yourdotcom

International Technical Support Organization and Authoring Services

2011 Parallel Sysplex and High Availability Update

ibm.com/redbooks

Frank Kyne, ITSO Poughkeepsie

Redbooks Workshop

Welcome

Thank you for coming to this year's update.

My background.....

My funny accent - please tell me when I start going too fast!

Questions?? PLEASE ask as I go along.

Please complete the evaluation forms.

- ESPECIALLY, PLEASE tell me if there is anything you would have liked to hear about that I didn't cover.

Want to apologize NOW for differences between your handouts and these slides - if you would like the latest PDF, please indicate so on your evaluation forms and I will email you the latest version

Redbooks Workshop

Acknowledgements

Thanks to Dave Petersen for GDPS/Active-Active material and help.

Thanks to Noshir Dhondy for STP material and help.

Thanks to residents of InfiniBand residency for all their hard work:

- Hua Bin Chu
- George Handera (Aetna)
- Marek Liedel
- Masaya Nakagawa
- Iain Neville
- Christian Zass

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Topics

Recap of sysplex enhancements in z/OS 1.12

What's new in z/OS 1.13 for sysplex

The latest addition to the GDPS family - GDPS/Active-Active

NEW

The latest and greatest in CF coupling connectivity

Latest enhancements in the area of IPL avoidance and MTTR

Miscellaneous

Redbooks Workshop

Timetable

Start	**09:00 (ish)**
Break	**10:30-10:45**
Lunch	**12:00-13:00**
Afternoon break	**14:30-14:45**
Finish	**17:00 (ish)**

yourdotcom

International Technical Support Organization and Authoring Services

z/OS 1.12 sysplex enhancements recap

ibm.com/redbooks

Sysplex Enhancements in z/OS 1.12

Enhancements to XCF exploiter monitoring and additional automatic actions for "Critical" XCF members

New capabilities in Sysplex Failure Management for hung structure processes

Enhancements to the REALLOCATE command

Sublist notification delay enhancement

Additional XCF FUNCTIONS parameters

New sysplex-related health checks

System Logger enhancements

Enhanced monitoring and Critical member support

Added the ability for an XCF exploiter to ask XCF to monitor a status field (that is maintained by the exploiter) and (optionally) call an exploiter-provided user exit in case the status is not updated:

- XCF monitors the contents of the identified status field.
- Exploiter can also provide an exit that can examine the address space (using the exploiter's knowledge of their own programs) and return a status to XCF - GRS exploits this capability.
- New messages IXC633I, IXC634I, IXC635E, and IXC636I in support of this monitoring.
  - Make sure you add these messages to your automation.

This monitoring potentially applies to any XCF exploiter

Enhanced monitoring and Critical member support

In *addition* to the extra monitoring (which can be used by ANY XCF exploiter), exploiters that provide crucial system services (like GRS) can:

- Identify themselves to XCF as being CRITICAL (note that it is the exploiter, not the customer, that makes this determination):
  - As part of the process of declaring themselves as CRITICAL=YES, the exploiter tells XCF "If I become unresponsive, this is what you should do". Options are to terminate the task, jobstep, address space, or the entire system.
  - This causes XCF to do extra monitoring (in addition to the new monitoring already described above).
  - Additionally, before it takes any terminating action, XCF will take a dump, attempting to include any address space that might be involved in the problem.
    - This is an attempt to ensure that IBM has all the information we need to debug the problem, avoiding the need to recreate the problem.

Neither of these enhancements are dependant on having an active SFM policy.

Enhanced SFM support for hung structures

As part of certain structure-related actions, all connectors to the structure are notified. The action cannot proceed until all connectors respond.

Prior to z/OS 1.12, if a structure connector did not respond to those requests within two minutes, messages IXL040E and IXL041E would be issued.

- However, if the operator didn't see the message, or didn't know how to react, the system would not take any further action, and the action would remain stalled.

Enhanced SFM support for hung structures

In z/OS 1.12, SFM was enhanced to give you the option to have SFM take action against a hung connector.

- SFM is enhanced to take the following recovery actions in an attempt to resolve the hang (depending on the type of problem):
  - Stopping the rebuild
  - Stopping signaling path (XCF signaling Structures only)
  - Forcing a disconnect (XCF signaling Structures only)
  - Terminating the connector task
  - Terminating the connector address space
  - Partitioning the connector system

Enabled by specifying new CFSTRHANGTIME keyword in SFM policy.

- Specification is at the system, not the individual structure, level.
- Recommended values are between 15 and 20 minutes.

XCF REALLOCATE command enhancements

z/OS 1.12 provided enhancements to the REALLOCATE capability.

DISPLAY XCF,REALLOCATE,TEST command simulates the actions that a SETXCF START,REALLOCATE command would take and shows you what the result *would* be before you actually do it.

- STRONGLY recommend using this command in preparation every time you are going to do a REALLOCATE

DISPLAY XCF,REALLOCATE,REPORT command provides a report, showing what action (or no action) was taken against each allocated structure the last time a REALLOCATE was run.

And remember, always use MAINTMODE to make CF outage management easier.

XCF REALLOCATE command enhancements

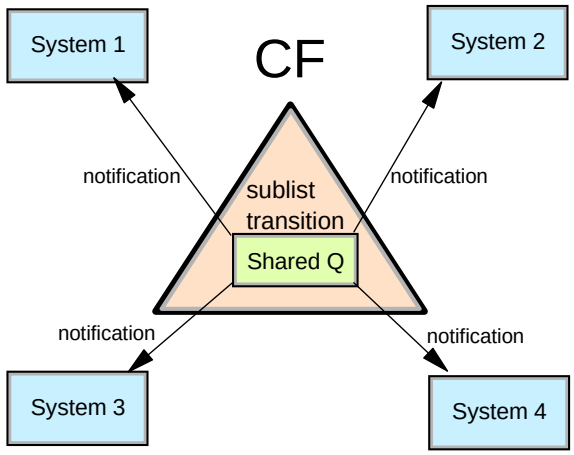
Sample D XCF,REALLOCATE,REPORT output:

```
D XCF,REALLOCATE,REPORT
IXC347I  15.05.43  DISPLAY XCF 450

THE REALLOCATE PROCESS STARTED ON 09/15/2011 AT 12:55:02.69.
THE REALLOCATE PROCESS ENDED ON 09/15/2011 AT 12:55:04.02.
------------------------------------------------------------------
STRUCTURE(S) WITH AN ERROR/EXCEPTION CONDITION

NONE
------------------------------------------------------------------
STRUCTURE(S) WITH A WARNING CONDITION

NONE
------------------------------------------------------------------
STRUCTURE(S) REALLOCATED SUCCESSFULLY

STRNAME: THRLCKDB2_1                              INDEX: 128
   1 REALLOCATE STEP(S): REBUILD
   COMPLETED ON SYSTEM #@$2 ON 09/15/2011 AT 12:55:03.91.
...
```


ibm.com yourdotcom International Technical Support Organization and Authoring Services

Sublist Notification Delay enhancement

The original implementation of shared message queues (used by MQ and IMS) was that *every* interested system would be informed every time a monitored list (or sublist) went non-empty.



Sublist Notification Delay enhancement

Why EVERY system?

Because the objective of shared queues is that work should flow to whichever system has the most spare capacity (and therefore should be able to deliver the best performance).

Over time, as that system pulls more work, it will get busier and its reaction time will increase, meaning that more messages will be retrieved by the other systems.

NOTE - The objective of Shared Queues is NOT to balance the transactions across all systems - it is to have the routing of transactions reflect the available capacity and performance of the members of the sysplex.

- If you had two systems, one with 10,000 MIPS and one with 5000 MIPS, would you want both of them to be sent the same number of transactions?

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Sublist Notification Delay enhancement

The downside of this implementation, especially for IMS, is that IMS has to do some amount of work before it decides whether it wants to retrieve a new message or not. This consumes some amount of capacity (and is reported by IMS as False Scheduling). And each message will only be retrieved by one system, so the more systems you have, the more False Scheduling you are likely to observe.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Sublist Notification Delay enhancement

To try to reduce this cost, CF Level 16 included an enhancement known as Sublist Notification Delay.

This changed the algorithm, so that instead of informing all interested parties, the CF will tell just one system. It then gives that system 5000 mics to retrieve the message.

- If the message is retrieved, there is no need to inform any of the other interested parties. This largely eliminates False Scheduling caused by CF notifications.
- If the message is NOT retrieved within 5000 mics (possibly because that system is having a problem or is over-utilized), all other interested parties will be informed.

Additionally, when the next message arrives, the next system in the list will be the first one to be informed.

- This effectively moves exploiters of the sublist notification process from using pull-based workload distribution to a round-robin mechanism.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Sublist Notification Delay enhancement

While some customers appreciated the decline in False Scheduling, others preferred the original pull-based workload distribution.

To accomodate THAT set of customers, z/OS 1.12 included the ability to control, at the structure level, how long the CF waits between informing the first interested party, and the others.

- New keyword is SUBNOTIFYDELAY in CFRM policy (this is specified at the structure level).
  - Default is 5000 mics unless you override it.

Smaller values tend to move the behavior back towards pull-based distribution. Larger values move more towards round-robin behavior (where every system gets the same number of transactions).

Note that, for most request types, MQ does NOT use sublist notification. This means that MQ Shared Queues still uses the original pull-based mechanism, and specifying a SUBNOTIFYDELAY for an MQ Shared Queues structure will probably have no noticeable impact.

Redbooks Workshop

Support for CF Level 17 nondisruptive dump capability

CF Level 17 provides a new capability to take a CF dump in a non-disruptive way.

- Prior to CF Level 17, a serialized dump (which is required to debug some problems) would reset the CF.
- CF Level 17 added the ability to take a serialized dump without resetting the CF. This is likely to be most beneficial in debugging System-Managed Duplexing problems.
  - This capability was subsequently rolled back to CF Level 16 - Service Level 4.0x, and z/OS APAR OA33723

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Support for CF Level 17 nondisruptive dump capability

Nondisruptive dumps can be initiated by operator command (on the CFCC console), by XES, and by CF link hardware.

- You can control the ability to the link hardware to initiate a dump by using the SETXCF FUNCTIONS,ENABLE|DISABLE=DUPLEXCFDIAG command.
  - Turns this capability on or off AT THE SYSTEM LEVEL.
- By default, this capability is turned OFF.
  - Recommend not changing this unless requested to do so by IBM Service.

Be aware that installing the MCLs that deliver this capability will increase the storage required for CFCC code by about 250MB compared to what it required previously.

- When moving from CF Level 16 (without Service Level 4.x) to CF Level 17, the increase would be about 350MB.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

New function in support of HyperSwap

There have been some instances of a HyperSwap not completing successfully because some of the programs or control blocks required to complete the HyperSwap were paged out.

z/OS 1.12 added the ability to identify certain critical address spaces as ones that the system should try to avoid paging.

- HyperSwap-related programs will automatically identify themselves as CRITICALPAGING=YES
- Not a guarantee, but they will only be paged as a last resort.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

New function in support of HyperSwap

The function is controlled by specifying FUNCTIONS ENABLE|DISABLE(CRITICALPAGING) in COUPLExx member

- Function can be turned OFF dynamically using SETXCF FUNCTIONS command, but cannot be turned ON without an IPL.
- Function is controlled on a system-by-system basis.

Recommend that this is ENABLED on every system that uses HyperSwap

For more information, refer to:

- WSC Flash
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10733
- White Paper
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101800

Redbooks Workshop

New sysplex-related health checks

- Use of dedicated CF Processors
- CF memory utilization
- Verify that CF structure policy SIZE is not more than 2x INITSIZE
- Verify that MAXSYSTEM value in every CDS is at least as large as MAXSYSTEM value in sysplex CDS
- Verify that MSGBASED CFRM processing is enabled
- Check that SFM Structure hang time value is between 900 and 1200 seconds

System Logger enhancements in z/OS 1.12

- Automatic adjustment for incorrect SHAREOPTIONS on *new* Logger offload and staging data set (will not change SHAREOPTIONS for existing data sets)
- Ability to include IDCAMS LISTC output in IXCMIAPU report (new LISTCAT keyword)
- 4GB Logstream data set support (increased from 2GB)
- New messages about offload data set allocation

```
IXG283I OFFLOAD DATASET IXGLOGR.IFASMF.STRIPE.TYPDFLT.A0000109
ALLOCATED NEW FOR LOGSTREAM IFASMF.STRIPE.TYPDFLT
CISIZE=24K, SIZE=491520000
```

- Virtual Storage Constraint Relief - move .5MB of Logger code out of EPLPA into Logger address space

Reminder for z/OS R11 customers

If you have not already done so, PLEASE enable z/OS BCPii and System Status Detection Partitioning Protocol

- Have received many positive reviews
- Applies to both GDPS and non-GDPS customers
- If you have not set it up yet, include a comment on your evaluation form and I will send you some information about how to do it.

yourdotcom

International Technical Support Organization and Authoring Services

z/OS 1.13 sysplex enhancements

ibm.com/redbooks

Sysplex enhancements in z/OS 1.13

- Enhanced problem determination information in output from D XCF,S commands
- XCF and XES CTRACE default buffer sizes increased
- Ability to turn off ALTER processing for specified structures
- Reason why your structure did not go where you thought you wanted it to go is now presented in..... ENGLISH!!!

- Ability for a program to use XCF services without having to join an XCF group
- ARM Timeout value can now be overridden

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Sysplex enhancements in z/OS 1.13

- Option to have all members *of the sysplex* automatically notified if the volume serial or the location of a VTOC are changed
- MQ is no longer required for SDSF to SDSF communication in a JES2 MAS
- Various enhancements in System Logger

Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

Enhanced D XCF,Sysplex commands

There are many aspects to delivering a high availability service....

One is to have a configuration that is resilient and flexible, so that it doesn't have many planned or unplanned outages.

But another is that if a system DOES fail, you want to:

- Detect the failure as quickly as possible.
- Gather as much diagnostic information as possible, so the cause of the outage can be determined and addressed and it is NOT necessary to suffer another outage to gather the required diagnostic information.

The System Status Detection Partitioning Protocol introduced in z/OS 1.11 is VERY effective at identifying and removing failed systems quicker than we could ever do before.

And the Auto IPL feature (delivered in z/OS 1.10) gives you the ability to automatically take a standalone dump when a system fails.

But there are also improvements that can be made to the operator interface, to help them gather more information, more easily....

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM.

Enhanced D XCF,Sysplex commands

To provide additional information in one place, the D XCF,SYSPLEX and D XCF,SYSPLEX,ALL commands have been enhanced.

z/OS 1.12 output

```
D XCF,S
IXC334I 01.55.14  DISPLAY XCF 260
   SYSPLEX PLEX75:     SC74              SC75
```

z/OS 1.13 output

```
D XCF,S
IXC336I 11.53.19  DISPLAY XCF 935
SYSPLEX PLEX75
SYSTEM   TYPE SERIAL LPAR STATUS TIME         SYSTEM STATUS
SC74     2817 3BD5   05   05/10/2011 11:53:19 ACTIVE         TM=STP
SC75     2097 DE50   2C   05/10/2011 11:53:16 ACTIVE         TM=STP

SYSTEM STATUS DETECTION PARTITIONING PROTOCOL CONNECTION EXCEPTIONS:
  SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE SSD PROTOCOL
```

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Enhanced D XCF,Sysplex commands

D XCF,S,ALL command can be used to get a list of the systems in the sysplex, their status (from XCF's perspective), and the time they last updated their heartbeat.

z/OS 1.12 output

```
D XCF,S,ALL
IXC335I  01.58.29  DISPLAY XCF 262
SYSPLEX PLEX75
SYSTEM   TYPE SERIAL LPAR STATUS TIME         SYSTEM STATUS
SC74     2817 3BD5   05   05/11/2011 01:58:28 ACTIVE         TM=STP
SC75     2097 DE50   2C   05/11/2011 01:58:26 ACTIVE         TM=STP
```


ibm.com yourdotcom International Technical Support Organization and Authoring Services

Enhanced D XCF,Sysplex commands

However there is more information that XCF maintains about the members of the sysplex that was not included on the output from that command. The D XCF,S,ALL command was enhanced in z/OS 1.13 to provide some of that information.

z/OS 1.13 output

```
D XCF,S,ALL
IXC337I  11.57.32  DISPLAY XCF 957
SYSPLEX PLEX75           MODE: MULTISYSTEM-CAPABLE

 SYSTEM SC74          STATUS: ACTIVE
                      TIMING: STP CTNID: ITSOPOK
                STATUS TIME: 05/10/2011 11:57:32.042885
                  JOIN TIME: 04/29/2011 07:08:07.839841
              SYSTEM NUMBER: 010001DB
            SYSTEM IDENTIFIER: 3BD52817 050001DB
              NODE DESCRIPTOR: N/A
                              PARTITION: N/A    CPCID: N/A
                     RELEASE: N/A

 SYSTEM SC75          STATUS: ACTIVE
 ....
```

XCF and XES CTRACE

As the use of CFs (and the size of sysplexes) continues to increase, there have been cases where the XCF and XES component trace buffers were not large enough to hold all the information required to debug some problems.

Also, additional default tracing is turned on in XCF and XES in z/OS 1.13.

Prior to z/OS 1.13, XCF CTRACE buffer size defaulted to 1MB and XES buffer space was 168KB.

Starting with z/OS 1.13, defaults changed to 4MB and 336KB. However, to pick up the new values, ensure that BUFSIZE is not specified in CTIXCF00 and CTIXES00 members.

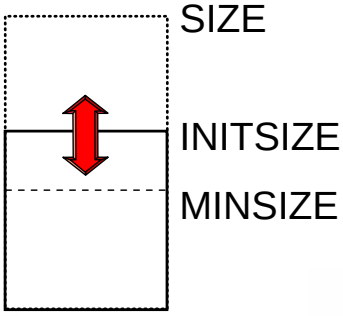
Generally, the BUFSIZE should only be changed if requested to do so by IBM Level 2 support.

- And remember to remove again after problem is resolved.

More granular control over structure ALTER

There are two ways to change the size of a CF structure:

- REBUILD the structure. However this quiesces access to the structure while the rebuild is taking place.
- ALTER the structure. This allows the size to be changed from the current size to any value between MINSIZE and SIZE WITHOUT quiescing access to the structure.
  - For this reason, the use of Alter is more flexible so it is generally preferred to the rebuild command if the objective is simply to change structure size.
  - In order to get the maximum benefit from ALTER, you should have a reasonable difference between SIZE and INITSIZE.



z/OS 1.12 introduced a healthcheck to ensure that SIZE is not >2x INITSIZE

More granular control over structure ALTER

There are three ways that a structure ALTER can be initiated:

- By the SETXCF START,ALTER,STRNM= operator command from console.
  - This command is limited to changing the overall size of the structure - you cannot use it to change the ratio of entries to elements, for example.
- Via Auto Alter (when ALLOWAUTOALT(YES) is specified in the structure definition in CFRM policy).
  - Automatically initiated by the system when it detects that some aspect of the structure has exceeded the FULLTHRESHOLD value.
    - Note that this is timer-driven. The threshold is not checked for every structure access.
  - Auto Alter IS able to change the ratio of objects in the structure as well as the overall size of the structure
- By the connected program issuing an IXLALTER command.
  - This can also change the structure size, and the allocation of objects in the structure.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

More granular control over structure ALTER

There are a few structures that do not support ALTER.

- RACF, JES2, GRS, some IMS

There are some structures that AUTO ALTER is not recommended for.

- Generally structures where the structure owner issues their own IXLALTER commands.
- For these structures, omit ALLOWAUTOALT keyword in CFRM policy.

There are some structures that AUTO ALTER IS recommended for.

AND there are some cases where ALTER would normally be fine, however due to special circumstances, the ALTER can run for a VERY long time, impacting performance during that time.

For those special cases, you now have the ability to disable ALL types of ALTER.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

More granular control over structure ALTER

Which structures does this apply to?

- Generally, follow the recommendations for which structures should be enabled for Auto Alter.
- Alter should only be disabled if you actually encounter a problem with long-running alters.
- How would I know?
  - Increased CF CPU utilization
  - Increased response time for some structure

If Auto Alter must be explicitly enabled in the CFRM policy, why not simply omit that keyword in the structure definitions?

- Because that won't stop a program-initiated Alter (IXLALTER)
- Also doesn't stop an operator issuing a SETXCF START,ALTER command...

ibm.com yourdotcom International Technical Support Organization and Authoring Services

More granular control over structure ALTER

New SETXCF MODIFY,STRNM=xxxx,ALTER=ENABLED|DISABLED command provides ability to ensure that NO Alter will run against the named structure, or to enable Alter again.

- Generic structure names (DUMMY_FRAN*) can be used.

```
D XCF,STR,STRNM=IXC_DEFAULT_1
IXC360I  11.47.00  DISPLAY XCF 487
STRNAME: IXC_DEFAULT_1
 STATUS: ALLOCATED
 EVENT MANAGEMENT: POLICY-BASED
 TYPE: LIST
 POLICY INFORMATION:
  POLICY SIZE    : 18 M
  POLICY INITSIZE: 9 M
  POLICY MINSIZE : 6912 K
  FULLTHRESHOLD  : 80
  ALLOWAUTOALT   : YES
  REBUILD PERCENT: N/A
  DUPLEX         : DISABLED
  ALLOWREALLOCATE: YES
  PREFERENCE LIST: FACIL03  FACIL04
  ENFORCEORDER   : NO
  EXCLUSION LIST IS EMPTY
```

```
SETXCF MODIFY,STRNM=IXC_DEFAULT*,ALTER=DISABLED
IXC556I SETXCF COMMAND COMPLETED: ALTER DISABLED FOR 3 STRUCTURE(S).
```

More granular control over structure ALTER

You can also get a list of the structures where Alter was disabled, by using the D XCF,STR,ALTER=DISABLED (or ENABLED) command.

```
D XCF,STR,STRNM=IXC_DEFAULT*,CONNM=ALL
IXC360I  11.54.17  DISPLAY XCF 487
STRNAME: IXC_DEFAULT_1
 STATUS: ALLOCATED
         START ALTER NOT PERMITTED
 EVENT MANAGEMENT: POLICY-BASED
 TYPE: LIST
 POLICY INFORMATION:
  POLICY SIZE    : 18 M
  POLICY INITSIZE: 9 M
  POLICY MINSIZE : 6912 K
  FULLTHRESHOLD  : 80
  ALLOWAUTOALT   : YES
  REBUILD PERCENT: N/A
  DUPLEX         : DISABLED
  ALLOWREALLOCATE: YES
  PREFERENCE LIST: FACIL03  FACIL04
  ENFORCEORDER   : NO
  EXCLUSION LIST IS EMPTY
```


ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM.

More granular control over structure ALTER

After the PTF has been applied, attempts to alter the structure will be rejected, and a console message will be issued

```
SETXCF START,ALTER,STRNM=IXC_DEFAULT_1,SIZE=50000
IXC531I SETXCF START ALTER REQUEST FOR STRUCTURE IXC_DEFAULT_1 489
REJECTED. REASON:
START ALTER NOT PERMITTED
```

Redbooks Workshop

More granular control over structure ALTER

New function is delivered (on top of z/OS 1/13) by APAR OA34579, available back to z/OS 1.10.

Command can be issued whether structure is allocated or not.

ALTER status is remembered even if structure gets deleted.

APAR must be installed on every system in the sysplex (via rolling IPL) in order to be fully effective.

- If fix is on SYSA, but not on SYSB, and ALTER is DISABLED on SYSA, an ALTER could still be initiated by SYSB.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Understandable structure placement messages!!

Did you ever find that you try to place a structure in a specific CF, but it won't move. AND, the message that you get is not quite as helpful as it might be?

```
STRNAME: IRRXCF00_B001                              INDEX: 67
   CFNAME     STATUS/FAILURE REASON
   --------   -----------------------------------------------------
   FACIL03    PREFERRED CF 1
                                 INFO110: 00000002 CC007800 00000010
   FACIL04    PREFERRED CF ALREADY SELECTED
                                 INFO110: 00000002 CC007800 00000011
```

Per Info APAR II14046, "The meaning of the 3rd fullword in the INFO110 are those attributes that the target CF didn't have that the current CF did. Thus the meanings are the exact opposite of those described for the 2nd fullword."

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Understandable structure placement messages!!

However, some strange people felt that this isn't obvious enough.... So, for those 2 customers, z/OS 1.13 provides the reasons in (even more) user-friendly text form....

