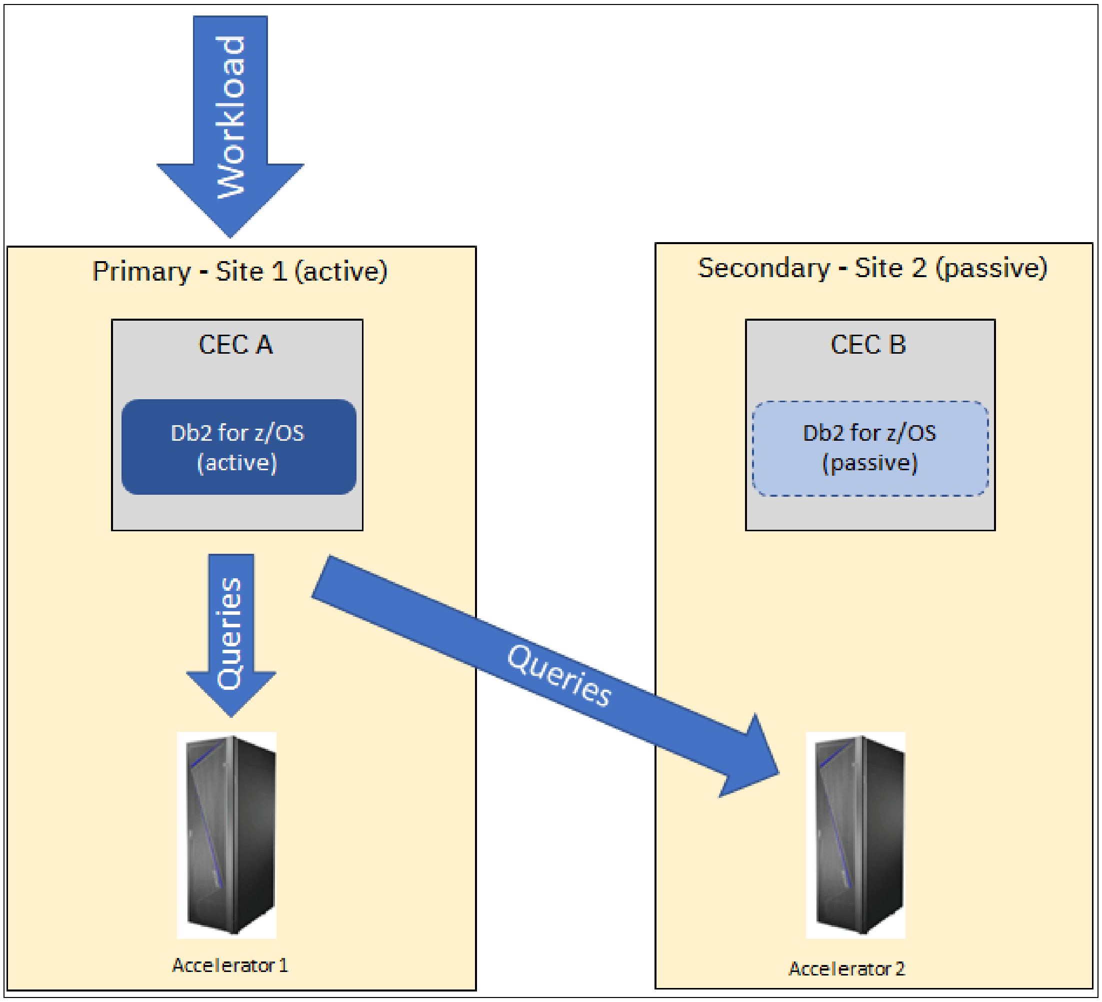


Figure 5-5 Workload flow in Db2 active-standby configuration with active IIAS accelerators

The picture changes after a failure of active Site 1. The passive Db2 for z/OS subsystems in Site 2 take over, which access the secondary DASD volumes. The shared accelerator definitions in the Db2 catalog tables ensure that the standby Db2 for z/OS subsystem can seamlessly access the previously configured accelerators.

If the Db2 for z/OS subsystem in Site 2 is a clone of the Db2 for z/OS subsystem in Site 1, all information to access both accelerators also is available after a failover to Site 2. Accelerator access information is stored in Db2 for z/OS catalog and pseudo-catalog tables. It is *not* mandatory to operate in data sharing mode to benefit from this concept.

Query workload is then accepted by only Db2 for z/OS subsystems in Site 2. The now activated Db2 for z/OS subsystem accesses and uses Accelerator 2. Depending on the scope of the outage in Site 1, it also can use Accelerator 1 (if it is not affected by the outage).

Accelerator data maintenance for load and incremental update is the same, as described with GDPS active-active in 5.4.1, “Data maintenance by using regular load cycles” on page 41, 5.4.2, “Data maintenance by using incremental update” on page 41, and 5.4.3, “High-performance storage saver” on page 41.

5.6 Db2 active-standby configuration with GDPS Global and active accelerators

The suggested configuration for this setup depends on available network bandwidth between both sites. In our scenario, we assume that network bandwidth is sufficient and that accelerators on both sites can be operated and maintained as active accelerators.

Scenarios in which bandwidth is insufficient and do not provide capabilities to operate an accelerator at the secondary site as part of the primary site's IT services are described in 5.7, "Db2 active-standby configuration with GDPS Global and limited bandwidth to remote accelerator" on page 46.

Figure 5-6 shows the recommended configuration with two accelerators in both sites that use a Db2 active-standby configuration with GDPS Global.

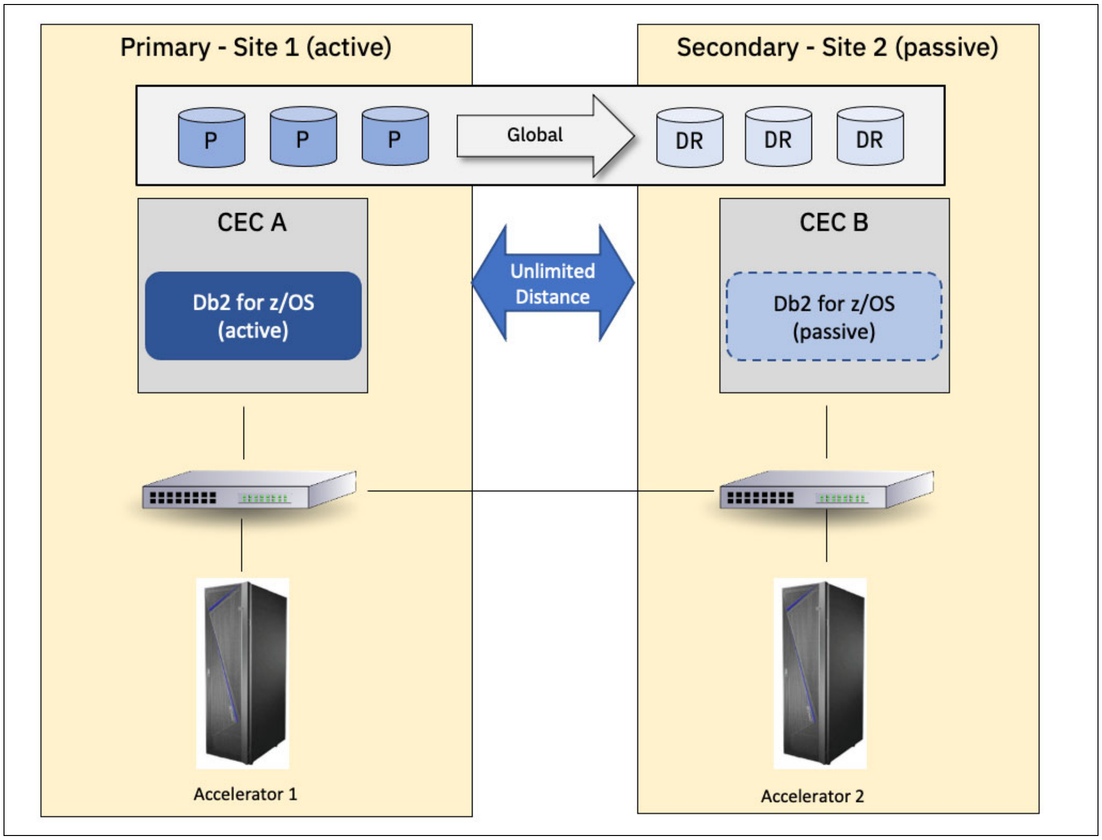


Figure 5-6 Db2 active-standby configuration with GDPS global and active IIAS accelerators

Incoming query requests in Site 1 use Accelerator 1 and Accelerator 2 because both accelerators can be reached from both sites. Even if the distance for GDPS Global exceeds supported distances for GDPS Metro architectures, bandwidth might be enough to allow for query requests and accelerator maintenance over long distances from Site 1.

