IBM System z
May 2006

GDPS/PPRC 100Km Distance Testing

Pierre Cassier
Pierre_Cassier@fr.ibm.com

Axel Ligny
Ligny@fr.ibm.com

David Raften
Raften@us.ibm.com

Table of Contents

Introduction . . . . . Page 3
Testing Environment . . . . . Page 4
Montpellier Benchmark . . . . . Page 5
DB2 Background Information . . . . . Page 9
Environment 1 Results - Single Site . . . . . Page 11
I/O for DBxMSTR address spaces . . . . . Page 12
I/O for DBxDBM1 address spaces . . . . . Page 14
GBP8 Access . . . . . Page 16
Impact on the response time of the transaction TPNO: . . . . . Page 16
Breakdown of the response time . . . . . Page 18
I/O for CICS address spaces: . . . . . Page 20
Conclusion . . . . . Page 20
Environment 2 - Split Coupling Facilities . . . . . Page 22
I/O for DBxMSTR address space . . . . . Page 23
I/O for DBxDBM1 address spaces . . . . . Page 24
GroupBufferPool Structure: . . . . . Page 24
Lock Structure Data . . . . . Page 28
I/O for CICS address spaces: . . . . . Page 29
Conclusion . . . . . Page 29
Environment 3 - Multi-Site Workload . . . . . Page 30
Adjusted Response times for TPNO transactions:* . . . . . Page 33
I/O for CICS address space: . . . . . Page 39
CF Link Utilization . . . . . Page 39
Conclusion (Environment 3) . . . . . Page 41
Summary . . . . . Page 42
Appendix A - Environment 3 Adjustments . . . . . Page 43
Appendix B - Subchannels Required . . . . . Page 46

Introduction

Today's technology allows ever-increasing separation of data center components across multiple locations. One can have a "tightly-coupled" Parallel Sysplex® cluster spanning 100km, with support provided by FICON® between server and disk, ISC-3 Coupling Links between server and CF, Server Time Protocol (STP) timing links between servers, and synchronous remote copy between control units using Metro Mirror. For each of these components, response time for single requests increases linearly. As incredible as it may seem, the issue facing real life situations is the speed of light.

In a vacuum, light travels at about 300,000 Km/second, or about 186,000 miles/second. This is fast enough to travel around the world about 7 ½ times each second. Through fiber optic glass, light slows down to about 2/3 this value, or about 200,000 Km/second. Despite this incredible speed, the slowdown has a very real and measurable effect on workloads with today's computers, especially when a transaction issues multiple I/O requests, causing this distance to be traversed many times.

This report documents the results of studies done to measure the effect of distance on a workload with the disk being synchronously remote copied and managed by GDPS®/PPRC.

Testing Environment

Workload transactions were injected by the TeleProcessing Network Simulation (TPNS) terminal simulator from an outside system with 100 simulated terminals on each of two z/OS® images, G1P1 and G1P2.

Although five types of transactions were scripted by TPNS, only transaction TPNO was studied as the other transactions had no significant response times or not enough occurrences.

Two IBM eServer™ zSeries® 990 (z990) 2084 Model 312 servers were used, one for each "site." Each z/OS LPAR was configured with eight logical CPs. Utilization on each LPAR usually stayed under 10% busy. This was done by design so that CPU delays would not be a factor. Coupling Facilities were configured as ICFs on the servers with two dedicated CPs each. IBM DS8000 disk was used.

Distances were simulated with DWDM (IBM 2029 or ADVA) equipment, and adding fiber optic cable to simulate the sites being further apart.

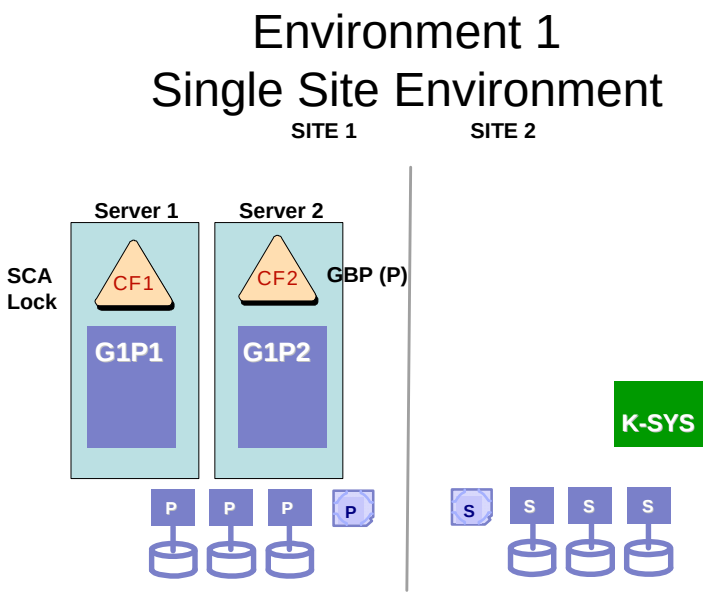
Montpellier Test Environments

There were three different environments conducted by Montpellier GDPS Test team.

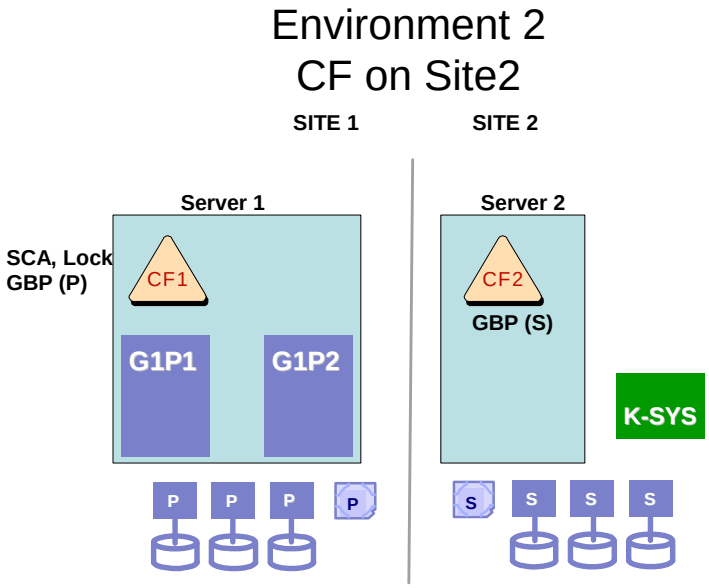
1. Environment 1 is configured with a single site CICS®/DB2® workload with two production LPARs (G1P1 and G1P2), both on "Site1". CF1 and CF2 are both local to the production systems in site1, and the remote site contains only secondary disk. The GDPS® controlling system (K-Sys) is on Site2.

   For Environments (2) and (3) described below, the group buffer pools were duplexed, with the primary GBPs in CF2, and the secondary GBPs in CF1. Environment (1) did not run with GBP duplexing. Note that this is not a recommended configuration as the purpose of having (DB2 managed) GBP duplexing is to improve recovery time by switching to use the duplexed copy of the GBP if the primary GBP structure is lost.

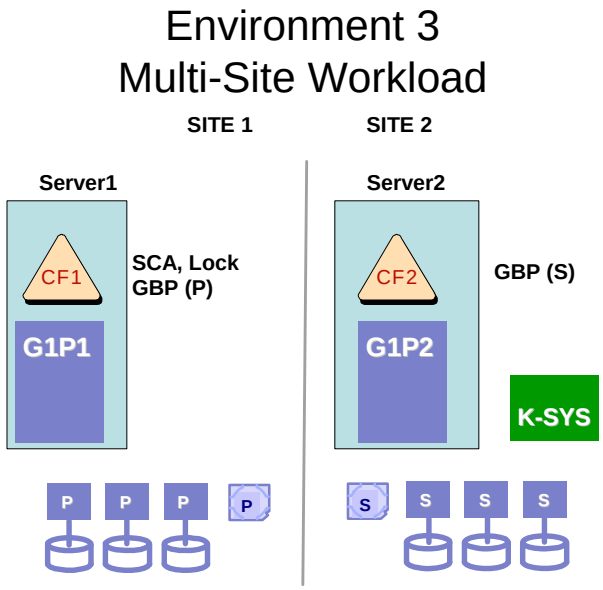
   The IRLM Lock structure (DSNDBR0_LOCK1) and DB2's SCA structure (DSNDBR0_SCA) are in CF1.



2. Environment 2 has CF1 in Site1 containing the DB2 primary group buffer pools, the IRLM lock structure, and the SCA structure. CF2 is in Site2 containing the secondary group buffer pools. This configuration is typically used by customers who wish to continue to use a CF structure at the secondary/recovery site after a primary site failure, avoiding the "log based" recovery procedures that are needed when there is a loss of CF structure data. This is supported by GDPS V3.3 “Enhanced Recovery Support”.



3. Environment 3 is a multi-site workload. One production LPAR is on Site1 while the other production LPAR is on Site2. The Coupling Facilities are split as described above. This test case corresponds to the GDPS Multiple Side workload. As seen from G1P1, the configuration is the same as environment 2. From G1P2's perspective, since that LPAR is running on the Site 2, each disk and CF access will be impacted by the distance between the two Sites.



For each environment, there were five sets of measurements:

1. Base environment, consisting of a Parallel Sysplex environment with the Coupling Facility near the servers. There was no data replication.
2. The servers and disk were near each other, but data was replicated using Metro Mirror (PPRC). The environment was managed by GDPS/PPRC.
3. Same as above, but the remote copy disk was "placed" 20km from the primary site. This was simulated by having the FICON and CF connections travel through 20 km of fiber optic cable looped within the computer room. DWDMs were used at both ends of the connections.
4. Same as above, but simulating the remote site 40 km away.
5. Same as above, but simulating the remote site 70 km away.
6. Same as above, but simulating the remote site 100 km away.

The expected result was for transaction response times to increase linearly as distances increase. Although an individual I/O or CF message would have its response time elongated by about 10 microseconds / km, an individual transactions can initiate several I/O and CF requests. For example, if the remote site is 70 km away, a single I/O (with synchronous remote copy) would be delayed by 10 microseconds/Km * 70 km = 700 microseconds = 0.7 millisecond (0.0007 seconds) for each back and forth communication (since FICON protocol requires one round trip for each I/O, response time is delayed by 0.7 milliseconds in this example). If, for example, there were 12 CF requests and four DB2 synchronous I/Os going across 70 km, then transaction response time would be elongated by (12 * .7 ms) + ((4*2) * .7 ms) = 14 ms.

