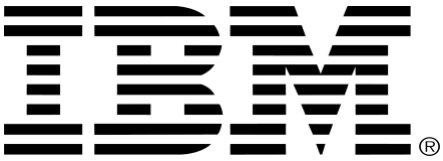


IBM Db2 Mirror for i: Performance Considerations

June 2022

IBM Corporation

Acknowledgements

We would like to thank the many people who made invaluable contributions to this document. Contributions included authoring, insights, ideas, reviews, critiques, and reference documents.

Our special thanks to key contributors from IBM i Development:

Susan Bestgen	IBM i Performance
Scott Forstie	IBM i Development
Craig Hogarth	IBM i Performance
Shawn Mohr	IBM i Development
Tim Rowe	IBM i Development
Dan Toft	IBM i Performance
Kris Whitney	IBM i Development
Ronald Young	IBM i Performance
Jia Tian Zhong	IBM i Performance

Our special thanks to key contributors from IBM Power Systems Performance:

Paul Nelsestuen	IBM i Performance
Rick Peterson	IBM i Performance

Our special thanks to key contributors from IBM Lab Services Power Systems Delivery Practice:

Eric Barsness	IBM i Performance

Table of Contents

Figures

1. Introduction

1.1. Purpose

The purpose of this document is to:

- Describe the Db2 Mirror implementation with enough detail that you can understand the performance considerations and case studies that follow.
- Help you prepare for and tune a Db2 Mirror deployment using hardware, configuration, and application performance considerations.
- Provide some performance expectations using case studies.

1.2. Document Responsibilities

The IBM i Development organization is responsible for editing and maintaining the IBM Db2 Mirror for i: Performance Considerations document in collaboration with IBM Power Systems Lab Services. Any contributions or suggestions for additions or edits should be forwarded to Eric Barsness, ericbar@us.ibm.com.

1.3. Updates in the latest version

Several new sections were added or changed in this version:

- Section 2.2: Sample data visualization of Collection Services data using iDoctor to understand potential Db2 Mirror overhead
- Section 3.4: RoCE adapters and RDMA protocols
- Section 7.4: Case Study #4: Native Database batch single row updates workload
- Section 7.5: Case Study #5: SR-IOV Exploration
- Section 7.6: Additional client scenarios documented
- Document links updated

2. What is Db2 Mirror for i?

2.1. Background

IBM Db2 Mirror for i is a major database enhancement introduced with IBM i 7.4. Db2 Mirror is an IBM i Licensed Program Product that enables near-continuous availability via an IBM i exclusive Db2 active-active two-system configuration.

Db2 Mirror, at its core, is a data-centric solution for continuous availability. Db2 Mirror includes synchronous replication of database files across a tightly coupled active-active configuration. Significant advances in networking technology are a key enabling element of Db2 Mirror. Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) provides far more than an acronym within an acronym. RoCE allows us to connect two IBM i partitions to establish something referred to as a node pair.

Db2 Mirror solves a very important requirement to ensure database objects and data within are identical (in sync) and available across the node pair. Data can be accessed and changed from either node. If there is a planned or unplanned outage, Db2 Mirror resynchronizes the data when the outage ends.

Db2 Mirror can be used to address business requirements such as:

- Avoiding downtime related to hardware or software upgrades
- Avoiding downtime related to maintenance
- Achieving an active-passive solution without having stale data in the mix
- Deploying a true active-active solution

2.2. Db2 Mirror Performance Overview

An important consideration of Db2 Mirror is the impact it can have on the performance of an application. While read-only workloads should not be affected, modify, or write (insert, update, and delete) activity on objects that are mirrored will include synchronous replication. This is due to the nature of Db2 Mirror and the requirement to keep objects completely in sync with each other on both nodes while Db2 Mirror is active. Modifications done on one node must also be done synchronously on the other node. The application cannot proceed until the work is complete on both nodes.

While the impact will vary with I/O intensity and the ratio of reads versus writes, the overhead of replicating database writes could be 2X or higher than non-replicated equivalent writes for a hypothetical 100% modify workload. The overhead is driven by performing the writes on both nodes and associated high-speed network traffic which coordinates the nodes. Most workloads are a mix of reads and writes and have other application activity and waits which means the overhead impact from Db2 Mirror will be much less than this worst-case scenario. In many cases application optimization and/or faster I/O subsystems can further reduce the impact of enabling Db2 Mirror.

Collection Services data can be used to determine the mix of read versus modify work done in a job, subsystem, or partition. For example, using IBM i 7.5 or older releases with the designated PTF applied (IBM i 7.4 SI75389, IBM i 7.3 SI75911, IBM i 7.2 SI75912) and IBM iDoctor for IBM i (client 1451 or newer) you can view graphs such as the following:

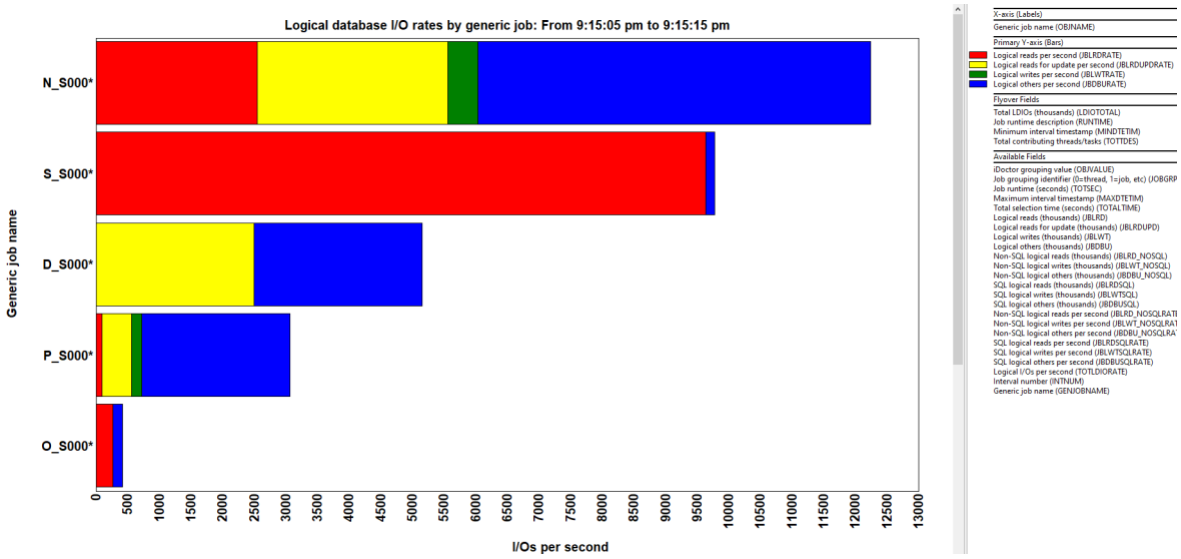


Figure 2-1. IBM iDoctor for IBM i visualization of Collection Services logical database I/O rates

The graph splits the activity for a group of jobs (jobs with the name first 6 characters) into read only (red), read for update (yellow), writes/inserts (green) and other (deletes, updates, etc. – blue). Jobs with mostly read-only activity will see less impact when running Db2 Mirror than jobs that have a higher ratio of modify activity. Reads-for-update also impact performance due to the need to acquire locks on both partitions.

2.3. Architecture

The Db2 Mirror architecture consists of two nodes that are paired together to create a synchronous environment. The nodes are independent, and both can access and update the data that is synchronously replicated in both directions. Db2 Mirror supports replication of data in SYSBAS and in Database independent auxiliary storage pools (IASPs). Applications can use either SQL or traditional record level access (RLA) to work with replicated data.

For example, Figure 2-2 shows separate instances of the same application running on each node using a synchronously replicated database file. The database file can exist either in SYSBAS or within an IASP. When Row A is changed on Node 1, it is synchronously written to the file on both Node 1 and Node 2. When Row B is changed on Node 2, it is synchronously written to the file on both Node 2 and Node 1.