Access information for both accelerators is stored in catalog and pseudo-catalog tables of the Db2 for z/OS subsystem in Site 1. Because of data duplication to Site 2, the same access information can be used if a failover occurs after the Db2 for z/OS subsystem is started in Site 2. The workload processing flow is the same, as shown in Figure 5-6.

If a disaster strikes Site 1, the Db2 for z/OS subsystem in Site 2 is brought up. Data that was not duplicated at DASD volume level between the last data duplication cycle and the disaster is lost. Because catalog and pseudo-catalog tables also are available to the Db2 for z/OS subsystem in Site 2, both accelerators can be used instantly.

If Site 1 is down, only Accelerator 2 is used for query routing after the failover. If the disaster affects only the Db2 subsystem and Accelerator 1 is still accessible, Accelerator 1 continues to serve query requests from Site 2.

5.6.1 Data maintenance by using regular load cycles

Consider maintaining data in Accelerator 2, as described in 5.4.1, “Data maintenance by using regular load cycles” on page 41.

5.6.2 Data maintenance by using incremental update

Consider maintaining data in Accelerator 2, as described in 5.4.2, “Data maintenance by using incremental update” on page 41.

With asynchronous copy, a situation can occur in which changes are applied to the accelerator with incremental update, while changes though GDPS Global were not yet written to the secondary disk at the disaster recovery site.

The recorded LRSN or RBA in the accelerator does not match the records in the Db2 subsystem. In such a situation, the incremental update component recognizes this inconsistency and the corresponding table on the accelerator must be reloaded.

5.6.3 High-performance storage saver

Consider archiving data in all active accelerators, as described in 3.4, “High-performance storage saver and multiple accelerators” on page 26. This configuration allows for continuous processing and for requests that access archived data.

5.7 Db2 active-standby configuration with GDPS Global and limited bandwidth to remote accelerator

If bandwidth is not sufficient to maintain data regularly on the accelerator in Site 2 (see Figure 5-7), query workload is processed only by Accelerator 1 during normal operations.

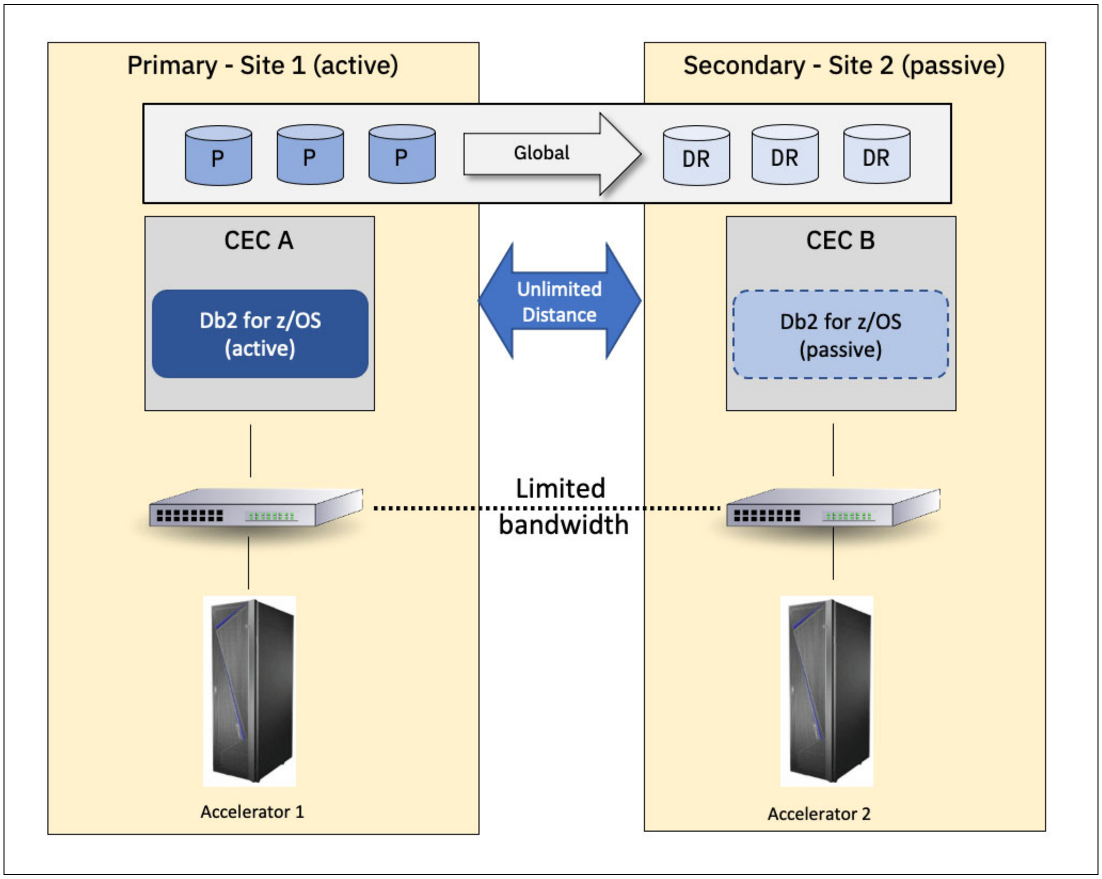


Figure 5-7 Db2 active-standby configuration with limited bandwidth to remote IIAS accelerator

Even with limited bandwidth, Accelerator 2 is paired with the Db2 for z/OS subsystem in Site 1. This configuration ensures that if a failover occurs, the Db2 for z/OS subsystem in Site 2 can be brought up and is ready to use accelerators in both sides.

If no TCP/IP connectivity is available from Site 1 to Accelerator 2, Accelerator 2 must be paired with the Db2 for z/OS subsystem in Site 2 after a failover was started.

If it is not possible to maintain data on Accelerator 2 from the active Db2 for z/OS subsystem in Site 1, Accelerator 2 is stopped during normal operations. Accelerator 2 must be loaded to satisfy incoming query requests if a disaster occurs. This issue leads to RTO > 0 from an accelerator perspective, as shown in Figure 5-8 on page 47.