Calculating the propagation delays on a DB2 transaction can be difficult. There are many types of I/O for different purposes. Local bufferpool updates to disk are typically asynchronous to the transactions. In addition, DB2 can have read-aheads of multiple pages which run together with a transaction. In addition, synchronous reads and many log updates are synchronous to the transaction. There are also different types of CF requests. Each transaction can issue many locks one at a time, but unlocks can be blocked. Data is updated to the CF at commit time, but data can be read from disk or CF depending upon if previous transactions updated that page recently. As response time increases, CF requests can be converted from "synchronous" to "asynchronous" to z/OS, but still being synchronous from the transactions point of view. However, a transaction would then need to be re-dispatched when the CF response is received, instead of just continuing execution. There is also the Cast-out processing to migrate GBP data to disk. Finally, these delays elongate the transaction response times, so the transactions hold on to resources longer, causing secondary effects on other transactions. With these considerations, it is often better to benchmark the environment instead of trying to rely on analysis.

The CICS®/DB2 workload was run using the TPNS terminal simulator.

DB2 Background Information

In the detailed analysis of transaction response times, this document references two specific DB2 address spaces. They are:

xxxxDBM1 The DBM1 address space performs database services and manipulates most of the structures in user-created databases. In DB2 Version 8, storage areas such as buffer pools reside above the 2 GB bar in the DBM1 address space. With 64-bit virtual addressing to access these storage areas, buffer pools scale to extremely large sizes. I/O activity includes synchronous reads of data, asynchronous reads (list prefetch, dynamic prefetch), and asynchronous writes (deferred database writes). There may also be some synchronous writes.

xxxxMSTR The MSTR (master) address space performs many system-related functions such as logging and archiving. Log entries are normally asynchronous, but are synchronous to the transaction when logs are forced to disk at commit, and when log buffers fill up. Reads are done for archiving and backouts.

"xxxx" is the subsystem name. "G1P1" is the subsystem name of the DB2 running on z/OS G1P1, and "G1P2" is the DB2 running on z/OS G1P2".

DB2 can contain several other address spaces such as to run stored procedures (SPAS), support distributed requests (DIST), manage locks (IRLM), as well as other address spaces for other types of work. These are described in DB2 Universal Database™ for z/OS Administration Guide Version 8 SC18-7413-02.

When DB2 data sharing is active, several structures are put into the Coupling Facilities. They include the DB2 owned Group Buffer Pools (GBP), one buffer pool for each "local" buffer pool that contains shared data, the DB2 owned Shared Communication Area (SCA) structure to pass recovery information between the DB2s, and the IRLM owned Lock structure, used to manage locks between the members of the DB2 data sharing group. The GBPs contain a copy of pages that were updated by one of the DB2s. If another DB2 wishes to read that page, it can be read from the CF, avoiding an I/O to disk. This report concentrates specifically on transaction TPNO. This uses GBP8 when updating its data.

Eventually, the GBPs fill up and data not referenced recently needs to be migrated to disk. Since the CF does not have direct connections to disk, it is the responsibility of one of the DB2s to read in old changed data, and write it out to disk. This is called the castout process. the "Castout owner" is the DB2 that performs the castout process. The first DB2 that opens for update a table or table partition is the castout owner for that table or partition. If transactions started on all DB2s at the same time, each DB2 would be a castout owner for

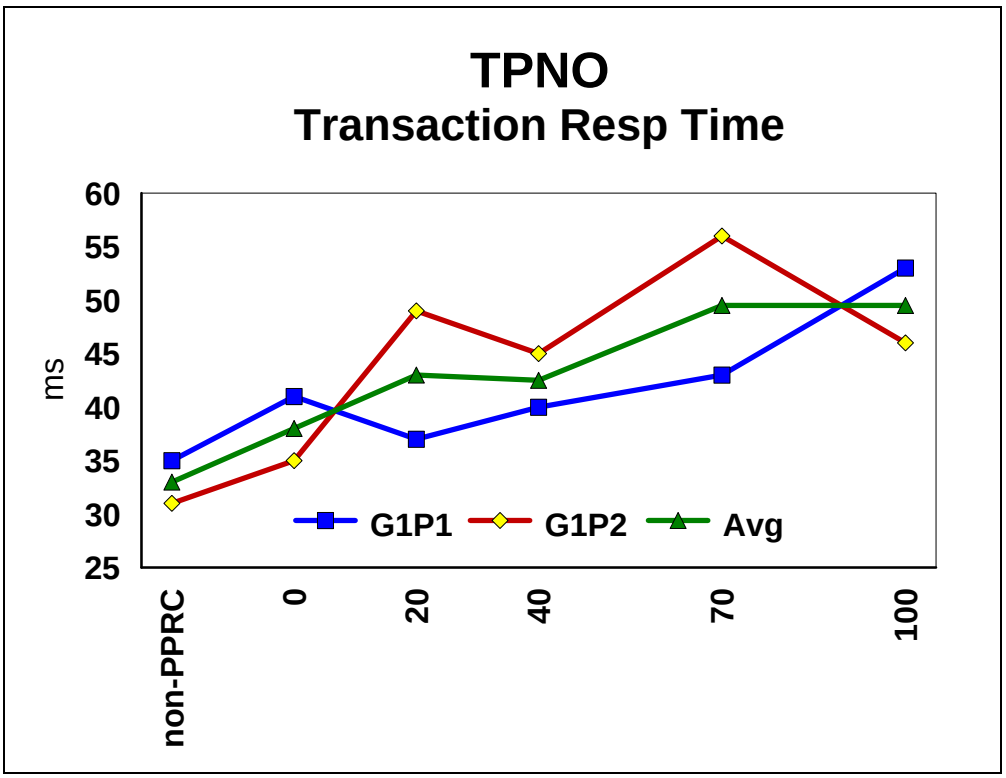
roughly an equal amount of data. In this environment, the workload started on one DB2 earlier than on the other DB2. The first DB2 became the castout owner for the bulk of the tables. This affected the amount of I/O that was done by that particular DB2. More information on castout processing is described in DB2 Universal Database for z/OS Data Sharing: Planning and Administration Version 8 SC18-7417-02.

In environments (2) and (3), the GBPs are (DB2 managed) duplexed. Instead of all updates to the GBPs copied to both structure instances, only the changed pages written to the primary GBP are duplexed. For these, DB2 initiates a write to the secondary GBP overlapped with a write to the primary GBP. DB2 then checks to ensure that the writes finished successfully. If there are several pages to write, in order to maximize performance, rather than writing one page at a time, DB2 schedules the writes for some number of pages to the secondary GBP and then writes those pages to the primary GBP, and finally checks for the completion of the writes to the secondary. To save on performance, DB2 "registering interest" in pages that are read from disk are not duplexed.

GBP8 is the busiest of the eight group buffer pools configured, with about 42% of all DB2 Group Buffer Pool requests. In addition, TPNO initiates about 78% of all GBP requests. From that, we looked at the number of GBP8 requests and service times as indications of TPNO performance.

Environment 1 Results - Single Site

Environment 1 is a typical single site environment with only the PPRC secondary disk residing on Site 2. This is a typical environment for a "Single Site" workload, where all production work is run on Site 1, with Site 2 used for disaster recovery. Site 2 can optionally be used to run development and test workloads as well.



As expected, as the secondary disk is moved further away from the primary, transaction response time increases linearly. Response time also increases slightly going from "non-PPRC" to a "PPRC 0 Km" configuration. This is because the Primary disk has to wait until the update is placed on the link to the secondary, it can process the request, and an acknowledgment is received. This takes about 0.4 ms per request. For example, if there were on the average seven I/O operations per transaction, that would translate to 2.8 ms response time increase due to PPRC. Since the two production LPARs are both configured on the same site and are configured similarly, it is not unreasonable to average the response times. There are subtle differences in the configuration, such as the location of the structures with respect to the z/OS images, and which is the owning DB2 for the GBPs. The effect of these differences are discussed below.

I/O for DBxMSTR address spaces

Since the G1P1 and G1P2 LPARs are both on the same site, average values for I/O response time for DB2 on G1P1 and G1P2 track closely. Merged data is shown below. Two of DB2's address spaces include the DB2MSTR and DB2DBM1 address spaces. The DB2MSTR address space is responsible of logging and archiving. With this workload, the I/Os from this address space were mostly Synchronous Writes. The DB2DBM1 address space is also in charge of asynchronous read (prefetch) and asynchronous write I/Os.

I/O for DBxMSTR address spaces in milliseconds (ms):

DBxMSTR	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	43.5	42.2	85.1	43	43.4	42.5
RESP	0.4	0.7	1	1.2	1.5	1.8
CONN	0.2	0.2	0.6	0.3	0.3	0.4
DISC	0	0.3	0.6	0.8	1.1	1.3
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0	0	0	0	0	0

- ***SSCH Rate***

  The rate at which start subchannel (SSCH) instructions to I/O devices completed successfully.

- ***RESP***

  Total disk response time in milliseconds required to complete an I/O request. This value reflects the total hardware service time which starts at the acceptance of a SSCH instruction and ends at the acceptance of the channel end, and the front end software queuing (IOSQ) time, which reflects front end queuing, involved for the average I/O request to the device.

  RESP Time = (Connect + Disconnect + Pending) + IOSQ Time

- ***CONN***

  Connect time is the average number of milliseconds the device was connected to a channel path and actually transferring data between the device and central storage. Typically, this value measures data transfer time but also includes the search time needed to maintain channel path, control unit, and device connection.

- ***DISC***

  Disconnect time is the average number of milliseconds the device was disconnected while processing an SSCH instruction. This value reflects the time when the device was in use but not transferring data. It includes the overhead time when a device might disconnect to perform positioning functions such as SEEK/SET SECTOR, as well as any reconnection delay.

- ***PEND***

  Pending Time is the average number of milliseconds an I/O request must wait in the hardware. This value reflects the time between acceptance of the SSCH function by the channel subsystem (SSCH-function pending) and acceptance of the first command associated with the SSCH function at the device (subchannel active). This value also includes the time waiting for an available channel path and control unit as well as the delay due to shared DASD contention.

- ***IOSQ***

  IOSQ time is the average number of milliseconds an I/O request must wait on an IOS queue before a SSCH instruction can be issued.

