SHARE
Technology • Connections • Results

z/OS Parallel Sysplex Update

Mark A. Brooks
mabrook@us.ibm.com
IBM

August 3, 2010
Session 7407

SHARE in Boston

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

IBM*
IBM (logo)*
ibm.com*
AIX*
BladeCenter*
DataPower*
CICS*
DB2*
DS4000*
ESCON®

FICON*
IMS
Parallel Sysplex®
POWER7
ProtecTIER*
RACF*
Rational*
Redbooks®
Sysplex Timer®
System Storage

System x*
System z*
System z9®
System z10
Tivoli*
WebSphere*
XIV*
zEnterprise
z/OS*
z/VM*
z9®

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

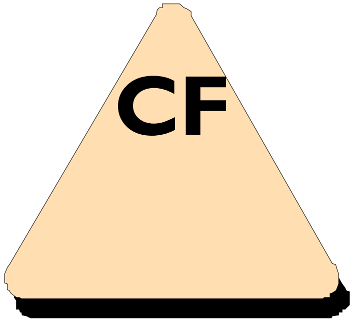
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda

- Hardware Updates
  - **CFCC Level 17**
  - CFCC Level 16
  - Parallel Sysplex InfiniBand Links
- z/OS Updates
  - Sysplex Failure Management
  - z/OS V1R12
- Summary



CFLEVEL 17

- **IBM zEnterprise™ 196 (z196),** Announced July 2010

- Up to 2047 structures
- Up to 255 connectors per structure

- Prerequisites
  - z/OS V1.10 or later with PTF for OA32807
  - z/VM V5.4 for guest virtual coupling

CFLEVEL 17 …

- CF Diagnostics
  - Non-disruptive dumping
  - Improved diagnostics (coordinated capture)

- Prerequisites
  - z/OS V1.12

CFLEVEL 17 …

Migration

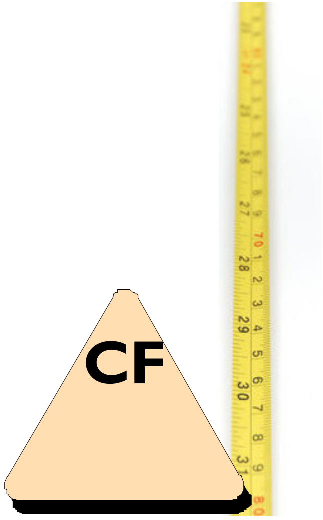
- z196 DR86 contains CFCC Level 17 support
  - In general, get to most current LIC levels

- Use CF Sizer website to check/update structure sizes:
  - CF structure sizes may increase when migrating to CFCC Level 17 from earlier levels due to additional CFCC controls
  - IBM's testers saw 0-4% growth from CFLEVEL=16

CF Sizer Enhancements

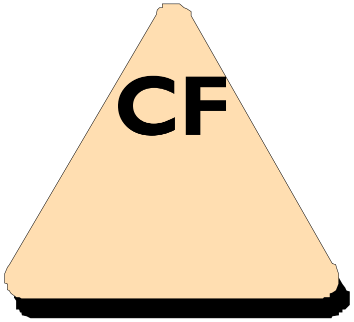
- Improved sizings
  - IMS, DB2, XCF
- Additional structures
  - IBM Session Manager, InfoSphere Classic
- Usability improvements

www.ibm.com/systems/support/z/cfsizer/



Agenda

- Hardware Updates
  - CFCC Level 17
  - **CFCC Level 16**
  - Parallel Sysplex InfiniBand Links
- z/OS Updates
  - Sysplex Failure Management
  - z/OS V1R12
- Summary