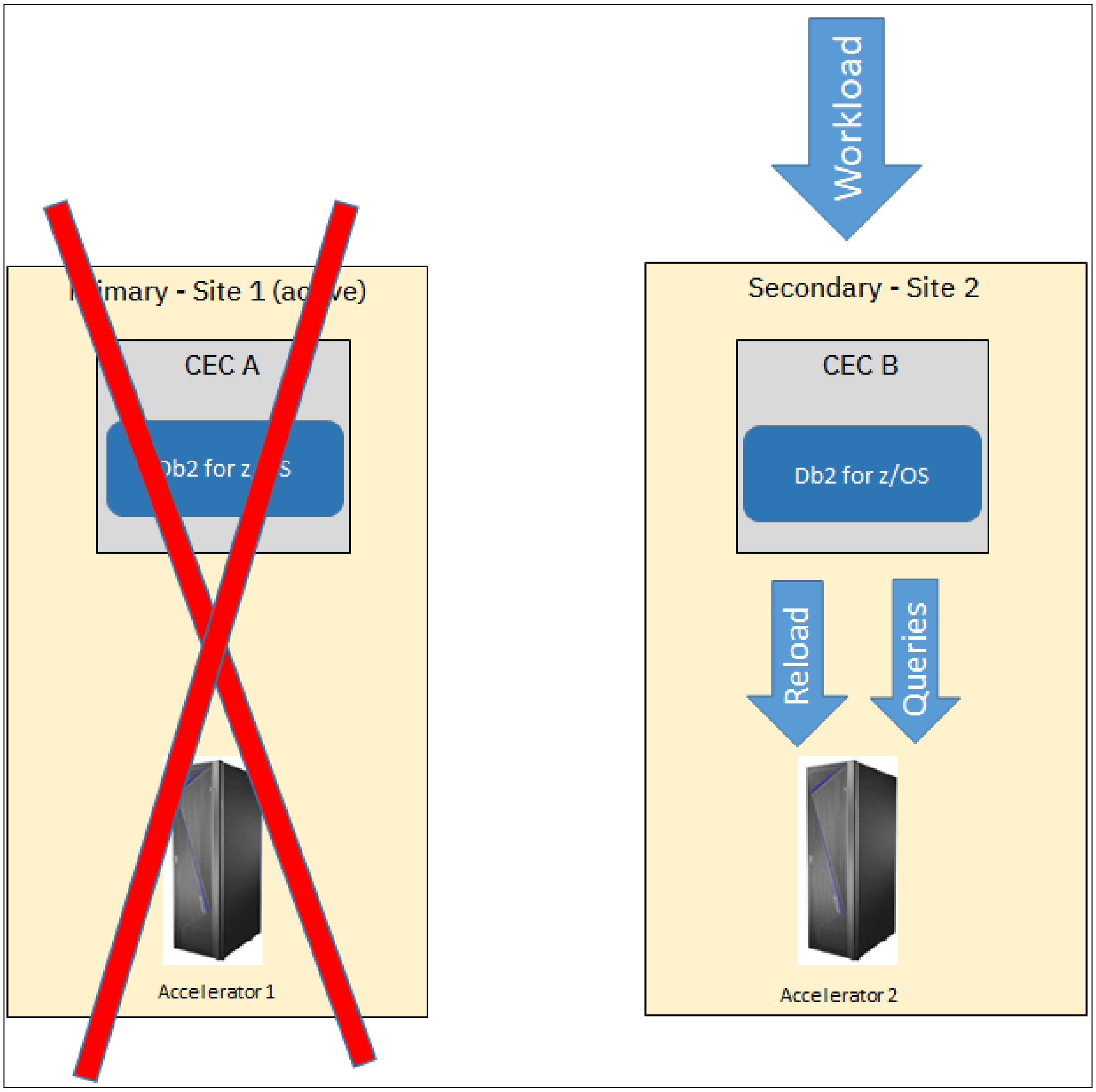


Figure 5-8 Accelerator processing after disaster if bandwidth is limited

After a failover, Accelerator 1 cannot be used for query workloads from the Db2 for z/OS subsystem in Site 2 because it cannot be maintained. Therefore, Accelerator 1 must be stopped from the Db2 for z/OS subsystem in Site 2 after a failover.

If available bandwidth does not permit maintaining data in Accelerator 2 through incremental update during ongoing operations, Accelerator 2 must undergo an initial load of all required tables after a disaster before reenabling incremental update for those tables.

Data that was archived on Accelerator 1 must be manually recovered only into Db2 for z/OS and archived again on Accelerator 2 after a failover.

Image copies that were automatically taken when data was archived on Accelerator 1 were also mirrored in the secondary DASD volume in Site 2. To recover previously archived partitions in Db2 for z/OS, persistent read only (PRO) state must be removed from partitions flagged as archived in Db2 for z/OS. After the PRO state is removed, RECOVER utility restores previously archived data in Db2 for z/OS. After data is restored in DB2 for z/OS, it can be archived again in Accelerator 2.

5.8 Combined Db2 active-active-standby environment with GDPS Metro and GDPS Global and accelerators

The concepts that are described in this IBM Redpaper publication can be combined to build a suitable configuration for even more complex setups.

An example in which GDPS Metro is used to implement an active-active solution with two systems on a campus (closely located to each other) and another remote site to cover a disaster recovery scenario is shown in Figure 5-9.

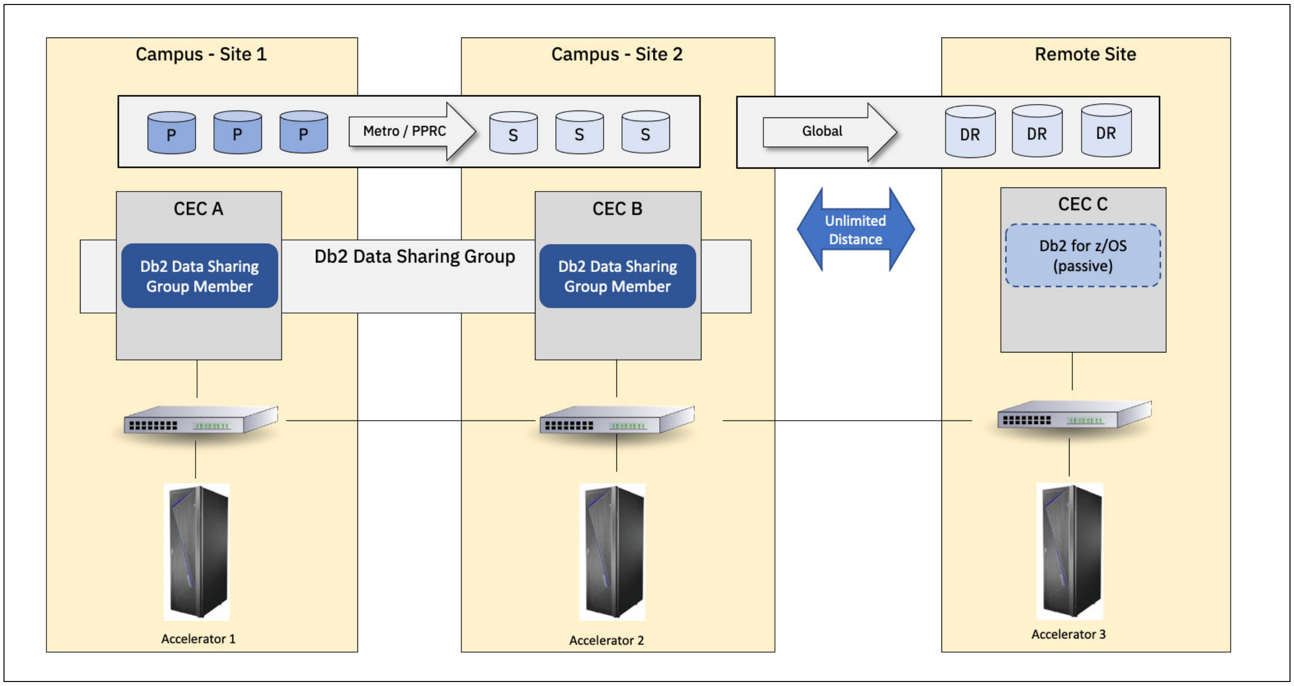


Figure 5-9 Combined Db2 active-active-standby environment with IIAS accelerators

The suggested setup is to have multiple accelerators (at least one per site) to address the same outage scenarios as with the Db2 subsystems. All accelerators can be maintained and active to serve query requests.