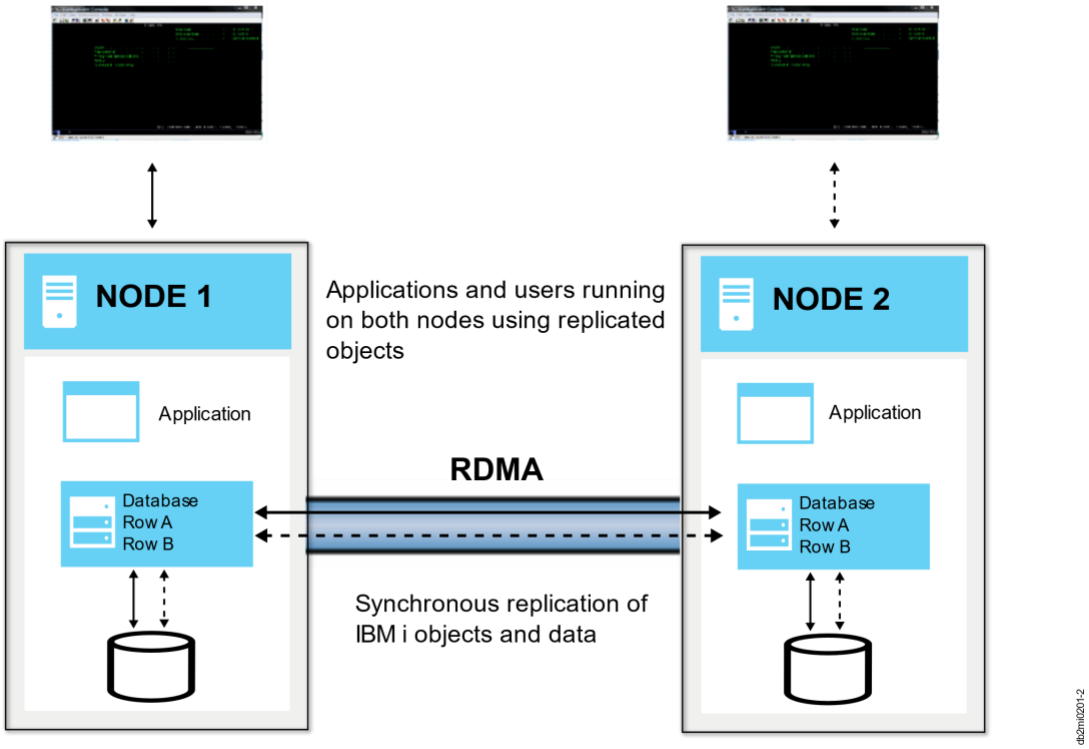


Figure 2-2. Db2 Mirror Architecture

2.4. RDMA over Converged Ethernet (RoCE)

The two nodes in a mirrored pair require a high-speed network connection to perform synchronous data operations. The high-speed connection interface being used is Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE). Adapters that support RoCE must be configured for both nodes.

The ports on the RoCE adapters can be cabled directly together or connected through a switch. IP addresses are configured on the line description associated with each port of the RoCE adapters. A pair of IP addresses, one from each node, is used to identify each physical RDMA link between the two nodes.

2.5. Mirrored Database Performance Overhead

The synchronous design for handling database objects ensures that data is always correctly replicated on both nodes. Overhead is incurred by RDMA communication flows to acquire and later release necessary row locks while accomplishing replication actions on each node.

Individual data modifications (insert, update, delete) effectively occur serially. A change on the source node may execute quickly, but the equivalent change must also be enacted via RDMA on the target node. The remote action is followed by feedback to the source node (again via RDMA) to indicate whether the remote change was successful or failed.

3. Db2 Mirror for i Hardware and I/O Considerations

3.1. Introduction

This section describes system level and I/O configuration considerations that can affect the performance of Db2 Mirror.

3.2. Partition Configuration

Dedicated processors are recommended to get the best performance for a partition. For more information on the technical reasons refer to the IBM i on Power Performance FAQ.

3.3. I/O Subsystem

The performance of I/O operations is critical for Db2 Mirror since the source side must wait for the target side to complete any replicated database operation including any required synchronous I/Os. Therefore, having a fast I/O subsystem is critical. Technology such as solid-state drives (SSDs) and flash storage as well as large write caches will help improve the performance of Db2 Mirror.

For example, writes to slower performing disks can degrade application response times and lead to increased Db2 Mirror CPU overhead. A sample case study which compares Db2 Mirror performance metrics when using solid state drives versus hard disk drives for an online transactional processing (OLTP) workload scenario is discussed in section 7.1 later in this document.

Similar concerns exist during both planned and unplanned outages. Slower performing I/O will extend the period that a node pair is out of sync and lengthen the amount of time required to complete resynchronization processing following an outage.

3.4. RoCE Adapters and RDMA Protocols

Additional RoCE adapters have been developed since prior publications of this document. Adapters available as of December 2021 can support up to 200Gb/s in both directions when using both ports (e.g., 400Gb/s full duplex). Specific supported adapters and their various features are noted in the network adapter section of hardware requirements in the Db2 Mirror book. The adapters provide the capability to run with several RDMA protocols:

- RoCE version 1 is a non-encrypted and non-routable RDMA protocol
- RoCE version 2 is a non-encrypted and routable RDMA protocol. The ports of RoCE v2 adapters can be connected through multiple switches or routers up to a maximum distance of 10 kilometers.
- Encrypted RoCE version 2 is an encrypted and routable RDMA protocol. It uses IPsec protocols to provide data authentication, integrity, and confidentiality. The ports of RoCE v2 adapters can be connected through multiple switches or routers up to a

maximum distance of 10 kilometers. Adapters that support encryption are available on Power9 or newer hardware.

Placing systems at greater distances apart adds new performance considerations. Side by side systems connected with 1- or 2-meter copper or fiber experience very small line transmission latencies (a few nanoseconds). Systems that are placed at the maximum distance of 10 kilometers apart experience greater latencies. The contributing factor is how long it takes for data to travel at the speed of light through fiber over the distance of up to 10 km. As a rough estimate it takes 5 microseconds for light to travel 1 kilometer one way through a fiber. For 10 kilometers the transmission latency would be 100 microseconds for a single round trip (send+acknowledge). Replicating a single data row change can involve 2 or more round trips between databases. A Db2 Mirror transaction that takes less than 1 microsecond total line latency with a 2-meter cable can experience over 200 microseconds additional latency with 10-kilometer lines. Not all database operations result in transmission between systems. Significant time and processing in addition to RDMA transmission is likely in applications using DB, but the additional replication latency due to long distance links between systems can become noticeable.

4. IBM i Configuration Considerations

4.1. Introduction

Each node in an IBM i Db2 Mirror configuration will be subject to the general performance considerations outlined in the IBM i on Power Performance FAQ. This document will, in some cases, reference pertinent sections of that Performance FAQ while also outlining specific aspects of Db2 Mirror which may impact performance. Ensuring each node is following the best practices outlined in the Performance FAQ will be important to the performance of your Db2 Mirror implementations.

The rest of this section describes IBM i operating system specific areas of configuration related to performance.

4.2. Prestart Job Configuration

A QDBMSRVR job is created on the target node and is associated with the job on the source node when the first database replication-eligible action is taken within the job. By default, the QDBMSRVR jobs run in the QUSRWRK subsystem. For nodes leveraging Db2 Mirror replication, the recommended prestart job settings for QDBMSRVR jobs are:

*INLJOBS(50) THRESHOLD(20) ADLJOBS(40) MAXUSE(*NOMAX)*

For example:

*CHGPJE SBSD(QUSRWRK) PGM(QDBMSRVR) STRJOBS(*YES) INLJOBS(50) THRESHOLD(20) ADLJOBS(40) MAXJOBS(*NOMAX) MAXUSE(*NOMAX)*

This will keep a pool of QDBMSRVR jobs available when source node jobs make the initial connection to the target node (during the first database replication-eligible action) and will optimally replenish the pool for workloads that have a spike in Db2 Mirror connections.

Routing QDBMSRVR jobs to a different subsystem enables the replication work of those jobs to occur a different memory pool to isolate that activity from other work on the target node.