There is a very linear growth of the Disconnect time. Connect time also increases, but not as much as DISC time. This growth is due to the extra time needed to communicate across the distance. All other values are staying flat

In order to check the values found for SSCH Rate, statistics from DB2 trace were produced. This showed the following data (for the non-prc measurement):

The DB2 Trace records included the following data for DB2 on G1P1:

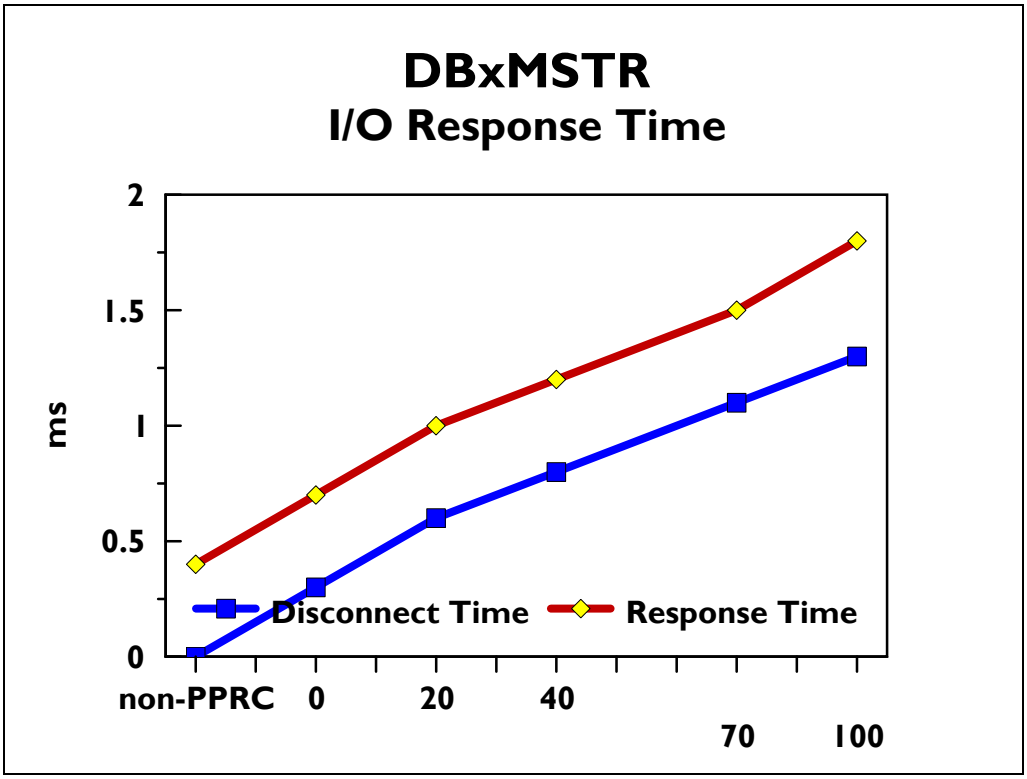
```
LOG ACTIVITY                  /SECOND
---------------------------   -------
LOG WRITE I/O REQ (LOG1&2)      42.03
```

And for the DB2 on G1P2:

```
LOG ACTIVITY                  /SECOND
---------------------------   -------
LOG WRITE I/O REQ (LOG1&2)      42.24
```

Since these match closely with the SSCH rates (43.2 and 43.7), it shows that almost all of the SSCH I/Os were Write-Only operations.

The chart below shows the effect of Disconnect Time on the total response times for the DBxMSTR address space. As expected, this grows linearly with distance.



I/O for DBxDBM1 address spaces

As with the MSTR address spaces, the I/O components for the DBxDBM1 address spaces were broken down. Since the two DB2s were in a symmetric configuration, the results were combined as shown below:

DBM1 G1P1 + G1P2	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	212.2	219.2	212.7	213.1	226.3	212.3
RESP	0.6	1.2	2.0	2.2	3.2	3.0
CONN	0.3	0.4	0.4	0.4	0.4	0.5
DISC	0.1	0.6	1	1.3	1.8	2.2
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0.1	0.1	0.5	0.4	0.8	0.2

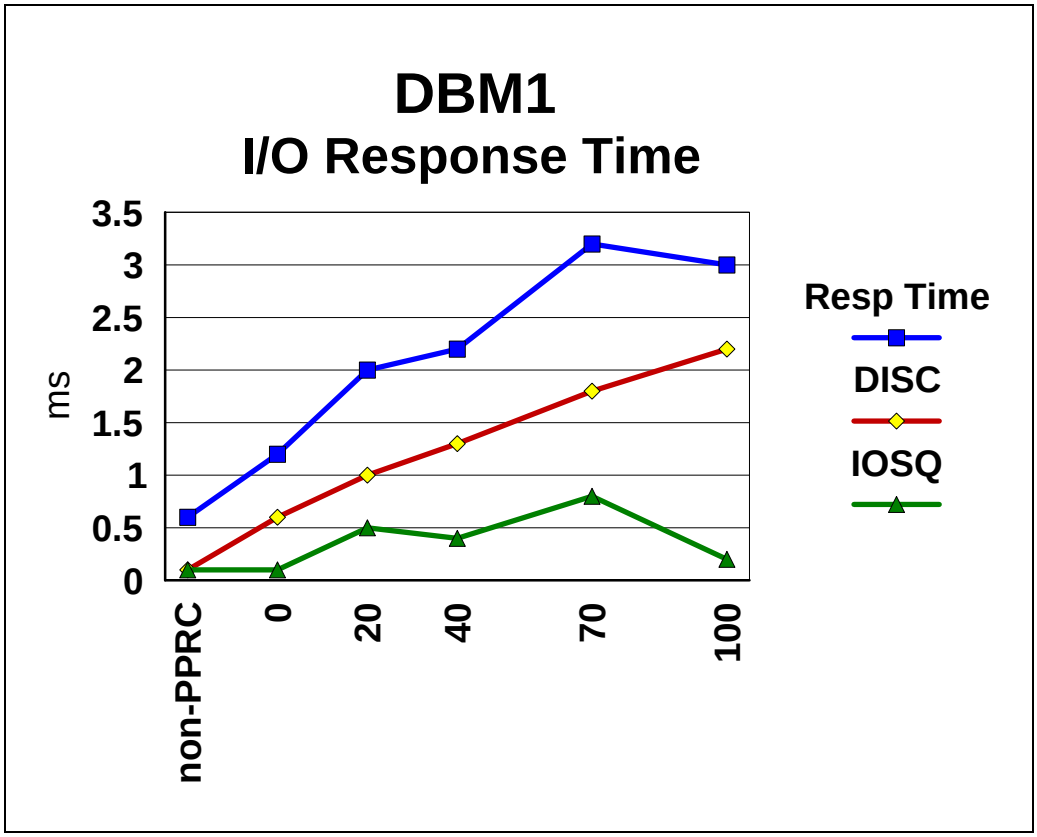
Since one of the DB2s (in this case, the DB2 on G1P1) is the "owning DB2" of the GBPs, it had different SSCH rates than the other DB2, but the I/O characteristics were the same.

The SSCH requests were compared against the Asynchronous Writes in the DB2 Statistics.

DB2	Total 4K Asynchronous Write Operations / Second	SSCH Rate
G1P1	176.49	178.5
G1P2	32.91	33.6

This close match confirms that in the environment, all the I/Os were DB2 updates.

The merged results for the DBM1 address space response times are shown as below.



The disconnect time growth is very linear for the DBxDBM1 Address Spaces, as it was for the DBxMSTR Address Spaces. But here, unfortunately, the total disk response time is impacted by the IOSQ time. This IOSQ time represents the I/O contentions on a given DASD volume. The DBM1 Address Spaces are accessing in parallel the same volume for both read I/O activity (Database files prefetch) and for Write activity (castout and asynchronous writes). In addition, the database synchronous read activity is not performed by the DBM1 Address Sapce, but is done by the "client" Address Space (in this case, the CICS region). The only solution to reduce the contention is to improve the access performance by PAV usage.

GBP8 Access

The table below shows the number of CF requests (in K) and the service time, synchronous or asynchronous from the two members: G1P1 and G1P2. Keep in mind that since z/OS V1.2, there has been code to convert synchronous CF messages to asynchronous if the response times are too long. The goal of this is to save CPU coupling costs by allowing the processor to perform other activity instead of just waiting for the response from the coupling facility.

GBP8	non PPRC		00 km		20 km		40 km		70 km		100 km	
	nb (k)	rt (µs)	nb (k)	rt (µs)	nb (k)	rt (µs)	nb (k)	rt (µs)	nb (k)	rt (µs)	nb (k)	rt (µs)
G1P1 s	32	44	28	45	29	44	26	43	30	44	25	44
G1P1 a	199	333	196	315	74	337	74	339	76	329	188	328
G1P2 s	100	11	101	11	218	11	218	11	226	11	102	11
G1P2 a												
Tot Req	331		325		321		318		332		315	

The Group Buffer Pools were in CF2 on the same server as where G1P2 was. This allowed the G1P2 LPAR to be able to use the fast internal IC links and get excellent CF response times. Because of that, all of G1P2's requests were all synchronous. None of the GBP structure messages were converted to asynchronous.

Impact on the response time of the transaction TPNO:

The Workload Manager (WLM) goal for the TPNO transaction is to have 90% of the transactions maintain an average response time of 0.125 sec., or 125 milliseconds (ms).

The average response time of the TPNO transaction (in milliseconds) for both z/OS members are shown below:

TPNO	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
Number	6882	6800	6785	6843	6869	6682
resp. time (ms)	33	38	43	43	50	50
percentile	98	97.2	95.7	96	94.4	93.7

Average response time of the TPNO transaction for both z/OS members:

- *Number: is the number of transactions executed during the 5 minutes measurement interval. For example, if there were 6882 transactions completed, then this translates to 6882 / 300 = 22.94 transactions / second.*
- *Response time: is the average response time (in millisecond) for the transaction*
- *Percentile: is percentage of transactions achieving the goal (125 ms)*

Details for each z/OS are shown below because response times are different between the two LPARs.

TPNO - G1P1	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
Transactions	3441	3427	3416	3423	3467	3433
Resp. time	**35**	**41**	37	40	43	**53**
Percentile	98.9	97.9	98.7	98.4	97.4	94.2

TPNO - G1P2	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
Transactions	3441	3373	3372	3420	3402	3449
Resp. time	31	35	**49**	**45**	**56**	46
Percentile	97.1	96.5	92.7	93.7	91.3	93.3

The DB2 group buffer pools used by TPNO transactions are located in the CF2 coupling facility. Since the G1P2 member is located in the same CEC as this CF, this DB2 member got better response time requests using IC links than the other DB2 member (located in the other CEC) which accesses this CF using ISC3 links.