- CONNECTIVITY REQUIREMENT MET BY PREFERRED CF:
- CFLEVEL REQUIREMENT MET BY PREFERRED CF
- FAILURE ISOLATION FOR DUPLEXING MET BY PREFERRED CF
- SPACE REQUIREMENT MET BY PREFERRED CF
- SPACE AVAILABLE FOR MINIMUM SIZE IN PREFERRED CF
- SPACE AVAILABLE FOR CHANGED DATA IN PREFERRED CF
- MORE SPACE AVAILABLE IN PREFERRED CF
- NON-VOLATILITY REQUIREMENT MET BY PREFERRED CF
- FAILURE ISOLATION REQUIREMENT MET BY PREFERRED CF
- STAND-ALONE REQUIREMENT MET BY PREFERRED CF
- EXCLLIST REQUIREMENT FULLY MET BY PREFERRED CF
- EXCLLIST REQUIREMENT MET BY PREFERRED CF
- Others.....

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Understandable structure placement messages

Messages IXL015I, IXC347I, and IXC574I are affected by this change.

New function *not* rolled back to previous releases.

Don't forget to use the D XCF,REALLOCATE,REPORT command to get a summary of the results of the last START,REALLOCATE command.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

New XCF Client/Server support

The traditional model of XCF signalling use was that all address spaces that wanted to talk to each other would join an XCF group. They could then ask XCF to monitor the members of the group, inform them when a new member joins the group, when an existing member leaves the group, and so on.

However the programming to use this is not trivial, and not all programs need all those capabilities.

- Some programs simply want to be able to easily communicate with other address spaces without having to worry about using TCP or SNA, or managing their own devices and their own recovery.

To help those potential exploiters, XCF in z/OS 1.13 provides a new model, known as Client/Server.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

New XCF Client/Server support

Programs can now register with XCF as server address spaces, providing a name that they can be addressed by.

- However they do NOT join an XCF group, so will not show up in the output from a D XCF,G command.
- But there IS a new D XCF,SERVER command

```
D XCF,SERVER
IXC395I  14.17.23  DISPLAY XCF 440
   SERVER NAME                          #INSTANCES
   ISFSRVR.SDSF                                 20
   SYSXCF.IXCREQ                                 2
```

Programs that wish to communicate with a server do not need to pre-connect to XCF in any way. They simply issue an IXCSEND command, naming the server(s) they wish the message to be sent to.

- Because they do not need to communicate with XCF prior to passing it the message, XCF has no prior knowledge of their existence, so there is no way to get information about them using a D XCF command.

Redbooks Workshop

New XCF Client/Server support

If they do not join an XCF group, how does their use of XCF resources get reported by RMF?

- All messages to and from a server program get reported by RMF in the SYSXCF group.
  - However, all the users of this service get grouped together and reported as one "member" (with the member name equal to the system name).
- There are no special considerations for transport classes. Assuming that you follow Best Practice guidelines and set up transport classes based on message size, these messages will automatically be assigned to the most appropriate transport class.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

New XCF Client/Server support

Only user of this service at the moment is SDSF.

For more information about SDSF use of this capability, refer to Paul Roger's z/OS BCP material.

The interface is documented, and a few customers have started looking at it for use by their own applications.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Automatic Restart Manager TIMEOUT control

If you have an active ARM policy, then:

- After a system failure, ARM waits up to two minutes for survivors to finish cleanup processing for the failed system
- If cleanup does not complete within two minutes, ARM proceeds to restart the failed work anyway

Problem: restart may fail if cleanup did not complete.

Issue: Two minutes may not be long enough for the applications to finish their cleanup processing, and you had no way to control how long ARM would wait for.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Automatic Restart Manager TIMEOUT control

CLEANUP_TIMEOUT

- New parameter for the ARM policy specifies the maximum number of seconds ARM should wait for survivors to cleanup for a failed system
- Specified in seconds, 120..86400 (2 min to 24 hours)

If parameter is <u>not</u> specified:

- TIMEOUT now defaults to 300 seconds (5 minutes, not 2)

If you want to continue to use the old TIMEOUT value, you must explicitly specify 120.

If you specify a value greater than 120:

- ARM issues msg IXC815I after 2 minutes to indicate that restart is being delayed
- If the timeout expires before cleanup has completed, ARM issues message IXC815I to indicate restart processing is continuing despite incomplete cleanup
- ARM restarts will commence as soon as cleanup is complete, if it completes in less than CLEANUP_TIMEOUT time.

Available for z/OS V1R10 and later with APAR OA35357

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Automated handling of VTOC and volser changes

Prior to z/OS 1.13, if you move a VTOC on a volume, or change the volser of a volume, you would need to do a VARY ONLINE UNCONDITIONAL on all other members of the sysplex in order to pick up the changes.

In z/OS 1.13, if DFSMSdss or DFSMShsm Fast Replication Backup and Recovery processing or ICKDSF REFORMAT NEWVTOC changes the volser or VTOC location AND ENABLE(REFUCB) is specified in the DEVSUPxx member, it is not necessary to issue the VARY commands.

- DSS and DSF use ENF to notify Device Support Services (DEVMAN address space) that the change has occurred.
  - If the device is online, DEVMAN will issue the VARY ONLINE UNCONDITIONAL.
  - If the device is offline, no action is taken.
  - Because XCF is used to transmit ENF signals, any systems outside the sysplex will *not* be made aware of the change.

For more info, see Paul's section on HSM Fast Replication

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM

Removal of need for MQ for SDSF MAS support

Prior to z/OS 1.13, certain SDSF functions required the use of MQ to obtain sysplex-wide information

Starting with 1.13, SDSF will use XCF by default instead of MQ for this communication

- However this requires that all systems in the sysplex are running z/OS 1.13 or later.

For more information, see Paul Roger's presentation (Part 17, slide 647)

Redbooks Workshop

Path busy conditions

Original design of Path Busy count (for CF requests) was that it was incremened by 1 every time all link buffers were found to be in use.

- So, if you have 1 CHPID (7 link buffers) and all are busy the first two times the system tries to find an available one, but one is available on the 3rd attempt, that would be counted as 2 path busy events.

However, as processors get faster, the same link buffer utilization will result in increasing path busy numbers (because the system can look over all buffers in less time).

- This can lead to concerns, because a processor upgrade may result in a higher path busy number, even though nothing has actually changed.

APAR OA35117 addresses this by changing the meaning of the Path Busy count.

- Now, the counter is incremented by 1 for every request that finds all link buffers busy, regardless of how many times it traverses the list.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Redbooks Workshop

System Logger enhancements

There are some functions within System Logger that may represent a bottleneck if they do not complete in a timely manner.

Back in z/OS 1.4, System Logger introduced monitoring of those functions, and associated messages if the current request does not complete within some amount of time:

- Offload data set allocation - warning message (IXG310I) issued after 30 seconds, and operator action messages (IXG311I and IXG312E) issued after 60 seconds.
- Offload data set recall - warning message issued after 60 seconds, operator action message issued after 120 seconds.

Naturally, these messages are only useful if someone sees them, so please ensure they are included in your automation tables.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger enhancements

However, some customers felt that, for their environment, the messages are issued too quickly (false positives).

And some customers felt that the messages were not issued quickly enough, and that there could be a service impact before the message was issued.

***One* customer was said to be "happy" with the values.**

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

System Logger enhancements

In order to let customers tailor System Logger monitoring to their environment, z/OS 1.13 includes a new IXGCNFxx Parmlib member.

The member lets you:

- Specify the Parmlib member containing the Trace options to be used when System Logger is started or restarted (CTRACE keyword).
- Specify the number of seconds that Logger should wait before issuing a warning message for an offload data set allocation delay (WARNALLOC(xx))
- Specify the number of seconds that Logger should wait before issuing an operator action message for an offload data set allocation delay (ACTIONALLOC(xx))
- Specify the number of seconds that Logger should wait before issuing a warning message for an offload data set RECALL delay (WARNRECALL(xx))
- Specify the number of seconds that Logger should wait before issuing an operator action message for an offload data set RECALL delay (ACTIONRECALL(xx))

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM

System Logger enhancements

There is a new keyword in IEASYSxx (IXGCNF=) that let's you point at the IXGCNF member you wish to use.

- Note that a default IXGCNFxx member (called IXGCNFXX) is shipped in SYS1.SAMPLIB, NOT SYS1.IBM.PARMLIB

You can dynamically switch from one IXGCNFxx member to another using the SET IXGCNF=xx command.....

- Note that access to the SAF profile for the following profiles may be needed:
  - MVS.DISPLAY.LOGGER
  - MVS.SET.IXGCNF
  - MVS.TRACE.CT
- These are documented in Setting Up a Sysplex, and will be added to the MVS System Commands manual.

Redbooks Workshop

System Logger enhancements

```
SET IXGCNF=00
IXG721I SET IXGCNF COMMAND ACCEPTED
IEE252I MEMBER IXGCNF00 FOUND IN SYS1.PARMLIB
IEF196I IEF285I   SYS1.PARMLIB                              KEPT
IEF196I IEF285I   VOL SER NOS= BH5CAT.
IEF196I IEF285I   CPAC.ZOSR1D.PARMLIB                       KEPT
IEF196I IEF285I   VOL SER NOS= BH5CAT.
IEF196I IEF285I   SYS1.IBM.PARMLIB                          KEPT
IEF196I IEF285I   VOL SER NOS= Z1DRC1.
TRACE CT,ON,COMP=SYSLOGR,PARM=CTILOG00
IEE536I IXGCNF00 VALUE 00 NOW IN EFFECT
DISPLAY LOGGER,IXGCNF
IXG731I LOGGER PARAMETER PROCESSING COMPLETED SUCCESSFULLY FOR SET
IXGCNF REQUEST
IXG607I   13.41.59  LOGGER DISPLAY 859
LOGGER PARAMETER OPTIONS
KEYWORD           SOURCE   VALUE
----------------- -------- ----------------------------------------
CTRACE            SET (00) CTILOG00
MONITOR OFFLOAD
 WARNALLOC        SET (00) 00020
 ACTIONALLOC      SET (00) 00040
 WARNRECALL       SET (00) 00030
 ACTIONRECALL     SET (00) 00060
IEE252I MEMBER CTILOG00 FOUND IN SYS1.IBM.PARMLIB
ITT038I ALL OF THE TRANSACTIONS REQUESTED VIA THE TRACE CT COMMAND
WERE SUCCESSFULLY EXECUTED.
IEE839I ST=(ON,0001M,00004M) AS=ON  BR=OFF EX=ON  MO=OFF MT=(ON,024K)
862
        ISSUE DISPLAY TRACE CMD FOR SYSTEM AND COMPONENT TRACE STATUS
        ISSUE DISPLAY TRACE,TT CMD FOR TRANSACTION TRACE STATUS
```

Issued "under the covers"

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger enhancements

You can also display the current settings AND where they came from...

This is the display from when the system was IPLed with no IXGCNFxx member:

```
D LOGGER,IXGCNF
IXG607I   13.41.49  LOGGER DISPLAY 845
LOGGER PARAMETER OPTIONS
KEYWORD           SOURCE   VALUE
----------------- -------- ------------------
CTRACE            DEFAULT  CTILOG00
MONITOR OFFLOAD
 WARNALLOC        DEFAULT  00030
 ACTIONALLOC      DEFAULT  00060
 WARNRECALL       DEFAULT  00060
 ACTIONRECALL     DEFAULT  00120
```

Redbooks Workshop

System Logger enhancements

And this is the result AFTER the SET IXGCNF=00 command was issued:

```
D LOGGER,IXGCNF
IXG607I   15.01.35  LOGGER DISPLAY 896
LOGGER PARAMETER OPTIONS
KEYWORD           SOURCE   VALUE
----------------- -------- --------------
CTRACE            SET (00) CTILOG00
MONITOR OFFLOAD
 WARNALLOC        SET (00) 00020
 ACTIONALLOC      SET (00) 00040
 WARNRECALL       SET (00) 00030
 ACTIONRECALL     SET (00) 00060
```


System Logger enhancements

You can also modify specific keywords without refreshing the member (similar to the SETSMF command)

```
SETLOGR MONITOR,OFFLOAD,WARNALLOC(15)
IXG651I SETLOGR MONITOR COMMAND ACCEPTED 905
FOR LOGGER SYSTEM CONFIGURATION CHANGE
DISPLAY LOGGER,IXGCNF
IXG731I LOGGER PARAMETER PROCESSING COMPLETED SUCCESSFULLY FOR
SETLOGR REQUEST
IXG607I   15.05.34  LOGGER DISPLAY 908
LOGGER PARAMETER OPTIONS
KEYWORD           SOURCE   VALUE
----------------- -------- ----------------------------------------
CTRACE            SET (00) CTILOG00
MONITOR OFFLOAD
 WARNALLOC        SETLOGR  00015
 ACTIONALLOC      SET (00) 00040
 WARNRECALL       SET (00) 00030
 ACTIONRECALL     SET (00) 00060
```

Note that default IXGCNF00 member is NOT shipped in Parmlib - sample is provided in IXGCNFXX in SAMPLIB

System Logger and IMS

IMS Shared Message Queue (a.k.a. Common Queue Server (CQS)) uses System Logger for its logging function

Conventional wisdom was that the impact of taking a structure checkpoint was so large that customers would only take one checkpoint a day.

- This meant that huge amounts of log records had to be moved from the CQS Logger structure to the offload data sets.
- If any delay was experienced with the offload (allocation delays, poor DASD response times, etc), the structure could potentially fill up and CQS would stop accepting new transactions.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger and IMS

However, enhancements to XES (message-based CFRM processing) and faster DASD and CFs have changed the dynamics. IMS Best Practice is now to take structure checkpoints much more frequently - for example, every 5 minutes (and ensure IMS APAR PK85568 is applied).

- This means that CQS only needs 10 minutes-worth of log records rather than 2 days.
- And THAT means that, it might be possible to size a structure so that CQS log records normally would not have to be moved to the offload data sets.

This would address many of the issues that can cause offload-related interruptions to CQS service.

So what are the considerations if you would like to do this?

Redbooks Workshop

System Logger and IMS

Some improvements were required to System Logger so that its calculations of when high and low thresholds have been reached are more accurate for log streams in structures that are larger than 4GB (you don't want to be doing offloads unless it is strictly necessary).

System Logger APAR OA36261 is available (back to z/OS 1.10) and addresses some issues that existed in this area.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger and IMS

If a log stream is full when an IXGWRITE is issued, the response that System Logger sends to the requestor indicates that the log stream is full. No subsequent notification will be sent until the log stream gets down to the low offload threshold.

- For a very large structure, this could take a VERY long time. Especially as the normal low offload threshold for structures like the CQS log stream is 0%.

System Logger APAR OA36175 (available back to z/OS 1.10) is available and changes the way this works.

- Rather than offloading all the way down to the LOWOFFLOAD threshold, Logger will only offload down to about 90% full. At that point it turns off the STRUCTURE FULL flag, stops the offload, and sends an ENF signal, notifying interested parties that the structure is no longer full.
- This means that programs that are waiting to be told that the structure is no longer full can start using it again much sooner.

Redbooks Workshop

System Logger and IMS

Previously, when the STRUCTURE FULL flag is turned on, IXGWRITE requests to that structure will be rejected until the offload completes.

System Logger APAR OA36662 is available (back to z/OS 1.10) and changes the way this works.

- Write requests from authorized callers (CQS, for example) will be accepted even when the STRUCTURE FULL flag is still on, as long as there is room in the structure (which there will be, shortly after the offload starts).

CQS APAR PM36652 exploits this capability by retrying the IXGWRITE even if the STRUCTURE FULL flag is turned on and before it receives the ENF48 signal that indicates the structure is no longer full.

System Logger and IMS

Previously, IXGWRITE requests would either complete successfully, or they would be rejected with a response indicating that the structure was full.

System Logger APAR OA36172 is available (back to z/OS 1.10) and changes the way this works:

- Now, the response to every IXGWRITE request provides information about how full the structure is. This allows the caller to take actions if the structure is starting to run out of space.

CQS APAR PM36659 exploits this capability:

- New keyword CHKNEARFULL on IMS PROCLIB STRUCTURE statement tells CQS to automatically initiate a structure checkpoint when the structure starts getting near full.
  - Note that this must be explicitly enabled - default is CHKNEARFULL=NO

System Logger and IMS

Note that in order to have all log records deleted before they are moved to an offload data set, you may require a very large structure (many GBs).

IF the usage of the structure changes so that the current entry-to-element ratio is not optimum, Logger may issue an IXLALTER for the structure:

- For very large structures, such requests may run for a very long time and impact response times.
- If you experience this, you can use the new (in z/OS 1.13) ability to disable alter for that structure
  - But be aware that this may result in sub-optimal use of the space in the structure if the current usage of entries and elements is significantly different to how the structure is allocated.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger and IMS

Also, providing a very large structure is not a guarantee that there will be zero movement to offload data sets:

- If a connector disconnects, Logger does an offload of everything in the structure to protect that connector's data.
- If a Logger structure is rebuilt, Logger does an offload.
- If the level of activity is much higher than expected, the structure may fill, prompting an offload
  - Especially if IMS APAR PM36659 is not applied.
- If staging data sets are being used and are not large enough, an offload may be triggered by a staging data set reaching the highoffload threshold before the structure reaches the highoffload threshold.

Summary - suggest that every IMS SMQ customer evaluate their options and decide on the best strategy for them - there is no one-size-fits-all.

Redbooks Workshop

System Logger enhancements

Whenever you issue an XCF ADD COUPLE command for a CDS, XCF passes information to the owner of that CDS and asks if the new CDS is acceptable to the owner (Logger, SFM, WLM, etc).

Prior to z/OS, if Logger rejected a candidate CDS, it would simply reject it, with no information provided about WHY it was rejected.

Starting with z/OS 1.13, information is returned by Logger to XCF so that the response to the SETXCF command not only tells you that the CDS was rejected, it also gives details about the CDS to help you understand the problem.

- Question - how would you get this information for a CDS that has not yet been successfully added to XCF?

Note: Ensure that Logger APAR OA37598 is applied

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger enhancements

Before:

```
SY1  IXC255I UNABLE TO USE DATA SET
SLC.FDSS2
AS THE ALTERNATE FOR LOGR:
ALLOWABLE SIZE OF LSR RECORDS IS LESS THAN CURRENT PRIMARY
```

After:

```
SY1  IXC255I UNABLE TO USE DATA SET
SLC.FDSS2
AS THE ALTERNATE FOR LOGR:
ALLOWABLE SIZE OF LSR RECORDS IS LESS THAN CURRENT PRIMARY
RELEVANT LOGR COUPLE DATA SET FORMAT INFORMATION
PRIMARY
  FORMAT LEVEL:    HBB7705
  FORMAT KEYWORDS:      LSR(25) LSTRR(25) DSEXTENT(15)
                        SMDUPLEX(1)
ALTERNATE
  FORMAT LEVEL:    HBB7705
  FORMAT KEYWORDS:      LSR(24) LSTRR(25) DSEXTENT(15)
                        SMDUPLEX(1)
```

Redbooks Workshop

System Logger offload considerations

There are basically two ways to configure a CF log stream:

- One copy of data in the CF, the other copy in a data space
- One copy of data in the CF, the other copy in a staging data set
  - Staging data sets result in elongated IXGWRITE response times because the staging data set write must complete before Logger replies to the requester.
  - However they protect data that might be lost in case of a CPC or power failure.
  - They are also required if you want to be able to restart workloads in a DR scenario

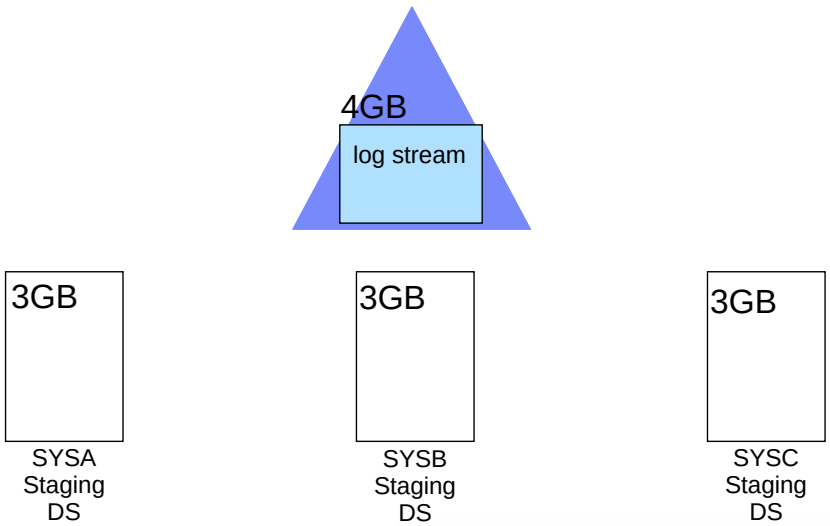
For CF log streams that use a staging data set, in order to get the optimum value from the CF resources, you want offloads to be triggered when the *structure* reaches the highoffload threshold, not when the *staging data set* does.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger offload considerations

What is most likely to trigger an offload in this configuration?

- Structure reaching HIGHOFFLOAD threshold?
- Staging Data Set reaching HIGHOFFLOAD threshold?



Redbooks Workshop

System Logger offload considerations

Answer:

- It Depends!

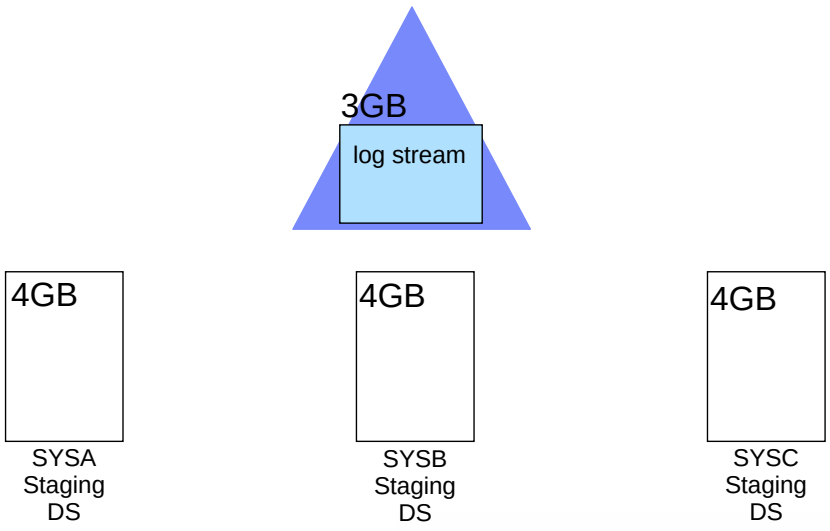
Not every system will create log records at the same rate - SYSA might generate 2.5 GB of log records while SYSB and SYSC only generate 50 MB.

So which will reach HIGHOFFLOAD first? Structure or staging data set?

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger offload considerations

How about in this configuration (EACH staging data set now larger than the structure)?



Redbooks Workshop

System Logger offload considerations

Surely in this case the structure will ALWAYS reach HIGHOFFLOAD before the staging data sets do?

Not necessarily....

Logger does not pack log blocks into a staging data set, as it does with log stream offload data sets.

Instead, to avoid having to buffer any data between IXGWRITE requests, it writes at most one log block to each 4K CI in the staging data set.

Therefore, the smaller the log block, the more space may be unused in the staging data set.

And because the %used for the staging data set is based on the number of used *CIs* (not used *bytes*), the combination of small log blocks and 4K CI size may result in the staging data set reaching HIGHOFFLOAD much sooner than the structure reaches it.

And remember - the default staging data size is equal to the corresponding structure size.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

System Logger offload considerations

So, how to know if this is a concern for you?

1) Monitor for non-zero values in the SMF88ETT field in the SMF 88 records.

2) Check the average buffer size for the log stream - is it much smaller than 4K?

- You can get this value:
  - Run an IXCMIAPU report and check the EFFECTIVE AVERAGE BUFFER SIZE for the structure associated with the log stream
  - From the Average Buffer Size field in an IXGRPT1 report

If you discover that this is happening, and the associated log stream is important to you, you can address is by making the staging data sets larger

- Staging data sets are unique for each system, and get reallocated when that system connects to the log stream.
- Max staging data set size is 4 GB

Redbooks Workshop

(yet more!!) System Logger enhancements

Sadly, we find that many customers do not put much time into monitoring or tuning their use of System Logger and log streams.

- This is a shame, because valuable performance and availability benefits can be achieved with just a little effort.

Part of the reason may be the lack of a strong toolset to help you analyze log stream activity.