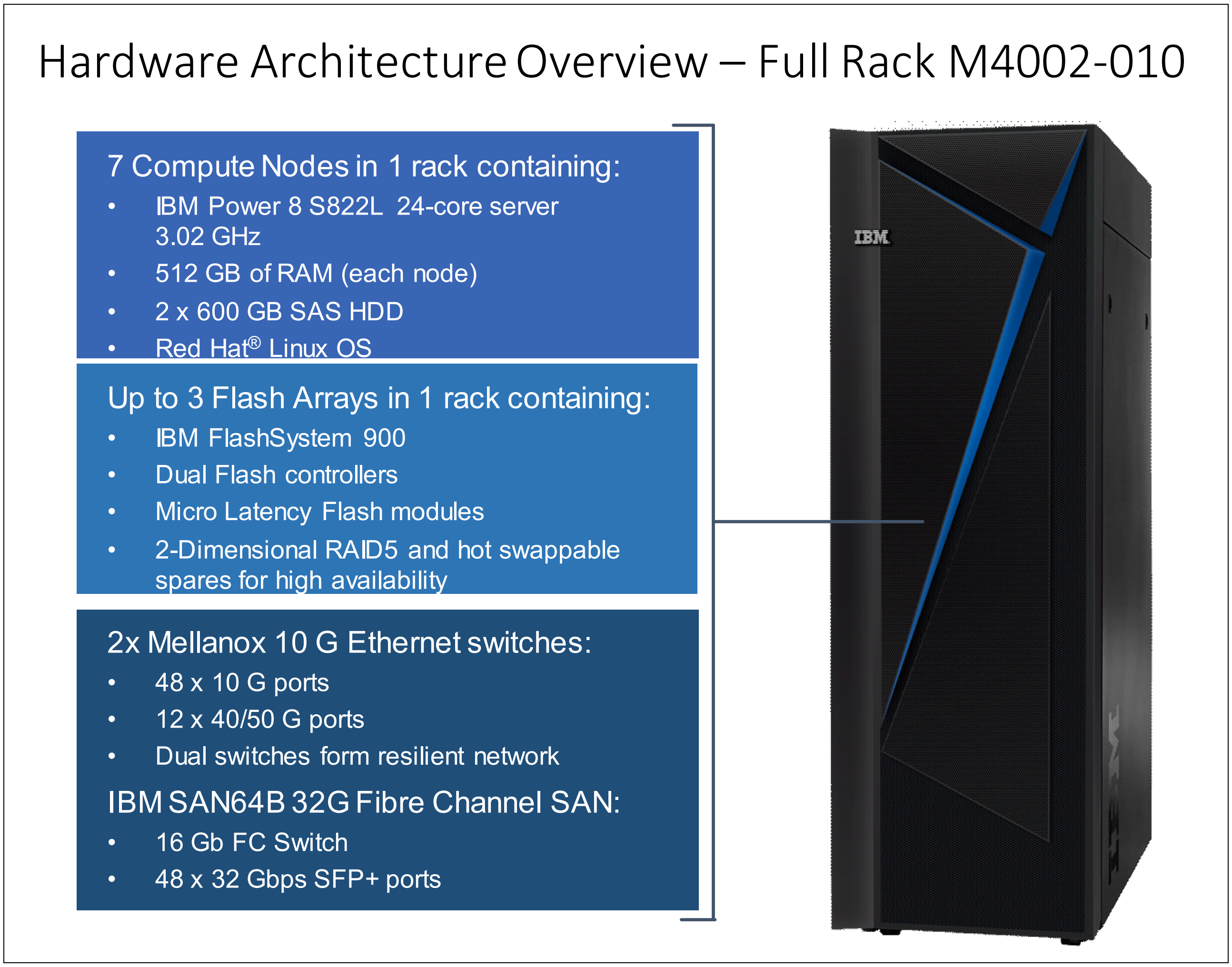
5.9 Built-in HA features within IBM Db2 Analytics Accelerator on IIAS

The IBM Db2 Analytics Accelerator on IIAS consists of multiple components that contribute to HA inside of the physical machine.

The following components are inherited from the underlying IBM Integrated Analytics System architecture (as shown in Figure 5-10), which was designed in a way to ensure that no single point of failure exists with redundancies and fault tolerance:

- Multiple compute nodes, consisting of IBM Power 8 servers
- Multiple IBM FlashSystem® storage arrays
- Multiple redundant networks:

- Data fabric network
- Management network
- Storage area network



Hardware Architecture Overview – Full Rack M4002-010

7 Compute Nodes in 1 rack containing:

- IBM Power 8 S822L 24-core server 3.02 GHz
- 512 GB of RAM (each node)
- 2 x 600 GB SAS HDD
- Red Hat® Linux OS

Up to 3 Flash Arrays in 1 rack containing:

- IBM FlashSystem 900
- Dual Flash controllers
- Micro Latency Flash modules
- 2-Dimensional RAID5 and hot swappable spares for high availability

2x Mellanox 10 G Ethernet switches:

- 48 x 10 G ports
- 12 x 40/50 G ports
- Dual switches form resilient network

IBM SAN64B 32G Fibre Channel SAN:

- 16 Gb FC Switch
- 48 x 32 Gbps SFP+ ports

Figure 5-10 Full rack IIAS hardware architecture

The compute nodes are responsible for running the incoming workload in a highly parallel fashion on multiple logical database nodes (MLNs). One node serves as the head node, which receives a query request and is responsible for returning the query result.

HA is accomplished through node redundancy. If a node becomes non-operational, the workload and MLNs are redistributed across the remaining nodes. An outage of some minutes occurs while the MLNs are redistributed.

Consider the following points:

- In a 1/3-rack appliance, a minimum of two nodes must be operating.
- In a 2/3-rack appliance, a minimum of three nodes must be operating.
- In a full-rack appliance, a minimum of four nodes must be operating.

If the head node fails, another node takes over this role. Because all nodes have network connectivity to Db2 for z/OS through a virtual IP address, this role change is seamlessly for the accelerator.

Multiple IBM Flash System 900 arrays are used for storage, each consisting of multiple IBM MicroLatency® Flash Storage Modules that provide a fully resilient storage subsystem for the accelerator. The storage arrays are protected with two load sharing power supplies, redundant fans, and two separate storage controllers,

A two-dimensional RAID5 layout is used to provide user data protection and redundancy. Two-dimensional (2D) flash RAID5 consists of IBM Variable Stripe RAID™ and system-wide RAID 5. Variable Stripe RAID technology helps reduce downtime and maintain performance and capacity if a partial or full flash chip failure occurs. With easily accessed hot swappable flash modules, System-wide RAID 5 helps prevent data loss and promote availability.

RAID 5 configurations provide a high degree of redundancy with Variable Stripe RAID and RAID 5 protection. RAID 5 data protection includes one IBM MicroLatency module that is dedicated as parity and another that is dedicated as a dedicated hot spare.

Multiple networks are contained in the IIAS system for different purposes. Each network uses a pair of switches providing full failover redundancy.

The data fabric network is used to transfer data from outside to the compute nodes. It also is used to transfer data between compute nodes during workload execution.

The management network is used to manage all components. The Fibre Channel SAN is used for fast data access and data transfer from the flash storage to the compute nodes.

6

Planning for HA and DR with IBM Db2 Analytics Accelerator on Z

Planning for HA and DR with the Accelerator involves multiple factors, including network configuration and storage planning.

This chapter includes the following topics:

© Copyright IBM Corp. 2019. All rights reserved.

6.1 IBM Z integration, network, and storage management

One key design point of Db2 Analytics Accelerator on Z is its integration in the IBM Z enterprise configuration. This integration includes CPs and memory and the option to take advantage of IBM Z enterprise storage capabilities. Furthermore, GDPS components were integrated into the Accelerator as part of a wider GDPS configuration.

This section explains how a defined Accelerator supports the move to a different IBM Z hardware environment, particularly regarding storage and network.

The Accelerator uses a set of defined disks to operate. These are determined initially by the boot disk (which is specified during image upload or LPAR startup in the SSC installer) and additional disks for runtime and data pools, as defined in the Accelerator configuration file.

If the Accelerator is brought up with a different boot disk, corresponding runtime and data disks must be identified and accessed by the Accelerator. This relationship can be defined as "storage environment" in the configuration file or definition can be delegated to the build-in GDPS agent which gets storage configuration from the defined GDPS controlling system.

GDPS mode and GDPS servers are specified in the configuration file as shown in Example 6-1.

Example 6-1 Configuration file

```
"gdps_mode": "true",
"gdps_servers": {
    "server1": {
        "ipv4": "10.2.1.11",
        "port": "1020"
    },
    "server2": {
        "ipv4": "10.2.2.11",
        "port": "1030"
    }
}
```

The GDPS components will then control automatically on which LPAR and with which boot disks the Accelerator will be started in case of a failure. Relationship between the boot disk and the corresponding data disks is defined and managed by GDPS so that the Accelerator can restart on another site with the right data pools in place. The following sections in this chapter explain in more detail how GDPS and the Accelerator operate in different configurations.

Definition of multiple storage environments is another option, if the use of GDPS is not possible or not desired. The configuration file section for "storage environments" looks like Example 6-2.