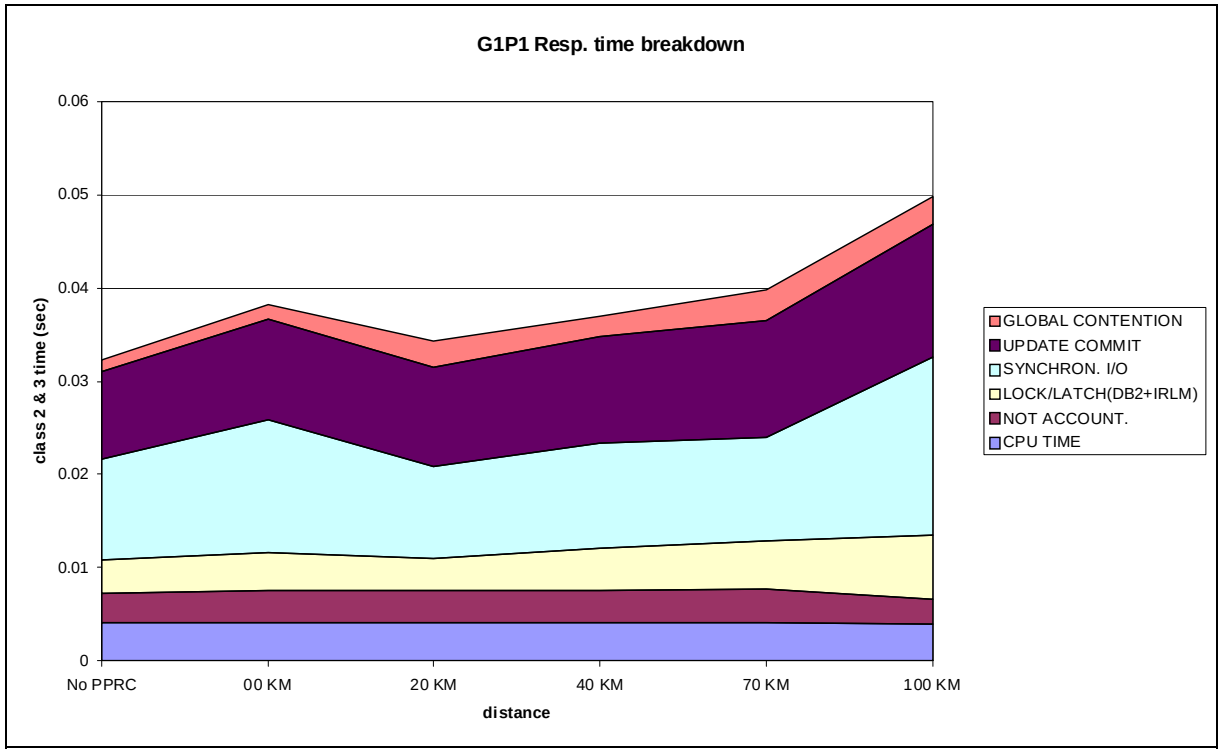
Depending on the owning member of the GBP, response times of requests can fluctuate. For the 20, 40 and 70 km cases, DB2 on G1P2 was the castout owner.

Bold response times are for the castout owning member. We see that the response time is better on the 'non castout member'. To be meaningful, we again have to look at the average for the two members of the DB2 datasharing group.

Since castout I/Os are asynchronous, they should not impact the transaction response time. It is important to focus on the breakdown of the response time in order to understand the fluctuation.

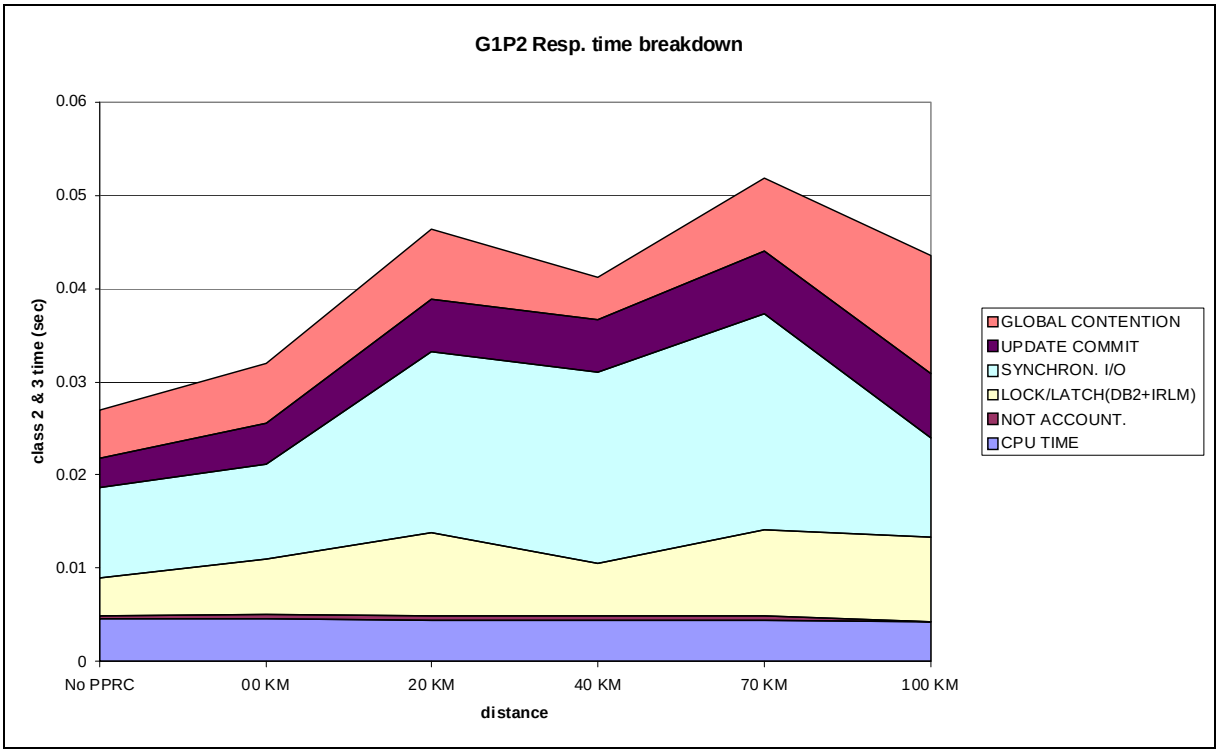
Breakdown of the response time

These breakdowns were obtained by extracting data from a DB2 accounting class 2/3 trace.



AVERAGE TIME GLOBAL CONTENTION

The total accumulated waiting time caused by the suspension of IRLM lock requests due to global lock contention in a data sharing environment that require intersystem communication to resolve.



AVERAGE TIME UPDATE COMMIT
The accumulated wait time due to a synchronous execution unit switch for DB2 commit, abort, or deallocation processing. This value is an average.

AVERAGE TIME SYNCHRONOUS I/O
The I/O elapsed time accumulated due to synchronous I/O suspensions. DB2 calculates this value by subtracting the store clock time when an agent begins waiting for a synchronous I/O from the time the agent is resumed.

AVERAGE TIME LOCK/LATCH (DB2+IRLM)
The accumulated lock and latch elapsed time. It indicates the elapsed time the allied agent waited for locks and latches in DB2. This value does not include suspensions due to group-level lock contentions in a data sharing environment. When the event completes, the ending time is used to calculate the total elapsed wait time.

The breakdown response time is different on the two members. There are two reasons:

1. The DB2 member that has the castout ownership. This has a direct relationship with the I/O response time.
2. The hardware access to the CFs is different. G1P1 is in the same physical box as CF1 (LOCK and SCA structures) and G1P2 is in the other box, as CF2 (Group buffer Pool structures).

In our case, synchronous I/Os, the most contributor of response time, are read I/O operations executed by CICS address spaces themselves. These the synchronous I/Os are performed by the 'client' address space. Keep in mind that Reads are only performed against the primary disk volumes, so are unaffected by distances.

I/O for CICS address spaces:

G1P1 CICS	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	327.6	317	315.9	312.4	328.2	318
RESP	0.4	0.6	0.3	0.4	0.3	0.8
CONN	0.1	0.1	0.1	0.1	0.1	0.1
DISC	0	0	0	0	0	0
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0.1	0.3	0	0.1	0	0.5

G1P2 CICS	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	312.2	311.8	315.1	317.2	306.1	317.2
RESP	0.3	0.3	0.8	0.9	1	0.3
CONN	0.1	0.1	0.1	0.2	0.1	0.1
DISC	0	0	0	0	0	0
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0	0	0.6	0.6	0.7	0

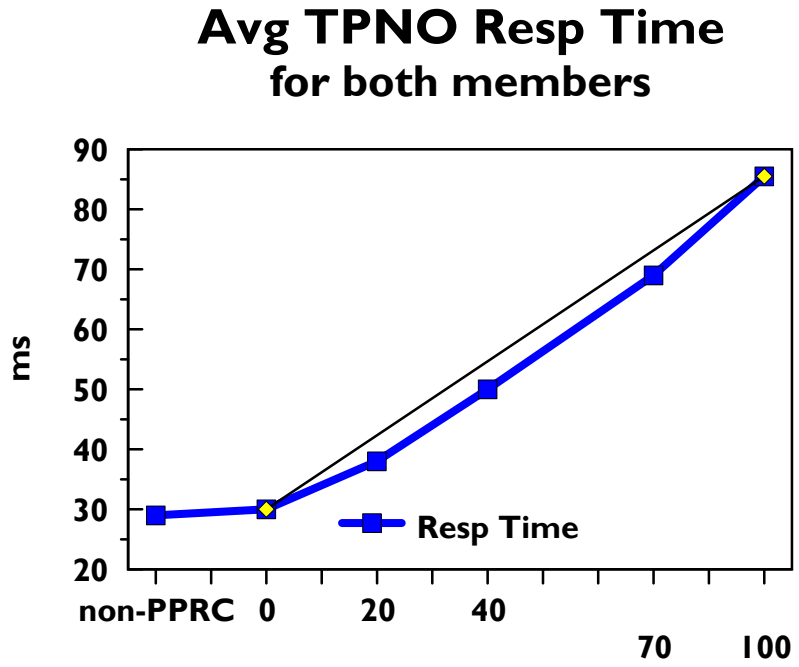
Conclusion

The graphics of the DASD resp. time for both logging activity and asynchronous write activity shows that the impact of PPRC distance is almost linear.

The study of the transactions response time (average of the two members) is almost linear too. A deeper analysis shows that the main contributor of the response time degradation is the synchronous I/O. Degradation of response time of those read I/Os is mainly due to IOSQ time. This should be improved by a heavier PAV usage. In the next test case, we have added aliases to verify this assumption.

Environment 2 - Split Coupling Facilities

Environment 2 is also a Single Site workload containing both the G1P1 and G1P2 LPARs, but with CF2 on the secondary site containing the DB2 secondary GBPs. This typical configuration is used by customers who wish to continue to use a CF structure at the secondary/recovery site after a primary site failure, avoiding the "log based" recovery procedures that are needed when there is a lost of CF structure data.



The DB2 Group BufferPools (GBP) are duplexed by (DB2 managed) CF duplexing. When transactions are COMMITed, if a page in a shared table is updated, a copy of this page (usually 4KB) is written to the primary as well as secondary instance of the GBP. Because this process is synchronous from the transaction point of view, response time is impacted more than in Environment 1 as distances increase.