In an effort to make it easier to report on log stream usage, System Logger in z/OS 1.13 provides a sample ICETOOL job in SYS1.SAMPLIB(IXGRPT2)

- You can run the job on lower level systems, however you will need to adjust the JCL to not use SYSIN inside a PROC

ibm.com yourdotcom International Technical Support Organization and Authoring Services

(yet more!!) System Logger enhancements

The provided sample does the following:

- COPIES LOGGER SMF TYPE 88 RECORDS FROM A DATA SET OR LOG STREAM TO A TEMPORARY DATA SET.
  - So the job supports both traditional SMF data sets AND SMF log streams.
- SORTS SMF TYPE 88 SUBTYPE 1 (log stream-level reporting) RECORDS
- SORTS SMF TYPE 88 SUBTYPE 11 (structure-level reporting) RECORDS
- PRODUCES A SUMMARY OF LOG STREAMS IN THE SMF TYPE 88 SUBTYPE 1 DATA
- PRODUCES A REPORT OF LOG STREAM ACTIVITY DURING EACH INTERVAL FOR SMF TYPE 88 SUBTYPE 1 RECORDS
- PRODUCES A SUMMARY OF STRUCTURES IN THE SMF TYPE 88 SUBTYPE 11 DATA
- PRODUCES A REPORT OF STRUCTURE ACTIVITY DURING EACH INTERVAL FOR SMF TYPE 88 SUBTYPE 11 RECORDS

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

(yet more!!) System Logger enhancements

SUMM01 report:

```
LOGSTREAM NAMES BY SYSTEM        09/13/11        00:45:04        - 1 -

LOGSTREAM                      SYSTEM NAME            RECORDS
--------------------------     -----------    ---------------
ATR.#@$#PLEX.DELAYED.UR        #@$A                        10
ATR.#@$#PLEX.MAIN.UR           #@$A                        10
ATR.#@$#PLEX.RESTART           #@$A                        10
ATR.#@$#PLEX.RM.DATA           #@$A                        10
HZS.HEALTH.CHECKER.HISTORY     #@$A                        10
IFASMF.#@$#PLEX.TYPALL         #@$A                        10
IGWTV010.IGWLOG.SYSLOG         #@$A                        10
IGWTV010.IGWSHUNT.SHUNTLOG     #@$A                        10
SYSPLEX.OPERLOG                #@$A                        10
```


ibm.com yourdotcom International Technical Support Organization and Authoring Services

(yet more!!) System Logger enhancements

REPORT01 report:

```
LOGGER 88-1 ACTIVITY REPORT        09/13/11        00:45:05        - 6 -

LOGSTREAM  IFASMF.#@$#PLEX.TYPALL

     TME          DTE   SYN                   LWI               LIB               LAB                       LWB
--------   ----------   --------   ---------------   ---------------   ---------------   -----------------------
05:46:00   2010/10/30   #@$A                    0        2147483647                 0                         0
05:47:00   2010/10/30   #@$A                 4636             33056             65412                 286688279
05:48:00   2010/10/30   #@$A               105810             33056             65528                6546914837
05:49:00   2010/10/30   #@$A               105796             32992             65528                6546149244
05:50:00   2010/10/30   #@$A               105710             32992             65528                6540827913
05:51:00   2010/10/30   #@$A               105844             32992             65528                6549000139
05:52:00   2010/10/30   #@$A                97014             32992             65528                6002133731
05:53:00   2010/10/30   #@$A                94835             32992             65528                5866973361
05:54:00   2010/10/30   #@$A                84984             32992             65528                5257515990
05:55:00   2010/10/30   #@$A                37893             32992             65528                2344396386

AVERAGE                                     74252         214778070             58963                4594059988

1LOGGER 88-1 ACTIVITY REPORT       09/13/11        00:45:05        - 10 -

     TME          DTE   SYN                   LWI               LIB               LAB                       LWB
--------   ----------   --------   ---------------   ---------------   ---------------   -----------------------

MAXIMUM                                    105844        2147483647             65528                6549000139

MINIMUM                                         0               176                 0                         0

AVERAGE                                      8255        1216910869             17476                 510455025

TOTALS FOR ALL INTERVALS                   743004       109521978293           1572846               45940952280
```

Plus lots more over here.....

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

System Logger

For guidance about the meaning of the various fields and the relationships between them, refer to the Redbook *System Programmer's Guide to: z/OS System Logger*, SG24-6898, available on the Web at:

- http://www.redbooks.ibm.com/abstracts/sg246898.html?Open

ibm.com yourdotcom International Technical Support Organization and Authoring Services

z/OS 1.13 sysplex enhancements summary

Enhanced D XCF command for improved data capture

XCF trace changes - remember to check CTIXCF00 and CTIXES00 members for BUFSIZE values

Long running alters - if this is NOT causing you a problem, you don't need to do anything. If long running alters ARE causing you a problem, you now have a way to address that

Improved structure placement messages (INFO110) make CF management easier

XCF client server support - provides additional information in SDSF if you were not using MQ previously, and provides simplified environment if you WERE using MQ

ARM TIMEOUT control helps if you use ARM and have issues with restarts failing because they start too soon

Redbooks Workshop

z/OS 1.13 sysplex enhancements summary

VTOC notifications - should make processes easier for storage management team.

Path busy reporting changes - make sure you keep a note of when that change became active - otherwise you may wonder what caused a large change in PATH BUSY numbers in RMF reports

Logger IXGCNFxx member - helpful if you have had problems with IXG31x messages. Also, will have additional function in the future

Logger and IMS - important changes for IMS Shared Message Queue customers (significant change in Best Practices recommendations)

Logger offload and staging data sets - check SMF88ETT for indication of inefficient use of CF storage

Logger reporting - provided job is only intended as a sample, to encourage people to do more monitoring.

yourdotcom

International Technical Support Organization and Authoring Services

IBM GDPS Active-Active Sites Overview

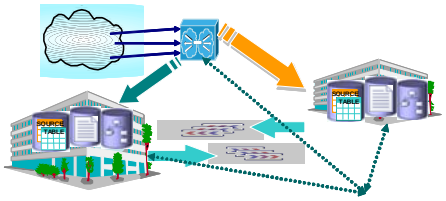
System z environments

ibm.com/redbooks

Courtesy of Dave Petersen
IBM Distinguished Engineer, GDPS Chief Architect

Active/Active Sites overview – agenda

- **Level set**
- **Active/Active Sites overview**
- **Preliminary testing results**
- **Components**



ibm.com yourdotcom International Technical Support Organization and Authoring Services

Active/Active Sites overview

⇨ **Level set**
Active/Active Sites overview
Preliminary testing results
Components

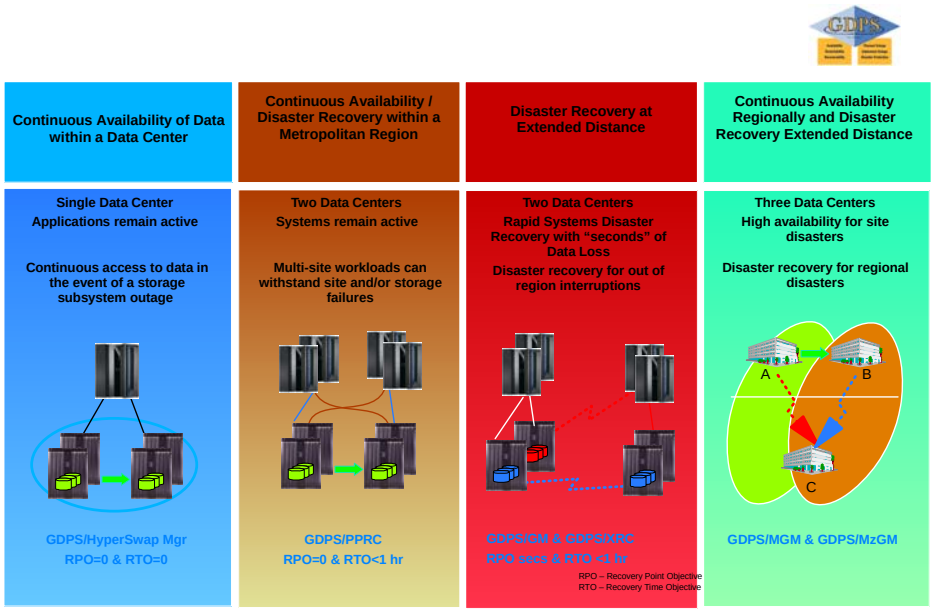
Redbooks Workshop

Interagency Paper on Sound Practices to Strengthen the Resilience of the U.S. Financial System [Docket No. R-1128] (April 7, 2003)

- **Identify clearing and settlement activities in support of critical financial markets**
- **Determine appropriate recovery and resumption objectives for clearing and settlement activities in support of critical markets**
  - …core clearing and settlement organizations should develop the capacity to recover and resume clearing and settlement activities within the business day on which the disruption occurs with the overall goal of achieving recovery and resumption within two hours after an event
- **Maintain sufficient geographically dispersed resources to meet recovery and resumption objectives**
  - Back-up arrangements should be as far away from the primary site as necessary to avoid being subject to the same set of risks as the primary location.
  - The effectiveness of back-up arrangements in recovering from a wide-scale disruption should be confirmed through testing
- **Routinely use or test recovery and resumption arrangements.**
  - One of the lessons learned from September 11, 2001 is that testing of business recovery arrangements should be expanded

ibm.com yourdotcom International Technical Support Organization and Authoring Services

What are companies doing today ?



Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

What are GDPS/PPRC customers doing today?

- **GDPS/PPRC, based upon a multi-site Parallel Sysplex and synchronous disk replication, is a metro-area Continuous Availability (CA), Disaster Recovery solution (DR)**
- **GDPS/PPRC supports two configurations:**
  - Active/standby
  - Active/active
- **Some customers have deployed GDPS/PPRC active/active configurations**
  - All critical data must be PPRCed and HyperSwap enabled
  - All critical CF structures must be duplexed
  - Applications must be Parallel Sysplex enabled
  - Both sites are in the same sysplex and logically are one operating environment
  - Signal latency will impact OLTP throughput and batch duration. Typically this means that the sites are separated by no more than a couple tens of KM (fiber)
- **Issue: the GDPS/PPRC active/active configuration does not provide enough site separation for some enterprises**

ibm.com yourdotcom International Technical Support Organization and Authoring Services

What are GDPS/XRC and GDPS/GM customers doing today?

- **GDPS/XRC and GDPS/GM, based upon asynchronous disk replication, are unlimited distance DR solutions**
- **The current GDPS async replication products require the failed site's workload to be restarted in the recovery site and this typically will take 30-60 min**
  - Power fail consistency
  - Transaction consistency
- **There are no identified extensions to the existing GDPS asynch replication products that will allow the RTO to be substantially reduced.**
- **Issue: GDPS/XRC and GDPS/GM will not achieve an RTO of seconds being requested by some enterprises**

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

How much interruption can your business tolerate?

Ensuring business continuity:

- **Disaster recovery**
  - Restore business after an unplanned outage
- **High availability**
  - Meet service availability objectives – e.g., 99.9% availability or 8.8 hours of down-time a year
- **Continuous availability**
  - No downtime (planned or not)

Global enterprises that operate across time zones no longer have any 'off-hours' window. Continuous availability is required.

What is the cost of one hour of downtime during core business hours?

Customer requirements

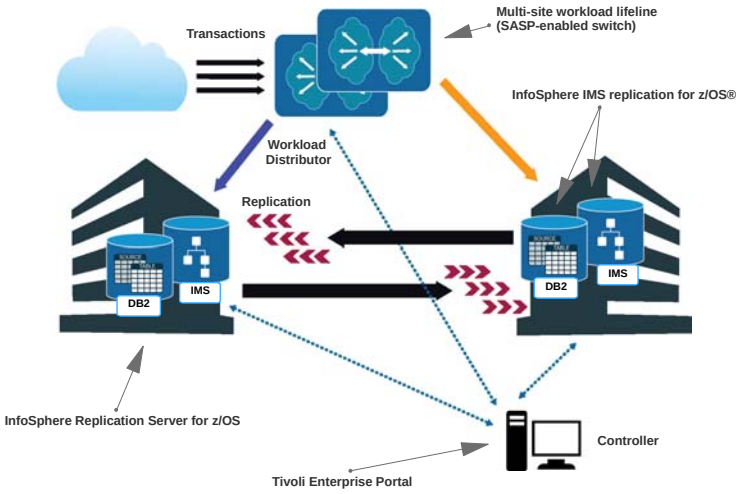
- **Want to shift focus from a failover model to a nearly-continuous availability model (RTO near zero)**
- **Access data from any site (unlimited distance between sites)**
- **No application changes**
- **Multi-sysplex, multi-platform solution**
  - "Recover my business rather than my platform technology"
- **Ensure successful recovery via automated processes (similar to GDPS technology today)**
  - Can be handled by less-skilled operators
- **Provide workload distribution between sites (route around failed sites, dynamically select sites based on ability of site to handle additional workload).**
- **Provide application-level granularity**
  - Some workloads may require immediate access from every site, other workloads may only need to update other sites every 24 hours (less critical data).
  - Current solutions employ an all-or-nothing approach (complete disk mirroring, requiring extra network capacity).
- **Replace "roll your own" solutions**

Active/Active Sites overview

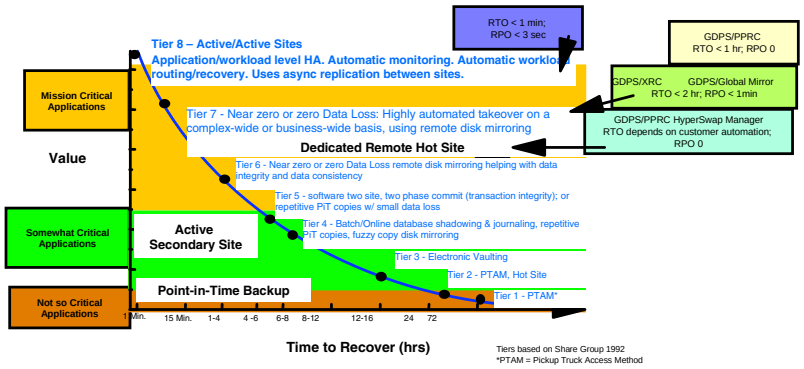
Level set
⇨ **Active/Active Sites overview**
Preliminary testing results
Components

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Active/Active Sites concept



Tiers of disaster recovery: Level-setting Active/Active Sites



Best D/R practice is blend tiers of solutions in order to maximize application coverage at lowest possible cost . One size, one technology, or one methodology does not fit all applications

IBM United States Services Announcement 611-023, dated May 24, 2011 - IBM GDPS active/active continuous availability

At a glance

- IBM® GDPS® active/active continuous availability is the next generation of GDPS and represents a fundamental paradigm shift for near continuous availability solutions.

Overview

- IBM GDPS active/active continuous availability is the next generation of GDPS and a fundamental paradigm shift from a failover model to a near continuous availability model. IBM GDPS active/active continuous availability combines the best attributes of the existing suite of GDPS services and expands them to allow you to achieve unlimited distances between your data center sites with recovery time objectives measured in seconds. IBM GDPS active/active continuous availability is a solution for an environment consisting of two sites, separated by unlimited distances, running the same applications and having the same data with cross-site workload monitoring, data replication, and balancing. IBM GDPS active/active continuous availability, as with previous GDPS solutions, provides a complete set of services to help achieve near continuous availability. This solution, which is an integration of IBM products and GDPS control software, is delivered through an IBM service engagement which includes project management throughout the implementation cycle.

Statement of direction

- IBM intends to deliver, over time, additional configurations that comprise GDPS active/active continuous availability. In addition to the Active/Standby configuration, IBM plans to make available the Active/Query configuration, which will provide the ability to selectively query data in either site.*

* This statement represents the current intention of IBM. IBM development plans are subject to change or withdrawal without further notice. Any reliance on this statement of direction is at the relying party's sole risk and does not create any liability or obligation for IBM.

Active/Active Sites configurations

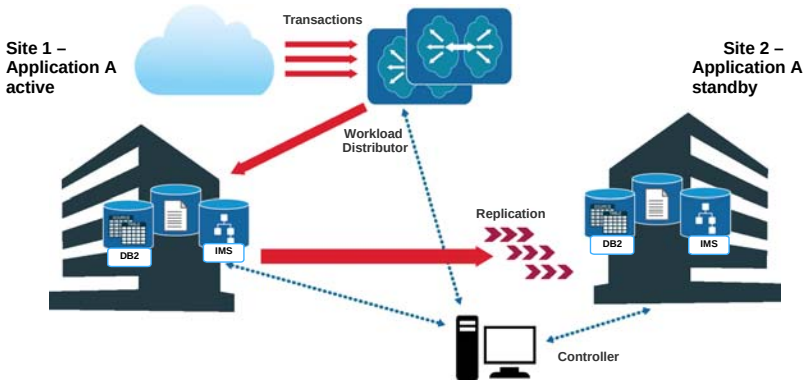
- **Configurations**
  - Active/Standby – general availability on June 30, 2011
  - Active/Query – statement of direction
  - …
- **A configuration is specified on an application basis**
- **An application is the aggregation of these components**
  - ***Software:*** user-written applications (e.g., COBOL program) and the middleware run-time environment (for example, CICS regions & DB2 subsystem)
  - ***Data:*** related set of objects that must preserve transactional consistency and optionally referential integrity constraints (for example, DB2 Tables)
  - ***Network connectivity*****:** one or more TCP/IP addresses & ports (for example, 10.10.10.1:80)

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Active/Standby configuration

GDPS active-standby configuration

- Static routing
- Automatic failover

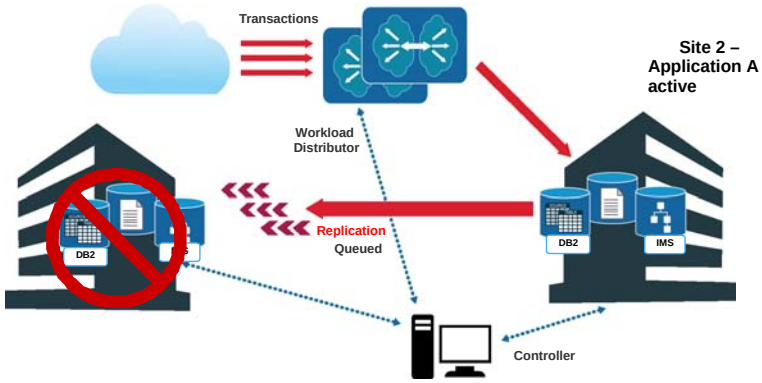


Redbooks Workshop

Active/Standby configuration (continued)

GDPS active-standby configuration

- Static routing
- Automatic failover



Active/Active Sites functions

- **Start/stop a controller** – start and stop an A/A Sites controller
- **Start/stop a site** – start and stop individual sysplexes (each sysplex maps to a site)
- **Stop/start a workload** – start and stop individual workloads
- **Monitoring** – monitor the A/A Sites configuration and, if any conditions that will potentially impact a workload and/or site switch, geerate an alert
- **Planned workload switch** – switch the workload site to the other site initiated by operator action
- **Unplanned workload switch** – switch a failed workload to the other site, either automatically or based upon operator prompt, after the workload failure detection interval
- **Planned site switch** – switch all workloads executing to the other site, initiated by operator action
- **Unplanned site switch** – switch the failed site's workloads to the other site, either automatically or based upon operator prompt, after the site failure detection interval

Active/Active Sites overview

Level set
Active/Active Sites overview
⇨ **Preliminary testing results**
Components

ibm.com yourdotcom International Technical Support Organization and Authoring Services

*Preliminary testing results**

- **Planned workload switch**
  - Operations-initiated switch from the active instance of a workload to the standby instance took 20 seconds
  - Not possible to swap a subset of the total workload with disk replication
- **Unplanned workload switch**
  - Automatic switch from the active instance of a workload to the standby instance took 120 seconds (workload failure detection interval is 60 seconds)
  - Not possible to swap a subset of the total workload with disk replication
- **Planned site switch (9 * CICS-DB2 and 1 * IMS workloads)**
  - Operations initiated switch from of the workloads in a site to the other site took 20 seconds
  - Current asynchronous GDPS and disk replication will take 1-2 hours
- **Unplanned workload switch (site failure detection interval is 60 seconds)**
  - Automatic switch of failed site workloads to the surviving site took 150 seconds (site failure detection interval is 60 seconds)
  - Current asynchronous GDPS and disk replication will take about one hour

* IBM laboratory results; actual results may vary.

Redbooks Workshop

Active/Active Sites overview

Level set
Active/Active Sites overview
Preliminary testing results
⇨ **Components**

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Minimum releases of required products installed on z/OS production and controller images for Active/Standby configuration

- **Operating system**
  - z/OS V1R11
- **Applications/Middleware**
  - DB2 for z/OS V9
  - IMS V10
  - WS MQ V7.0 (for DB2 replication)
- **Replication**
  - InfoSphere Replication Server (DB2) V10
  - InfoSphere IMS Replication for z/OS V10.1 (new product)
- **Management and monitoring**
  - GDPS/Active-Active V1.1 (new product)
  - NetView for z/OS V6.1
  - System Automation for z/OS V3.3
  - IBM Multi-site Workload Lifeline V1.1 (new product)
  - IBM Tivoli Monitoring V6.2.2
  - Optional OMEGAMON products (required only if the customer wants to monitor the behavior of the respective products/resources that they deal with (DB2, CICS, storage, etc.)
    - OMEGAMON XE on z/OS V4.2.0
    - OMEGAMON XE for Mainframe Networks V4.2.0
    - OMEGAMON XE for Storage V4.2.0
    - OMEGAMON XE for DB2 Performance Expert (or Performance Monitor) on z/OS V4.2.0 (if DB2 is running)
    - OMEGAMON XE on CICS for z/OS V4.2.0 (if CICS is running)
    - OMEGAMON XE on IMS V4.2.0 (if IMS is running)
    - OMEGAMON XE for Messaging V7.0 (if MQ is running)

Redbooks Workshop

GDPS/Active-Active

Key points:

- This is being driven by customer (and, indirectly, government) requirements for the highest levels of availability that will survive just about any type of disaster.
- GDPS has already addressed the CA/DR requirements for metro distance sites.
- This is COMPLETELY different to all GDPS offerings to date. They were all based on DASD mirroring with a layer of automation and processes on top. GDPS/Active-Active doesn't use mirroring at all - all the replication is done at the data level, by DBM-specific products. And the management is much more granular and therefore more involved and more complex.
  - However, the team that are building Active-Active are the same team that have years of experience with creating the existing GDPS offerings.
- This is going to be a journey, and we are only at step 1. However the name of the offering (Active-Active) indicates where IBM plans to take it.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Thank you

- **Current GDPS family of offerings**
  - Over 12-year history of disaster recovery and continuous availability for System z customers
  - A proven track record of success, with almost 600 clients worldwide and growing
  - Ongoing investment and updated annually – currently up to GDPS V3.8
- **NEW GDPS/Active-Active family of offerings**
  - The next generation of GDPS
  - Concept: Active applications, transactional integrity, shared data, replication, and automation over global distances for true continuous availability worldwide
    - First configuration is Active/Standby
    - Statement of Direction on Active Query configuration
    - Additional configurations planned for the future

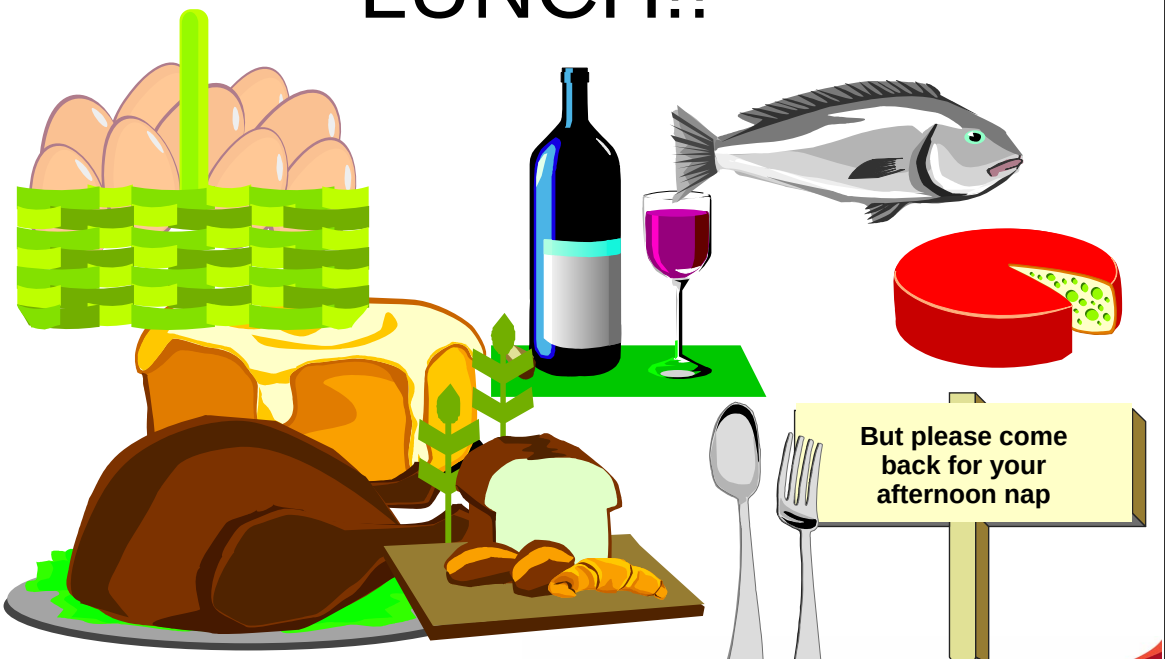
Redbooks Workshop

Questions

Please direct follow-up questions to tdallman@us.ibm.com.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

LUNCH!!



Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

yourdotcom

International Technical Support Organization and Authoring Services

InfiniBand
The Latest and greatest (really!) in coupling technology

ibm.com/redbooks

Redbooks Workshop

Infiniband

- Little history lesson
- Advantages of InfiniBand compared to other coupling types
- What's new with z196 GA2/z114
- Planning considerations
- Performance information
- Operation and management
- STP enhancements with new InfiniBand adapters

There are a lot of new concepts and terms in here, so PLEASE ask questions as I go along - it is very important that everyone understands this stuff

InfiniBand - A Brief Anthology on Coupling

When IBM originally announced Parallel Sysplex, the only type of link that was available was ISC (non-peer mode) running at 100 MB/sec.

- "Good" response times were over 100 mics.

In 1998 IBM introduced ICB2 copper links on 9672 G5 processors - higher bandwidth (250 MB/sec), shorter response times, but much smaller supported distance (10 meters).

With z900, IBM introduced:

- Peer mode links - each end could act as both sender (z/OS) and receiver (CF) concurrently, so could connect to both z/OS and CF LPARs.
- ISC3 links - double the bandwidth of ISC (200MB/sec).
- ICB3 links - superceded ICB2 links. Better performance, more bandwidth (500 MB/sec), same distance restrictions as ICB2.

InfiniBand

ICB4 links were announced with z990 family and superceded ICB3 links. Better performance, more bandwidth (1500 MB/sec), same distance restrictions as ICB2.

Parallel Sysplex over InfiniBand (PSIFB) links were introduced with z10 with limited support rolled back to z9.

- 12X links supported up to 150 meters. Logical follow-on to ICB4 links. Design bandwidth of 6000 MB/sec.
- 1X links supported up to 10km unrepeated. Logical follow-on to ISC3 links. 100km with DWDM (200km with RPQ). Design bandwidth of 500 MB/sec.
- Both link types support multiple CHPIDs per physical link, allowing you to have more subchannels per link (although not more subchannels per CHPID).

z196 GA2 includes new IFB3 mode 12X links and 4-port 1X cards.

- Same bandwidth as original InfiniBand links, but better performance for 12X links and more flexibility for 1X links.

NEW!!

SOD that z196 is last generation that will support ordering of ISC links

InfiniBand

The net of the most recent announcements is that:

- We are now able to deliver ICB-levels of performance with InfiniBand.
- We can support the same distances with InfiniBand that we support with ISC.

With the SOD about ISC no longer being orderable on the next generation, it is clear that everyone soon will have to replace their pre-InfiniBand coupling infrastructure with an InfiniBand-based one:

- This provides a fantastic opportunity to go back and apply your years of coupling experience (and STP migration) and migrate to a configuration that is ideal for your environment.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Infiniband

Little history lesson

Advantages of InfiniBand compared to other coupling types

What's new with InfiniBand on z196 GA2/z114

Planning considerations

Performance information

Operation and management

STP enhancements with new InfiniBand adapters

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

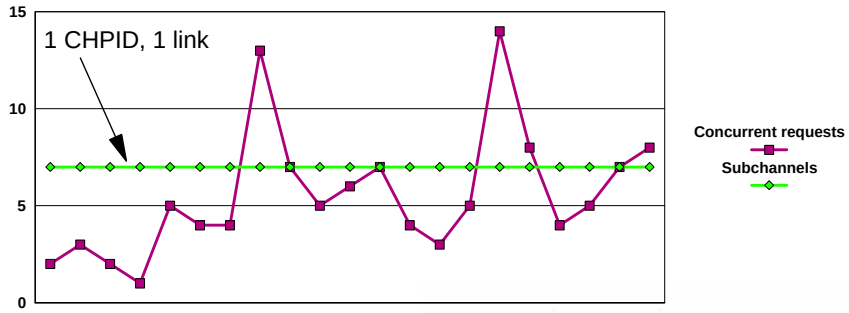
InfiniBand advantages

Cost effective handling of peak loads.....

Prior to InfiniBand, there was a one-to-one relationship between coupling link CHPIDs and physical connections.

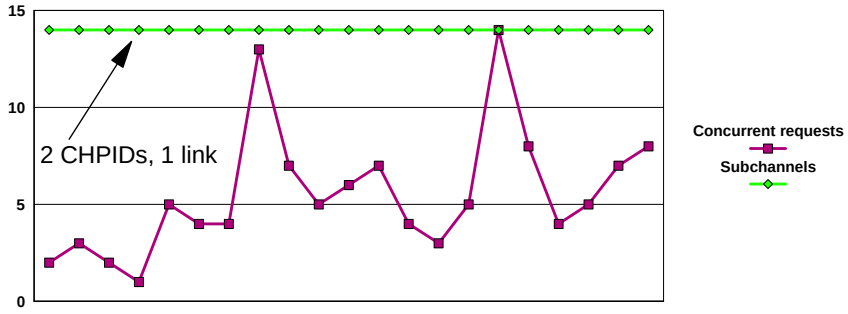
Many customers have more coupling link bandwidth than they actually require. However because of the spiky nature of their workload, they needed additional links to control the amount of Path Busy conditions.

If you wanted more CHPIDs, you had to purchase more links.



InfiniBand advantages

With InfiniBand, you can address the Path Busy situation by defining more CHPIDs and associating them with existing InfiniBand links. There is no financial cost for adding more CHPIDs



This delivers better response times, avoids the CPU cost of Path Busies, and provides more efficient use of coupling investment

InfiniBand advantages

Prior to InfiniBand, there were basically two types of external coupling links:

- ICB
  - Very fast, but very limited distance (10 meters cable length)
- ISC
  - Support up to 200km (with RPQ), but VERY slow, especially for current CPC technology

If you were unable to meet the distance limitations of ICB, you had no choice but to live with the performance impact and CPU cost of ISC links

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM

Impact of coupling link technology on z/OS coupling cost

CF/Host	z9 BC	z9 EC	z10 BC	z10 EC	z114	z196
z9 BC ISC3	14%	15%	17%	19%	18%	23%
z9 BC ICB4	9%	10%	11%	12%	NA	NA
z9 BC 12X IFB	NA	NA	13%	14%	13%	16%
z9 EC ISC3	13%	14%	15%	18%	17%	22%
z9 EC 12X IFB	NA	NA	13%	14%	13%	16%
z9 EC ICB4	9%	9%	10%	11%	NA	NA
z10 BC ISC3	13%	14%	16%	18%	17%	22%
z10 BC 12X IFB	11%	12%	13%	14%	13%	15%
z10 BC ICB4	9%	9%	10%	11%	NA	NA
z10 EC ISC3	12%	13%	15%	17%	17%	22%
z10 EC 12X IFB	10%	11%	12%	13%	12%	15%
z10 EC ICB4	7%	8%	9%	10%	NA	NA
z114 ISC3	14%	14%	16%	18%	17%	21%
z114 12X IFB	10%	10%	12%	13%	12%	15%
z114 12X IFB3	NA	NA	NA	NA	12%	15%
z196 ISC3	11%	12%	14%	16%	17%	21%
z196 12X IFB	9%	10%	11%	12%	11%	14%
z196 12X IFB3	NA	NA	NA	NA	9%	11%

z/OS CPU cost, based on 9 CF requests/MIPS/second

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand advantages

12X InfiniBand links deliver ICB-class performance, but at distances up to 150 meters.

- This provides far greater flexibility for data center physical planning.
- For customers that are forced to use ISC links today because of proximity limitations, moving to InfiniBand can improve performance and reduce coupling cost.

Even 1X InfiniBand links (which support the same distances as ISC) provide better performance than ISC links.

- At distances of more than a few kms, the speed of light is likely to dominate the response time, so the performance difference between ISC and 1X links will be less obvious.
- However InfiniBand still has many advantages over ISC for long=distance sysplexes.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand advantages

For long distance sysplexes, the bottleneck in nearly every case is the utilization of the coupling link subchannels:

- CF subchannels can only be used by one CF request at a time. So the longer it takes to complete the request, the higher will be the subchannel utilization, and the longer other CF requests will have to queue, waiting for an available subchannel.

Traditionally, the only way to address this was by adding more ISC links

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand advantages

Infiniband helps you address the subchannel bottleneck by letting you add CHPIDs to existing links at no additional financial cost.

This is REALLY significant for customers with large distance sysplexes:

- Can quickly react to subchannel bottlenecks by adding more CHPIDs.
- Reduces the number of coupling link adapters required on the CPCs
- Reduces the number of coupling ports on DWDMs
- May reduce the inter-site cabling requirement.

AND, starting with z196 Driver 93, HCA2 LR and HCA3 LR links support 32 subchannels per CHPID.

- So, by moving to Driver 93, you can increase number of subchannels more than 4X without financial cost or even without having to use up a CHPID.

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM

InfiniBand advantages

Flexibility in multi-sysplex environments.....

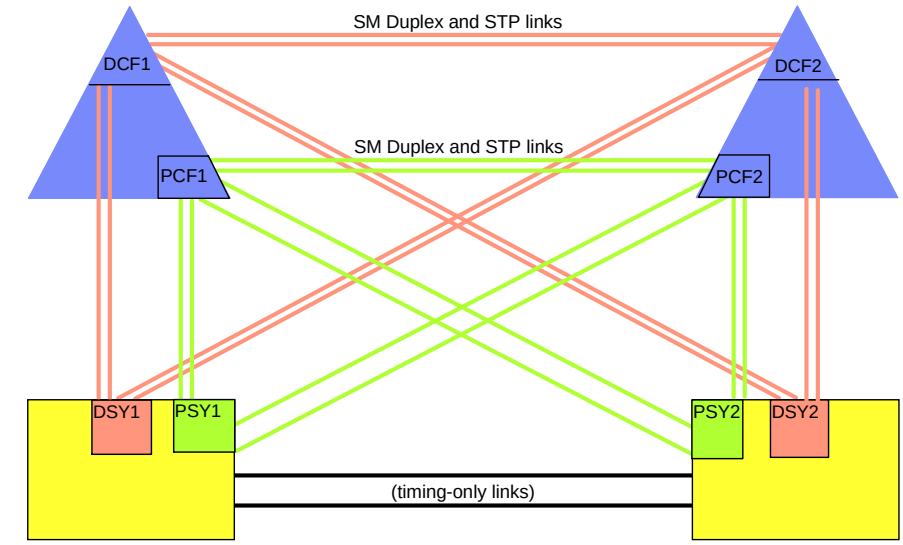
Prior to InfiniBand, coupling links could not be shared across sysplexes.

- So customers with many sysplexes probably had more links (to provide the needed connectivity) than they actually needed from a capacity perspective.

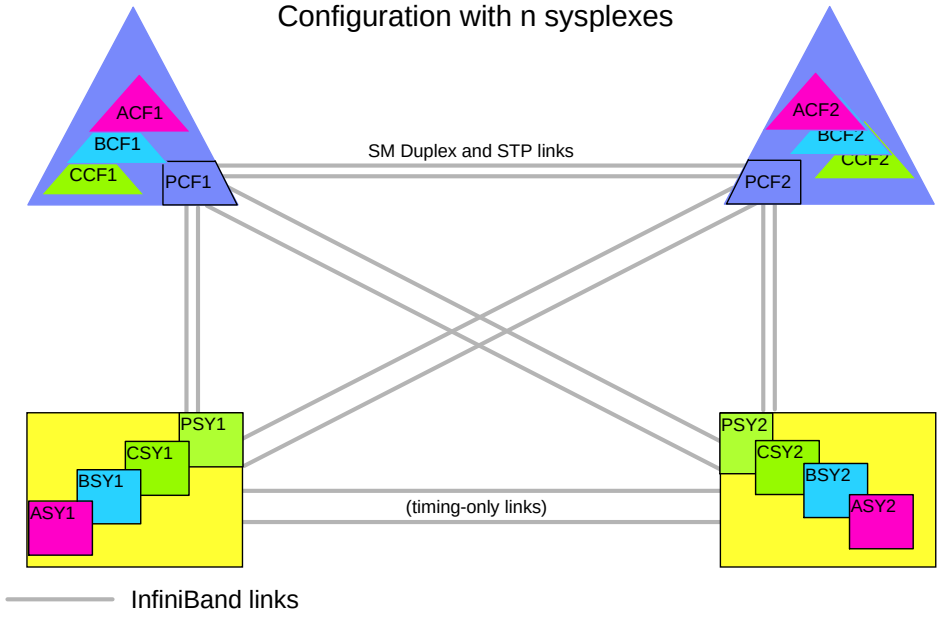
InfiniBand links can be shared by multiple sysplexes.

Redbooks Workshop

Configuration with 2 sysplexes



Configuration with n sysplexes



InfiniBand advantages

z196 and z114 support larger numbers of coupling link CHPIDs than previous generations - 128 vs. 64.

However, to get the maximum value from that capability, you want to be able to assign multiple CHPIDs to a single coupling link - only InfiniBand provides that capability.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Infiniband

Little history lesson

Advantages of InfiniBand compared to other coupling types

What's new with z196 GA2/z114

Planning considerations

Performance information

Operation and management

STP enhancements with new InfiniBand adapters

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

What's new with InfiniBand on z196 GA2/z114?

Driver 93 on z196 and z114 provides support for a new type of InfiniBand adapter - HCA3.

- Like HCA2, HCA3 comes in 12X and 1X varieties

HCA3 12X supports a new protocol mode called IFB3 (previous mode is now referred to as IFB).

- IFB3 supports same bandwidth as IFB (controlled by the InfiniBand standard)
- But it delivers better response times due to a more efficient protocol

HCA3 1X:

- Now comes with 4 ports per adapter (instead of 2 before).
- Has 32 link buffers per CHPID, rather than 7.

Support for new "Going Away Signal" for STP

- Requires HCA3 12X to HCA3 12X (IFB or IFB3 mode) or
- HCA3 1X to HCA3 1X (IFB mode) connection

HCA2 adapters no longer orderable on z196 (effective GA2)

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM.

Infiniband

Little history lesson

Advantages of InfiniBand compared to other coupling types

What's new with z196 GA2/z114

Planning considerations

Performance information

Operation and management

STP enhancements with new InfiniBand adapters

Redbooks Workshop

Planning for InfiniBand implementation

First a quick summary of some InfiniBand terminology...

- CIB - type used to define InfiniBand links to HCD (all InfiniBand link types are defined as CIB - HCD doesn't understand 1X or 12X, or HCA2 or HCA3)
- AID - Adapter ID. Equivalent to PCHID for FICON channels - used to associate CHPID with physical adapter in HCD.
- VCHID - Like a PCHID, but for ICP, PSIFB, and QDIO links. **1 per CHPID**
- PSIFB - Parallel Sysplex InfiniBand links.
- Lanes - each link consists of 2 fibres per lane. PSIFB supports 1 lane (1X) or 12 lanes (12X). System z10 and later support Double Data Rate, meaning that 1X has a bandwidth of 5.0 Gbps, and 12X has a bandwidth of 60 Gbps. All requests are striped across the number of available physical lanes, regardless of the number of CHPIDs defined to that port.
- HCA - Host Channel Adapter - name of the adapter that is used for InfiniBand links.
  - HCA1 (12X only) were only supported on z9
  - HCA2 (1X and 12X) were supported on z10 and z196
  - HCA3 (1X and 12X) only supported on z196 or z114 at Driver 93 or later

ibm.com yourdotcom International Technical Support Organization and Authoring Services

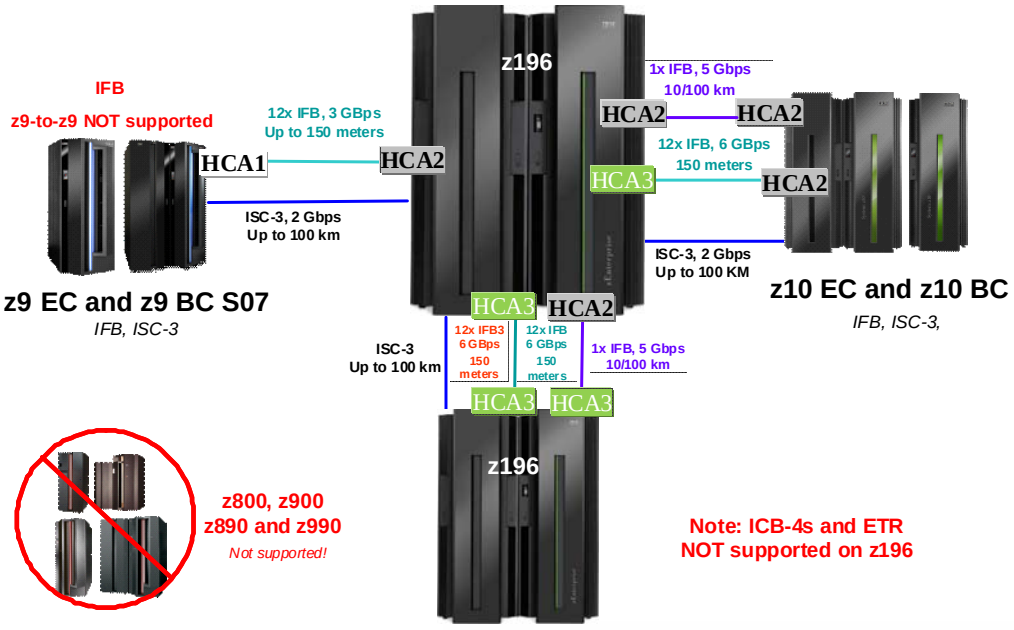
Planning for InfiniBand implementation

Supported InfiniBand link types by processor type

	HCA1	HCA2-12X	HCA2-1X	HCA3-12X	HCA3-1X
z9	Yes				
z10 BC		Yes	Yes		
z10 EC		Yes	Yes		
z196 GA1		Yes	Yes		
z196 GA2		Yes (1)	Yes (1)	Yes	Yes
z114		(2)		Yes	Yes
		1) Carried forward 2) Only for connection to z9			

Redbooks Workshop

z196 coexistence with other CPCs



ibm.com yourdotcom International Technical Support Organization and Authoring Services

Planning for InfiniBand implementation

Supported coupling inter-connectivity options

	HCA1	HCA2-12X	HCA2-1X	HCA3-12X	HCA3-1X
HCA1	No	Yes	No	No	No
HCA2 12X	Yes	Yes	No	Yes	No
HCA2 1X	No	No	Yes	No	Yes
HCA3 12X	No	Yes (IFB)	No	Yes (IFB3)	No
HCA3 1X	No	No	Yes	No	Yes

InfiniBand planning

ICB4 is supported on z10 but not on z196 or z114.

InfiniBand HCA2 is supported on z10 and z196.

InfiniBand HCA3 is supported on z196 and z114, but not on z10.

For customers with ICB4 links, this probably means installing InfiniBand HCA2 before install of first z196/z114, migrating to z196/z114, and then (possibly) migrating to HCA3.

Statement of Direction:

z196 and z114 are the last generation that will support ordering of ISC links on new builds.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Server	1x IFB (HCA3-O LR)	12x IFB & 12x IFB3 (HCA3-O)	1x IFB (HCA2-O LR)	12x IFB (HCA2-O)	IC	ICB-4	ICB-3	ISC-3	Max External Links	Max Coupling CHPIDs
z196	48 M15 – 32*	32 M15 – 16*	32 M15 – 16* CF only	32 M15 – 16*	32	N/A	N/A	48	104 (1)	128
z114	M10 – 32* M05 – 16*	M10 – 16* M05 – 8*	M10 – 12 M05 – 8* CF only	M10 – 16* M05 – 8*	32	N/A	N/A	48	M10 (2) M05 (3)	128
z10 EC	N/A	N/A	32 E12 – 16*	32 E12 – 16*	32	16 (32/RPQ)	N/A	48	64	64
z10 BC	N/A	N/A	12	12	32	12	N/A	48	64	64
z9 EC	N/A	N/A	N/A	HCA1-O 16 S08 - 12	32	16	16	48	64	64
z9 BC	N/A	N/A	N/A	HCA1-O 12	32	16	16	48	64	64

1) A z196 M49, M66 or M80 supports a maximum 104 extended distance links (48 1x IFB and 48 ISC-3) plus 8 12x IFB links.
A z196 M32 supports a maximum 96 extended distance links (48 1x IFB and 48 ISC-3) plus 4 12x IFB links*.
A z196 M15 supports a maximum 72 extended distance links (24 1x IFB and 48 ISC-3) with no 12x IFB links*.
2) z114 M10 supports a maximum of 72 extended distance links (24 1x IFB and 48 ISC-3) with no 12x IFB links*.
3) z114 M05 supports a maximum of 56 extended distance links (8 1x IFB and 48 ISC-3) with no 12x IFB links*.
*** Uses all available fanout slots. Allows no other I/O or coupling.**

Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

InfiniBand planning

First step is to ensure that you have all InfiniBand-related software service.

Use SMP/E FIXCAT function to retrieve and check the following FIXCATs:

- IBM.Device.Server.z114-2818.ParallelSysplexInfiniBandCoupling
- IBM.Device.Server.z196-2817.ParallelSysplexInfiniBandCoupling
- IBM.Device.Server.z114-2818.ServerTimeProtocol
- IBM.Device.Server.z196-2817.ServerTimeProtocol
- Equivalent buckets for all your other CPUs. For a list of all fix categories, see http://www-03.ibm.com/systems/z/os/zos/smpe/fixcategory.html

Apply any service and roll around the entire sysplex.

- This typically will take some time, so start now.

Note that fixes ARE required to deliver z/OS support of Driver 93.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

Any CPCs connected using InfiniBand must be at current hardware service levels:

- z196 and z114 should be at Driver 93 for the latest enhancements
- System z10 should be at Driver 79
- System z9 should be at Diver 67L

Earlier processors are not supported in the same sysplex as z196.

For physical planning information (cables, fanouts, Adapter IDs, etc) refer to material from z196 day of ITSO workshops.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand planning

Strongly recommend that you do NOT blindly replace existing coupling links with an equivalent number of InfiniBand links:

- For most customers, even very large ones, two InfiniBand 12X links between z/OS and CF should provide sufficient bandwidth/capacity.
  - If your workload is spiky, adding CHPIDs should help reduce path or subchannel busy
- Need to consider connectivity, especially for STP, and allowing for STP role re-assignment.
- Recommend that you work with IBM account team to use zCP3000 to validate plans.
- Always want at least two HCA adapters per CPC, to avoid single point of failure.
  - Note that CHPID mapping tool does not provide SPOF analysis for InfiniBand links
- The performance difference between IFB3 and IFB mode is so significant that you really want to avoid having more than 4 CHPIDs per link if you are doing many CF requests.
  - If the load is low, test sysplexes for example, having more than 4 CHPIDs per link should be fine.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

Link planning.....

- Maximum number of CHPIDs that can be associated with an AID (a.k.a. HCA adapter) is 16. These can be distributed across the ports as you wish.
  - With 12X HCA2 links, optimum throughput is reached with 8 CHPIDs per AID (hence the recommendation of not more than 4 CHPIDs per port for optimum throughput).
  - With 12X HCA3 links, port will operate in IFB3 mode IF:
    - The port is connected to another HCA3 port, AND
    - Not more than 4 CHPIDs are DEFINED on the port.
  - With HCA3 1X adaptor, there are 4 ports
    - But still only 16 CHPIDs on the adapter
    - However, 1X adapters support 32 subchannels per CHPID, so for long distance, there is less need to define large numbers of CHPIDs.