Example 6-2 Configuration file

```
"storage_environments": [
{
    "boot_device": {
        "type": "dasd", "device": "0.0.5e29"
    },
    "runtime_devices": {
        "type": "dasd",
        "devices": ["0.0.5e25", "0.0.5e26"]
    },
    "data_devices": {
        "type": "dasd",
        "devices": ["0.0.5e14",["0.0.5e80","0.0.5e8f"]]
    }
},
{
    "boot_device": {
        "type": "dasd", "device": "0.0.1b11"
    },
    "runtime_devices": {
        "type": "dasd",
        "devices": ["0.0.1b25", "0.0.1b26"]
    },
    "data_devices": {
        "type": "dasd",
        "devices": ["0.0.1c14",["0.0.1d00","0.0.1d0f"]]
    }
}
]
```

The configuration basically defines the runtime and data disks that should be used when the Accelerator is started, based on the boot disk device ID. Figure 6-1 shows two sites (CEC A/LPAR A1 and CEC B/LPAR B1) with boot disk device 0.05e29 or 0.0.1b11 respectively. If the Accelerator is started on LPAR B1 with boot disk 0.0.1b11, the configuration tells the system which disks to use for data and runtime data pools.

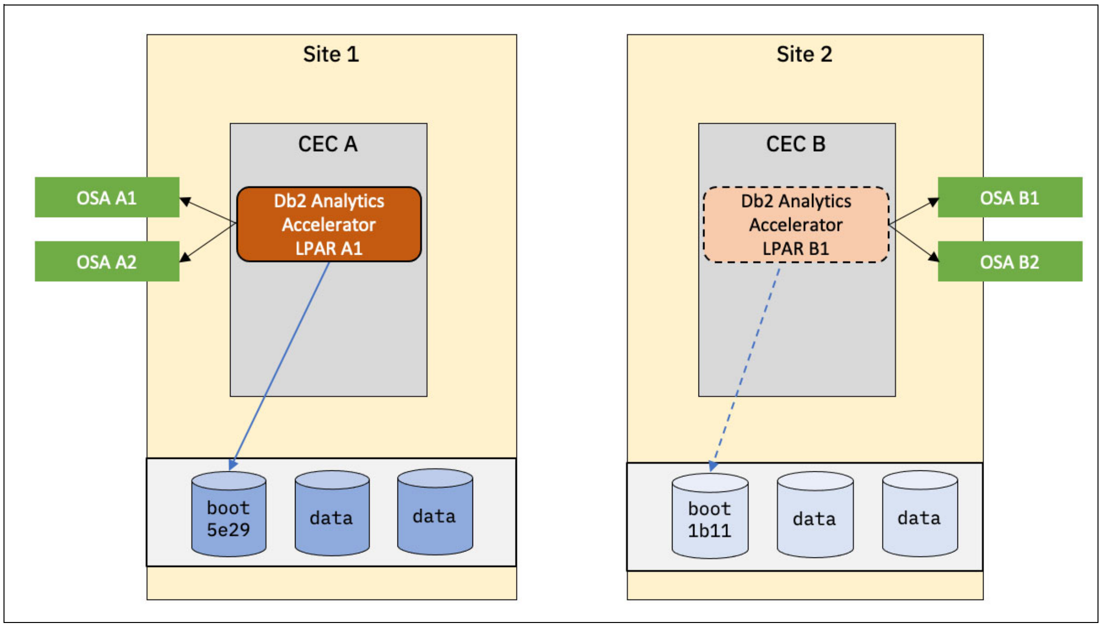


Figure 6-1 Storage and network definitions per site

In contrast to GDPS integration, defining multiple storage environments does not handle disk mirroring, LPAR management and boot disk assignment. But even if LPAR restart and boot disk assignment must be managed manually, this configuration option opens a large set of options:

- Manual HA/DR scenarios (where disks are mirrored by another component and manual LPAR start is acceptable or automation is implemented differently)
- Tests on a flash copy of the data
- Storage subsystem migration with new device IDs

Accelerator configurations can be changed dynamically by uploading a modified configuration file. This allows to add storage environments in preparation for a planned switch to another set of disks.

A planned storage subsystem migration could add the details of the new disks, then bring the Accelerator down, copy the data from the old to the new environment and then start the Accelerator with the new boot disk in the new target environment.

In case storage subsystems use the same device IDs, the specification of the optional storage_uuid attribute helps to address the desired subsystem disks.

Network configuration has two parts. First the network configuration for an SSC LPAR in the HMC activation profile defines an initial network interface, which might be used as a management network for the Accelerator. Second, the Accelerator configuration file allows definition of additional network interfaces, particularly one that is used to connect to Db2.

In a multi-site setup, these extra network connections are also configured per “site” (uniquely defined by CPC name and LPAR name). When an Accelerator starts on a specific site the network configuration from the SSC LPAR definition is applied and then, the extra connections are applied as defined for this particular site in the configuration file.

Example 6-3 assumes that the data network (for Db2 connection) is defined as "db2osa" and has different OSA card device ID definitions for CEC A/LPAR A1 and CEC B/LPAR B1.

Example 6-3 Configuration file

```
"runtime_environments": [
    {
      "cpc_name": "CECA",
      "lpar_name": "LPARA1",
      "network_interfaces": [
        {
          "name": "db2osa",
          "device": "0.0.4b00",
          "port": "0"
        }
      "cpc_name": "CECB",
      "lpar_name": "LPARB1",
      "network_interfaces": [
        {
          "name": "db2osa",
          "device": "0.0.4300",
          "port": "0"
        }
      ]
    }
    ],
```

6.2 Network setup options

As with Db2 Analytics Accelerator on IIAS, network connections from the Accelerator to Db2 subsystems can be implemented in a redundant way to address component failures.

This process is achieved by using multiple OSA cards, multiple switches, and redundant cables between components, as shown in Figure 6-2. The Accelerator allows definition of bonding interfaces over two defined Ethernet connections, which provides a single Accelerator IP address that is based on two physical interfaces.

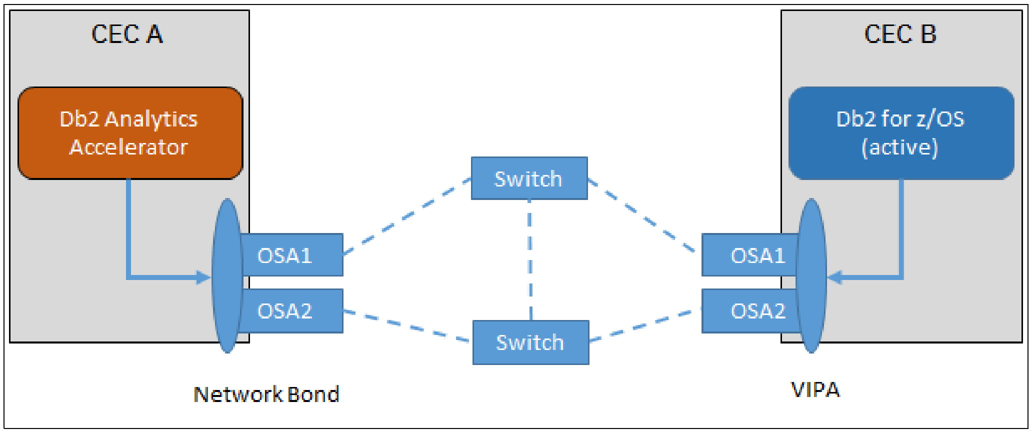


Figure 6-2 Accelerator on Z network connection using OSA cards

From a (continuous) operational perspective, the network interface between the Accelerator and connected Db2 for z/OS subsystems is the most important interface.

In case of a failover scenario where another Accelerator on a different CEC takes over, this Accelerator has some flexibility in defining different network interfaces (also by using different devices) but must define the same IP address for connectivity to Db2 and CDC components. Otherwise, a reconnect attempt from Db2 or CDC is unsuccessful after a failover.

6.3 Latency and bandwidth requirements

Network interfaces for Db2 Analytics Accelerator on IBM Z are defined for the following purposes:

- Data network connection for loading, replicating, and querying data between Db2 for z/OS and the Accelerator
- Management network connection for accessing the Admin User Interface and managing the solution
- If integration with GDPS is used, GDPS network connection for interaction with a GDPS controlling system

The data network connection is the most important connection in terms of data volume, latency, and bandwidth. As with V5 Accelerators and deployments on IIAS, latency is often not an issue and a connection between Accelerator and Db2 subsystem can span a larger distance.