The plot of the average response time from the two sites are shown, but this is not a straight line. As shown by the dotted line, the penalty per km is increasing slightly as the distances increase. This is due to the fact that as transactions start to take longer to run, they hold on to database locks and other resources longer. Other transactions are now waiting for these resources to free up before they can continue processing. This secondary effect is causing an exponential slope on the response time graph. In this environment, although noticeable, the secondary effects are not significant.

I/O for DBxMSTR address space

The I/O response times for the DBxMSTR address spaces in milliseconds (ms) are shown below:

G1P1 MSTR	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	43.2	42.7	41.7	41.6	40.7	42
RESP	0.6	0.7	0.9	1.6	1.5	2.1
CONN	0.3	0.2	0.2	0.4	0.2	0.3
DISC	0.2	0.3	0.6	1.1	1.1	1.7
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0	0	0	0	0	0

G1P2 MSTR	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	45.3	44.7	43.6	42.8	42.5	40.6
RESP	0.4	0.8	1	1.2	1.5	1.7
CONN	0.2	0.3	0.3	0.3	0.3	0.3
DISC	0	0.3	0.6	0.8	1.1	1.3
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0	0	0	0	0	0

The summary for both DB2s gives the following:

Both MSTR	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	88.5	87.4	85.3	84.4	83.2	82.7
RESP	0.5	0.7	1	1.4	1.5	1.9
CONN	0.2	0.3	0.3	0.3	0.3	0.3
DISC	0.1	0.3	0.6	1	1.1	1.5
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0	0	0	0	0	0

For the system level I/O performed by the DB2 Master address space, there are no significant differences between Environment 1 and Environment 2. Logging and archiving of data is related to disk location, so moving CF2 with the Group Buffer Pools to Site 2 has no effect.

I/O for DBxDBM1 address spaces

The I/O response times for the DBxDBM1 address spaces in milliseconds (ms) is shown below. This is the sum of the two DB2 members:

Both DBM1	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	227.1	219.4	210.2	216.5	213.2	211.8
RESP	0.7	1.3	1.5	1.9	2.3	2.7
CONN	0.3	0.4	0.4	0.4	0.4	0.4
DISC	0.1	0.6	1	1.3	1.7	2.2
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0.2	0.1	0	0.1	0.1	0

Again, the results are very close to the first study. Moreover, as we improved access to the volumes thanks to the usage of more aliases, we don't see here the impact of the IOSQ time anymore.

GroupBufferPool Structure:

In Environment 2, the Group Buffer Pools are duplexed, with the "Secondary" GBP located on Site 2. The number (n) of CF requests to each GBP in thousands (k) as well as the response times (rt) in microseconds (µs) is shown for both the primary and secondary GBPs for each z/OS image. The table also breaks up the GBP messages by synchronous and asynchronous CF operations.

GBP8 primary (located in CF1)

GBP8 was looked at because it best shows the effect of distances on TPNO.

GBP8 prim	no pprc		00 km		20 km		40 km		70 km		100 km	
	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)
G1P1 s	104	9.4	102	9.4	99	9.3	101	9.4	99	9.5	220	9.4
G1P1 a												
G1P2 s	229	8.9	227	8.9	218	8.9	226	9	218	9.1	102	9.5
G1P2 a												

GBP8 Secondary (located in CF2)

GBP8 Sec	no pprc		00 km		20 km		40 km		70 km		100 km	
	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)
G1P1 s												
G1P1 a	35	153	34	151	34	538	35	826	34	1146	82	1436
G1P2 s												
G1P2 a	87	233	84	210	82	538	84	777	82	1053	34	1625

DB2 user managed duplexing for the Group Buffer Pools differs from z/OS System Managed Duplexing in that not all updates are duplexed (copied) to both structures. While any transaction generated updates to database pages are duplexed to both structures, the "Register Interest" CF notifications after a page is read only get propagated to the primary GBP (in the directory part of the cache structure). For that reason, the secondary GBP is significantly less busy than the primary, saving on overall performance.

The secondary GBPs see a much larger response time than the primary GBPs. This is due to three reasons.

1. The average amounts of bits being transferred to the secondary is much less. Operations to the primary GBP include registering interest after every read of a new page, writing updated 4K data pages, and reading any 4K data pages that the other DB2 may have written (GBP cache hit). Since the read/write ratio is normally around 5:1 or more, and the "Register Interest" operation only sends a relatively small amount of bits compared to the 4K data pages, and there are relatively few GBP cache hits, the amount of data written the primary GBP is less per CF message.
2. The CF containing the primary GBP structure is on the same server as the z/OS images. Fast IC coupling links can be used instead of the much slower ISC3 coupling links to connect to the CF holding the secondary GBPs.
3. The CF holding the secondary GBPs is much further away and there are propagation delays.

Because of the long response times, CF requests to the secondary structure on CF02 are converted to asynchronous.

Only changed pages written to the primary GBPs are duplexed. DB2 initiates an asynchronous write to the secondary GBP, followed by a synchronous write to the primary GBP. DB2 then checks to ensure that the asynchronous write finished successfully. Write operations to both the primary and secondary GBP are overlapped for better performance. If there are several pages to write, in order to maximize performance, rather than writing one page at a time, DB2 schedules the writes for some number of pages to the secondary GBP and then writes those pages to the primary GBP, and finally checks for the completion of the writes to the secondary.

As expected, the service time is increasing as the distance growths.

TPNO Response Times

The average response time for both z/OS systems for the TPNO transactions is seen in the table below, with the percentile of transactions meeting the WLM goal of 125 ms.:

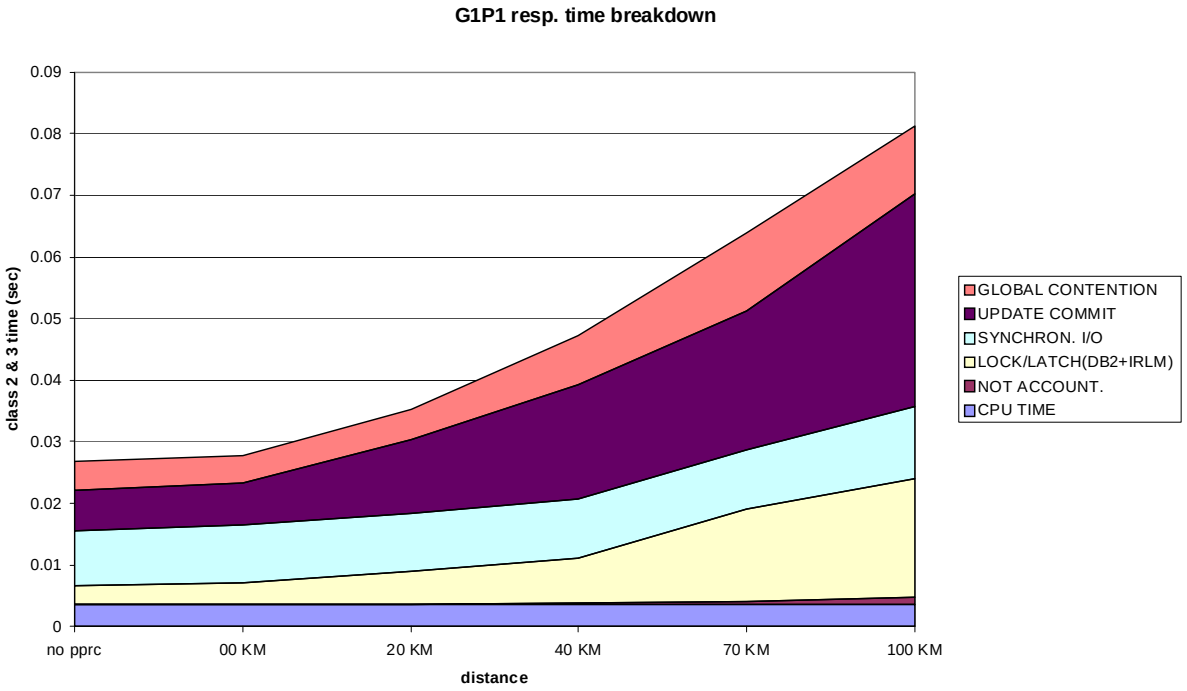
TPNO	Non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
Number	6880	6820	6731	6841	6746	6795
resp. time	32	34	43	52	69	87
percentile	97	96.6	95.9	94.3	90.7	85

This chart shows the response time of the transaction TPNO for the five simulated distances.

As described in the beginning of this chapter, the effect of distances is more pronounced and the curve is starting to show an exponential shape.

Breaking down the components of the response times:

TPNO P1	no-pprc	pprc 0 km	20 km	40 km	70 km	100 km
Number	3425	3401	3355	3428	3364	3411
resp. time	29	30	38	50	69	85.5
percentile	97.7	97.7	97.2	95.3	91.4	81.9



The two main contributors at long distance are the "update commit" and "lock/latch". These two components are probably linked: the longer the update takes, the longer the locks will be kept). The "update commit" depends of the CF access time to the group buffer pools structures.

In those breakdowns, we see that the two main contributors at long distance are the "update commit" and "lock/latch". These two components are probably linked: the longer the update takes, the longer the locks will be kept). The "update commit" depends of the CF access time to the group buffer pools structures.

Lock Structure Data

The following table shows service time for the IRLM structure DSNDBR0_LOCK1, located in CF1 on Site 1.

LOCK1	no pprc		00 km		20 km		40 km		70 km		100 km	
	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)
G1P1 s	186	7	185	6.8	206	6.6	202	6.8	196	6.9	188	7.2
G1P1 a												
G1P2 s	202	6.5	216	6.3	181	6.5	185	6.4	177	6.8	211	6.5
G1P2 a												

This service time is very good, close to the best we can obtain on a lock structure with this configuration of an IBM eServer zSeries 990 (z990) server with IC internal links.

The 'automatic' synchronous to asynchronous CF conversion has a direct impact on the UPDATE COMMIT time as well as the DB2 'not account' values.

I/O for CICS address spaces:

CICS (synchronous I/Os) disk response times are shown below. The data shows that service times are stable. Since these are Read I/Os performed against the primary disk, they are unaffected by distances.

G1P1 CICS	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	316.8	317.2	315.4	319.8	314.7	321.2
RESP	0.4	0.3	0.3	0.3	0.3	0.4
CONN	0.1	0.1	0.1	0.1	0.1	0.1
DISC	0	0	0	0	0	0
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0.1	0	0	0	0	0.1

G1P2 CICS	non-pprc	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	315.8	324.4	311.2	322.7	315.6	322.3
RESP	0.4	0.5	0.3	0.4	0.4	0.3
CONN	0.1	0.1	0.1	0.1	0.1	0.1
DISC	0	0	0	0	0	0
Q+PEND	0.1	0.1	0.1	0.1	0.1	0.1
IOSQ	0.1	0.2	0	0.1	0.1	0

The PPRC impact for CICS is comparable to what we got in the first case.