CFLEVEL 16

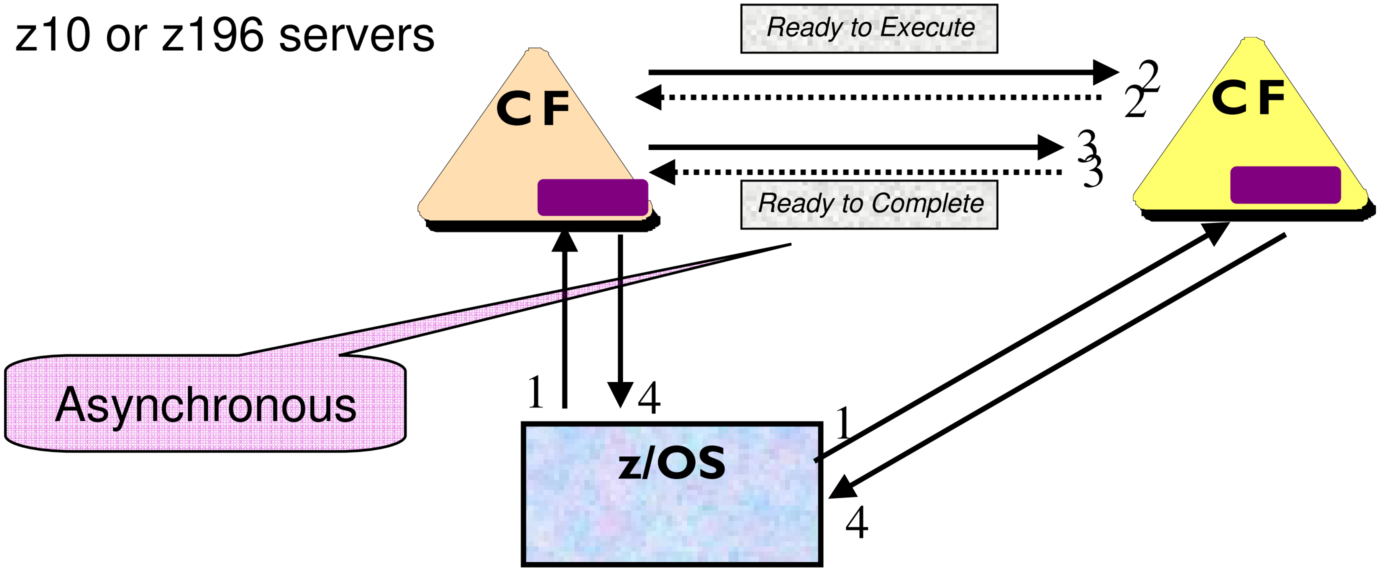
- System z10, October 2008

- CF Duplexing Protocol Enhancements for improved duplex response time
- CF Notification Enhancements to avoid false schedules for Shared Message Queue exploiters
- CF Storage increment size increase 512KB → 1 MB

- Prerequisite:
  - z/OS V1R6 or later with PTFs for APAR OA25130
  - z/OS V1R11 or later