4.3. Journal Configuration

IBM i journals can have their *attributes* replicated between source and target nodes by adding them to the Replication Criteria List (RCL), but updates to journal receivers and their entries are never replicated. Each node manages its journal receivers and deposits independently. Thus, when mirroring is active and objects are journaled, each transaction will be journaled by both nodes. Ensuring optimal journal performance on both nodes for journaled objects is critical to overall performance of your Db2 Mirror implementation.

Many helpful techniques can be found in the Striving for Optimal Journal Performance Redbook, and also in the Journal management and system performance section of IBM i documentation. Particularly, consider omitting the *Open/Close* and *Before* journal entries, as they are not generally useful.

5. Db2 Mirror for i Configuration Considerations

5.1. Introduction

This section describes topics and concepts related to configuring Db2 Mirror which can have an impact on the performance experience for your implementation.

5.2. Considering What to Mirror

Mirroring changes to replicated objects consumes system resources and will add overhead to the applications driving those changes. It is important to carefully consider and control what objects will be replicated for your Db2 Mirror node pair. The possible performance impact that mirroring can have increases accordingly with the number of objects replicated and their aggregate storage consumption.

The definition of the list of objects to be replicated is determined by three criteria:

1. The default inclusion state, which is defined only at initial set up.
2. System rules.
3. Replication Criteria List (RCL), which can be user defined for eligible objects.

More information on configuring the replication rules for your installation can be found in the Replication Criteria List configuration wizard online documentation.

5.3. IFS Implementation and Performance Characteristics

Integrated File System (IFS) objects are made accessible on both Db2 Mirror nodes by using a different technology than the replication eligible IBM i object types. To be accessible from the secondary node, the IFS objects must be contained within an Independent Auxiliary Storage Pool (IASP), and the IASP must be part of a cluster resource group (CRG).

Simultaneous IFS access from both nodes is implemented by using an IFS client/server technology that uses a mutable file system architecture.

For IFS IASPs, the performance might differ depending upon from which node a file system operation is initiated. Users on the server node where the IASP is varied on might experience faster response times than users on the client node.

For more details on the Db2 Mirror implementation for IFS IASPs see the IBM i Db2 Mirror Getting Started Redbook.

5.4. Tracking and Resync Considerations

5.4.1. Performance Impact of Tracking

When mirroring is suspended, the primary node tracks change operations to mirrored objects. Tracking overhead is very minimal. Newly generated or modified mirrored objects (now out of sync) are added to the Object Tracking List (OTL) and all row level modifications of tracked database files are noted as they occur.

5.4.2. Resynchronization Considerations

As part of resuming replication, the secondary node must be brought up to date with any changes that occurred on the primary node. Resynchronization is the process by which tracked changes are applied to the secondary node.

The processing involved to accomplish resynchronization depends upon the amount of tracked information that must be propagated to the secondary node to bring all mirrored objects back into sync. Keep in mind that this resynchronization work is additional processing overhead while the primary node handles on-going business needs. Strategies to reduce the elapsed time required to complete resynchronization include:

- Isolate replication activity to a different link than the one(s) used for normal database traffic when configuring Network Redundancy Groups. This prevents replication traffic from negatively affecting network latency on the link used for normal database traffic and avoids the possibility that the combined traffic could saturate a single link.
- Specify a parallel degree to improve the performance of resynchronization if Db2 Symmetric Multiprocessing (SMP) is installed. However, since the primary node is still running applications, you should consider the performance impact to those applications. Parallel degree for resynchronization is set using the QSYS2.CHANGE_MIRROR procedure.
- Dynamically allocate spare CPU and memory resources to the primary node during resynchronization.

Additional resynchronization best practices recommended to improve performance and reduce the complexity and problems during resynchronization are documented in more detail in the IBM i Db2 Mirror book.

6. Db2 Mirror for i Application Considerations

6.1. Introduction

This section highlights topics for consideration within an application environment which can have an impact on performance when implementing Db2 Mirror for IBM i.

6.2. Reduce Full Opens

High rates of database file full opens can lead to poor performance and unnecessary resource consumption on any IBM i partition — even without Db2 Mirror. Applications with excessive rates of full opens where the file is open for update will suffer additional overhead and reduced throughput with Db2 Mirror. Reducing overall database file full open rates is a long-standing best practice for IBM i.

Section 8.16 of the IBM i on Power Performance FAQ has general performance guidance on this topic and how to analyze your application for full opens.

6.3. Open Files for Update Only When Necessary

Opening files for update will cause file opens and any subsequent record locks to be mirrored to the target node. Read operations can remain local and avoid mirror overhead within a program. Therefore, if processing is only reading rows of the file, but the open indicates updateability, there will be additional open as well as row processing work for mirroring that could be avoided.

6.4. Commitment Control and Locking

Changes to database objects grouped under commitment control will be mirrored with the same isolation level on the target node. Appropriate locks will be acquired on both nodes during the transaction. This includes read for update locks. Using the lowest level of commitment control required for application transactional integrity can help ensure optimal performance in general and for Db2 Mirror.

6.5. Batch Considerations

Long running serial write-intensive batch applications will experience longer run times when Db2 Mirror is active. Splitting the processing into multiple jobs or threads that each process a subset of the data in parallel can greatly reduce the run time of time sensitive sections of batch processing.

If there are situations during batch processing where new or temporary files are created and the contents of the files are not needed on both nodes while the batch process is running, consider excluding these files from replication while the batch process is running. Add the appropriate files for replication after batch processing has completed and they will then be available on both nodes while avoiding mirror overhead within the batch run.

7. Db2 Mirror for i Case Studies

7.1. Case Study #1: Native Database OLTP

7.1.1. Workload Description

This scenario is an OLTP COBOL workload with embedded native database functions running under commitment control with the workload running on one side of the node pair (active-passive). Five concurrent transactions perform database functions configured to approximately 60% read and 40% write (insert/update/delete). Multiple job sets service the configurable number of virtual users to achieve a desired activity threshold.

7.1.2. Configuration

Two pairs of nodes were tested:

Pair 1: The source node and target node were both configured with 2 cores and memory assigned at 0.5MB per user. A Storwize V7000 configured with solid state drives (SSD) was used as storage for both nodes.

Pair 2: The source node and target node were both configured with 2 cores and memory assigned at 0.5MB per user. A Storwize V7000 configured with hard disk drives (HDD) was used as storage for both nodes.

7.1.3. Baseline

Pair 1 (SSD): Simulation of 32,000 users at a total system CPU utilization of 45.6.

Pair 2 (HDD): Simulation of 32,000 users at a total system CPU utilization of 56.1.

7.1.4. First Scenario

Pair 1 (SSD): Simulation of these same 32,000 users after configuring Db2 Mirror resulted in a CPU utilization of 81.4 on the source node and 56.5 on the target node. The users were able to complete comparable throughput, but response time went up.

Pair 2 (HDD): Simulation of these same 32,000 users after configuring Db2 Mirror resulted in a CPU utilization of 85.2 on the source node and 64.3 on the target node. The users were able to complete comparable throughput, but response time went up.

7.1.5. Second Scenario

We kept all other parameters constant but increased the CPU to 4 cores on all nodes and repeated the Db2 Mirror tests, with the following results (NET: doubling the cores brought CPU utilizations back in line with baselines):

Pair 1 (SSD): Simulation of these same 32,000 users with replication and additional cores resulted in a CPU utilization of 46.7 on the source node and 29.9 on the target node. The users were able to complete comparable throughput with response times generally like the 2 core Db2 Mirror results from the first scenario.

Pair 2 (HDD): Simulation of these same 32,000 users with replication and additional cores resulted in a CPU utilization of 52.9 on the source node and 32.9 on the target node. The users were able to complete comparable throughput with response times generally like the 2 core Db2 Mirror results from the first scenario.

7.1.6. Third Scenario

The final test kept all nodes configured at 2 cores but simulated 48,000 users to examine Db2 Mirror characteristics at higher CPU and/or disk utilizations (NET: I/O speed can become a factor at high utilizations):

Pair 1 (SSD):

- Baseline: Simulation of 48,000 users at a total system CPU utilization of 64.8.
- Db2 Mirror: Simulation of 48,000 users at system CPU utilization of 100.0 on the source node and 77.2 on the target node. The users were able to complete comparable throughput, but response times were significantly higher than the first scenario of 32,000 users.