So far, these measurements have proved that the address spaces that aren't impacted by the CF2 distance show an I/O response time close to what we had measured in Environment 1 (with local CF2). In fact, I/O response time is better thanks to the PAV usage.

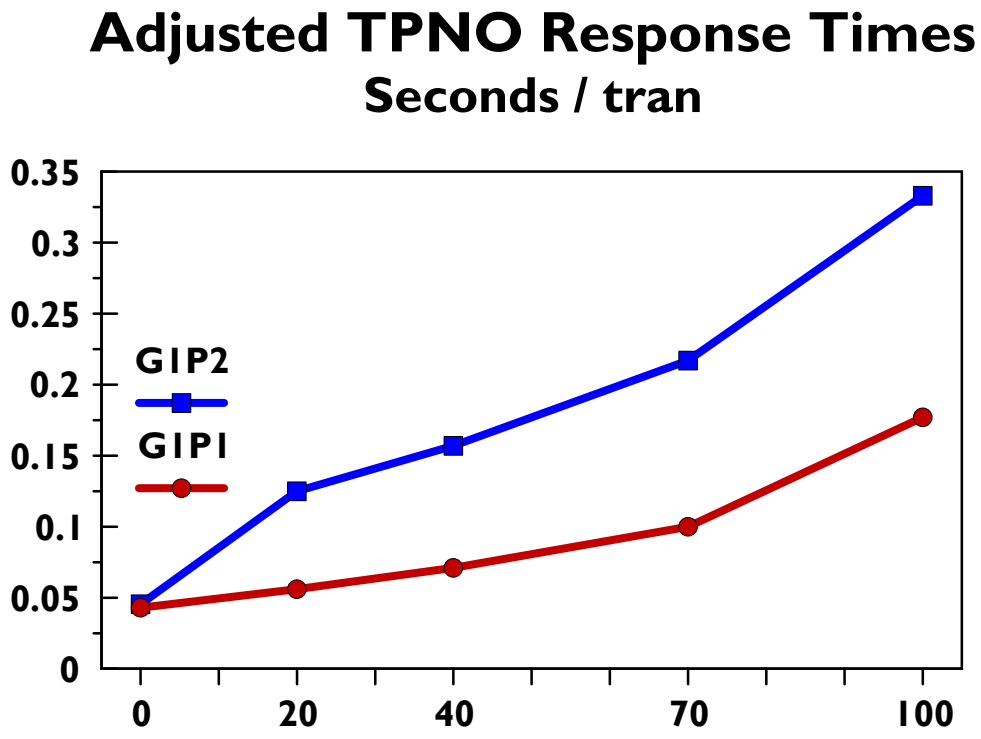
CONCLUSION

In this test case, we see the impact of the CF duplexing, on top of the PPRC impact. As we use the PAV for the disk, the I/O response time is more linear than in the test case 1. The CF duplexing impacts directly the transaction response time. We can estimate the 'cost', in term of transaction response time, of the 'CF enhanced recovery support" configuration.

Environment 3 - Multi-Site Workload

This simulates a multi-site workload environment. LPAR G1P1 is on Site1, with LPAR G1P2 on Site2. Site1 also has the primary disk and CF1 holding the LOCK, SCA, and GBP(P) structures. Site2 has the secondary disk with CF2 holding the GBP(S) structure. The GDPS controlling system is in Site2.

This test case corresponds to the GDPS Multiple Side workload. As seen from G1P1, the configuration is the same as for the environment 2. But, as the G1P2 LPAR is running on the second site, each access (disk and CF, read and write) will be impacted by the distance between the two sites.



*In this environment, there was a bottleneck on the CF subchannels. The response times shows are adjusted to remove the effects of this bottleneck. The methodology used is shown in **Appendix A.**

I/O for G1P1 DBxMSTR address space:

MSTR	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	41.4	42.6	41.4	40	37.4
RESP	1	1.5	1.7	1.9	2.2
CONN	0.3	0.4	0.3	0.3	0.4
DISC	0.1	0.5	0.7	1	1.3
Q+PEND	0.6	0.6	0.6	0.6	0.6
IOSQ	0	0	0	0	0

Note that in this measurements, Q+PEND time was observed (contention with an other system). But these I/Os are asynchronous and therefore do not affect transactions response time.

I/O for G1P2 DBxMSTR address space:

MSTR	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	44.1	41.9	41.1	39.7	34.9
RESP	0.5	1.4	1.7	2.4	2.9
CONN	0.2	0.4	0.4	0.4	0.4
DISC	0.1	0.6	0.8	1.1	1.3
Q+PEND	0.1	0.4	0.6	0.9	1.2
IOSQ	0	0	0	0	0

Here, we notice a double effect Q+PEND and DISC. Q+PEND represents the impact of distance for ‘normal’ I/O operation between the P2 system and the DASD in SITE1, and DISC time represents the PPRC effect.

I/O for G1P1 DBxDBM1 address space:

DBM1	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	14.8	21.1	22.6	16.5	15.5
RESP	1.5	3	2.8	3.8	4.4
CONN	0.6	0.8	0.7	0.8	0.8
DISC	0.4	1.5	1.5	2.5	3
Q+PEND	0.6	0.6	0.6	0.6	0.6
IOSQ	0	0	0	0	0

I/O for G1P2 DBxDBM1 address space:

DBM1	pprc 0 km	20 km	40 km	70 km	100 km
SSCH Rate	197.5	199.9	196.3	177	132.9
RESP	0.8	1.9	2.4	3.3	4.6
CONN	0.4	0.5	0.5	0.5	0.6
DISC	0.2	1	1.3	1.9	2.8
Q+PEND	0.2	0.4	0.6	0.9	1.2
IOSQ	0	0	0	0	0

The response time here increases significantly. The DISC time and Q+PEND are increasing linearly too. Unfortunately, the castout is managed by this DB2 member, as we see a large number of I/Os.

For G1P2, the distance has an impact on Q+PENDing time on the I/Os for CICS, MSTR and DBM1 address spaces (almost the same values for all the Address Spaces).

Adjusted Response times for TPNO transactions:*

G1P1:

TPNO	pprc 0 km	20 km	40 km	70 km	100 km
Number	3388	3393	3404	3381	3321
resp. time*	.043	.056	.071	.100	.177

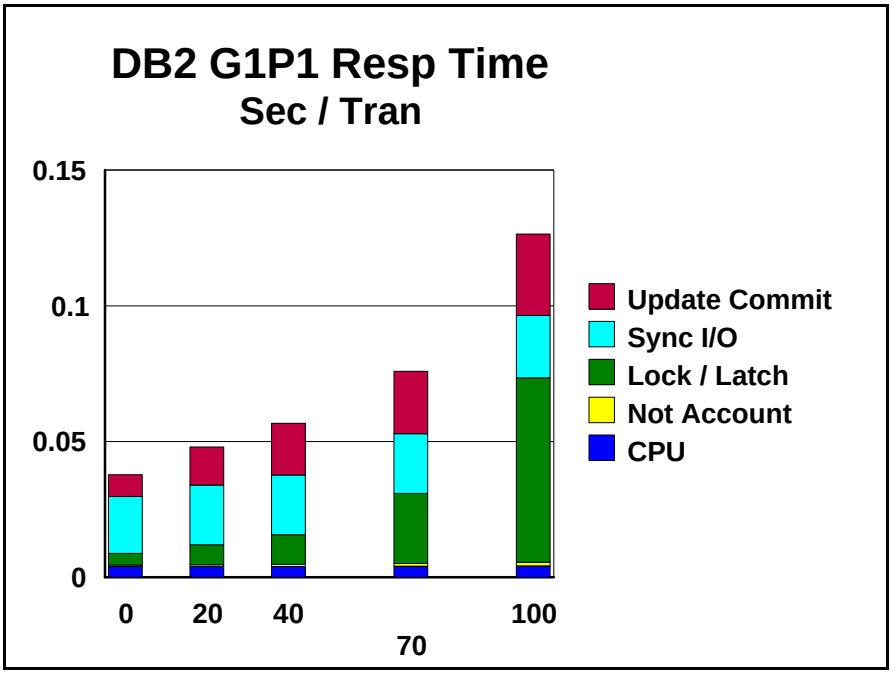
G1P2:

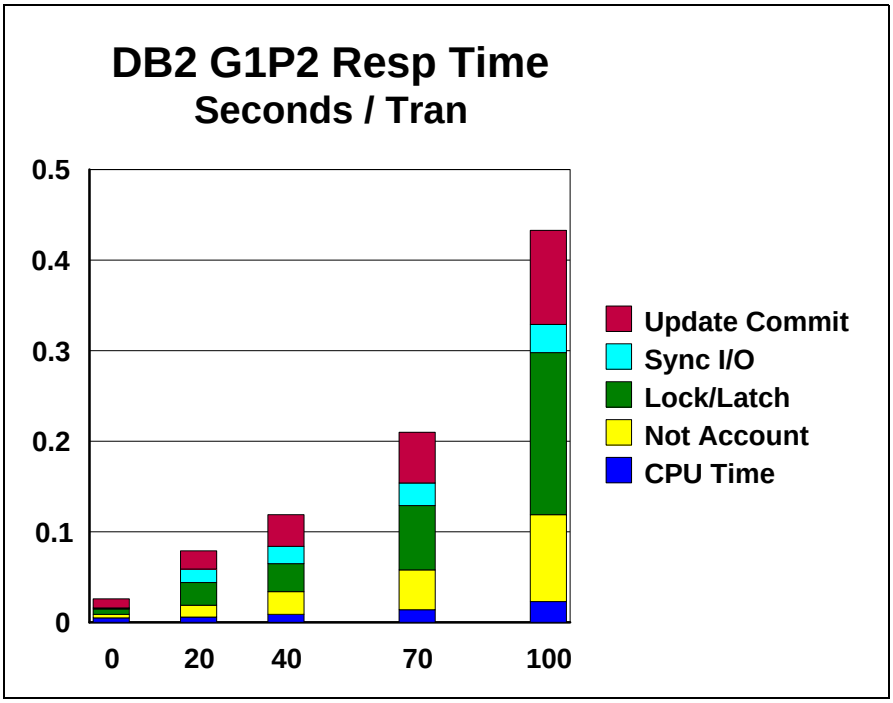
TPNO	pprc 0 km	20 km	40 km	70 km	100 km
Number	3394	3361	3322	3241	3079
resp. time*	.045	.125	.157	.217	.333

For G1P2, the primary cause of the response time increase over G1P1 is that all of the I/O and CF requests must traverse the distance back and forth. At a cost of .01 ms/request, this adds to 1 ms for each round trip. A FICON I/O has two protocol exchanges, plus one needs to add the delays for CF lock and GBP requests. This by itself would cause the elongation in transaction response time of about 86 ms for CF propagation delay, and another 35 ms for I/O delays, giving 121 ms growth, just due to distance propagation delays.