However, bandwidth is critical and directly affects load rates and replication capabilities. Therefore, we recommend at least a 10 GbE connection for the data network connection or the use of IBM HiperSockets™ (if both components are on the same CEC).

Although HiperSockets provide good performance and are highly reliable, secure, and do not need more physical network hardware, they are implemented in software. Therefore, they incur extra CPU costs.

The management network connection is used to upload and update Accelerator images that are several GB in size. Therefore, remote access and image transfer through a low-bandwidth ADSL line is not recommended.

For efficient image transfer, use a PC or server with a 1 GbE connection to the Appliance. As an alternative to the web-based User Interface, a set of Python sample scripts can be used on an attached Linux computer for Accelerator image handling.

For GDPS Metro or GDPS Global disk mirroring, the underlying network latency directly affects I/O latency and performance of the storage subsystem. This issue is not specific to Db2 Analytics Accelerator; rather, it is a general consideration when designing a GDPS based HA/DR solution. GDPS guidelines and recommendations explain reasonable distances (some kilometers) between primary and secondary storage sites.

6.4 Storage failures and HyperSwap

The Accelerator cannot provide HyperSwap capabilities. Therefore, a “swap” of storage from primary to secondary causes a temporary service outage.

The GDPS controlling system detects I/O problems and can then trigger a swap operation. The Accelerator is informed and restarts with new (secondary) disk device IDs assigned. Restart times for an accelerator vary dependent on system size and system state, but often takes several minutes.

This swap scenario to secondary disks for GDPS Metro mirrored Accelerator storage is shown in Figure 6-3.

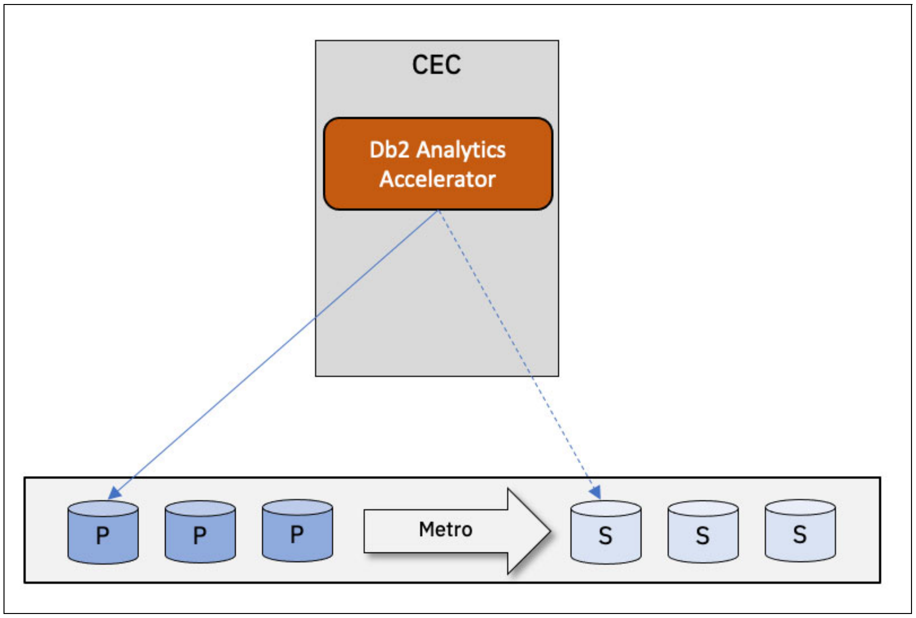


Figure 6-3 Accelerator on Z with GDPS Metro mirrored storage for swap purposes

6.5 Active-passive configuration with standby Accelerator and GDPS Metro

Unlike Db2 for z/OS, the Accelerator is not based on a “shared everything” data model. Therefore, two active Accelerators cannot access the same data on the same device concurrently. Each active Accelerator maintains its own data on the assigned disks.

One option to address local failures is the implementation of a standby Accelerator. This Accelerator runs in another LPAR (potentially on a different CEC) and is not active in normal operation mode.

If the active Accelerator fails, GDPS detects this issue and activates the standby Accelerator and pointing to the same disks that were used by the previously active Accelerator.

If the entire site fails, another storage swap is triggered so that the standby Accelerator uses the secondary disks.

This integration with GDPS provides a unique advantage for the IBM Z deployment option for the accelerator. Because all data (program, state, and user data) is stored on enterprise storage and is mirrored, the standby accelerator starts on an identical set of data as the previously active accelerator.

This environment is shown in Figure 6-4.

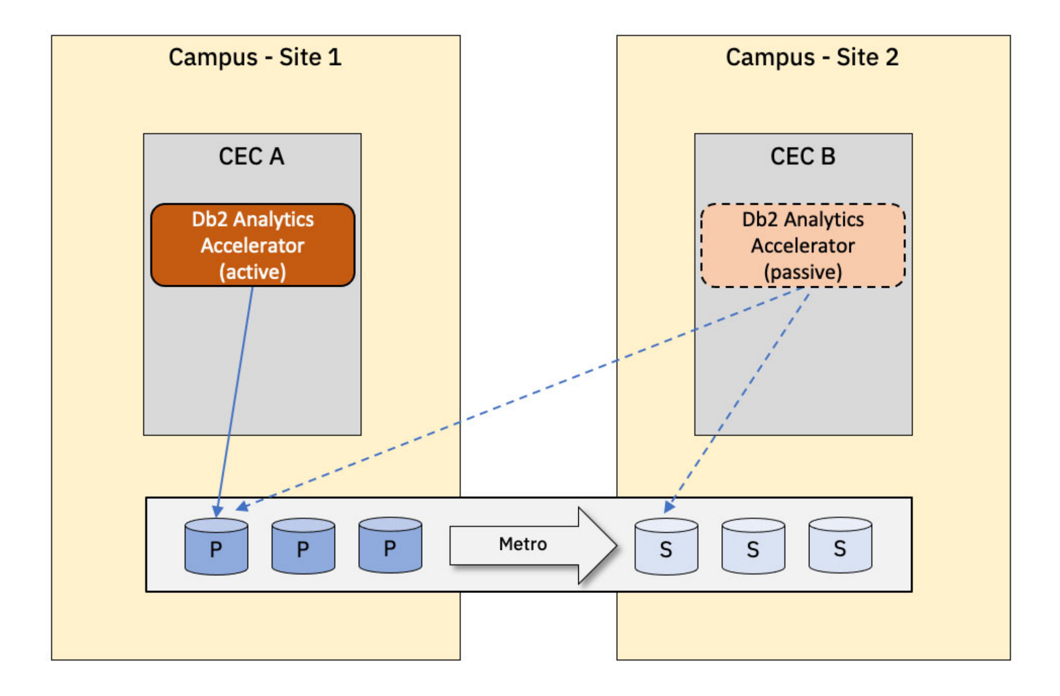

Campus - Site 1
Campus - Site 2
CEC A
CEC B
Db2 Analytics Accelerator (active)
Db2 Analytics Accelerator (passive)
P
P
P
Metro
S
S
S

Figure 6-4 Active-passive configuration with standby Accelerator on Z and GDPS Metro

The definition and setup for such a multi-site environment includes the following aspects:

- GDPS definitions for storage and Accelerator targets
- Additional network definitions for the Accelerator to access the GDPS controlling systems

For more information about GDPS-related aspects, see *GDPS Metro V4R3: Planning and Implementation*.

For the additional network definitions for the Accelerator consider the following. The IP addresses and ports to the GDPS controlling systems must be defined. A GDPS agent component in the accelerator uses this network information to exchange state and data with the controlling systems. In the JSON configuration file use the gdps_servers tag to define these IP addresses and ports and additionally use the gdps_nw tag to define the network devices for each participating Accelerator to be used to access the GDPS controlling systems from the Accelerators.