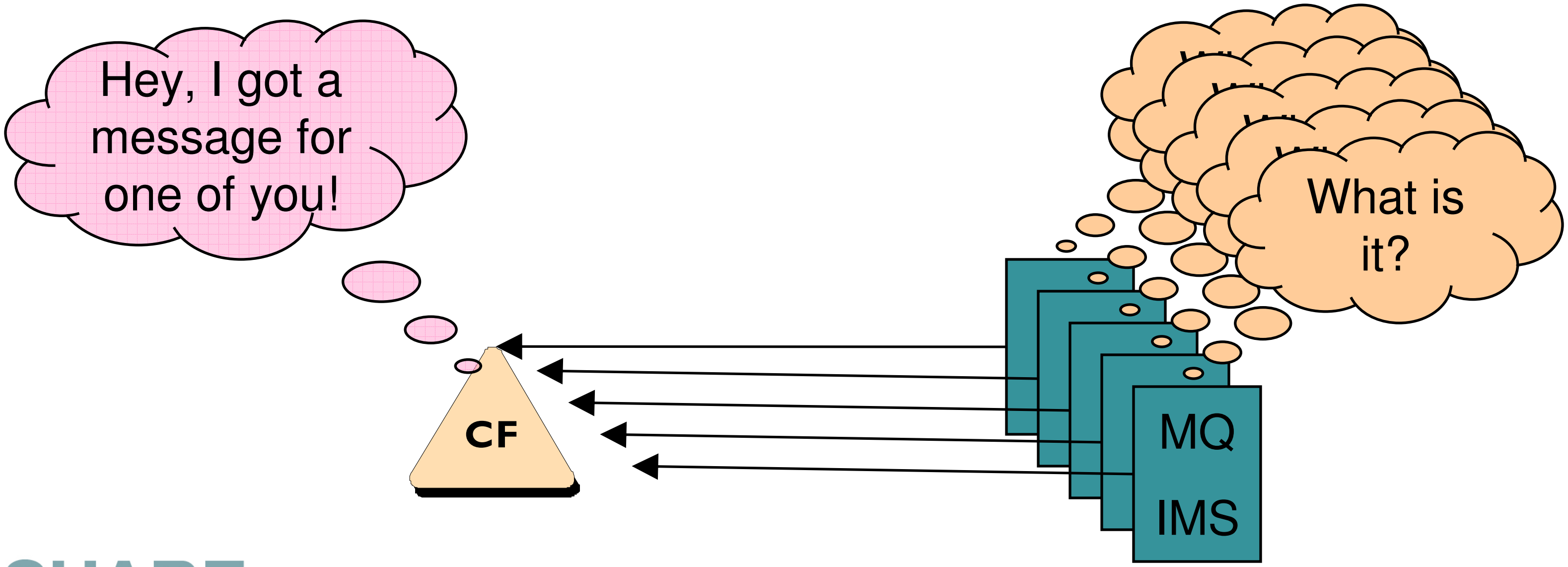
CFLEVEL 16 – Asynchronous RTC

- Designed to improve duplexed request response time
  - Depends on structure's usage of duplexed CF requests
  - Improvements vary with distance
- Requires pairs of CFs
  - CFCC Level 16 or later
  - z10 or z196 servers



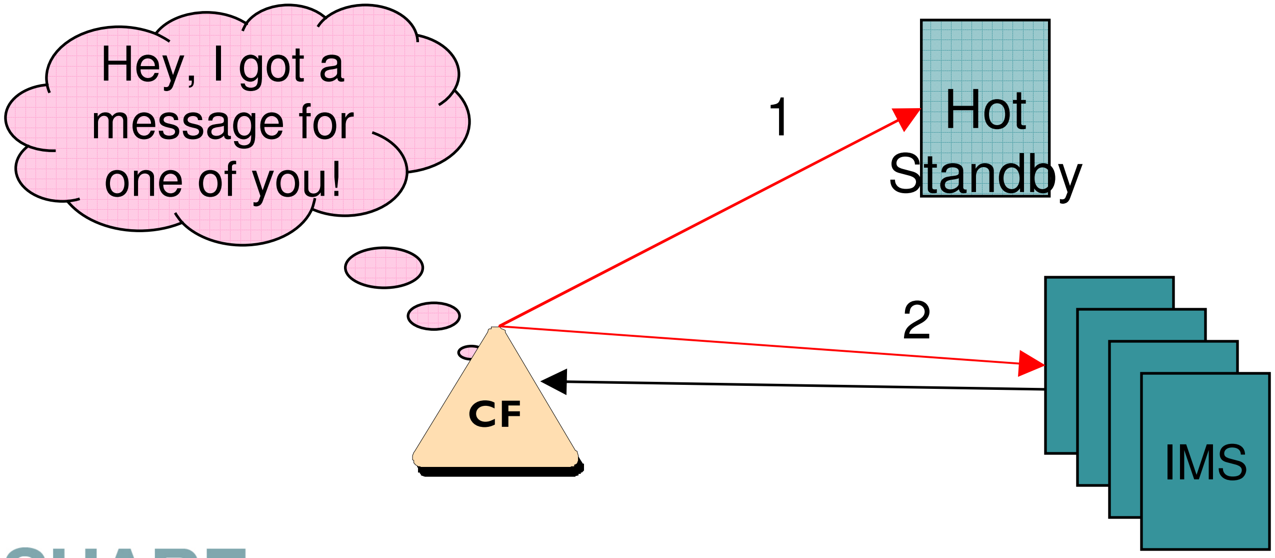
CFLEVEL 16 – Sublist Notification

- Avoid false schedules for shared queue exploiters
  - IMS Shared Message Queue
  - MQ Shared Queues
- Empty → non-empty state change notification sent to one connector
  - Round Robin
- If no response in (time period), then send to next connector



CFLEVEL 16 – Sublist Notification …

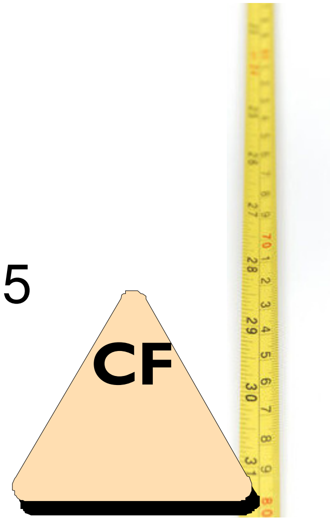
- But … in hindsight, round robin notification may not be so hot for a Hot Standby environment
  - Standby does not process the notifications
  - CF eventually notifies next in line
  - But the delay may be intolerable
- APAR OA30994 provides new controls to either tune or disable the protocol, on a structure by structure basis