In addition, the secondary effects of other transactions waiting for resources starts to become felt at the longer distances.

An analysis of where the delays come from confirm this. For both G1P1 and G2P2 on Site1 and Site2, DB2/PM reports the largest contributor of response time growth outside of CF delays is waiting for Lock/Latch. Note that the scale between the two graphs below are different.


DB2 G1P1 Resp Time
Sec / Tran
0.15
0.1
0.05
0
0
20
40
70
100
Update Commit
Sync I/O
Lock / Latch
Not Account
CPU



AVERAGE TIME UPDATE COMMIT - The accumulated wait time due to a synchronous execution unit switch for DB2 commit, abort, or deallocation processing. This value is an average.

AVERAGE TIME SYNCHRONOUS I/O - The I/O elapsed time accumulated due to synchronous I/O suspensions. DB2 calculates this value by subtracting the store clock time when an agent begins waiting for a synchronous I/O from the time the agent is resumed.

AVERAGE TIME LOCK/LATCH (DB2+IRLM) - The accumulated lock and latch elapsed time. It indicates the elapsed time the allied agent waited for locks and latches in DB2. This value does not include suspensions due to group-level lock contentions in a data sharing environment. When the event completes, the ending time is used to calculate the total elapsed wait time.

DB2 NOT ACCOUNT - CF requests getting changed from synch to asynch because of high CF utilization.

The main contributors is Lock/latch contention. The impact is exponential, as we suffer from a cascading effect. The time needed to get the lock from P2 is increasing with the distance,

as the time to read/write the data itself. Therefore, total duration of this lock kept by P2 is increasing dramatically. As a direct effect, the data isn't accessible to P1 during that period. This increases the value of the global contention.

The synchronous I/O effect isn't very important, this represents the time needed to read data from DASD in SITE1 from P2.

Update commit

The UPDATE COMMIT portion is due to the Group buffer pool duplexing effect (as seen in test case 2) AND, a new effect: as the Pri group buffer pool structure is located in SITE1, the delay to access it from P2 is bigger, and therefore, the Synchronous access are converted to asynchronous.

Contributors to DB2 NOT ACCOUNT:

- *CF requests getting changed from synch to asynch because of high CF utilization (V5 and earlier)*
- *Synch write to secondary GBP does not complete before synch write to primary GBP with GBP duplexing.*

CPU TIME:

On the response breakdown graphs, the CPU time increase isn't very sensible. But we've tried to deeper check what was the effect of distance on this parameter.

Only G1P2 CPU time is increasing:

Impact on global CPU consumption:

We notice that the LPAR busy percentage is more than two times the busy time of the 0 km measurement!

A deeper analysis gives the following data.

CPU TIME - Time spent waiting for CPU. For G1P2 at 0 km, RMF™ report shows:

```
CICSERV   APPL %       12.3
DB2DBM    APPL %        3.7
DB2MSTR   APPL %        1.6
SYSSTC    APPL %        2.6
SYSTEM    APPL %        2.3

TOTAL     APPL %       23.3
```

For G1P2 100 km , RMF report shows:

```
CICSERV   APPL %       34.2
DB2DBM    APPL %        2.9
DB2MSTR   APPL %        5.1
SYSSTC    APPL %        7.9
SYSTEM    APPL %        7.3

TOTAL     APPL %       58.5
```

CICSERV is the CICS address space. These values are expressed in percentage of one processor

The CPU time didn't have any effect on our transaction response time because our system wasn't CPU constrained. But we see that we use 23% of a CP at 0 km, but up to 58% of a CP at 100 km.

The main contributors are the Global contention and Lock/latch. The impact is exponential, as we suffer from a cascading effect. The time needed to get the lock from G1P2 is increasing with the distance, as the time to read/write the data itself. Therefore, total duration of this lock kept by G1P2 is increasing dramatically. As a direct effect, the data isn't accessible to G1P1 during that period. This increases the value of the global contention.

The UPDATE COMMIT portion is due to two reasons:

1. the Group buffer pool duplexing effect as seen in Environment 2
2. since the Primary group buffer pool structure is located in Site1, the delay to access it from G1P2 is bigger, and therefore, synchronous access are converted to asynchronous (since z/OS V1.2).

The CPU time didn't have any effect on our transaction response time because our system wasn't CPU constrained. But we see that we use 23% of a CP at 0 km (with eight Logical CPs on the LPAR, this results in about 4% CPU Busy), but up to 58% of a CP at 100 Km (or about 7% CPU Busy).

GBP8:

GBP8 prim	0 km		20 km		40 km		70 km		100 km	
	nb (k)	st (µs)	nb (k)	st (µs)	nb (k)	st (µs)	nb (k)	st (µs)	nb (k)	st (µs)
G1P1 s	99	11.7	81	17	76	19	77	20.5	71	21.2
G1P1 a			18	328	23	373	26	393	28	422
G1P2 s	21	46								
G1P2 a	199	332	221	517	213	731	204	954	186	1249
Converted*			56		97		125		158	

(*) number of synchronous requests converted to asynchronous (included in asynch. number).

GBP8 secondary:

GBP8 sec.	00 km		20 km		40 km		70 km		100 km	
	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)	n (k)	rt (µs)
G1P1 s										
G1P1 a	33	205	33	493	34	736	34	984	33	1335
G1P2 s										
G1P2 a	83	100	81	97	78	97	75	113	61	142

DSNDBR0_LOCK1

LOCK1	0 km		20 km		40 km		70 km		100 km	
	nb (k)	st (µs)	nb (k)	st (µs)	nb (k)	st (µs)	nb (k)	st (µs)	nb (k)	st (µs)
G1P1 s	196	7.6	199	7.6	210	7.3	206	7.6	216	7.4
G1P1 a										
G1P2 s	190	26.1								
G1P2 a			194	460	201	680	184	933	172	1221

The 'automatic' synchronous to asynchronous CF access has a direct impact on the UPDATE COMMIT time, but, also on the 'not account' values.

CF1 contains SCA, LOCK and primary Group Buffer Pools structures.

CF2 contains secondary (duplexed) Group Buffer Pools structures.

This test case corresponds to the GDPS Multiple Site workload. Seen from P1, the configuration is the same as for the test case 2. But, as the P2 is running on the second Site, each access (DASD and CF, read and write) will be impacted by the distance between the 2 sites.

I/O for CICS address space:

First measure: We'll have to take into consideration the read IO operations because of SITE1 dasd remotely accessed by G1P2 – these are the synchronous reads of the database.

CICS P2	0 km	20 km	40 km	70 km	100 km
SSCH Rate	307.7	308.1	314.1	296.6	288.1
RESP	0.3	0.6	0.8	1.1	1.4
CONN	0.2	0.2	0.2	0.2	0.2
DISC	0	0	0	0	0
Q+PEND	0.1	0.4	0.6	0.8	1.2
IOSQ	0	0	0	0	0

As we can see, those I/Os being read only, we are only affected once by the distance (From Site2 to Site1). The value of the Q + PEND represents this access time elongation.

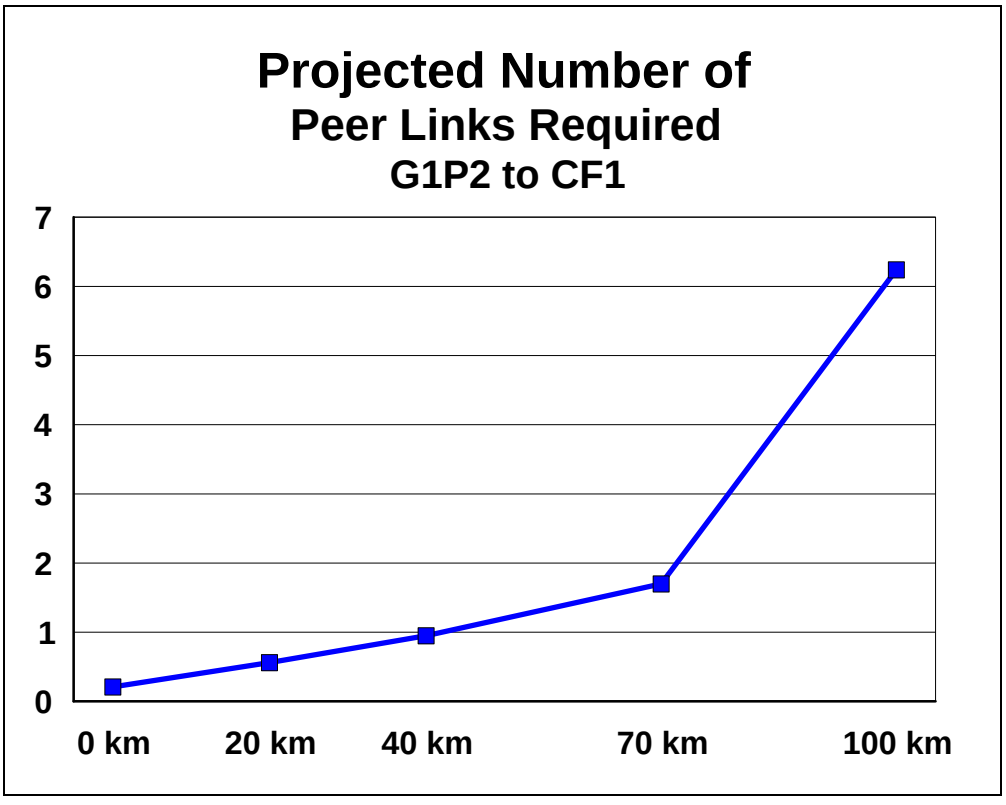
We expect the write I/Os to be affected twice by the distance: Q+PEND for the normal I/O (site 2 to site 1): this time is growing with distance and DISC time for the PPRC elongation.

From G1P2, for CICS, MSTR and DBM1 I/Os, distance impact is Q+PENDing time (almost strictly the same values for all the Address Spaces).

CF Link Utilization

A CF link subchannel is allocated for the life of the CF request. With two subchannels, there can be two simultaneous CF requests from each z/OS to a CF. Peer links support seven subchannels per link, supporting up to seven simultaneous CF operations.