Pair 2 (HDD):

- Baseline: Simulation of 48,000 users at a total system CPU utilization of 79.1.
- Db2 Mirror: Simulation of 48,000 users at system CPU utilization of 72.1 on the source node and 54.7 on the target node. Both throughput and response times were significantly impacted as disk could not keep up with transaction rates.

The performance summary for all test scenarios is found in Table 7-1. Charts of comparable results include Figure 7-1 for 32,000 users across scenarios, Figure 7-2 for 48,000 users across scenarios, and Figure 7-3 for average response times across all user loads and associated scenarios.

Table 7-1. Case Study #1 Performance Metrics

32,000 users	Source CPU%	Target CPU%	Throughput	Average Response Time (Sec)
Pair 1 Baseline	45.6		38990	0.004
Pair 2 Baseline	56.1		38974	0.012
Pair 1 Scenario #1	81.4	56.5	38994	0.011
Pair 2 Scenario #1	85.2	64.3	38686	0.452
Pair 1 Scenario #2	46.7	29.9	39000	0.007
Pair 2 Scenario #2	52.9	32.9	39074	0.455

48,000 users	Source CPU%	Target CPU%	Throughput	Average Response Time (Sec)
Pair 1 Baseline	64.8		58485	0.006
Pair 2 Baseline	79.1		58454	0.021
Pair 1 Scenario #3	100	77.2	58485	0.525
Pair 2 Scenario #3	72.1	54.7	30433	2.476

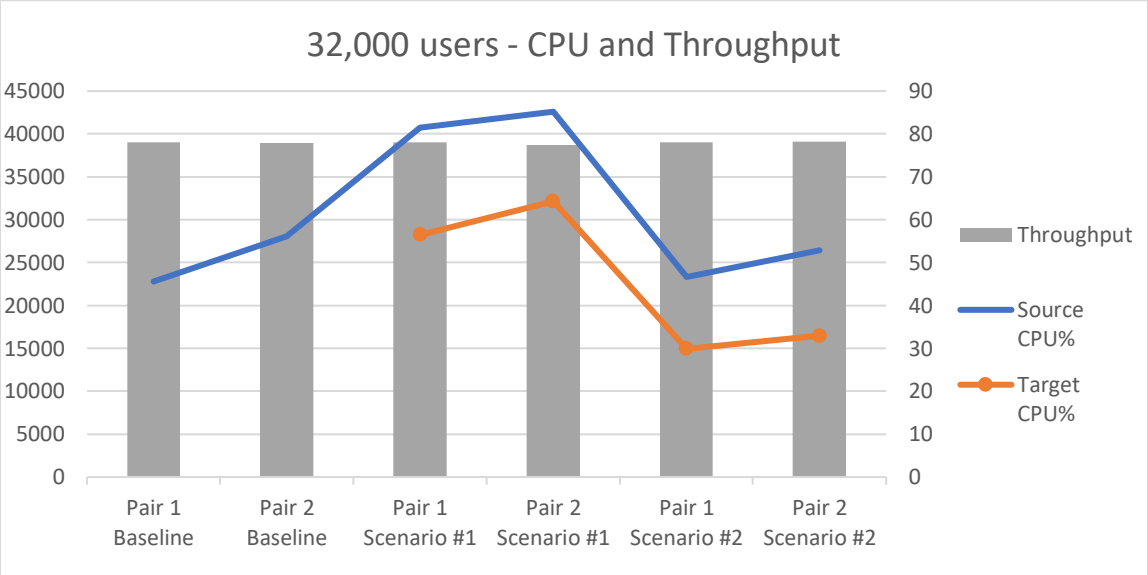


Figure 7-1. Case study #1: CPU and Throughput for 32,000 users across scenarios

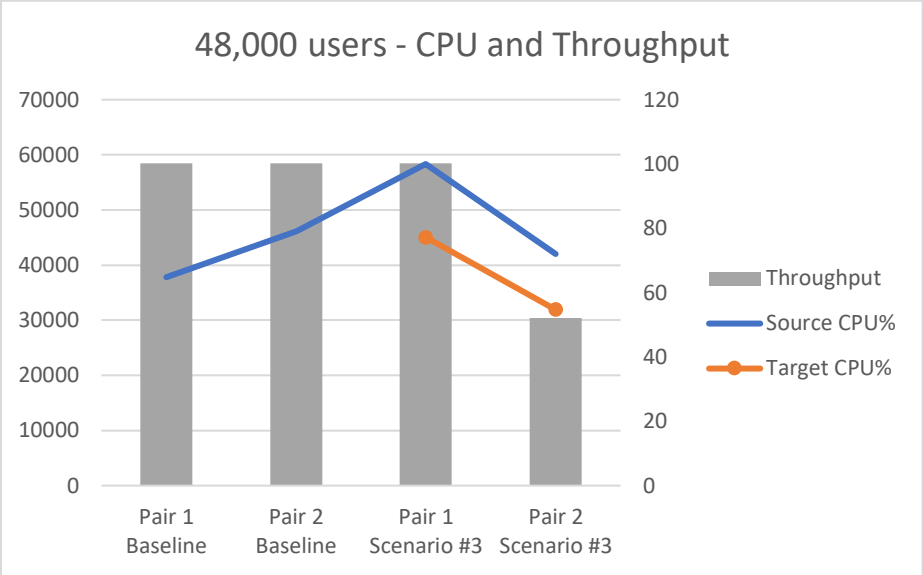
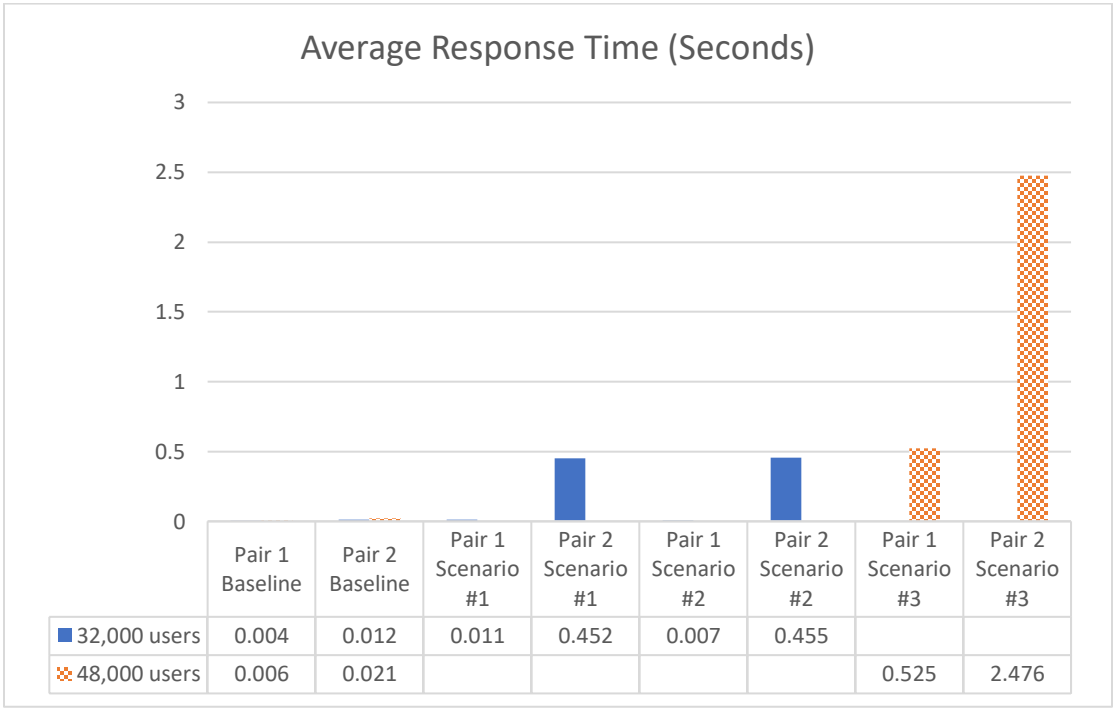


Figure 7-2. Case study #1: CPU and Throughput for 48,000 users across scenarios



	Pair 1 Baseline	Pair 2 Baseline	Pair 1 Scenario #1	Pair 2 Scenario #1	Pair 1 Scenario #2	Pair 2 Scenario #2	Pair 1 Scenario #3	Pair 2 Scenario #3
32,000 users	0.004	0.012	0.011	0.452	0.007	0.455		
48,000 users	0.006	0.021					0.525	2.476

Figure 7-3. Case study #1: Average response times across scenarios

7.1.7 Summary

This case study shows that for our OLTP COBOL workload the overhead incurred by Db2 Mirror replication in terms of application response time and CPU usage increase could be compensated by a faster disk I/O subsystem and by doubling CPU resources. With CPU resources held constant on the SSD-Flash I/O pair, the workload was able to fully utilize the available CPU on the source node but response time was significantly impacted. Holding the CPU resources constant on the HDD pair resulted in substantial impact to throughput and response times.