Note - it is the number of DEFINED CHPIDs that determines the link mode, NOT the number of online CHPIDs

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand planning

NOTE: If you do a dynamic reconfiguration that increases the number of CHPIDs on an HCA3 port above 4, ALL CHPIDs ON THAT PORT WILL GO OFFLINE FOR about 10 SECONDS, then come back online automatically.

IF those CHPIDs were the last ones connecting to a CF, they WILL go offline and you WILL NOT get a WTOR asking if it is OK to take the last path offline.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

Coupling Link Capacity Planning

- For customers wanting to perform self analysis of their coupling links capacity requirements, the currently available documentation may be inadequate.
- Information APAR II14483 was released to help with proper capacity planning for coupling links.
- CF link % busy alone is not a valid indicator of use, it does not tell the complete story for CF link capacity planning and can be misleading.
  - Review RMF CF reports to determine causes for link contention.
- Strongly recommend working with your IBMer to use zCP3000 for coupling link capacity planning.

Additional reference information can be found in:

- z/OS RMF Report Analysis, SC33-7991
- InfiniBand Coupling Links on System z, SG24-7539 (currently being updated)

Redbooks Workshop

InfiniBand planning

"Over-configuring" coupling links.....

z/OS supports dynamic reconfiguration, so you can add coupling links without a POR of the z/OS processor.

Standalone CFs do NOT currently have a dynamic reconfiguration capability.

Prior to z/OS 1.13, every coupling CHPID had to have an associated AID/Port and be coupled with a coupling CHPID on the other "end".

This means that every time you need to add a new link OR a new coupling CHPID, you must POR the CF processor.

- Not the end of the world, because you can empty and repopulate a CF non-disruptively.
- However it is not ideal.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

"Over-configuring" coupling links..... (a.k.a. planahead)

Starting with z/OS 1.13, HCD provides the ability to define placeholder coupling links, pointing to an AID of "*".

- This is a similar concept to placeholder LPARs.

You still need to pre-install the adapters on the CF end, and plug the correct AID/port into the definition, however on the z/OS end, you can use an AID of * (but valid port number must be assigned), and then perform a dynamic reconfiguration to fill in the real AID later, when the adapter is installed.

- This allows you to add an adapter in the future and start using it WITHOUT doing a POR of the standalone CF processor.

Redbooks Workshop

InfiniBand planning

Driver 93 supports 32 link buffers for HCA2 1X and HCA3 1X links.

For ALL coupling link types, to get the best performance and value from the available link buffers, each z/OS sharing a link should have a number of subchannels that matches the number of link buffers in the hardware.

So, HCD now supports either 7 or 32 subchannels for CHPID type CIB.

HOWEVER... 12X links still support just 7 link buffers (and therefore 7 subchannels).

But HCD doesn't know the difference between 12X and 1X links.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

Because it must default to *some* value, HCD now defaults to 32 subchannels per CIB CHPID.

```
  Goto  Filter  Backup  Query  Help
-                 Add CF Control Unit and Devices                    -

C
    Confirm or revise the CF control unit number and device numbers
S   for the CF control unit and devices to be defined.

S   Processor ID . . . . . . . : SCZP301
S   Channel subsystem ID . . . : 2
S   Channel path ID  . . . . . : 9E          Operation mode . . : SHR
    Channel path type  . . . . : CIB

/   Control unit number  . . . . FFF2  +
_
_   Device number  . . . . . . . FD3C
/   Number of devices  . . . . . 8
_
_
_    F1=Help      F2=Split     F3=Exit      F4=Prompt    F5=Reset
_    F6=Previous  F9=Swap     F12=Cancel
_
_
_   Invalid number of 8 devices owned by channel path 2.9E of type CIB on
_   processor SCZP301. 7 or 32 devices expected.
_
_ A9    Y  CIB  SHR  N      SCZP301.2  A5    N  CIB  SHR     CFP   7
_ AA    Y  CIB  SHR  N      SCZP301.2  A6    N  CIB  SHR     CFP   7
******************************* Bottom of data ********************************


 F1=Help      F2=Split     F3=Exit      F4=Prompt    F5=Reset     F7=Backward
 F8=Forward   F9=Swap     F10=Actions  F12=Cancel   F13=Instruct F22=Command
```

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand planning

Considerations for subchannel numbers:

- When connecting two CIB CHPIDs on z196 at Driver 93 or later, number of subchannels will default to 32.
  - If channels are 1X (either HCA2 or HCA3), accept the default.
  - If channels are 12X, you should override the default and select 7.
    - If you do NOT do this, HCD will not complain, but now you have more subchannels than there are link buffers, and this may result in increased Path Busy conditions. This is NOT recommended.
- If you move an existing z196 to Driver 93, the number of link buffers in the hardware will increase, but the number of subchannels will NOT increase unless you explicitly make this change in HCD.
- When connecting a z196 to some earlier generation, HCD knows that that generation does not support 32 subchannels, and therefore defaults to 7.
  - When that other processor is upgraded to z196, remember to go into HCD and change the number of subchannels to 32 for any 1X links.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

When you create your configuration in HCD, you need to:

- Define the CHPIDs, associating each one with an AID/Port
- Couple the CHPIDs

Is the following configuration valid?

AID/Port	CHPID	Coupled to	CHPID	AID/Port
08.1	00		80	18.1
08.1	01		81	18.1
08.1	02		82	18.2
08.2	03		83	18.2

Redbooks Workshop

InfiniBand planning

HCD checks that all the coupled CHPIDs are using the same AID/Port pair.

If it finds a mismatch, it issues a WARNING:

CBDG542I The following CIB channel paths of processor SCZP301 connect the same HCA port 1B.2 with different target HCA ports: 2.8D, 2.8A

Note, however, that this warning will NOT stop HCD from creating a production IODF.

The CHPIDs with the incorrect AID/Port will not be usable, AND you may see a (misleading) status of Loss of Signal for those VCHIDs on the SE.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

CPATH and CSYSTEM.....

When HCD creates IOCP statements to define your configuration, it uses CSYSTEM and CPATH statements to identify the processor and CHPID that each coupling CHPID is to connect to..

The CSYSTEM value comes from the LSYSTEM value that you specify for a processor. If you don't specify an LSYSTEM value, CSYSTEM is the CPC name of the processor (2nd part of SNA name).

What happens when you install a new CPC?

- It normally gets a new SNA name.

What happens if you have not specified an LSYSTEM value, and you install a new system?

- All the coupling definitions have to be changed.

THEREFORE, we strongly recommend specifying an LSYSTEM value that will be carried from one processor over to its replacement.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand planning

Because z/OS currently is unaware of the InfiniBand infrastructure that sits below the CHPIDs, any functions that look for single points of failure in a coupling configuration will NOT be able to know if two CHPIDs are on the same InfiniBand port, or the same InfiniBand adapter, or different adapters.

Therefore, you must be extra careful to provide a configuration that does not contain any single points of failure:

- Two physical InfiniBand links between every pair of connected processors
- Links should be connected to TWO adapters
- If possible, use CHPID naming that gives an indication of the AID/Port that is being used (to make it easier for operators or automation to determine if two online ports share a SPOF).

ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand planning

The performance of System Managed Duplexing on HCA2 is noticeably worse than on ICB4.

- This improves with HCA3 in IFB3 mode
  - However you can't go directly from ICB4 to HCA3/IFB3 unless you upgrade all CPCs at the same time.
- This might be an opportune time to re-evaluate your Coupling configuration, and see if a standalone CF (z114, for example) would not be more cost effective overall (bearing in mind cost of SM Duplexing in terms of CF CPU capacity, z/OS CPU capacity, software costs for z/OS MIPS, impact on batch job elapsed times)

Redbooks Workshop

InfiniBand planning

REMINDER FOR ANYONE WITH ICP (internal) LINKS

The z/OS LPARs AND the CF LPAR should be defined on BOTH ends of the link.

- Remember that peer mode links act as both sender and receiver, so both CF and z/OS systems should be on both ends of the link.
- It is very common to see that systems are NOT configured in this way.

Generally speaking, two ICP links (4 CHPIDs) should be enough for most sysplexes.

- One exception is if you are using SM Duplexing, especially over large distances

See the PR/SM Planning Guide for the latest recommendations on the number of ICP links for a given configuration

Infiniband

Little history lesson

Advantages of InfiniBand compared to other coupling types

What's new with z196 GA2/z114

Planning considerations

Performance information

Operation and management

STP enhancements with new InfiniBand adapters

Objectives

To provide comparative (not benchmark) measurements, to help customers understand the benefit of faster links (in general) and HCA3 specifically

Help customers understand the impact of moving off ICB4 connections to HCA2 and on to HCA3

Investigate relative performance of ICB4, HCA2, and HCA3 for SM Duplexing

Physical configuration

z10 E26 712
ICB4
PSIFB 12X
4 CF LPARs
  2 DED ICFs each
2 z/OS LPARs
  3 DED CPs each

z196 M32 716
ISC
PSIFB HCA2 1X
PSIFB HCA2 12X
PSIFB HCA3 1X
  8 ports over 2 cards
PSIFB HCA3 12X
  4 ports over 2 cards
2 CF LPARs
  1 DED ICFs each
2 z/OS LPARs
  3 DED CPs each

Logical configuration

z/OS 1.12

CF - Level 16 on z10, Level 17 on z196. Current service levels on each box

z10 running Driver 79

z196 running Driver 93 (GA2)

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Logical configuration

Workload consists of "Acme" thrasher jobs used by Poughkeepsie performance team:

- Objective is purely to generate CF requests designed to mimic those created by DB2, IMS, MQ, Logger, and GRS.
  - There is no subsystem or other processing of the requests or their responses (no DB2, IMS, MQ, or whatever)
- Keywords let you control the target number of requests per second, request size, read/write ratio, and more.
- Produce more repeatable (and therefore more comparable) results than using "real" DB2, IMS, or MQ workloads.
- Programs support System Managed Duplexing, so simplex and duplex can be compared.
- Each program drives requests serially - next one not sent until previous one completed.

Redbooks Workshop

Workload

Structure	Target request rate / system
IMS Shared Message Queue	1000
DB2 GBP 4KB heavy write bias	2500
DB2 GBP 32KB heavy read bias	2500
DB2 GBP 4KB heavy read bias	2500
DB2 GBP 4KB read bias	2500
DB2 GBP 4KB read bias	2500
DB2 Lock	10000
GRS Lock	5000
IMS Cache 4 KB read bias	3000
IMS Cache 4KB read boas	3000
IMS Lock	5000

Structure	Target request rate / system
Logger 1KB mainly writes	1000
Logger 4KB only writes	1000
MQ	500
MQ	500
MQ	500
MQ	500
MQ	500
Total	44000

Measurements

Runs

- z10:
  - ICB4 on z10
  - ICB4 on z10 with SM Duplex for IMS and DB2 lock structures
  - 12X PSIFB on z10 with 4 CHPIDs/port, 2 ports
  - 12X PSIFB on z10 with 4 CHPIDs/port, 2 ports, SM Duplexing
- z196
  - ISC3 on z196
  - HCA2 12X on z196
  - HCA2 12X on z196, SM Duplexing
  - HCA2 1X on z196
  - HCA3/IFB3 12X on z196
  - HCA3/IFB3 12X on z196, SM Duplexing
  - HCA3/IFB 12X on z196
  - HCA3 1X on z196

Measurements

Runs

- All these measurements were taken with 2 z/OS systems.
- The z/OS systems and CFs were in the same CEC in all cases.
- In all cases, the CF link CHPIDs were shared between the z/OS LPARs.
- In nearly all cases, we had 2 InfiniBand CHPIDs per link, and two links shared between the z/OS LPARs.
  - ICB4 measurements used 2 ICB4 CHPIDs shared between the z/OS LPARs
  - ISC measurements used 8 ISC links, with 4 links dedicated to each z/OS LPAR.
- In all SM Duplexing measurements, only the lock structures were duplexed.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

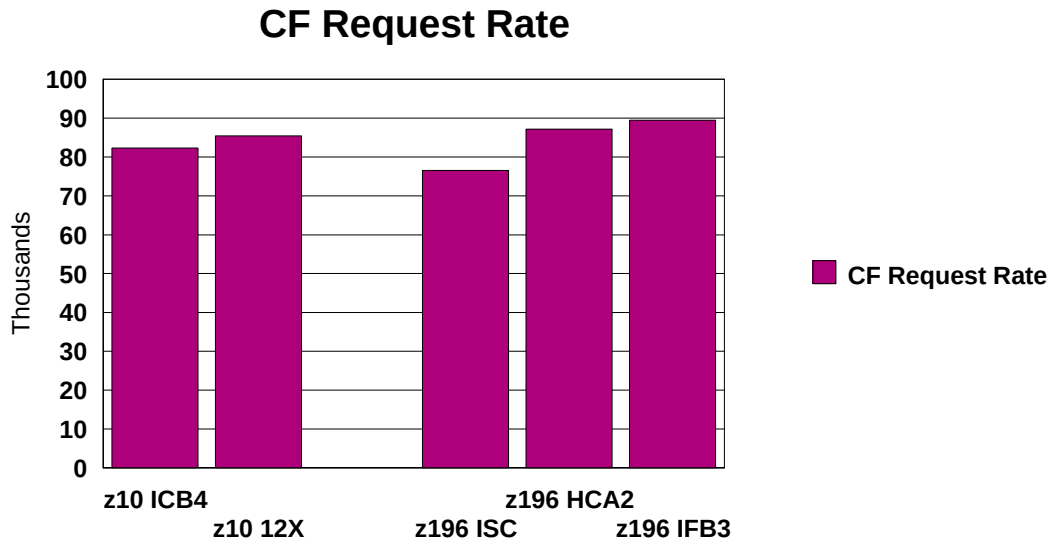
Metrics

What we used to understand the differences:

- Request rate
- % of requests that were synchronous (the more, the better)
- Average synchronous response time
- CF CPU busy
- CF CPU per request

Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

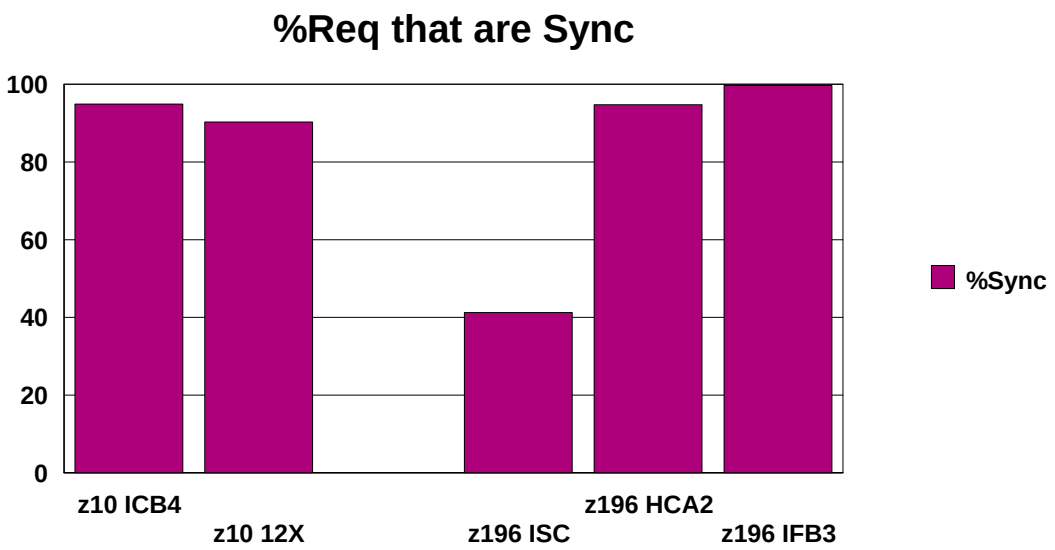
Results



z196 IFB3 represents HCA3 link running in IFB3 mode

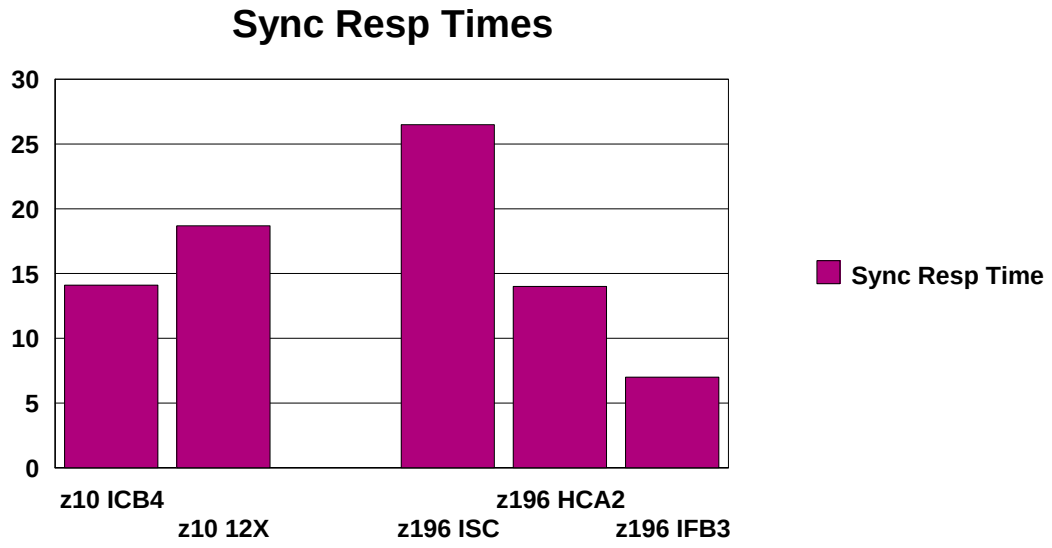
ibm.com yourdotcom International Technical Support Organization and Authoring Services

Results



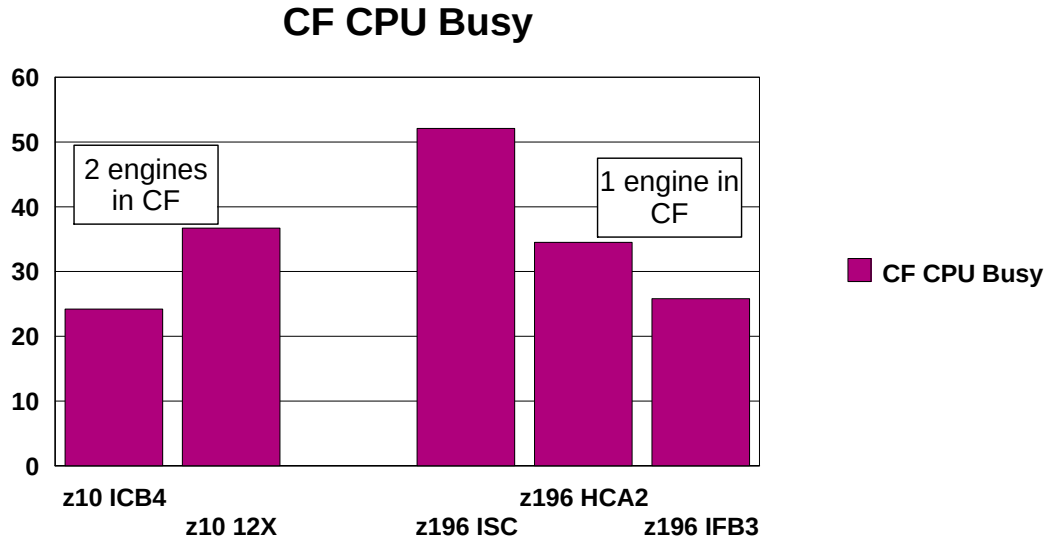
Remember that, due to the XES heuristic algorithm, a sync request will nearly always take equal or less z/OS CPU than an asynch one.

Results



ibm.com yourdotcom International Technical Support Organization and Authoring Services

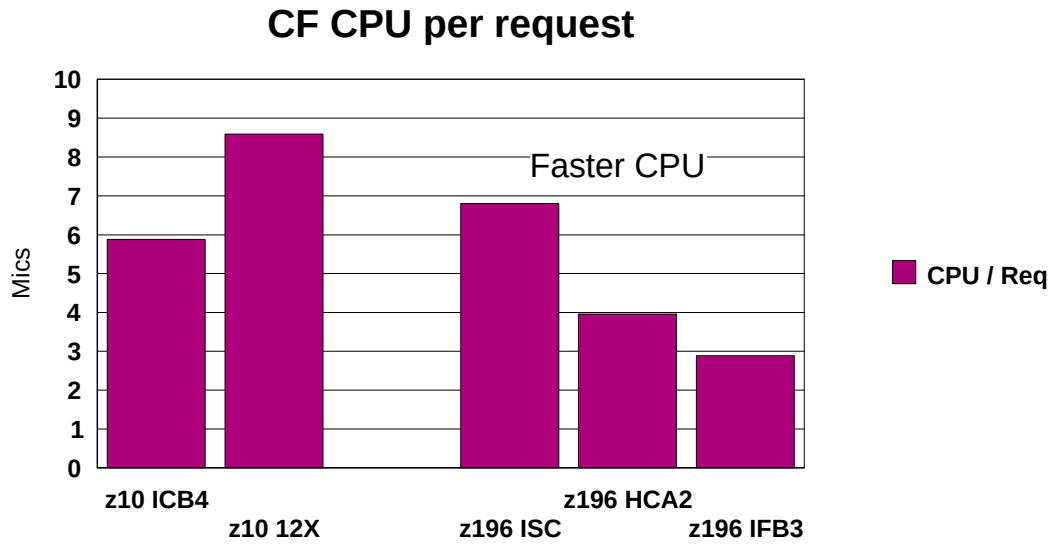
Results



CF utiizations are NOT normalized

Redbooks Workshop

Results



CF CPU times are NOT normalized

ibm.com yourdotcom International Technical Support Organization and Authoring Services

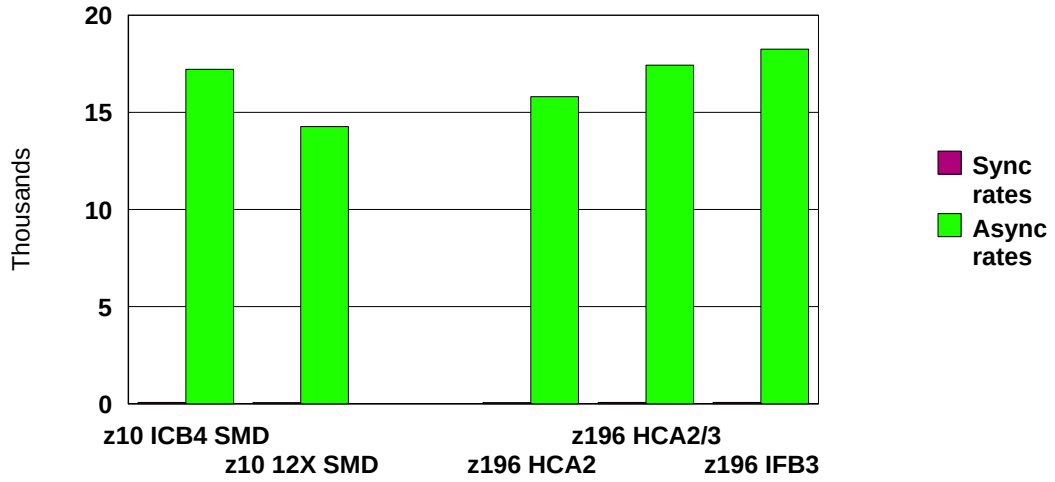
Results

Now let's look more closely at the impact of System Managed Duplexing

Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

SM Duplexing Results

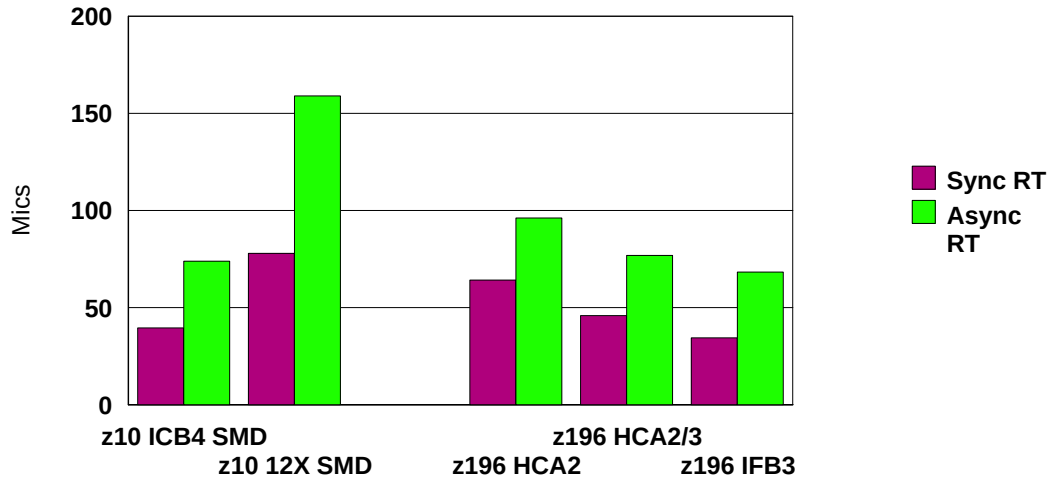
Duplexed Structure Request Rates



ibm.com yourdotcom International Technical Support Organization and Authoring Services