The subchannel is allocated for the life of the CF operation. If a message is sent to a CF, the subchannel gets allocated, the message is placed on the link, travels across the fiber cable to the remote CF, and into the CF's coupling link. The Coupling Facility processes the request and sends a response and the processed is reversed. The link is busy only as bits travel across the fiber cable, but the subchannel is allocated (busy) for the duration of the entire process. If any structures are duplexed using System-Managed CF Duplexing, response times will be longer as there are protocol exchanges between the CFs before the final response is sent back to the sender. In the original measurements, there was an insufficient amount of links configured, so many CF requests were delayed, waiting for the CF subchannel. Based upon the "Total Requests Delayed", the number of links that should have been configured is shown in the graph below.



The calculations to obtain these projections are described in Appendix B.

Conclusion (Environment 3)

The initial hypothesis that response time grows linearly with distance was only partially correct. This was true for environment (1) where only the PPRCed disk is on Site 2, and response times also stayed "low". Environment (2) with DB2's GBP(S) structure also in Site 2 had more messages that needed to cross the separation, thus affecting response time more. This started to cause secondary effects and the response time curve is starting to show an exponential shape. This effect is magnified greatly with environment (3) in a multi-site workload, especially with the transactions running on Site 2 itself. These secondary effects include transactions holding on to resources longer such as database locks, delaying other transactions that are waiting for these resources.

Although in a real environment CP/SM with the z/OS Workload Manager would try to direct the transactions to where they could get better responses, the curve would still be there, although less steep.

Because of the complexities of the different types of I/Os and the different database designs and transaction accesses that influence the secondary effects, it is recommended that a benchmark be performed for each environment, although this document can be used as a guide on what to look for.

Summary

Elongation of disk response time due to PPRC wasn't a problem for the test OLTP application. This is mainly due to the nature of the workload, as most of the write I/Os are done by DB2 and are asynchronous. The G1P2 asynchronous write I/Os response time are multiplied by a factor of about six, but as it's asynchronous, that doesn't have a huge impact.

On the other hand, the longer distance between the production systems and the CF has a huge effect. After 40 km, the transaction response time degradation is more important.

In a real environment, the CPU consumption should be closely checked.

In this test case, the DB2 member in the G1P2 system was in charge of the castout. This is bad luck as it has to access both the Primary CF GBP structures remotely, and access the Primary disk remotely too. The results would have probably been better if the DB2 on G1P1 had been the castout owner.

In real life, CICSPlex/System Manager (CP/SM) would probably be routing most of the transactions to the CICS region with the best response time.

The OLTP workload that we used didn't require high GRS activity for data set serialization. The access to the ISGLOCK structure (remote for the G1P2 system) didn't impact our results. In batch workloads this effect could have been much more important. This shows again that pure analysis isn't valid enough to estimate what the impact of a large distance Multiple Site Workload configuration would be.

Appendix A - Environment 3 Adjustments

The results for Environment 3 (Multi-Site Workload) as presented is based upon projections of actual measurements. In the real measurements, we were limited by:

- *Lack of CF links spanning the sites. Where a normal Parallel Sysplex configuration has two CF links even in a normal single-site configuration to provide redundancy, the test environment only had a single CF link.*
- *Lack of availability of current DWDMs. Current DWDMs are able to support CF Peer links with seven subchannels each. The hardware available to the test team was only able to support two subchannels per CF link (as with pre-Peer Mode links).*

This gave us only two subchannels from each z/OS to each Coupling Facility.

A CF link subchannel is allocated for the life of the CF request. With two subchannels, there can be only two simultaneous CF requests from each z/OS to a CF. Most of the time, the CF CHANNEL has a very low utilization while transactions were waiting for the SUBCHANNEL. This artificially skewed the initial results, especially at longer distances when the response times were longer. To project what the response times would be had we the proper hardware, we needed to calculate the time delays waiting for the subchannels for Lock and Cache (GBP) structures, then subtract this from the measured response times. A significant part of the GBP accesses was due to castout processing. We also had to figure out what percent of the CF accesses was castout related, and ignore those requests in the calculations. In addition, since TPNO transaction accounted to about 78% of all GBP accesses, we needed to factor that in as well.

The methodology used to project response times for G1P1 and G1P2 in the multi-system configuration (Environment 3) is shown below. All this is based upon RMF CF Report data as well as DB2 Performance reports. We assume that transaction response time is based upon:

- *CPU*
- *I/O service time*
- *CF service time (GBP and Lock requests)*
- *Delays waiting for CF link subchannel*
- *Delays waiting for database locks, I/O contention, etc.*

Note that while the RMF report shows CF response times on a CF request basis, we need to normalize this on a transaction basis.

1. Multiply the total number of GBP (cache) delays by their duration, multiply by 0.78 (TPNO accounts for 78% of all GBP accesses), then divide by the number of transactions to get the average GBP delays/Tran

SUBCHANNEL ACTIVITY

SYSTEM NAME	£ REQ TOTAL AVG/SEC	TYPE	LINKS GEN	LINKS USE	PTH BUSY		REQUESTS £ REQ	-SERVICE TIME(MIC)- AVG	STD_DEV		DELAYED REQUESTS £ REQ	% OF REQ	AVG TIME(MIC) /DEL	AVG TIME(MIC) /ALL
G1P1	403643	ICP	1	1	0	SYNC	355285	18.7	66.9	LIST/CACHE	0	0.0	0.0	0.0
	1345.5	SUBCH	7	7		ASYNC	51549	454.6	347.5	LOCK	0	0.0	0.0	0.0
						CHANGED	0	INCLUDED IN ASYNC		TOTAL	0	0.0		
						UNSUCC	0	0.0	0.0					
G1P2	534859	CFS	2	1	0	SYNC	1525	477.5	23.0	LIST/CACHE	**165K**	50.1	**1091**	546.5
	1782.9	SUBCH	4	2		ASYNC	529128	709.4	261.4	LOCK	102K	50.6	234.1	118.4
						CHANGED	166916	INCLUDED IN ASYNC		TOTAL	267K	50.3		

In this case, for G1P2 it is 160,000 * 4028 uSec / 3079 trans = 209 ms/tran

2. The same thing is done for Lock requests.
3. Service time per transaction for GBP requests were calculated in a similar manor: (Service time) * (# requests) * 0.78 / (# trans)
4. Service time per transaction for Lock structure requests were calculated using the same formula.
5. I/O time and CPU time for each transaction was obtained from DB2 accounting performance reports. These non-CF related response time components were totalled.
6. The sum of I/O and CPU time (step 5) is subtracted from the observed response time, giving just the CF + "other" component.
7. The sum of the GBP8 and Lock delays and service times are added, giving the measured CF component of response time.
8. The results from (6) and (7) do not match. Much of this difference is reported by DB2/PM as Lock/Latch contention, and Update Commit. The ratio between the results from step 6 and step 7 was calculated. This ratio stays fairly consistent, reflecting the delays for database locks, I/O contention, etc.
9. The CF Service Times (sum of steps 3 and 4) is multiplied by the ratio (step 8) and added to the transaction I/O and CPU time to give the projected transaction response times if there was no CF subchannel contention.

The results from step 9 are used to generate the charts for “Environment 3” for the non-castout owner.

Because of the hardware configuration, the actual Delayed Requests as reported by RMF in Environment 3 was:

% Delayed Requests to CF1	G1P1	G1P2
0 km	0	4.4
20 km	0	31.1
40 km	0	50.3
70 km	0	67.7
100 km	0	90

% Delayed Requests to CF2	G1P1	G1P2
0 km	21.7	0
20 km	14.9	0
40 km	16.5	0
70 km	22	0
100 km	29	0

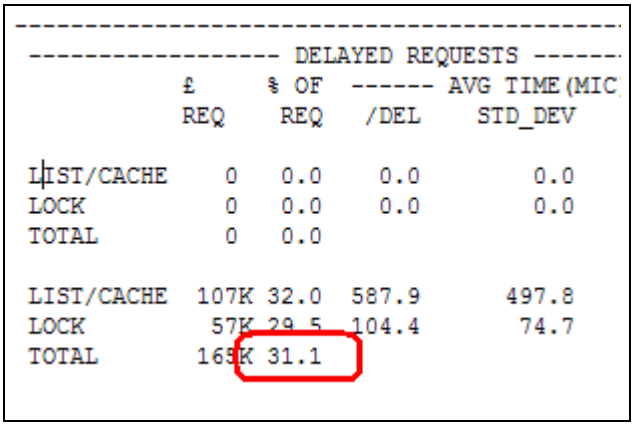
For the castout owner, we need to again determine the amount of Lock and GBP requests/transaction. This is done by using the same number requests/tran as in the non-castout owner, then apply the same methodology. For example:

System	G1P1	G1P2
Transactions	3,321	3,079
Lock	218,000	173,000
GBP	177,960	283,381
Locks / Tran	65.6	56.2
GBP / Tran	53.6	92.1
non-castout GBP/tran	53.6	53.6
Adjusted GBP requests		165,034

Appendix B - Subchannels Required

The amount of subchannels required for Environment 3 (Multi-Site Workload) is based upon projections of actual measurements.

The Subchannel Activity contains the Total % of Requests delayed. In the example below, this is 31.1%. The rule of thumb is to keep this below 10%.



	DELAYED REQUESTS			
	£ REQ	% OF REQ	------ AVG TIME(MIC /DEL	STD_DEV
LIST/CACHE	0	0.0	0.0	0.0
LOCK	0	0.0	0.0	0.0
TOTAL	0	0.0		
LIST/CACHE	107K	32.0	587.9	497.8
LOCK	57K	29.5	104.4	74.7
TOTAL	165K	31.1		

Since the experiments were done with only two, instead of the normal seven subchannels per link, and with only a single subchannel configured, the delay shown needs to be projected for 7 subchannels, then calculate how many links are needed to get the number below 10%.

1. The first step was to convert this to a number if there were only a single subchannel. For example, if one subchannel was 55% busy, the chance that two subchannels were both busy is 0.55 * 0.55 = .303. Similarly, the chance of all seven subchannels (in one link) being busy is .55 ** 7 = .01522, or about 1.5% busy.
2. Under this assumption,
   sch = amount of subchannels required
   del = Measured delayed requests
   sqrdel = SQRT(del) = delayed requests if there was a single subchannel
   sqrdel sch = 10%
3. Taking the Log of both sides of the equation:

   sch * LOG(sqrdel) = LOG(0.1)
   sch * LOG(sqrdel) = -1

   sch = -1 / LOG(sqrdel)
4. In this case, we would need -1 / LOG(.558) = -1 / (- 0.23) = 4.3 subchannels, or 0.62 links to get the % delayed under the 10% recommendation.

IBM Corporation
New Orchard Road
Armonk, N.Y. 10504

Produced in the United States of America
05/06

ZSW03119-USEN-00