CFCC Level 16 Migration

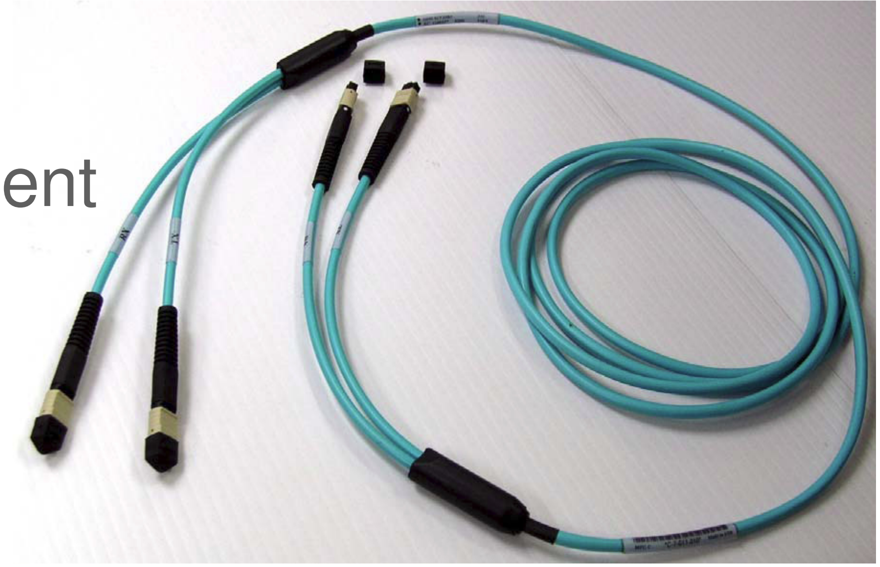
- z10 DR76 and DR79 contains CFCC Level 16 support
  - In general, get to most current LIC levels (**02.25**)
  - **See APAR OA31960 regarding service level 02.12 to 02.22**
- Use CF Sizer website to check/update structure sizes:
  - Many CF structure sizes will increase when migrating to CFCC Level 16 from earlier levels due to:
    - Increase in storage increment size to 1MB
    - Additional CFCC controls
  - Info APAR II14431
    - Recommend using CF Sizer
    - IBM's testers saw 5-10% growth from CFLEVEL=15

www.ibm.com/systems/support/z/cfsizer/index.html



Agenda

- Hardware Updates
  - CFCC Level 17
  - CFCC Level 16
  - **Parallel Sysplex InfiniBand Links**
- z/OS Updates
  - Sysplex Failure Management
  - z/OS V1R12
- Summary



Infiniband Coupling (PSIFB)

- New CF link type (CIB) for all IB coupling links
  - System z9, z10, and z196
  - 7 subchannels per CHPID applies to CIB links
- Simplifies physical connectivity
  - Multiple CIB CHPIDs per physical IB Coupling link
  - A single physical link can “share” CHPIDs across multiple CF images, within same sysplex, or across different ones
  - Additional subchannels, without additional physical links
- Additional flexibility for physical configuration
  - 150 meters (vs ICB4 limit of 10 meters)

PSIFB = PSIB = IFB = IB Coupling = CIB

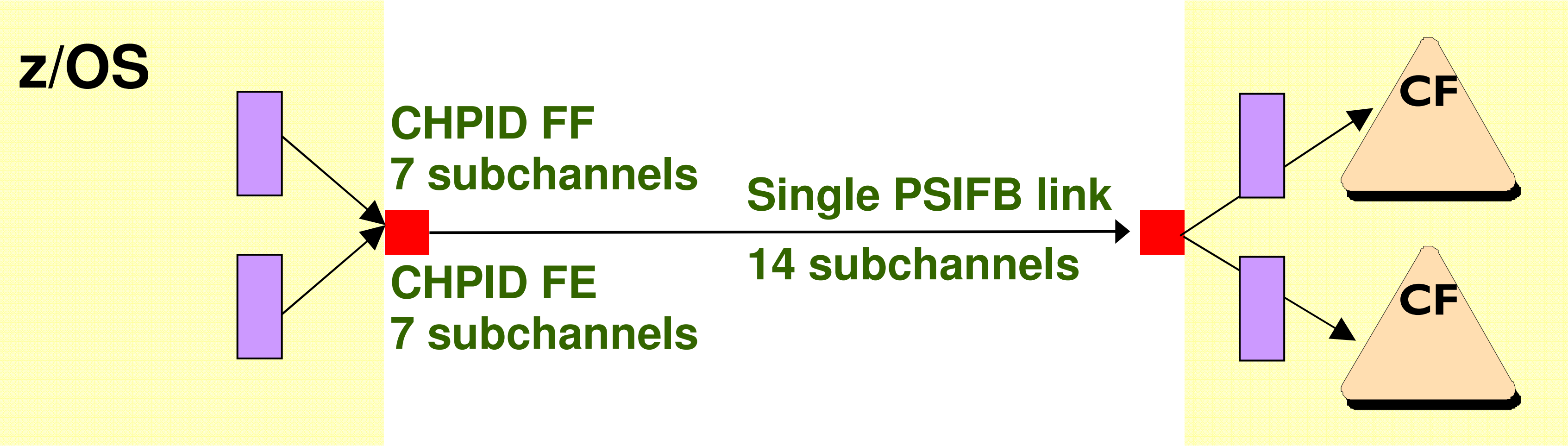
Infiniband Coupling (PSIFB)