6.6 Active-active configuration with capacity growth

With the build-in workload balancing feature of the Accelerator, two active Accelerators can be deployed by using independent LPARs and an independent set of disks. This configuration is similar to the proposed active-active setup for Db2 Analytics Accelerator for IIAS and Db2 Analytics Accelerator Version 5. Both active Accelerators must be managed separately and data maintenance must be implemented for both individually.

All of the data management concepts that are explained in 5.4, “Db2 active-active configuration with GDPS Metro and active accelerators” on page 39 apply to an active-active setup with an Accelerator on IBM Z as well.

However, Z can dynamically assign resources (such as CPU or memory) to LPARs and Db2 Analytics Accelerator on IBM Z can dynamically respond to this change.

If an Accelerator fails on site 1, the remaining Accelerator on site 2 can be assigned more resources to cope with the higher workload.

Today, this resource assignment requires manual action that uses the HMC. In the future, this assignment also might be automated.

If the entire site fails, storage swap can triggered and secondary disks might be used.

The roles and implication of primary and secondary storage must be considered. If the second active Accelerator on site 2 also uses primary storage disks from site 1 (as shown in Figure 6-5 on page 59), an entire site failure triggers a storage swap and causes the Accelerator to be restarted with new disk device definitions. With this configuration, a service interruption occurs and the value of active-active is diminished.

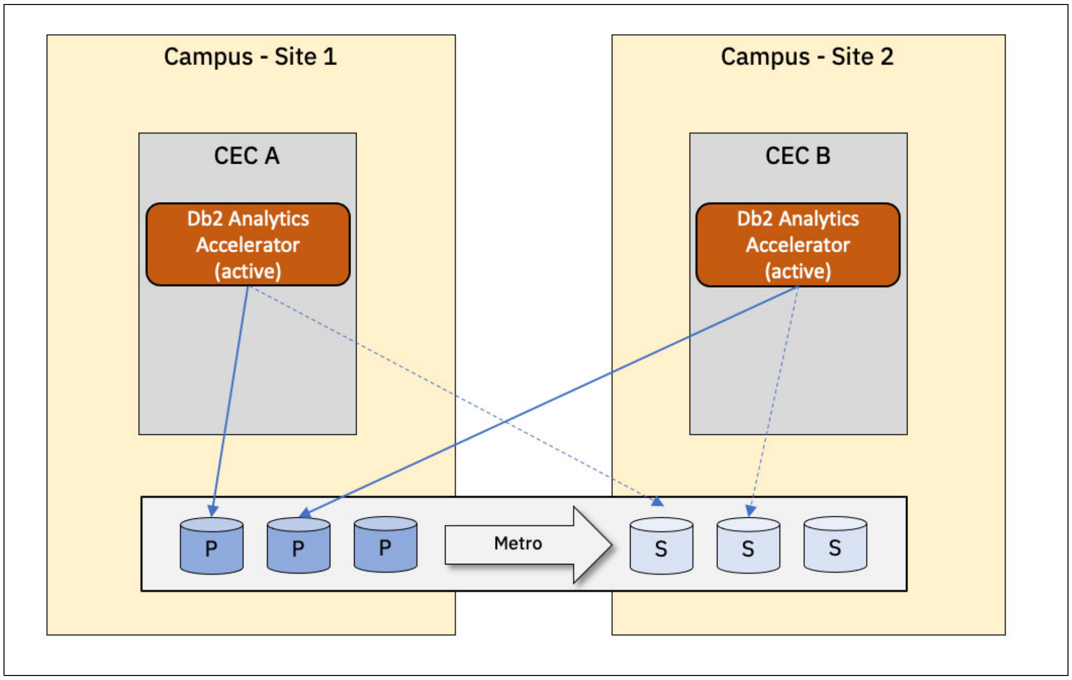

Campus - Site 1
Campus - Site 2
CEC A
CEC B
Db2 Analytics Accelerator (active)
Db2 Analytics Accelerator (active)
P
P
P
Metro
S
S
S

Figure 6-5 Active accelerators on Z with GDPS Metro mirrored storage

With the alternative layout that is shown in Figure 6-6, local primary disks are assigned (the Accelerators are operating on distinct disks and a concept, such as data sharing, is not supported for two active Accelerators).

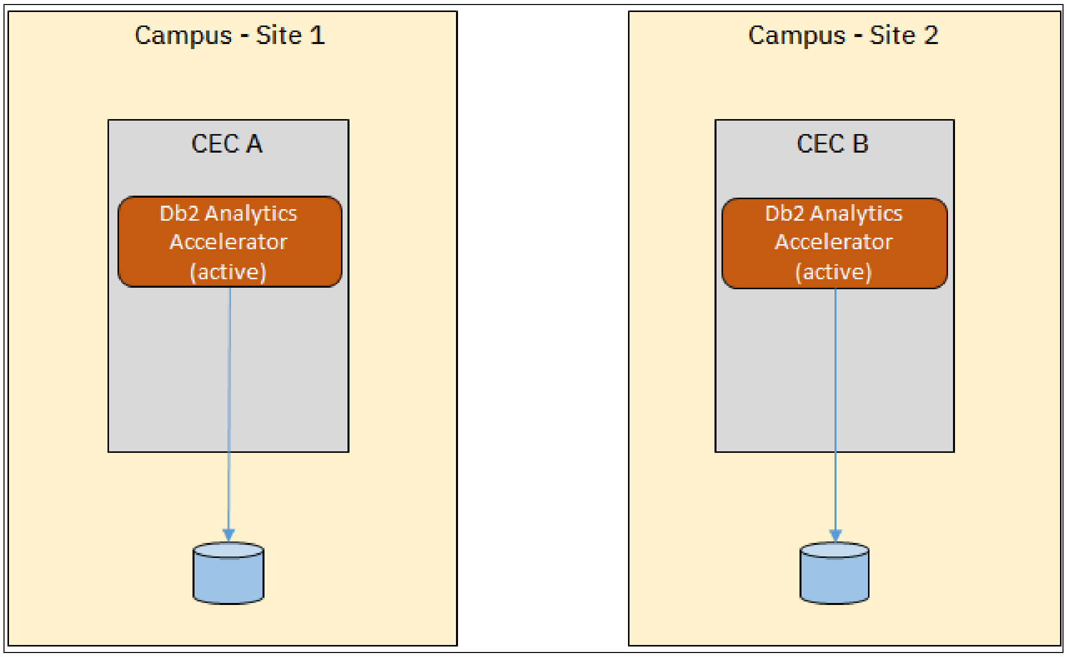

Campus - Site 1
CEC A
Db2 Analytics Accelerator (active)
Campus - Site 2
CEC B
Db2 Analytics Accelerator (active)

Figure 6-6 Active Accelerators on Z with local storage

In this case, Accelerator service continues, even with a complete failure of one site.

6.7 Multi-site configurations with active and passive accelerators and GDPS

In a multi-site configuration the accelerator active-passive or active-active configurations are extended by a third site resulting in a 3-site configuration or even by a fourth site resulting in a 4-site configuration. The sites 3 and 4 have additional passive accelerators defined Db2 Analytics Accelerator on IBM Z is now also supported by GDPS Metro and GDPS Global GM in three-site and four-site setups in cascaded or multi-target topologies.

Figure 6-7 on page 61 shows an example of a 3-site setup for Db2 Analytics Accelerator on Z managed by GDPS Metro and GDPS Global GM in a cascaded topology. Storage mirroring and failover is managed by GDPS Metro (between site 1 and site 2 in Region A) and GDPS Global GM (between site 2 in Region A and site 3 in Region B),

Within region A the failover and storage swap scenarios are the same as described in 6.5, “Active-passive configuration with standby Accelerator and GDPS Metro” on page 57. If the active Accelerator fails, GDPS detects this issue and activates the standby Accelerator and pointing to the same disks that were used by the previously active Accelerator. If the entire site fails, another storage swap is triggered so that the standby Accelerator uses the secondary disks.