7.2. Case Study #2: SQE OLTP

7.2.1. Workload Description

This is an SQL based OLTP workload consisting of 23 different transactions implemented via SQL stored procedures which execute multiple SQL statements per transaction under commitment control. The workload runs actively on one side of the node pair (active-passive). Ten of the transactions result in database changes (Insert/Update/Deletes) as well as reads, while the other 13 are read-only transactions.

The workload is configurable by the number of jobs sets executing and a specific ratio of transactions to achieve a consistent transaction rate and mix for each test.

7.2.2. Configuration

The source node had 22 cores with 964 GB of memory in the application pool.

The target node had 11 cores also with 964 GB of memory, but in the *BASE pool where the QDBMSRVR replication jobs run by default.

7.2.3. Baseline

Eighty-two job sets were executed collectively at a rate of ~29,900 transactions/second (TPS).

7.2.4. First Scenario

The same eighty-two job sets were executed again after configuring Db2 Mirror which resulted in a decrease in throughput to ~23,770 TPS. The CPU utilization was approximately the same, but response time went up.

7.2.5. Second Scenario

The job set quantity was increased to 104 job sets to achieve approximately the same transaction rate as the baseline test, while keeping the number of cores on the source node the same at 22 cores. This test saw a transaction rate of ~29,870 TPS, which was like the baseline. CPU increased from the baseline as well as response time.

7.2.6. Third Scenario

For this test we increased the number of cores in the source node and adjusted the number of job sets so that the transaction rate and CPU utilization were like the baseline. By using 92 jobs sets and 25 cores on the source node, we saw a transaction rate of 29,820 TPS with a response time closer to the baseline.

The results summary for all test scenarios is found in Table 7-2. Charts of comparable results are in Figure 7-4.

Table 7-2. Case study #2 Performance metrics

	Source CPU%	Target CPU%	TPS	Response Time (msec)	Cores Src/Tgt	Job Sets
Baseline	74.4	N/A	29,900	3.24	22 / NA	82
Scenario #1	74.3	20.9	23,770	4.24	22 / 11	82
Scenario #2	83.9	26.6	29,870	4.34	22 / 11	104
Scenario #2	75.5	26.0	29,820	3.88	25 / 11	92

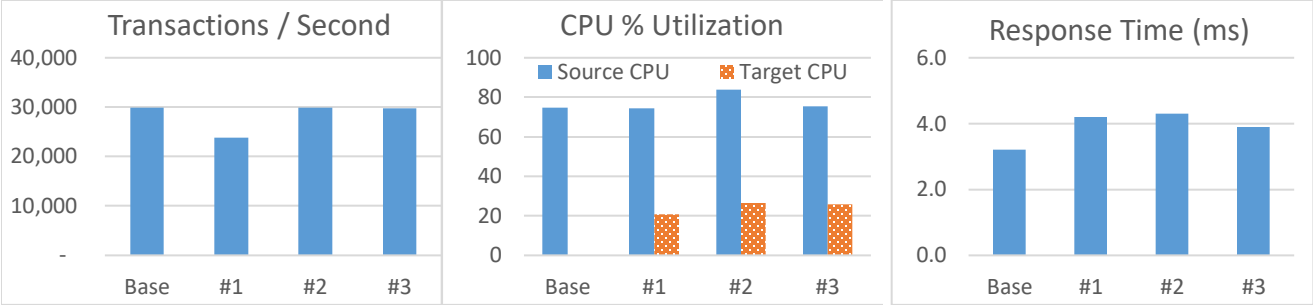


Figure 7-4. Case study #2: Summary charts across scenarios

7.2.7. Summary

What we found with this case study is ~14% additional cores were required on the source node to maintain about the same overall system CPU utilization and throughput. This workload also required additional jobs to maintain the same throughput due to increased response time. Workloads that have a set of client jobs processing work that are always busy and not typically waiting for work will likely require additional jobs to maintain the same throughput. For a workload that uses connection pooling, this could mean an increased value for maximum jobs in the pool.

7.3. Case Study #3: SQE OLTP: active-passive vs active-active

7.3.1. Workload Description

This is the same SQL based OLTP workload as Case study #2. In this study, we ran the workload without Db2 Mirror, with Db2 Mirror in both active-passive mode and active-active mode.

7.3.2. Configurations

When running without Db2 Mirror active, the source node had 32 cores with 487 GB of memory in the application pool.

When running in active-passive mode, the target node varied from 6-8 cores also with 487 GB of memory but in the *BASE pool where the replication jobs run by default.

When running active-active mode, the source and target nodes varied from 21-22 cores memory ranged from 487-584 GB.

We executed a varying number of job sets for each test attempting to achieve ~44,500 transactions/second for each test and keeping CPU utilization around 79%.

7.3.3. Baseline

This test achieved 44,500 TPS (Transactions per Second) with CPU running at 78.9% of the 32 cores with an application average response time of 3.6 milliseconds.

7.3.4. Active-Passive Scenarios

The first version ran with 38 cores on the source node and 8 cores on the target node achieving 44,600 TPS at an average response time of 4.7 milliseconds. CPU utilization on the source node was 78.5% of the 38 cores and 56.32% of the 8 cores on the target node.

The second version ran with 38 cores on the source node and 6 cores on the target node achieving 44,700 TPS at an average response time of 5.0 milliseconds. CPU utilization on the source node was 79.38% of the 38 cores and 78.74% of the 6 cores on the target node.

7.3.5. Active-Active Scenarios

The first version ran with 22 cores on each node and used 487 GB of memory on each node, but split this memory 80/20%, with 390 GB in the application pool and 97 GB in *BASE where the replication jobs run. This test saw 43,500 TPS at 7.2 and 6.5 millisecond response times on the primary and secondary nodes respectively. CPU utilization was 80.16% on the primary node and 81.31% on the secondary node. To achieve the same TPS as other tests would have required an additional core on both nodes.

The second version also ran with 22 cores on each node and used 487 GB of memory on each node again but had the application work and replication jobs share the same memory pool. This test achieved 44,600 TPS at a 5.2 millisecond response time on each node. CPU utilization was 79.62% on the primary node and 79.04% on the secondary node.

The third version again ran with 22 cores on each node but used 584 GB of memory (120% of baseline), with 487 GB in the application memory pool and 97 GB in *BASE for the replication jobs on each node. This test achieved 44,300 TPS and 5.6 / 5.1 millisecond response times with CPU utilization of 77.21% on the primary node and 77.30% on the secondary node.

The fourth active-active version also ran with 22 cores on each node and used 584 GB of memory, but with the memory in one pool shared by the application and replication jobs. This test achieved 44,400 TPS at an average response time of 4.3 / 4.2 milliseconds. CPU utilization on the primary node was 76.08% and 75.80% on the secondary node.

The last active-active test ran with 21 cores on each node and 584 GB of memory again with the application jobs and replication jobs sharing a memory pool. This test achieved 44,600 TPS at an average response time of 4.5 milliseconds on each node. CPU utilization was 79.32% on the primary node and 78.99% on the secondary node.