SM Duplexing Results

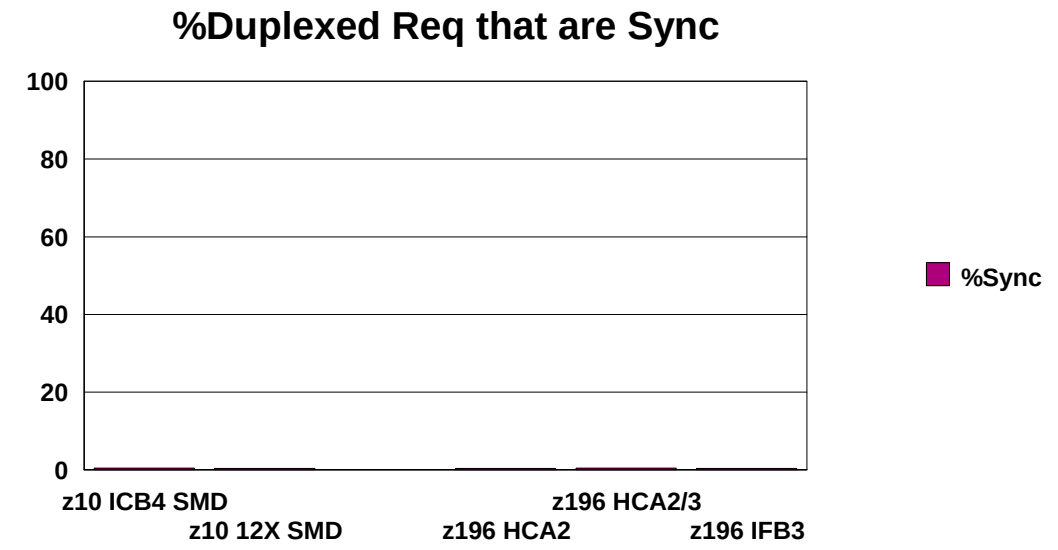
Duplexed Structure Response Times



Response times for primary structures

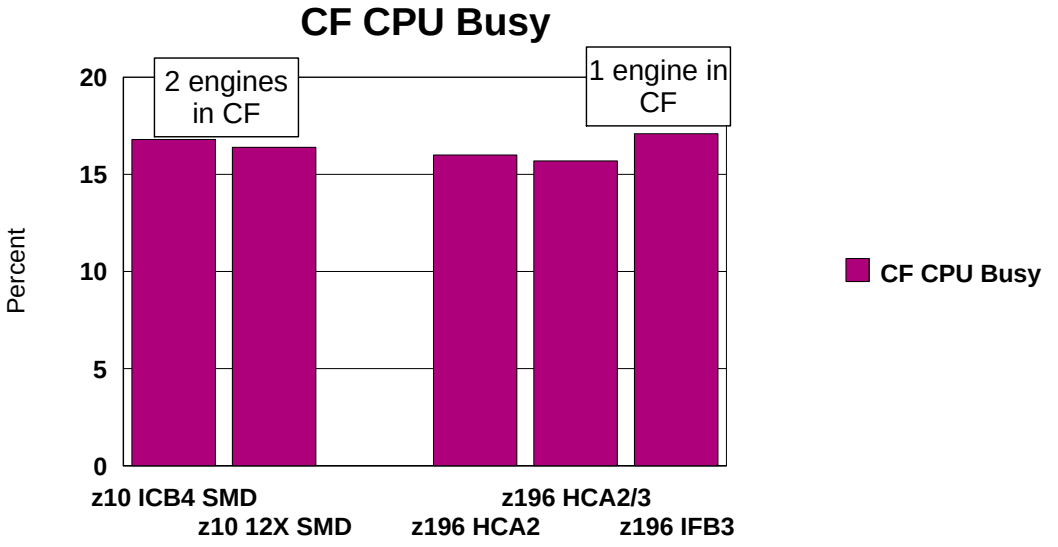
Redbooks Workshop

SM Duplexing Results



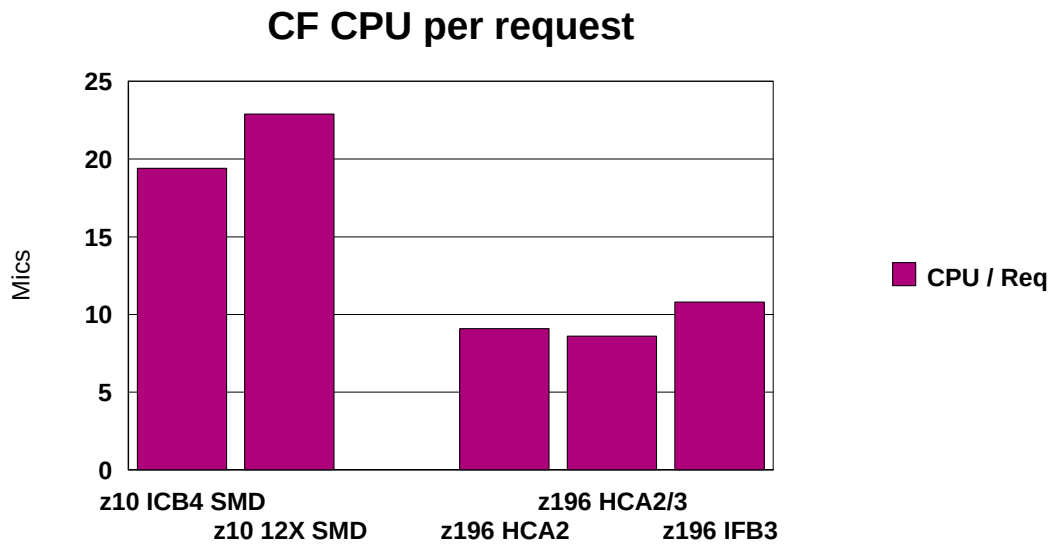
ibm.com yourdotcom International Technical Support Organization and Authoring Services

SM Duplexing Results



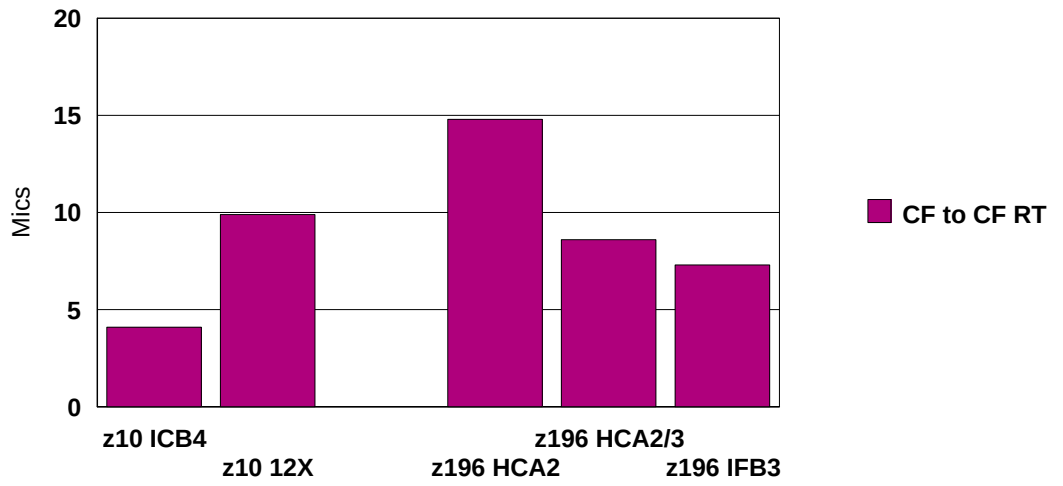
Redbooks Workshop

SM Duplexing Results



ibm.com yourdotcom International Technical Support Organization and Authoring Services

SM Duplexing Results



Redbooks Workshop

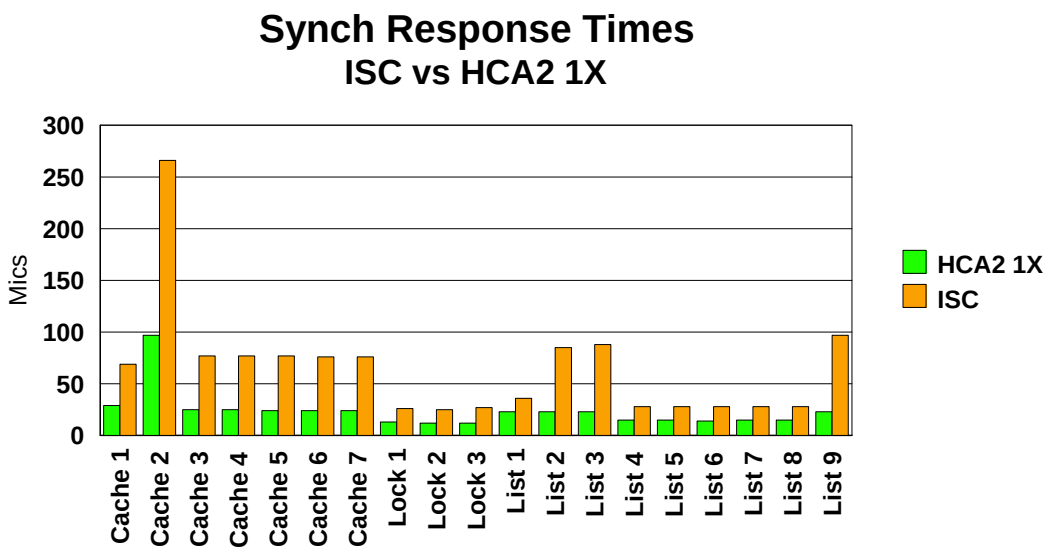
InfiniBand measurement results

What about 1X InfiniBand links?

- How do they compare to ISC links?
- How do they compare to HCA2 12X links?

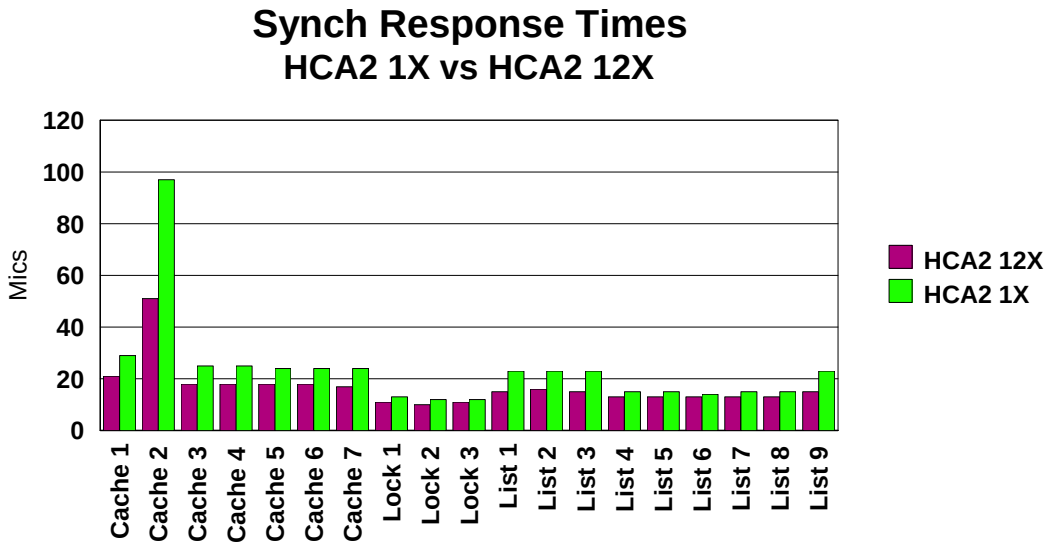
ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand 1X links



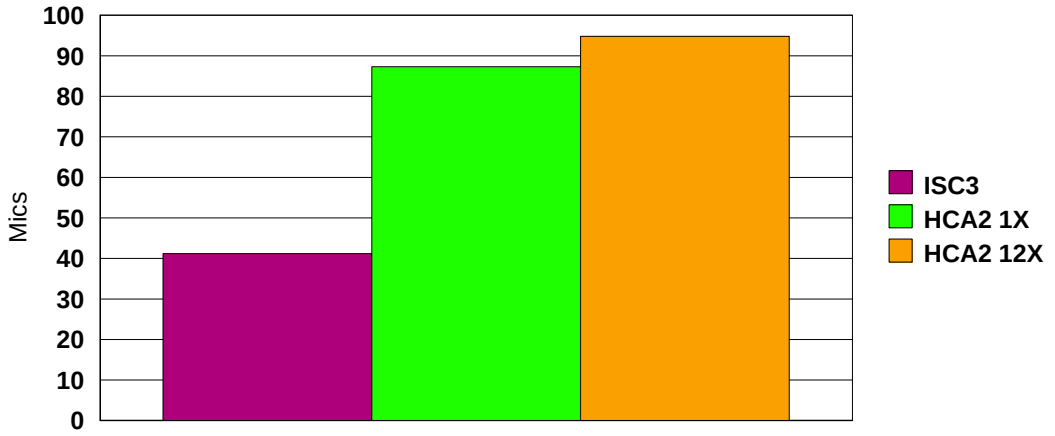
Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

InfiniBand 1X links

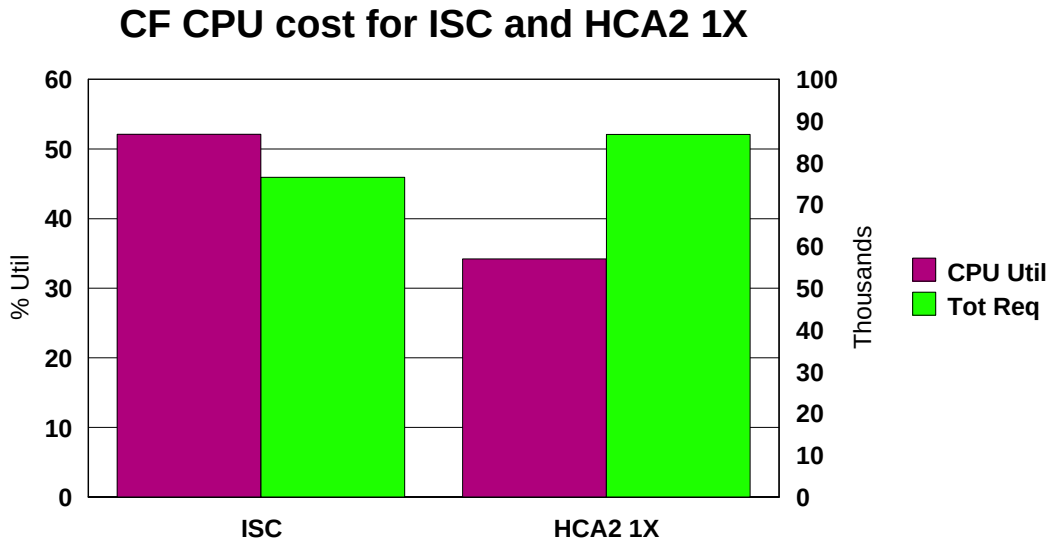


ibm.com yourdotcom International Technical Support Organization and Authoring Services

InfiniBand 1X links



InfiniBand 1X links



InfiniBand Performance Measurements

Summary of performance comparisons:

- Response time of HCA3 12X links in IFB3 mode is a significant improvement over IFB mode:
  - May deliver noticeable improvements for large batch jobs
  - For large CF users, IFB3 may result in significant z/OS CPU savings
    - In our case (about 90K requests a second and an average of 7 mics improvement in response time), IFB3 would have saved us about .6 of a z/OS CP.
- For large customers, the benefit of IFB3 mode is so significant that it may be cost-effective to buy more HCA3 adapters rather than defining more than 4 CHPIDs per HCA3 port.
- Faster links consistently resulted in noticeably reduced CF CPU consumption.
- Performance of System-Managed Duplexing on HCA2 12X is not as good as ICB4. However performance of HCA3 IFB3 mode is similar to ICB4.

Infiniband

Little history lesson

Advantages of InfiniBand compared to other coupling types

What's new with z196 GA2/z114

Planning considerations

Performance information

Operation and management

STP enhancements with new InfiniBand adapters

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Managing an InfiniBand infrastructure

Infiniband brings many advantages, but there is no free lunch. Along with the flexibility comes greater complexity.

In ISC and ICB, there is only one CHPID per port.

If you display an ICB link, you see something like this:

```
D CF
IXL150I  16.35.29  DISPLAY CF 450
COUPLING FACILITY 002097.IBM.02.00000001DE50
                  PARTITION: 0D  CPCID: 00
                  CONTROL UNIT ID: FFCF
NAMED CF04
COUPLING FACILITY SPACE UTILIZATION
 ALLOCATED SPACE                 DUMP SPACE UTILIZATION
  STRUCTURES:         832 M       STRUCTURE DUMP TABLES:          0 M
  DUMP SPACE:           2 M                 TABLE COUNT:          0
 FREE SPACE:          863 M      FREE DUMP SPACE:                 2 M
TOTAL SPACE:         1697 M     TOTAL DUMP SPACE:                 2 M
                                MAX REQUESTED DUMP SPACE:         0 M
...
SENDER PATH       PHYSICAL              LOGICAL      CHANNEL TYPE
  BA / 0019        ONLINE                ONLINE       CBP
  BB / 001B        ONLINE                ONLINE       CBP
```

Target CF LPAR

PCHID

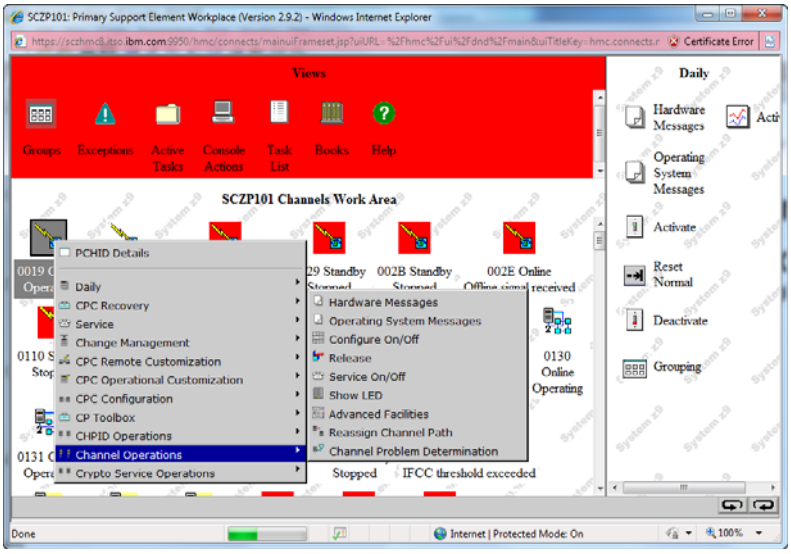
Managing an InfiniBand infrastructure

And coming from the hardware side, if you need to understand who is using a physical link or adapter, you can use the SE to determine which CHPID is associated with a PCHID, and the LPARs that are sharing that CHPID....

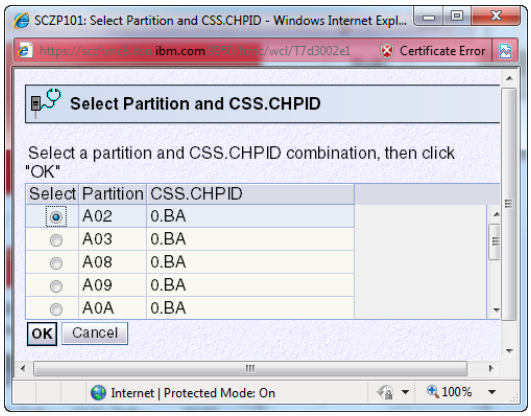
ibm.com yourdotcom International Technical Support Organization and Authoring Services

Managing an InfiniBand infrastructure

Using the SE (Single Object Operations), find out who is using a given PCHID.....



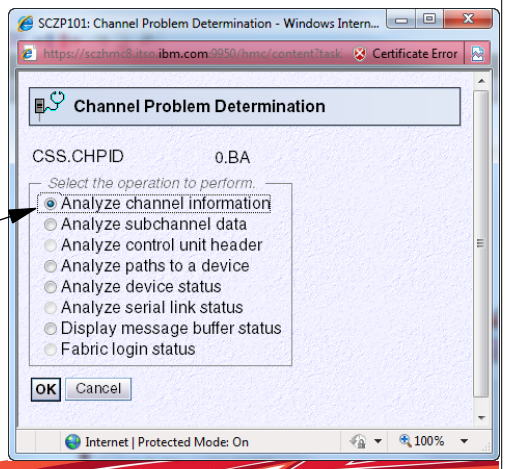
Redbooks Workshop
©2011 IBM Corporation. All rights reserved.



List shows us the list of LPARs that are sharing this CHPID, as well as the CHPID associated with this PCHID.

For more information, select any of the LPARs

Then select Analyze channel information

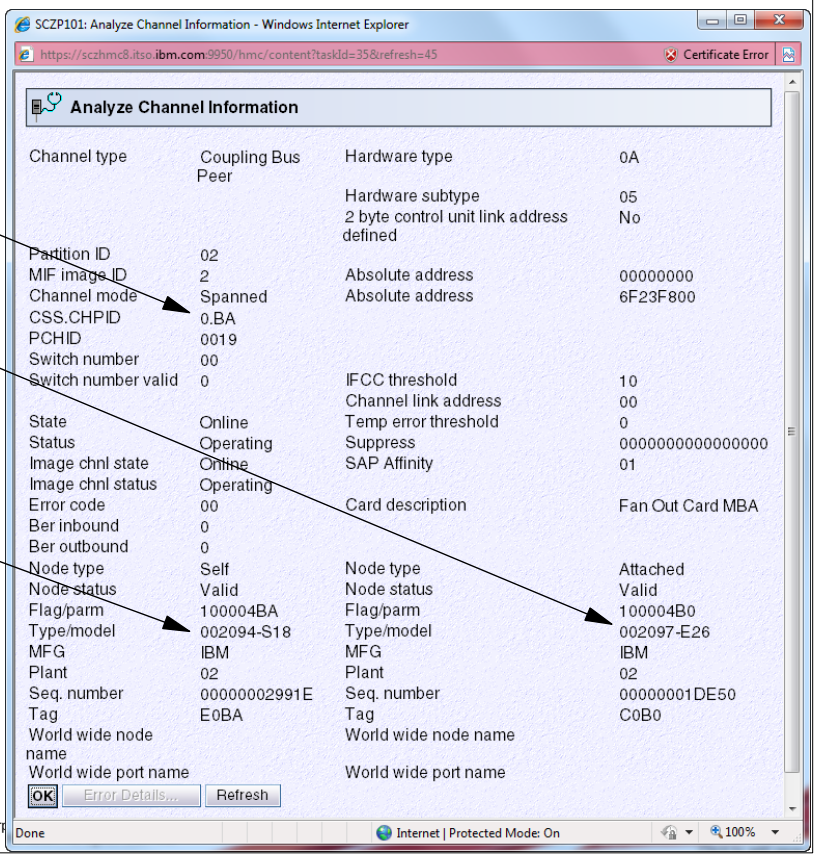


ibm.com yourdotcom International Technical Support Organization and Authoring Services

This CSS and CHPID

What CPC is at the other end of this link

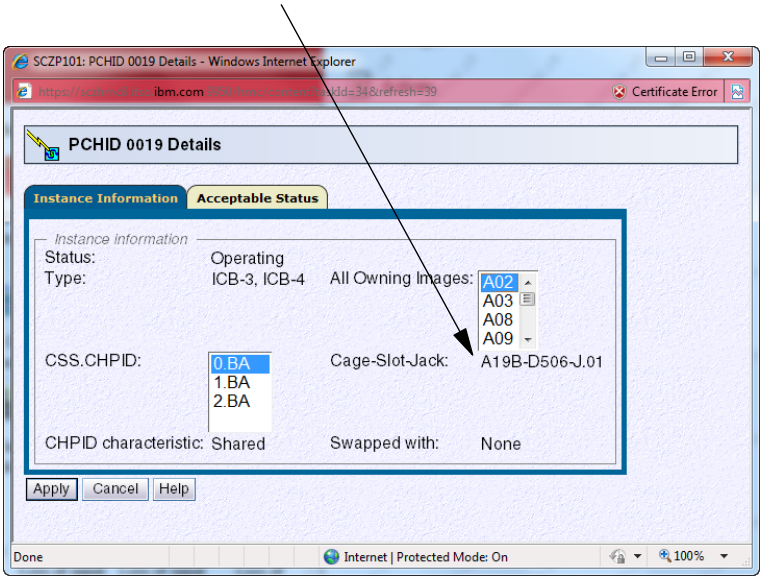
This CPC



Redbooks Workshop

Managing an InfiniBand infrastructure

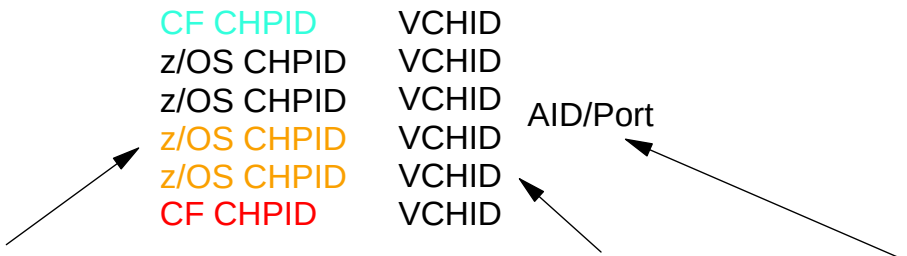
Also on the SE, get a list of the LPARs associated with the PCHID and the physical location of the port



Managing an InfiniBand infrastructure

You now know exactly which LPARs will be affected by any changes you need to make.