In addition, the following scenarios between both regions are supported in a GDPS managed 3-site configuration with Db2 Analytics Accelerator:

- Region switch to CEC C in DR Site as planned action
- Region switch to CEC C in DR Site as an unplanned action
- Return home from CEC C in DR Site in Region B to CEC A in Region A as planned action

- DR testing in CEC C in DR Site using GM secondary or flash copy devices

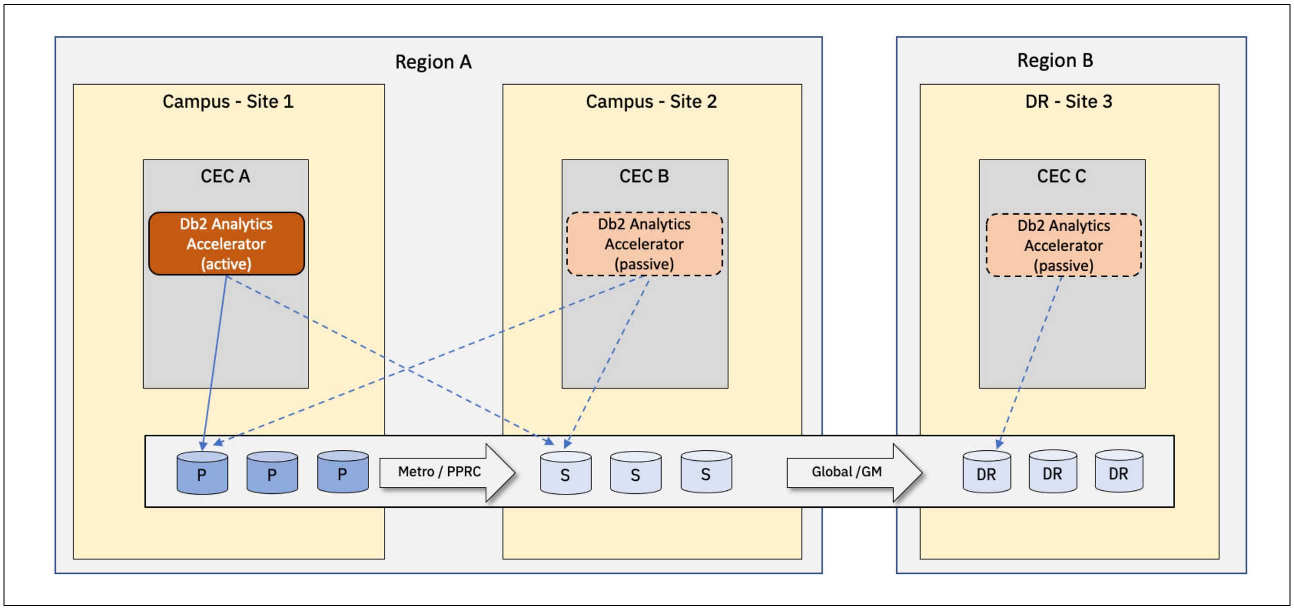


Figure 6-7 Accelerator 3-site configuration with GDPS Metro and Global GM

The 4-site configuration is an extension of the 3-Site configuration. The difference between the 3-site solution and the 4-site solution is that, with the 4-site solution, a second copy of data is available in the recovery region (Region B). This fourth copy of data is created using asynchronous Global Copy that can be switched to synchronous-mode (that is, Metro Mirror) during a planned or unplanned region switch, which provides the HA copy in that region.

An additional difference is that on site 4 an additional passive accelerator is set up. Figure 6-8 on page 62 depicts a 4-site configuration with Db2 Analytics Accelerator on Z in a cascaded topology.

This means that the 4-site configuration is a symmetrical configuration because from a data high-availability perspective, the same capabilities are available whether you are running your production services in Region A or Region B.

Within Region A or Region B the failover and storage swap scenarios are the same as described for the 3-site configuration.

In addition, the following scenarios between both regions are supported in a GDPS managed 4-site configuration with Db2 Analytics Accelerator:

- Region switch to CEC C and CEC D in DR Site as planned action
- Region switch to CEC C and CEC D in DR Site as an unplanned action
- Return home from CEC C (running the active accelerator) in DR Site in Region B to CEC A (running the active accelerator) in Region A as planned action
- DR testing in CEC C and CEC D in DR Site using GM secondary or flash copy devices

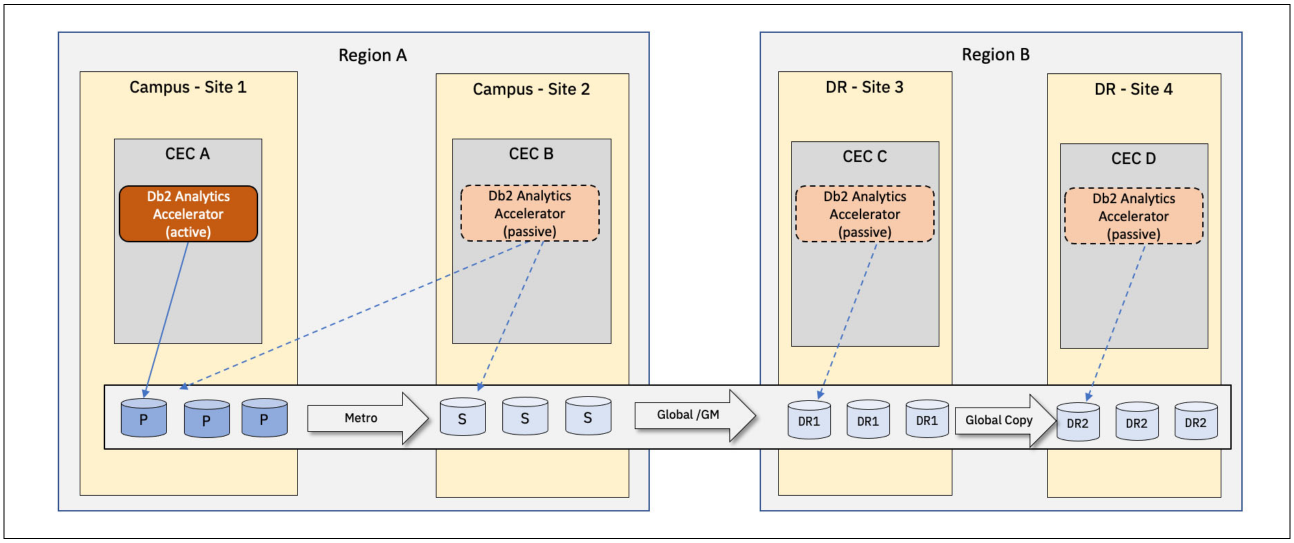

Region A
Campus - Site 1
Campus - Site 2
CEC A
Db2 Analytics Accelerator (active)
CEC B
Db2 Analytics Accelerator (passive)
Region B
DR - Site 3
DR - Site 4
CEC C
Db2 Analytics Accelerator (passive)
CEC D
Db2 Analytics Accelerator (passive)
P
P
P
Metro
S
S
S
Global /GM
DR1
DR1
DR1
Global Copy
DR2
DR2
DR2

Figure 6-8 Accelerator 4-site configuration with GDPS Metro and Global GM/Copy

7

Summary

IBM Db2 Analytics Accelerator can integrate well into HA and DR environments and extend the strength of IBM Z servers and Db2 for z/OS. The suggested options and their characteristics are listed in Table 7-1.

Table 7-1 Available options

GDPS option DB2 operation	Accelerator option	RPO DB2	RTO DB2	RPO Accelerator	RTO Accelerator
Metro active/active	IIAS: 2 active	0	0 (HyperSwap)	No data loss	0
Metro active/active	Z: 2 active	0	0 (HyperSwap)	No data loss	0 With disk swapping >0 (needs restart)
Metro active/standby	IIAS: 2 active	No data loss	>0	No data loss	0
Metro active/standby	Z: Metro/PPRC active/standby	No data loss	>0	No data loss	>0 (needs restart)
Global active/standby	2 active	A few seconds (potential data loss in case of disaster)	1 - 2 hours	No data loss (reload if out of synch)	0
Global active/standby	1 active, 1 standby	A few seconds (potential data loss in case of disaster)	1 - 2 hours	>0	>0 (needs reload)

© Copyright IBM Corp. 2019. All rights reserved.

REDP-5536-00

ISBN 073845768x

Printed in U.S.A.

Get connected

Redbooks ®

ibm.com/redbooks