The results summary for all test scenarios is found in Table 7-3. Charts of comparable results are in Figure 7-5.

Table 7-3. Case study #3 Performance metrics

	Source CPU%	Target CPU%	TPS	Response Time (msec)	Cores Src/Tgt	Job Sets
Baseline	78.90	N/A	44,500	3.7	32 / NA	140
Act/Pas #1	78.50	56.32	44,600	4.7	38 / 8	164
Act/Pas #2	79.38	78.74	44,700	5.0	38 / 6	172

	Primary CPU%	Secondary CPU%	TPS	Response Time (msec)	Cores Prim/Sec	Job Sets
Act/Act #1	80.16	81.31	43,500	7.2 / 6.5	22 / 22	108 / 108
Act/Act #2	79.62	79.04	44,600	5.2 / 5.2	22 / 22	90 / 90
Act/Act #3	77.21	77.30	44,300	5.6 / 5.1	22 / 22	86 / 86
Act/Act #4	76.08	75.80	44,400	4.3 / 4.2	22 / 22	72 / 72
Act/Act #5	79.32	78.99	44,600	4.5 / 4.5	21 / 21	76 / 76

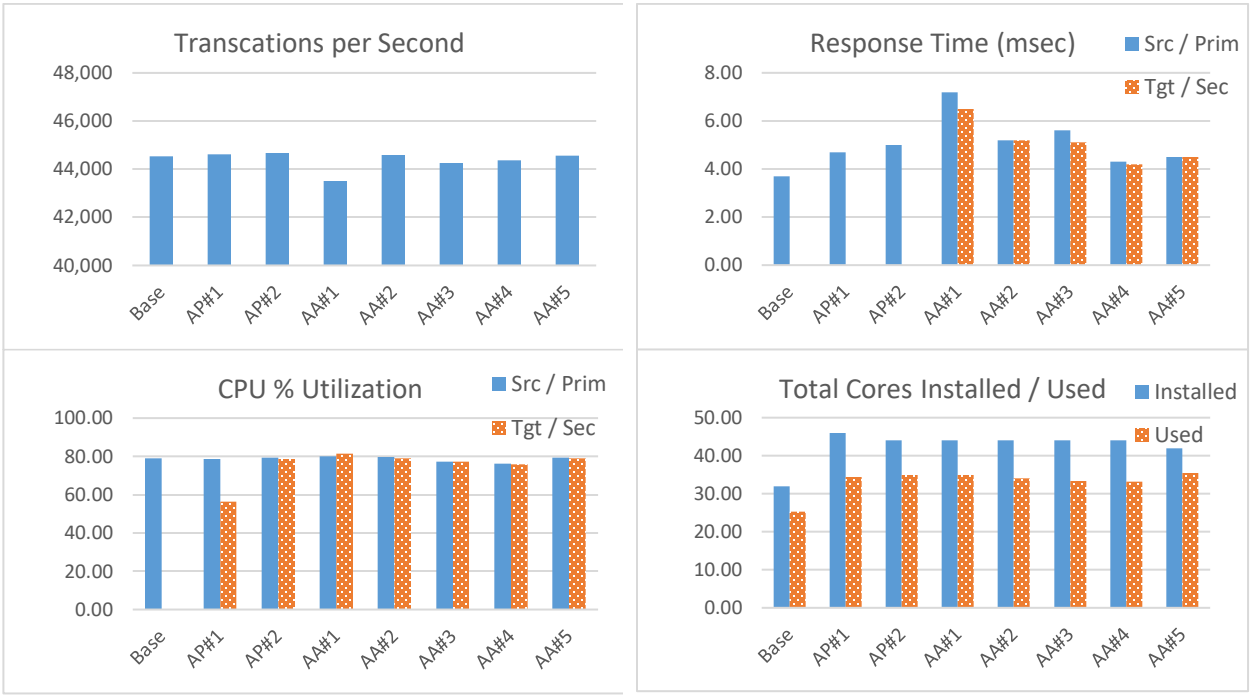


Figure 7-5. Case study #3: Summary charts across scenarios

7.3.6. Summary

The findings for the active-passive scenarios compared to the baseline was very similar to the previous case study. For the active-active scenarios, having the QDBMSRVR replication jobs share the same memory pool as the application resulted in the best performance. Having additional memory in the pool that is again shared provided even slightly better performance.

7.4. Case Study #4: Native Database Batch Single Row Updates

7.4.1. Workload Description

This scenario is a simple batch update workload designed to maximize native database updates in a 2-minute period. A specified number of jobs each run against their own individual database file consisting of approximately 1 million 161-byte records. Each job iterates sequentially over its database file reading each record for update and then updating the record. If a job reaches the end of a database file it loops back to the beginning of the file and continues the process. The database files are kept memory resident to minimize paging activity as much as possible. All scenarios are run active-passive for this workload.

7.4.2. Configuration

The source and target node were configured with 10 cores and 480 GB of memory each. The systems were configured with a pair of 100GbE RoCE with Crypto Adapter x16 adapters (Feature Code EC77). The adapters were directly connected with a single 2m RoCE copper cable as the active link.

7.4.3. Scenario

We ran multiple measurements with job counts ranging from 10 up to 100 in 10 job increments. The Db2Mirror NRG was configured with a single link (load balance = 1) for database replication. Each set of runs used one of the three supported RDMA protocols: RDMA v1, RDMA v2 or RDMA v2 with Encryption (v2SEC).

7.4.4. Summary of Results

As seen in Figure 7-6, measurements show there are minimal performance differences between the three RDMA protocols: v1, v2, and v2 with encryption.