However, with InfiniBand links, there are more levels in the path...



Multiple sysplexes, multiple systems, multiple VCHIDs, but just one cable

So how do you understand the relationship between cables/ports and who is using them?

Managing an InfiniBand infrastructure

Coming from the top (z/OS) down is not too bad...

Display CF information

```
D CF,CFNM=FACIL04
IXL150I  21.05.54  DISPLAY CF 708
COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                  PARTITION: 2F  CPCID: 00
                  CONTROL UNIT ID: FFF2
NAMED FACIL04
COUPLING FACILITY SPACE UTILIZATION
...

SENDER PATH       PHYSICAL              LOGICAL       CHANNEL TYPE
  81 / 071D        ONLINE                ONLINE        CIB
  82 / 0700        ONLINE                ONLINE        CIB
  83 / 071E        ONLINE                ONLINE        CIB
  84 / 0701        ONLINE                ONLINE        CIB
  87 / 0702        ONLINE                ONLINE        CIB
  91 / 072B        ONLINE                ONLINE        CIB
  92 / 072C        ONLINE                ONLINE        CIB
  97 / 072F        ONLINE                ONLINE        CIB
```

CHPIDs

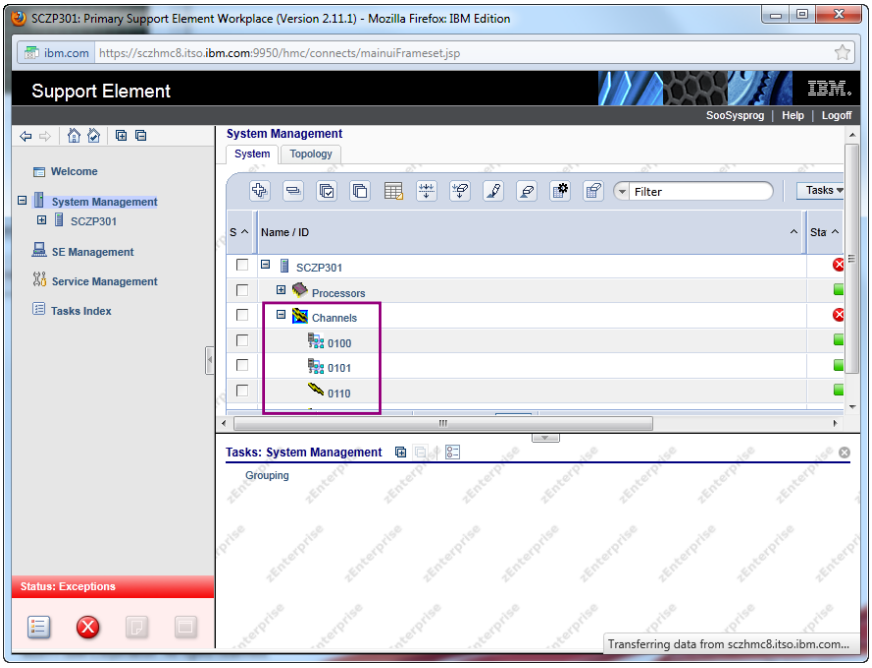
VCHIDs

Note that z/OS doesn't know the exact link type or the AID or Port

ibm.com yourdotcom International Technical Support Organization and Authoring Services

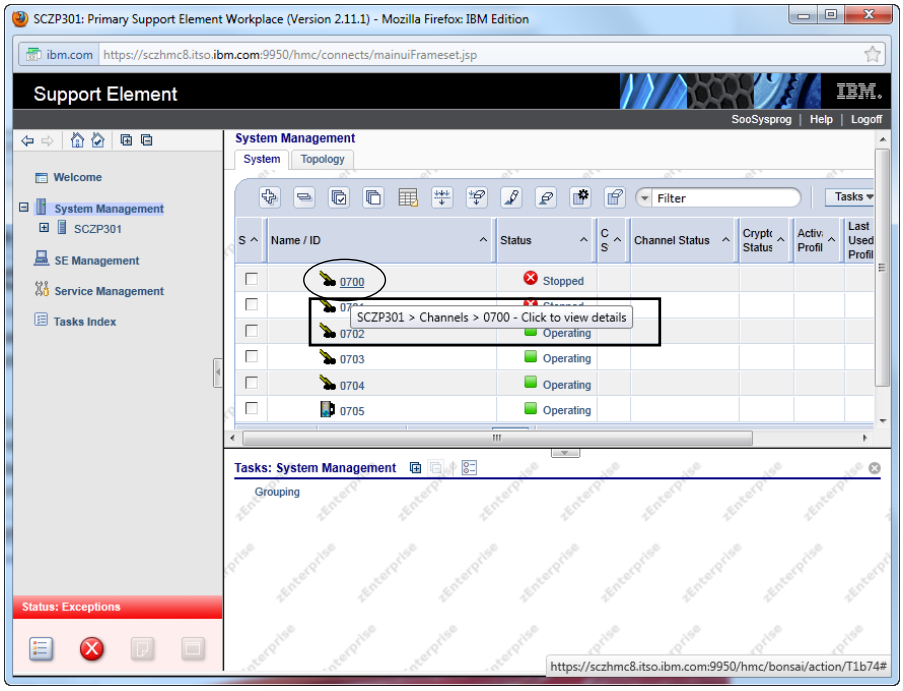
To get information about the underlying links, logon to SE, select System Management, expand CPC, expand Channels

PCHIDs and VCHIDs are listed



Scroll down to find VCHID listed on D CF command

Then click on that VCHID to get details

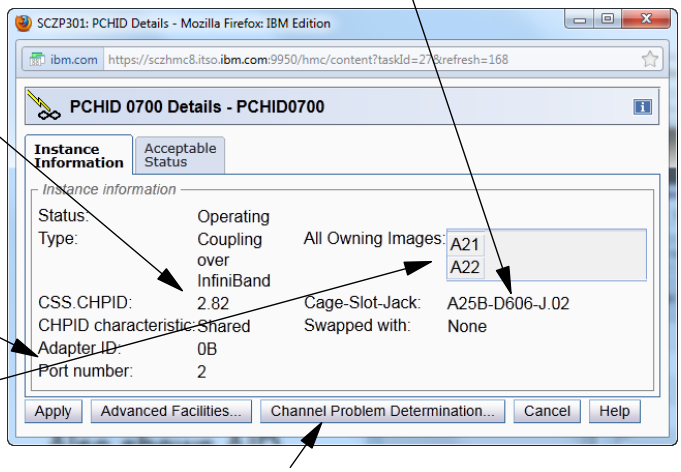


This provides info about the VCHID - the CSS and CHPID that is assigned to it (note that there is only one CHPID per VCHID).

Also shows the AID and port that this CHPID is using.

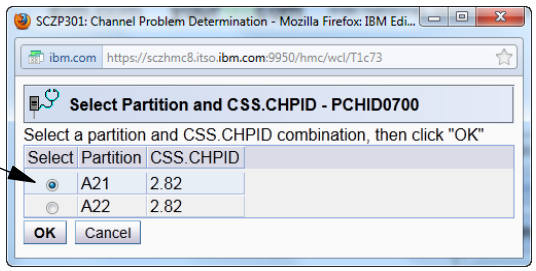
Also has a list of the LPARs that can use that VCHID

And the physical location of this AID and port

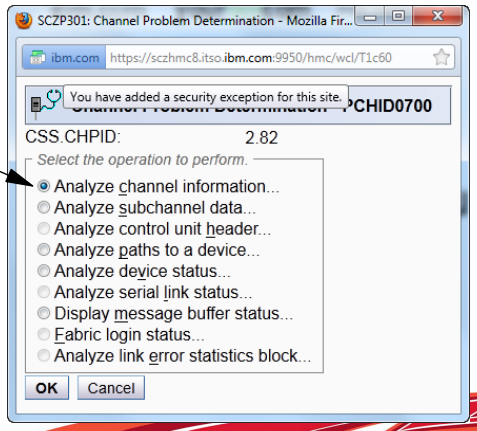


For more information, click Channel Problem Determination

Select the LPAR you are interested in



Then select Analyze Channel Information



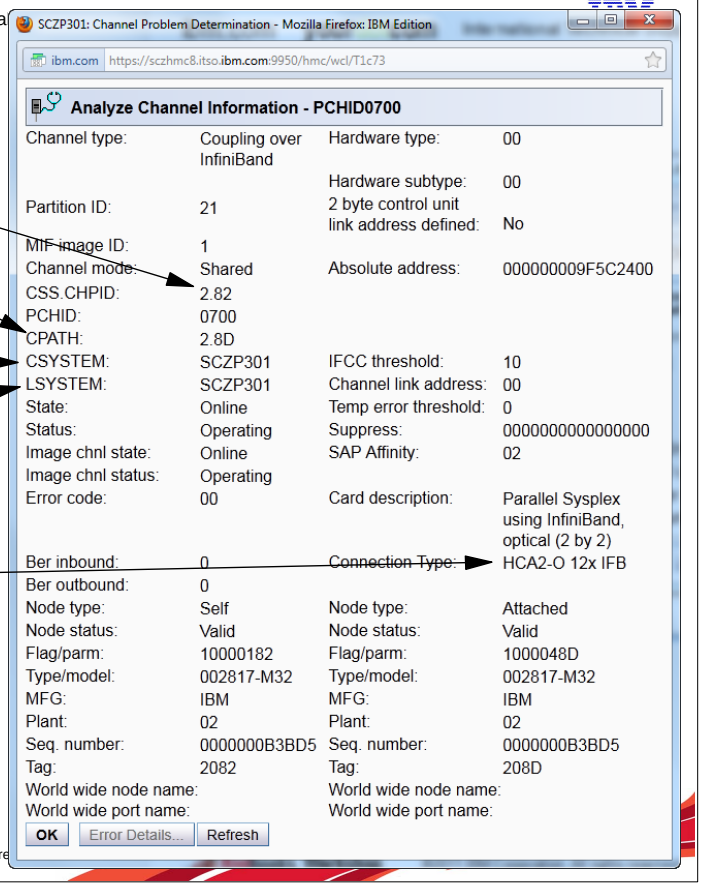
CSS and CHPID associated with this VCHID

Target CSS and CHPID

Target CPC

This CPC

HCA type and mode (this is the only place you will get this information)



Managing an InfiniBand infrastructure

This showed the flow from MVS, to the VCHID, and down to the AID and port, and on to the connected system and CHPID.

However, coming from the HW up is not so easy....

The lowest level display on the SE is the VCHID - there is no way to display an adaptor to see what VCHIDs or LPARs are using it.

If you want to find out what is assigned to a port, the only way to do this currently is via HCD...

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Managing an InfiniBand infrastructure

On the channel path list panel:

```
  Goto  Filter  Backup  Query  Help
--------------------------------------------------------------------------------
                           Channel Path List     Row 1 of 127 More:        >
Command ===> _                                              Scroll ===> CSR

Select one or more channel paths, then press Enter. To add use F11.

Processor ID . . . . : SCZP301
Configuration mode . : LPAR
Channel Subsystem ID : 2

                    DynEntry Entry +
/ CHPID Type+ Mode+ Switch + Sw Port Con Mngd Description
_ 00    OSD   SPAN  __       __ __      No   Exp3 1KBaseT All LPARs 9.12.4 #1
_ 01    OSC   SPAN  __       __ __      No   Exp3 1KBaseT All LPARs OSC #1
_ 06    OSD   SPAN  __       __ __      No   Exp3 1KBaseT
_ 08    OSD   SPAN  __       __ __      No   Exp3 1KBaseT
_ 09    OSD   SPAN  __       __ __      No   Exp3 1KBaseT Yellow zone
_ 0A    OSM   SPAN  __       __ __      No   Exp3 1KBaseT
_ 0B    OSM   SPAN  __       __ __      No   Exp3 1KBaseT
_ 0C    OSD   SPAN  __       __ __      No   Exp3 1KBaseT All LPARs 9.12.4 #2
_ 0D    OSC   SPAN  __       __ __      No   Exp3 1KBaseT All LPARs OSC #2
_ 0E    OSD   SPAN  __       __ __      No   Exp3 1KBaseT Yellow zone
_ 10    OSD   SPAN  __       __ __      No   Exp3 GbE SX
_ 11    OSD   SPAN  __       __ __      No   Exp3 GbE SX
_ 12    OSD   SPAN  __       __ __      No   Exp3 GbE SX
_ 13    OSD   SPAN  __       __ __      No   Exp3 GbE SX
_ 18    OSX   SPAN  __       __ __      No   Exp3 10GbE SR
 F1=Help       F2=Split      F3=Exit      F4=Prompt     F5=Reset      F7=Backward
 F8=Forward    F9=Swap      F10=Actions   F11=Add       F12=Cancel    F13=Instruct
F20=Right     F22=Command
```

Select Filter

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Managing an InfiniBand infrastructure

In the AID/Port field, enter the AID and port that you are interested in:

```
  Goto  Filter  Backup  Query  Help
--------------------------------------------------------------------------------
                              Channel Path List
Command ===> ____________________________________________ Scroll ===> CSR

Select one or more channel paths, then press Enter. To add use F11.

Processor ID . . . . : SCZP301
Configuration mode . : LPAR
Channel Subsystem ID : 2
                        Filter Channel Path List
/ C
_ 0   Specify or revise the following filter criteria.
_ 0
_ 0   Channel path type  . ___
_ 0   Operation mode . . . _____  +
_ 0   Managed  . . . . . . _     (Y = Yes; N = No)   I/O Cluster _________  +
_ 0   Dynamic entry switch __     +
_ 0   Entry switch . . . . __     +
_ 0   CF connected . . . . _     (Y = Connected; N = Not connected)
_ 0   PCHID or AID/P . . . 08/1
_ 0
_ 1   Description  . . . . _____________________________
_ 1
_ 1   Partition  . . . . . ________  +
_ 1   Connected to CUs . . _     (Y = Connected; N = Not connected)
_ 1
_ 1
 F1   F1=Help      F2=Split     F3=Exit      F4=Prompt    F5=Reset     F9=Swap
 F8   F12=Cancel
```

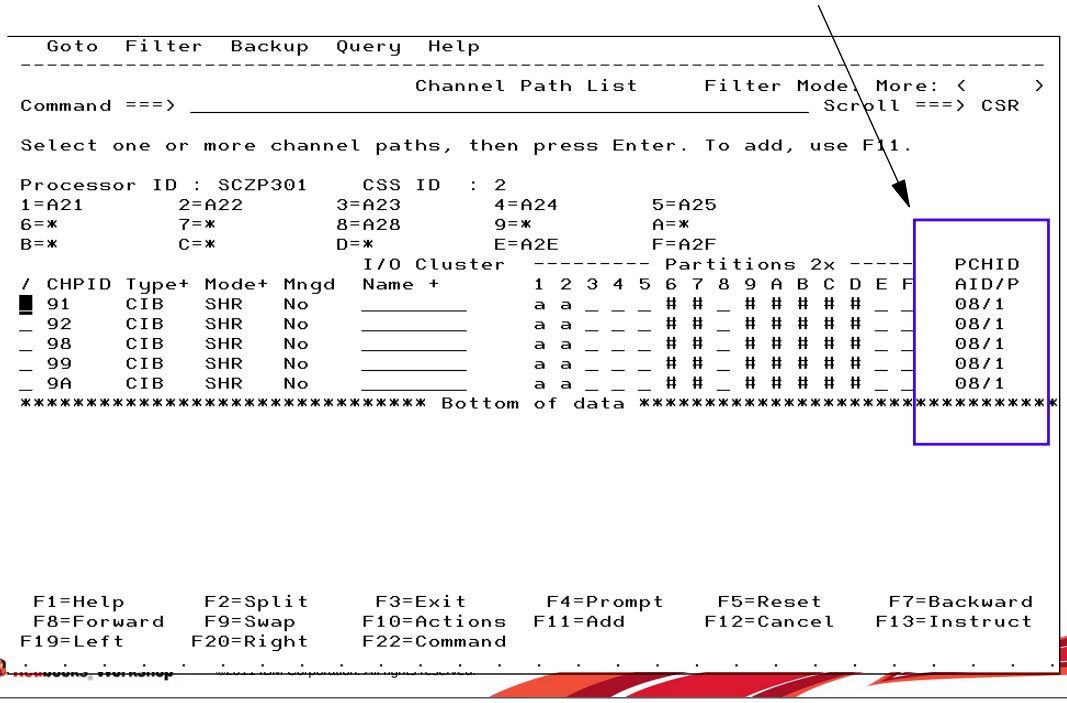

Managing an InfiniBand infrastructure

This will present you with a list of all the CHPIDs in this CSS that are assigned to the named port.....

```
  Goto  Filter  Backup  Query  Help
--------------------------------------------------------------------------------
                              Channel Path List     Filter Mode. More:       >
Command ===> ____________________________________________ Scroll ===> CSR

Select one or more channel paths, then press Enter. To add use F11.

Processor ID . . . . : SCZP301
Configuration mode . : LPAR
Channel Subsystem ID : 2

                     DynEntry Entry +
/ CHPID Type+ Mode+ Switch + Sw Port Con Mngd Description
_ 91    CIB   SHR    __        __ __   Y   No   IB 12x-3 Trainer loop
_ 92    CIB   SHR    __        __ __   Y   No   IB 12x-3 Trainer loop
_ 98    CIB   SHR    __        __ __   Y   No   IB 12x-3 Trainer loop
_ 99    CIB   SHR    __        __ __   Y   No   IB 12x-3 Trainer loop
_ 9A    CIB   SHR    __        __ __   Y   No   IB 12X-3 Trainer loop
******************************** Bottom of data ********************************

 F1=Help      F2=Split     F3=Exit      F4=Prompt    F5=Reset     F7=Backward
 F8=Forward   F9=Swap      F10=Actions  F11=Add      F12=Cancel   F13=Instruct
F20=Right     F22=Command
```

Managing an InfiniBand infrastructure

If you scroll to the right (PF20), you will see the AID/Port



Switching IFB modes

As stated previously, in order to run in IFB3 mode, HCA3 must be connected to HCA3, AND the port must not have more than 4 CHPIDs DEFINED to it.

If the number of CHPIDs goes from below 5 to above 4, all CHPIDs associated with that port will AUTOMATICALLY be taken offline to change modes, then online again.

Switching IFB modes

Starting configuration:

LPAR	CHPID	AID/Port	Target
$2/$3	87	18.2	FACIL04
$2/$3	91	08.1	FACIL04
$2/$3	92	08.1	FACIL04
$2/$3	98	08.1	FACIL05
$2/$3	99	08.1	FACIL05
$2/$3	9A	18.2	FACIL05

- Note there are 4 CHPIDs defined on 08.1. All those CHPIDs were in IFB3 mode

For the test, we moved CHPID 9A from port 18.2 to 08.1, bringing the number of defined CHPIDs on that port above 4.....

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Switching IFB modes

Update IODF

Take CHPID 9A offline on both z/OS images

Take CHPID corresponding CHPID offline on FACIL05

Activate new IODF

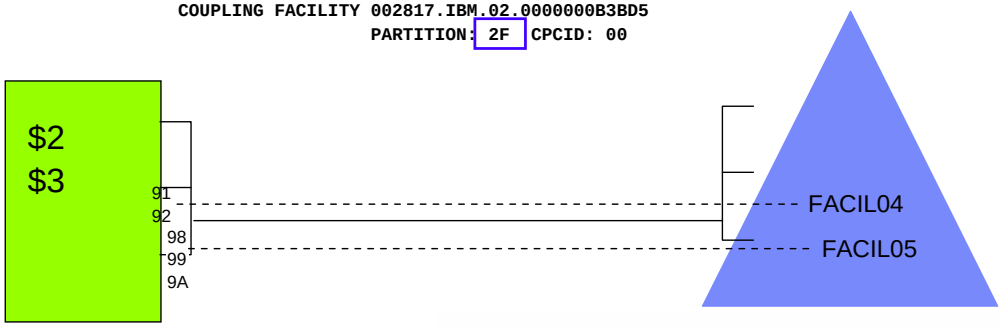
Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

Switching IFB modes

ALL CHPIDs associated with that port go offline....

```
*IXL158I PATH 98 IS NOW NOT-OPERATIONAL TO CUID: FFF9 956
                 COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                   PARTITION: 2E  CPCID: 00
*IXL158I PATH 99 IS NOW NOT-OPERATIONAL TO CUID: FFF9 957
                 COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                   PARTITION: 2E  CPCID: 00
*IXL158I PATH 92 IS NOW NOT-OPERATIONAL TO CUID: FFF2 958
                 COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                   PARTITION: 2F  CPCID: 00
*IXL158I PATH 91 IS NOW NOT-OPERATIONAL TO CUID: FFF2 959
                 COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                   PARTITION: 2F  CPCID: 00
```

Note, connections to TWO CFs were affected by the change



ibm.com yourdotcom International Technical Support Organization and Authoring Services

Switching IFB modes

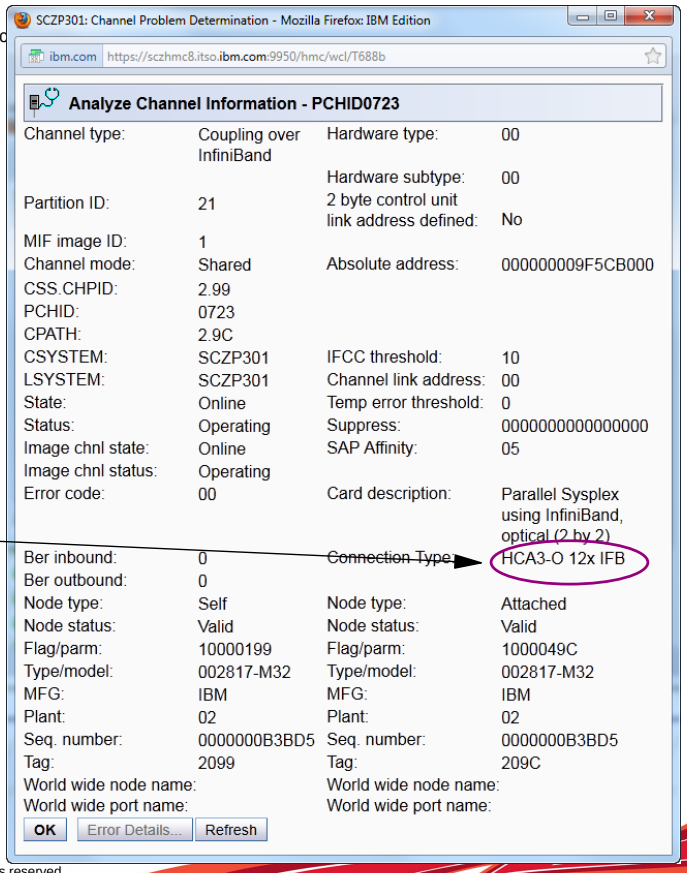
And then back online (10 seconds later)

```
IXL157I PATH 98 IS NOW OPERATIONAL TO CUID: FFF9 972
                COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                  PARTITION: 2E  CPCID: 00
IXL157I PATH 92 IS NOW OPERATIONAL TO CUID: FFF2 973
                COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                  PARTITION: 2F  CPCID: 00
IXL157I PATH 91 IS NOW OPERATIONAL TO CUID: FFF2 974
                COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                  PARTITION: 2F  CPCID: 00
IXL157I PATH 99 IS NOW OPERATIONAL TO CUID: FFF9 975
                COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                                  PARTITION: 2E  CPCID: 00
```

Remember that you have to manually take CHPID 9A and the corresponding CF CHPID back online manually

Redbooks Workshop

How do we KNOW they are in IFB mode?



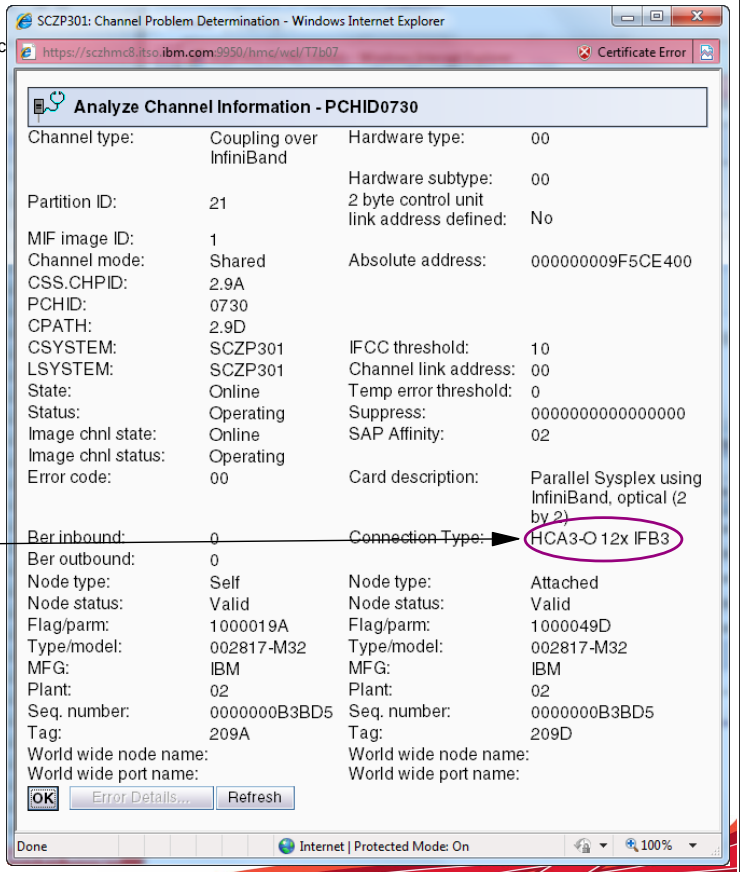
Switching IFB modes

Then we moved CHPID 9A back to AID/Port 18.2

- Take CHPID 9A offline on all z/OS
- Take the corresponding CHPID offline on the CF
- Do the soft ACTIVATE
- Do the HW ACTIVATE

Going from IFB to IFB3 mode, the other CHPIDs on that port did NOT toggle offline

To be sure the CHPID went back to IFB3 mode, we checked SE:



Switching IFB modes

We also tried this, with a configuration where the ONLY links to the CF were on the port that was affected by the change..

In that case, the CHPIDs were again taken offline, removing the last path to the CF.

Be VERY CAREFUL - no WTOR or message is issued to warn you that the change you are trying to make will remove the last path to the CF

Switching IFB modes

Summary.....

- Remember that any change to add or remove CHPIDs to a HCA3 12X port has the potential to be disruptive.
- Need to check the number of CHPIDs on ALL ports that are affected by the change, and verify if the change will increase the number of CHPIDs above 4, or decrease the number below 5.
  - HCM or HCD can be used to check the number of CHPIDs on the ports.
  - Should also be used to check which LPARs are using those CHPIDs
- As long as the change is planned, carried out at an off-peak time, and you ensur that alternate CHPIDs on other ports are online, there should not be any problems.
- Also, this only affects HCA3 12X ports that are connected to other HCA3 ports - HCA3 to HCA2, or HCA3 1X ports are not affected.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Infiniband

Little history lesson

Advantages of InfiniBand compared to other coupling types

What's new with z196 GA2/z114

Planning considerations

Performance information

Operation and management

STP enhancements with new InfiniBand adapters

Redbooks Workshop

STP-related InfiniBand enhancement

Summary of STP recovery rules (pre-Driver 93):

- CANNOT have two Stratum 1 servers in timing network
- Backup Time Server (BTS) can take over as Current Time Server (CTS), active Stratum 1, only if either:
  - Preferred Time Server (PTS) can indicate it has “failed”
  - BTS can unambiguously determine the PTS has “failed”
- In a Coordinated Timing Network (CTN) with only 2 servers, the state of the "old" CTS is
  - Based on combination of:
    - Server Offline Signal (OLS- Channel going away signal) and
    - Console Assisted Recovery (CAR)
  - Server Offline signal (OLS) is transmitted on a channel by the server to indicate that the channel is going offline
    - However OLS was never intended to play a role in STP
  - CAR is initiated when OLS not received but signals from the "old" CTS are lost
    - Uses HMC/SE LAN path to determine if BTS can take over as CTS
    - If HMC/SE on target CEC doesn't/is unable to answer, state of old CTS is ambiguous

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM

STP-related InfiniBand enhancement

Summary of STP recovery rules (pre-Driver 93):

- In a Coordinated Timing Network (CTN) with 3 or more servers:
  - If BTS loses communication on all established paths to CTS
    - BTS and Arbiter communicate to establish if Arbiter has also lost communication on all established paths to CTS
  - If both BTS and Arbiter cannot communicate with CTS
    - BTS takes over as CTS (S1)
  - If CTS loses communication with both BTS and Aribter (not necessarily at the same instant), it knows that they will assume that it is dead, so it goes to stratum 0.

Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

STP-related InfiniBand enhancement

Going Away Signal is a reliable unambiguous signal to indicate that the CPC is about to enter a check stopped state.

When a GOSIG from the CTS is received by the BTS:

- BTS safely takes over as CTS
- GOSIG has priority over OLS in a 2 server CTN
- BTS can also use GOSIG to take over as CTS for CTNs with 3 or more servers without communicating with Arbiter

Dependencies on OLS and CAR removed in a 2 server CTN

Dependency on BTS>Arbiter communication removed in CTNs with 3 or more servers

ibm.com yourdotcom International Technical Support Organization and Authoring Services

STP-related InfiniBand enhancement

Prerequisites:

- InfiniBand (IFB) links using
  - HCA3-O to HCA3-O
  - Operating in either IFB mode or IFB3 mode
- HCA3-O LR to HCA3-O LR
  - 1x IFB

Current recovery design still used when GOSIG not received by BTS and for other failure types

There are other STP-related enhancements in z196 GA/z114, however this is the only one that is related to the new InfiniBand links.

Other enhancements were covered in z196 session.

Redbooks Workshop
©2011 IBM Corporation. All rights reserved.

yourdotcom

International Technical Support Organization and Authoring Services

Latest developments in IPL Avoidance and MTTR improvements

ibm.com/redbooks

Redbooks Workshop

IPL Avoidance

Historically, customers have run many applications in a single z/OS system.

- Compared to the UNIX environment, where there tends to be one application per system

Many applications means many diverse users means that finding a time when NO one needs the system is very challenging.

Parallel Sysplex data sharing and dynamic workload balancing are the primary means of meeting the conflicting requirements of keeping systems up to date while not impacting application availability.

However, just because you CAN IPL a system doesn't mean that you WANT TO IPL it.

To give you the flexibility to extend the time between IPLs, IBM is continually trying to address situations where a change requires an IPL to implement it.

IPL Avoidance

The ITSO produced a Redbook, z/OS Planned Outage Avoidance Checklist, SG24-7328, in 2006 to document things that previously required an IPL, but now could be changed dynamically.

We have not had a chance to update that book since then, however this section lists some of the enhancements in this area since that book.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

IPL Avoidance

z/OS 1.8 Enhancements:

- Ability to move GRS Contention Notification System without an IPL
- Ability to dynamically change size of SMSPDSE1 hiperspace
- New SET DEVSUP=xx command to dynamically activate changes to DEVSUP member
- Dynamically add TCP NJE nodes to JES3

Redbooks Workshop

IPL Avoidance

z/OS 1.9 Enhancements:

- Sample RACF ICHPWX11 (password phrase) exit updated to call System Rexx - allowing you to update the function of the exit without an IPL
- New Healthchecks to monitor for pending shortages of linkage indexes and non-reusable address spaces
- SETPROG LNKLST command enhanced to make it more flexible
- REUSEASID parm added to DIAGxx
- Ability to restart system rexx address space - AXRPSTRT
- SETOMVS AUTOCVT command lets you dynamically modify the AUTOCVT setting in BPXPRMxx
- New option on START command, to specify that named STC should use a reuseable ASID. Initially for LLA, DLF, and VLF.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

IPL Avoidance

z/OS 1.10 Enhancements:

- Dynamic JES2 exit support
- Ability to change sysplex root data set without sysplex IPL
- Ability to move from GRSRNL=EXCLUDE to full RNLs without a sysplex IPL
- z/OS UNIX RESOLVER address space, TCP/IP address spaces, DFSMSrmm address space, and the TN3270 address spaces now support ASID reuse.
- New SETRRS ARCHIVELOGGING lets you turn RRS archiving on and off without restarting RRS
- Basic HyperSwap lets you swap from primary to secondary DASD without an IPL
- Ability to dynamically add a CP (DYNCPADD in LOADxx) on z10

Redbooks Workshop

IPL Avoidance

z/OS 1.11 Enhancements:

- Ability to point at specific parmlib AXR members when you use the AXRPSTRT proc to restart System REXX
- SETALLOC command changes values in ALLOCxx member without an IPL
- System Status Detection Partitioning Protocol may improve the chances of spin loop recovery completing successfully (thereby avoiding an IPL)
- ALTROOT statement lets you specify alternate sysplex root file system to dynamically switch to in case current sysplex root becomes unavailable
- Ability to specify maximum time that the system is set to be non-dispatchable during a dump - MAXSNDSP
- Enhancements to make dynamic LPA exit (CSVDYLPA) more usable

ibm.com yourdotcom International Technical Support Organization and Authoring Services

IPL Avoidance

z/OS 1.12 Enhancements:

- CRITICALPAGING function for HyperSwap environments
  - Note, however, that an IPL is required to ENABLE this feature
- Ability to specify NOBUFFS action (SMF) at the log stream level
- VSAM CA Reclaim for KSDSs
- Support for non-disruptive CF Dump
- If a broken PDSE is encountered in LNKLST during IPL, a message is now produced identifying the bad data set and IPL continues without that data set in the Link list.

Redbooks Workshop

IPL Avoidance

(more) z/OS 1.12 Enhancements:

- Enhancements to LLA and PROGxx processing
- CSVLLIX1 and CSVLLIX2 (LLA exits) added to dynamic exits facility
- Extended addressability support for catalogs
- New DEFERTND option to delay making address spaces non-dispatchable during an operator-initiated dump
- Ability to specify a hot-standby Sysplex Distributor
- Able to change number of Common Inet ports without OMVS restart
- HIS detects change in CPU speed without an IPL

ibm.com yourdotcom International Technical Support Organization and Authoring Services

IPL Avoidance

z/OS 1.13 Enhancements:

- Ability to stop a JES2 job at the end of the current step
- Dynamically discontinue use of a JES2 spool volume or increase spool volume size
- Ability to dynamically add spool volumes to JES3
- Ability to change spool-related JES3 parms without an IPL
- DADSM and CVAF support for concurrent service
- Dynamic support for DADSM IGGPRE00 and IGGPOST0 exits
- New FORCE option of CMDS command
- New UNALLOC parameter for the SPIN keyword on the DD statement, to allow you to specify that output data set should be spun off without stopping and starting address space
- DEVMAN added to CATALOG, LLA, VLF, RESOLVER, TCP/IP, DFSMSrmm, and TN3270 to mark address spaces as reusable

Redbooks Workshop

IPL Avoidance

Other Enhancements?

If you know of any other enhancements in this area that I have missed, please come and talk to me, or send me an email (kyne@us.ibm.com)

Also, I am trying to compile a list of items that still require an IPL, so if you would like, please send me that list as well.

ibm.com yourdotcom International Technical Support Organization and Authoring Services

MTTR enhancements

MTTR is Mean Time To Recovery - objective of this journey is to reduce the elapsed time between when you start taking a system down for a planned IPL, and when all the applications are available again.

The journey started with z/OS V1R10 and will continue into the foreseeable future.

z/OS 1.12 delivered a lot of improvements that were especially beneficial for large DB2s

- See last year's ITSO material for more information.

Redbooks Workshop

MTTR enhancements

z/OS 1.13 continues to deliver enhancements to help you reduce MTTR times.

- Enhancements to VSAM close for VSAM linear data sets - APAR OA36390 (R13 only). Can reduce DB2 utility time by as much as 1/3 for large DB2s.
- APAR OA36354 for Media Manager to reduce overhead associated with new XTIOT support.
- New message to warn you when used TCTIOT reaches 95% - I *think* it is IEFA050, but you should check that.

yourdotcom

International Technical Support Organization and Authoring Services

Miscellaneous Bits and Pieces

ibm.com/redbooks

Sending RMF exceptions to MVS console

RMF Monitor III has many powerful capabilities. However most people tend to use it in reactive mode - something doesn't appear to be behaving as you expect, so you look in Monitor III.

Wouldn't it be nice if you could have "someone" looking at *all* the Monitor III displays *all* the time? So that if something starts to go wrong, you can start to address it *before* the users start complaining?

Well now you can, with the new(ish) handy-dandy RMF Client/Server support! You define thresholds, and RMF issues WTO.

Requires about 30 minutes of set up work - for details and examples, refer to:

- ftp://public.dhe.ibm.com/eserver/zseries/zos/rmf/RMF2WTO.pdf

RMF

Other cool RMF stuff

- Do you use the VSAM data set support in RMF Mon III for after-the-fact problem analysis? Lets you do everything afterwards that you are able to do at the time of the problem.
  - RMF provides execs (ERBV2S and ERBS2V) to save and reload the Monitor III data sets.
- RMF Speadsheet Reporter will be enhanced to add information about XCF Group and member usage.
- RMF feeds zOSMF. Can also consolidate information from multiple sysplexes in a single zOSMF instance.
- For a load of interesting presentations on RMF, see:
  - http://www-03.ibm.com/systems/z/os/zos/features/rmf/presentations/rmfpres.html

Insights into VSAM/RLS performance

VSAM/RLS performance and tuning are something of a dark art...

- Doesn't have its own dedicated performance monitors like DB2 and IMS
- Most VSAM/RLS performance information is written to SMF Type 42 records
  - But there is no IBM-provided product to post-process those records
- RMF Monitor III provides VSAM/RLS performance information
  - But that information is not written to RMF SMF records, so can't be produced using RMF PostProcessor

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Insights into VSAM/RLS performance

VSAM/RLS performance information is written to:

- SMF Type 64 records
- SMF Type 42 Subtype 15 records
- SMF Type 42 Subtype 16 records
- SMF Type 42 Subtype 17 records
- SMF Type 42 Subtype 18 records
- SMF Type 42 Subtype 19 records

These records provide an abundance of information... the description of just the subtype 15-19 records takes up 68 pages in the SMF manual!

So, where to start..... The following are Terri Menendez's favorite SMF fields.....

Redbooks Workshop

Insights into VSAM/RLS performance

If you are using 31-bit buffers, precede these field names with SMF42. If using 64-bit buffer, precede them with SMF2A

- There will be a set of the following records for each STOCLAS in use with RLS Data Sets. So to find the sysplex total, we need to add up each of the subtype 15 records cut in the interval for the whole sysplex:
  - FAC - Storage Class (SC) name
  -
  - FCB - Total number of direct requests for this SC
  - FEB - Total number of sequential requests for this SC
  - ----------
  - FCB + FEB = Total number of requests for this SC
  - FCE - Total number direct writes for this SC
  - FEE - Total number of sequential writes for this SC
  - ----------
  - FCE + FEE = Total Number of writes for this SC

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM.

Insights into VSAM/RLS performance

- FCC - Total number of direct reads (NRI) for this SC
- FCD - Total number of direct reads (CR) for this SC
- FEC - Total numer of sequential reads (NRI) for this SC
- FED - Total number of sequential reads (CR) for this SC
- ----------
- FCC + FCD + FEC + FED = Total number of reads for this SC
-
- FCX - Average dir request response time for this SC
- FEX - Average seq response time for this SC
-
- FCF - Total BMF direct-access requests for SC
- FEF - Total BMF sequential-access requests for SC
- FCK - BMF dir false invalids for this SC
- FEK - BMF seq false invalids for this SC
- ----------
- (FCK + FEK) / (FCF+FEF) * 100 = percentage of false invalids for this SC

Insights into VSAM/RLS performance

- FOA - Total number of lock request for this SC
- FOB - Lock requests in true contention for this SC
- FOC - Lock request in false contention for this SC
- ----------
- (FOB + FOC) / FOA * 100 = Percent of contention
-
- FOE - Total DIWA lock accesses for this SC
- FOI - Total DIWA true contention for this SC
- FOE / (FCE+FEE) * 100 = Percentage of splits per write for this SC
-
- FOL - Total upgrade lock accesses for this SC
- FOM - Total upgrade lock true contention for this SC
-
- FOT - Total ESDS Add to End lock accesses for this SC
- FOU - Total ESDS ATE contention for this SC
- FOT / (FCE+FEE)*100 = percentage of ATEs per writes for this SC

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Insights into VSAM/RLS performance

- The following are already SYSPLEX wide stats (using below the bar only):
  - JN7 - Total write requests
  - JNL - Total number of BMF requests
  - JNN - Total BMF hits
  - JNO - Percentage of BMF hits
  - JNT - CF hits
  - JNU - Percentage of CF hits
  - JTO - CF reads
  - JNZ - DASD hits
  - JOA - Percentage DASD hits
  - JTB - Castout contentions
  - JTB / JTO * 100 percentate of castout contentions
  - JRC - Castout retries
  - JRC /. JTB * 100 Percentage of castout contentions retried
  - JTJ - Redos
  - JTJ / (FCB+FEB) * 100 = Percentage of requests which had to redo
  - JTL - recursive redos
  - JTQ / JTJ * 100 = Percentage of redo requests which were recursive
  - JUA - Ave LRU intervals over goal
  - JNG - Ave LRU intervals
  - JUA / JNG * 100 = percentage of avg intervals over the goal

Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

Insights into VSAM/RLS performance

So, what to do with this invaluable information?

1. Pray that you have SAS/MXG
2. Print a copy of the ICETOOL User's Guide for bedside reading...

Look at the following online introduction to VSAM/RLS performance:

1. http://publib.boulder.ibm.com/infocenter/ieduasst/stgv1r0/index.jsp?topic=/com.ibm.iea.zos/zos/1.0/DFSMS/zOSV1R0_DFSMS_RLSPerformance_Tuning/player.html

Come back next year for the next thrilling installment

And watch out for an upcoming update to the (extremely popular) VSAM Demystified Redbook.

ibm.com yourdotcom International Technical Support Organization and Authoring Services IBM

zCP3000

zCP3000 is IBM's capacity planning tool

- Availability limited to IBMers and Business Partners

Looks at many aspects of capacity - sysplex is just one of them.

Over the last year, the developer has made a special attempt to make the sysplex part of zCP3000 even more helpful and has added new function in response to a number of customer upgrade scenarios.

- These enhancements make zCP3000 very powerful to help you model the impact of moving structures, turning SM Duplexing on or off, changing link types, changing CF CEC model....

zCP3000

- Updated support to allow CF link type (ISC-3 and InfiniBand -1x) distance to be specified in units less than 1km. The maximum distance for InfiniBand -1x is 100km.
- Main CF window and CF Summary Report redesigned for clarity and readability.
- Separate algorithms for estimating service time for linktype change: one for simplex and duplex-cache requests, a different one for other duplex requests.
- Red triangle flags CFs that are incompletely defined.
- Move structures between CFs. At this time, the "move" is enabled only when the linktypes for the CFs are identical.
- Improved CF support to handle subchannels, when some are varied off due to channel path busy.
- Easier, 2-step migration for z10 to z196 and ICB4 to InfiniBand.
- New IBM zEnterprise 114 processors (2818) have been added.
- Support for new HCA3 links, including IFB3 mode and 32 subchannels on 1X
- New support to add an unassigned CF (not defined to an existing sysplex and no activity in an unknown sysplex) to an existing sysplex via "Add New CF to Sysplex" (right click on unassigned CF) and select the existing sysplex.
- New support to move structures between CFs even if their hardware configuration is different; sync and async service times are adjusted for different CF engine speeds and linktypes.

zPDT

Facility to let you run a special version of z/OS under Linux

- Was originally limited to IBMers and IBM Business Partners
- Intended for education purposes and as a development environment for software developers

zPDT is now packaged as part of an offering called Rational Developer for System z

Additionally, zPDT now has full Parallel Sysplex support - so you can run a data sharing environment with z/OS under z/VM under Linux.

- Very intersting alternative for application development and system programmer education.

Synchronous WTORs

We have seen an increasing number of situations where some system problems results in a synchronous WTOR being issued, and no one responds successfully to the message.

- Unless you specify otherwise, the default is that Synchronous WTORs will only be issued to the HMC console.
  - In many installations, the HMC console is not located beside the operator's work area, so they are not even aware that a message has been issued - all they see is that the system appears to be dead.
- If a synchronous console group is defined, the WTOR will be sent to each console, two minutes apart, until a reply is received.
- If no reply is received, the WTOR will be issued to the HMC. Once the WTOR is sent to the HMC, replies from other consoles will not be accepted.
- On the HMC, the reply will *only* be accepted if the "Priority message" box is checked.

For more information, please see recent WSC Flash10671

System Programmer tools

RACF profile to let you rename duplicate data sets when there is an ENQ on the DSN.

Profile name is STGADMIN.DPDSRN.dsn and profile length is limited to 23 characters.

Can be exploited using either ISPF 3.2 or API

- Data set cannot be SMS-managed

To use the ISPF interface:

- Specify the data set name and VOLSER
- Ignore the warning saying that you specified a DSN that is cataloged and you specified a VOLSER
- Enter the new name, press Enter, and you will get a “Rename Failed” message followed by three asterisks
- Press Enter again. If you have access to the STGADMIN... profile, a new “Rename Data Set In Use” panel appears. Press Enter again to complete the rename.

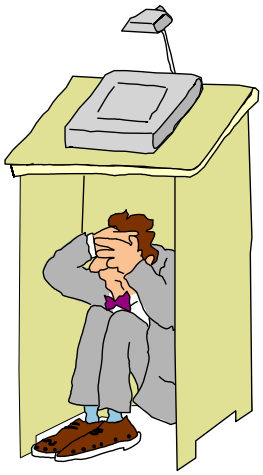
System Programmer tools

SDSF:

- To enter a command that is longer than the normal input space, enter a slash (/) on its own.
- If you are entering a command and run out of space, place a + at the end of the command - this will place you in System Command Extension panel.
- To issue a command against a range of objects, place "//n" (where n is the action to be carried out) on the first line, and "//" on the last line.
- To edit JCL (so you can submit job again), enter SJ beside the job output.
- To edit job output (with all the capabilities of ISPF Edit), enter SE beside the job.
- To easily save job output to a data set, enter XDC beside job.
- To get a list of just the return codes in a job, enter SE beside job output and then ISFESUM edit macro
- To see all MVS commands you entered and the responses, enter ULOG
- To change the layout of YOUR SDSF screen, enter "ARR ?"

ibm.com yourdotcom International Technical Support Organization and Authoring Services

Questions?



Redbooks Workshop
®2011 IBM Corporation. All rights reserved.

Thanks!

And please remember to hand in your evaluations