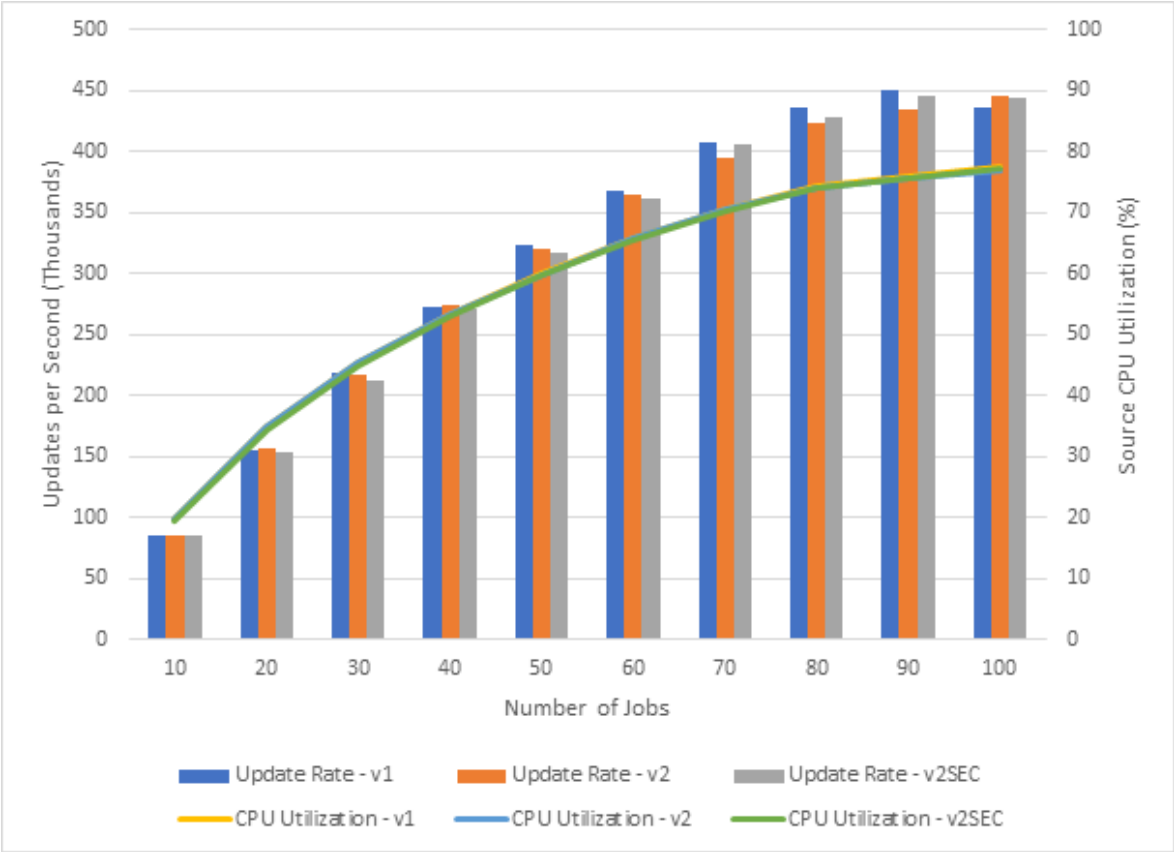


Figure 7-6. Case study #4 RDMA protocol performance comparison

7.5. Case Study #5: SR-IOV Exploration

7.5.1. Workload Description

This case study explores the performance behavior of Db2 Mirror RDMA network traffic using hardware virtualized RoCE adapters. Single Root Input/Output Virtualization (SR-IOV) lets the Power hypervisor manage the RoCE adapters directly. SR-IOV allows logical ports to be configured and associated with each physical port which enables multiple partitions to logically share an adapter's physical connections. These logical ports on a RoCE adapter can handle both Ethernet and RoCE communication flows on behalf of multiple logical partitions. PowerVM® SR-IOV FAQs is a useful resource for more in-depth knowledge regarding implementation details for SR-IOV scenarios on Power hardware.

7.5.2. Configuration

The case study was performed on a pair of mid-range Power9 servers (9009-42G) each configured with PCIe4 2-port 100GbE RoCE x16 adapters (Feature Code EC76) for data mirroring and NVMe internal storage as a fast I/O configuration. Two logical partitions were created on each system. The first partition was configured for Db2 Mirror use. The second partition was configured to

drive additional intense network RDMA traffic to compete for aggregate bandwidth on the hardware virtualized RoCE adapters.

7.5.3. Scenario

Each experiment in the case study consisted of two sets of competing RDMA communication flows. The Db2 Mirror activity involved a consistent set of 80 database jobs each running a rapid loop iterating over individual files to sequentially read and update individual rows in their respective files. Upon reaching the end of the file, the read/update loop would continue processing by again starting with the first row of the file associated with each job. Each database file contained approximately 1 million rows of data with a uniform record size of 1300 bytes.

The second RDMA activity was implemented as a set of threads continuously sending 64K RDMA packets from the second partition on one machine to the second partition on the other machine. This interfering logic was configured to run continuously before, during and after the corresponding database activity on the mirrored partitions. Individual experiments varied the number of threads processing the continuous one-way 64K RDMA sends to vary the total amount of competing RDMA data sent over the shared physical link. This case study summarizes data from runs with between 8 and 11 threads of competing continuous 64K RDMA sends. At the higher thread counts logical resource constraints were demonstrated which resulted in observed throughput impacts on the mirrored database activity whose partitions were assigned significantly less capacity on the RoCE adapters.

Initial experiments used one logical port configured with 3% of capacity (limit) on each end to handle Db2M traffic between the mirrored partitions. The remaining 97% of capacity was allocated to a second logical port on each end to handle competing RDMA traffic between the second pair of partitions. Follow-up measurements adjusted the percentages with 6% of capacity (limit) assigned for Db2M traffic on each mirrored partition with the remaining 94% of capacity again assigned to the second partition on each machine for the competing RDMA traffic. Note that under normal operation all partitions with logical ports assigned have full access to the available bandwidth of the hardware virtualized adapter. The capacity limits (allocated to individual partitions in 1% increments with a maximum per adapter sum of 100%) only take effect when the adapter is very heavily utilized and approaching bandwidth limits. These capacity limits are then enforced accordingly by the hypervisor which yields the observed reduction in throughput rates for these experiments.

High level performance counter metrics were harvested during each experiment by using the Db2 Mirror SQL service QSYS2.NRG_INFO. Cumulative metrics for inbound and outbound RDMA bandwidth rates and message tallies were captured on each partition before, during and after each run. Deltas were summarized to show the varying performance characteristics when the 80 looping database jobs on the mirrored partition pair ran alongside between 8 and 11 threads of competing RDMA traffic on the second partition pair.

Table 7-4 shows the replicated database workload achieved an unconstrained throughput rate of approximately 416 thousand row updates per second. As the aggregate RDMA comm flows scaled up, there was a clear shift in the performance signature with 10 or more interfering RDMA threads. The replicated database throughput rates dropped to 177 thousand row updates per

second and 353 thousand row updates per second respectively when the hypervisor limited those logical ports to 3% and 6% of total capacity for the RoCE adapters. Table 7-5 shows more detailed information about aggregate inbound and outbound comm activity as well as message tallies for the Db2 Mirror activity, interfering RDMA activity and combined metrics. The key observation from the detailed metrics is that **the hypervisor capacity metrics were enforced in terms of aggregate data bandwidth and not by message count ratios**. If RDMA comm bandwidth used by Db2 Mirror is ever observed to be capped by the hypervisor during heavy RDMA activity, an administrator should consider increasing the capacity percentage assigned to the associated RoCE cards on the affected partitions running Db2 Mirror.

Table 7-4. Case study #5: SR-IOV Capacity performance metrics

SR-IOV Capacity Settings		**Mirrored Database Jobs**	**Interfering RDMA Threads**		**Average Updated Rows/Sec**	**Updates MB/sec**
3% Db2M; 97% Interference		80	8		416,049	516.20
3% Db2M; 97% Interference		80	9		415,318	515.30
3% Db2M; 97% Interference		80	10		177,259	219.93
3% Db2M; 97% Interference		80	11		177,339	220.03
6% Db2M; 94% Interference		80	8		417,753	518.32
6% Db2M; 94% Interference		80	9		416,458	516.71
6% Db2M; 94% Interference		80	10		352,851	437.79
6% Db2M; 94% Interference		80	11		352,804	437.73

Table 7-5. Case study #5: SR-IOV Capacity network traffic metrics

SR-IOV Capacity Settings	Sum Db2M RDMA	Sum Db2M RDMA	Sum Db2M RDMA	Sum Inter-fering RDMA	Sum Inter-fering RDMA	Sum Interfering RDMA	Sum All RDMA	Sum All RDMA	Sum All RDMA
	OUT Gb/sec	IN Gb/sec	Total Msg/sec	OUT Gb/sec	IN Gb/sec	Total Msg/sec	OUT Gb/sec	IN Gb/sec	Total Msg/sec
3% Db2M; 97% Interference @ 8 threads	4.49	4.57	4,278,626	83.47	83.96	686,563	87.96	88.53	4,965,189
3% Db2M; 97% Interference @ 9 threads	5.70	4.75	4,613,528	83.53	83.98	686,806	89.23	88.73	5,300,334
3% Db2M; 97% Interference @ 10 threads	1.86	1.90	1,759,392	87.29	87.89	718,553	89.15	89.79	2,477,946
3% Db2M; 97% Interference @ 11 threads	2.03	1.98	1,819,434	87.28	87.89	718,534	89.31	89.87	2,537,969
6% Db2M; 94% Interference @ 8 threads	4.93	4.67	4,525,618	83.51	83.93	686,409	88.44	88.60	5,212,027
6% Db2M; 94% Interference @ 9 threads	4.38	4.83	4,339,698	83.57	83.95	686,743	87.95	88.79	5,026,441
6% Db2M; 94% Interference @ 10 threads	4.45	3.79	3,824,974	84.54	89.06	703,502	88.98	92.84	4,528,476
6% Db2M; 94% Interference @ 11 threads	3.71	3.93	3,545,200	84.54	85.02	695,284	88.25	88.96	4,240,484

7.6. Client Scenarios

This section includes workloads that are based on actual client requirements or client-based workloads. The summaries are intended to give real-world performance results without sharing any sensitive information. Therefore, these results are much less detailed than the internal workload case studies.

7.6.1. Interactive Workload

This workload simulated interactive users in a retail banking environment and was run active-passive. The application was written in RPG and uses native I/O to access database files. The workload is 62% reads and 38% writes. The client worked with IBM Lab Services to analyze the

performance of their applications while running with Db2 Mirror. After a series of application changes were implemented, the results were:

- Throughput 19% lower (using Db2 Mirror compared to baseline behavior)
- Response time 24% longer
- CPU utilization 1% higher

Additional improvements to the application were identified but not implemented because the results had achieved the client's performance requirements.

7.6.2. End of Day Batch – Example #1

This workload simulated end of day batch and was run active-passive. It was a more write intensive workload, consisting of 37% reads and 63% writes. No application changes were made to improve performance. The Set Object Access (SETOBJACC) command was used on the target partition to pre-load objects into memory. The results were:

- Run time 34% longer
- CPU consumption 19% higher

7.6.3. End of Day Batch – Example #2

This workload simulated a batch banking application and was run active-passive. It used RPG programs and native I/O to access the database. While the workload was read intensive (82% reads and 18% writes), most of the read activity was read for update, which caused read locks to be acquired on both partitions. Therefore, the workload was nearly 100% writes. The results were:

- Run time 86% longer
- CPU utilization 0.8% lower

No application changes were made to optimize performance of this workload. Changing the read for updates activity to be read-only would significantly improve the performance of this application running with Db2 Mirror enabled.

7.6.4. End of Day Batch – Example #3

This workload simulated a batch banking application but was for a different banking solution than the previous two examples. The workload ran in dozens of jobs with a mix of read/write ratios. The database used was large and included tables with hundreds of millions of records. The workload was run active-active.

While the initial test ran 62% slower with Db2 Mirror enabled, implementing application best practices, and enabling journal cache greatly improved the performance of the batch process. With those changes the batch runtime increase was 25% with Db2 Mirror enabled. Notably, the optimized batch process ran faster than the original batch runs, even with Db2 Mirror enabled.

7.6.5. End of Day Batch – Example #4

This workload simulated a batch application using the Copy File (CPYF) command. The workload ran one job to copy data from one table to another table. The runtime increased as the number of

records in the table increased. Normalized throughput rates (records copied per second) were used to summarize the performance results consistently. The workload was run active-passive.

As data record lengths increased, the throughput of CPYF decreased because fewer records were processed in each individual block. The historical default I/O block size of CPYF was 8 KB, which had been used for many years. The default I/O block size for CPYF has been increased to 256 KB as a recent 7.4 enhancement. This has improved the performance of CPYF significantly regardless of whether CPYF runs in a replicated or non-replicated environment.

No application changes were made to optimize performance of this workload. The results were:

- CPYF throughput rates decrease as record length increases. This would be also applicable for insert/update scenarios.
- Customers upgrading from IBM i 7.3 or older IBM i 7.4 code levels prior to the CPYF block size enhancement and then adopting Db2Mirror along with the CPYF enhancement (blue bars versus golden bars) would experience a slight performance degradation with smaller record sizes. Net performance improvement would be expected when running replicated CPYF scenarios with larger record sizes (e.g., record length >= 1024 bytes). A significant enabler of CPYF performance when processing replicated data is the underlying use of blocked insert support to efficiently process rows in large quantities rather than individually.
- The performance improvement of CPYF itself when processing **non-replicated** data (blue bars versus green bars) is also significant. These experiments demonstrate that the new, larger I/O block size improves throughput rates in the 3-4x range depending on record size.

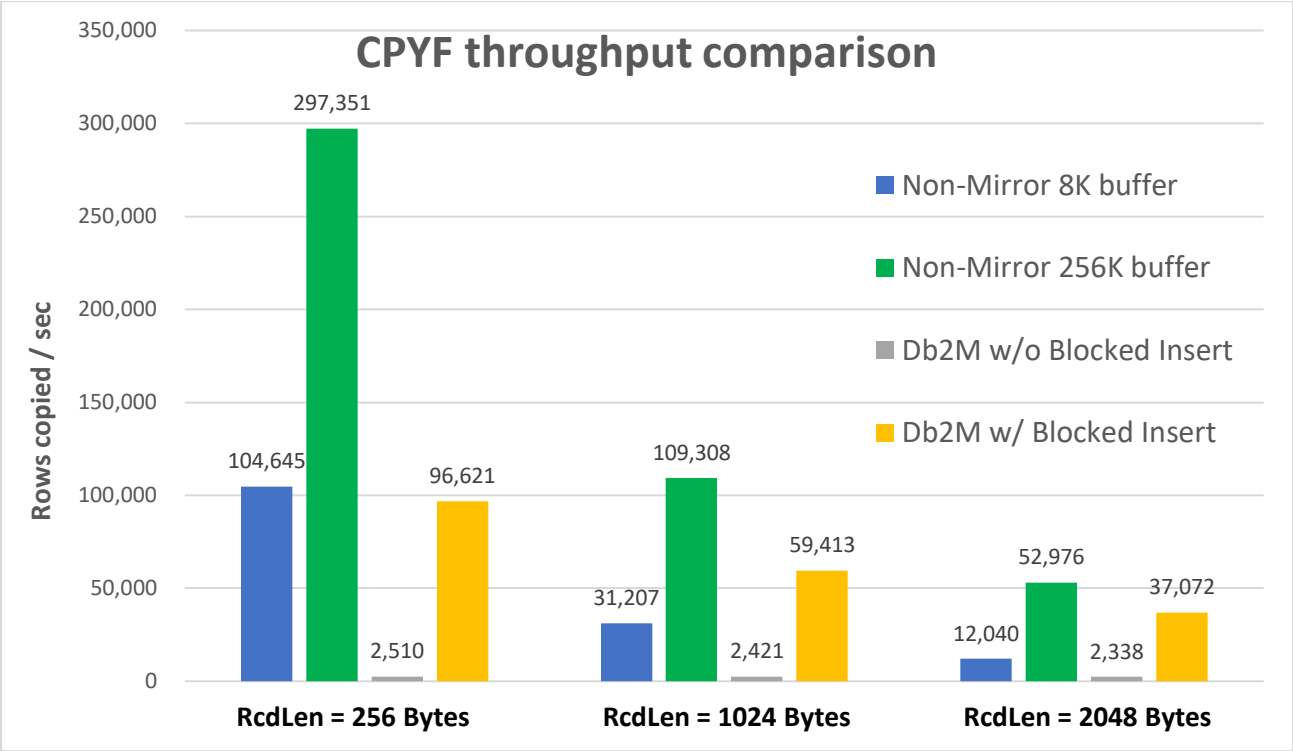


Figure 7-7. CPYF throughput comparison

8. Conclusion

Performance is critical to successfully deploying Db2 Mirror. This document describes ways to configure your system, partition, and Db2 Mirror to help improve performance. It also includes potential application changes to reduce the impact of running with Db2 Mirror.

References for more information are included in the next section. If you need additional help in assessing the potential impact of implementing Db2 Mirror or identifying ways to improve the performance of Db2 Mirror in your environment, IBM Power Systems Lab Services can help either through general consulting or through our Db2 Mirror workshop.

The Lab Services' Db2 Mirror for i Readiness Assessment Workshop is a two week, hands-on workshop where IBM provides the expertise and test environment and you provide your application(s) and test data. The first week covers planning, implementation, and setting up libraries in a Db2 Mirror environment. The second week focuses on application and performance requirements and testing.

You can contact IBM Power Systems Lab Services at ibmsls@us.ibm.com.

9. References

The following is a list of IBM documents that are good references:

IBM i on Power – Performance FAQ

IBM i Db2 Mirror Book

IBM Db2 Mirror for i Getting Started Redbook

Db2 Mirror for IBM i Readiness Assessment

Monitoring the Db2 Mirror environment

Striving for Optimal Journal Performance Redbook

Journal Management and System Performance

RoCE Adapters supported for use with Db2 Mirror

Disclaimer – IBM Db2 Mirror for i: Performance Considerations

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information may include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT.

IBM shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

The performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere.

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

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

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

IBM Corporation
Systems and Technology Group
Route 100
Somers, New York 10589

Produced in the United States of America May 2022

10032410USEN-00