- Up to 16 CHPIDs can use same physical link
  - More subchannels / physical link
  - NOT more subchannels / CHPID
- Can connect to multiple CF LPARs
- CF Receiver CHPIDs can share link

7 SubChannels / path

Up to 16 channel paths / link

Up to 112 subchannels / link



- MIF uses same address, 7 subchannels / CHPID

What does it mean to you?

- Fewer physical links
  - Easier to configure since less cabling is needed
  - Potential cost savings if allows fewer books on the machine
- Can define more CHPIDs to a physical link to get more subchannels
  - Beneficial for long links as can get more bandwidth
  - Reduce subchannel busy conditions
- Can have more CF's on a box if links were limiting factor

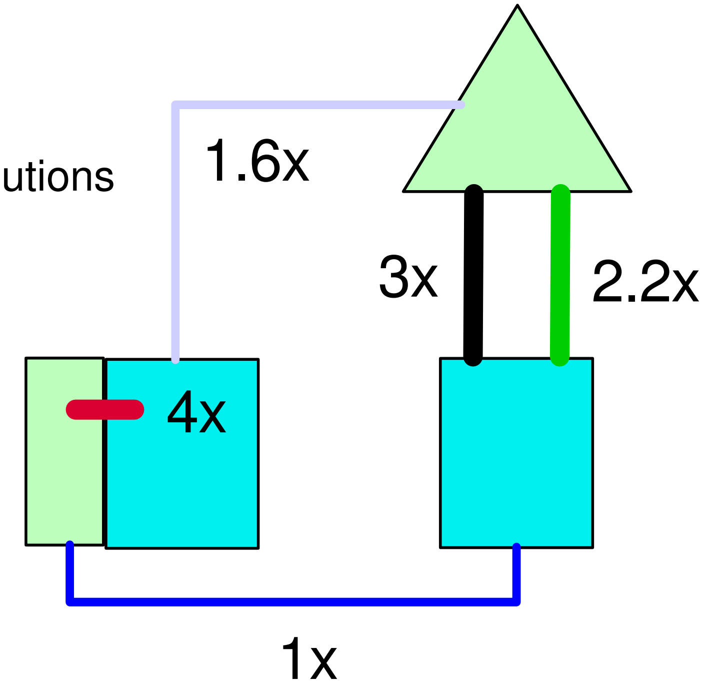
- Note that the z196
  - Doubles number of coupling CHPIDs to 128
  - Increases number of physical links to 80

PSIFB Configuration Considerations

- Pure Capacity
  - One 12x PSIFB replaces one ICB4
  - One 12x PSIFB replaces four ISC3
- Eliminating Subchannel and Path Delays
  - Extra ICB4 links might be configured to get additional subchannels/paths to eliminate delays caused by busy conditions
  - One 12x PSIFB link with multiple CHPIDs could replace multiple ICB4s in this case
- Multiple sysplexes sharing hardware
  - Production, development, and test sysplexes can share hardware, but they each need their own ICB4 and ISC3 links
  - One PSIFB link with multiple CHPIDs could replace multiple links in this case

Coupling Link Choices - Overview

- ISC (Inter-System Channel)
  - Fiber optics
  - I/O Adapter card
  - 10km and longer distances with qualified WDM solutions
- PSIFB (1x IFB)
  - Fiber optics – uses same cabling as ISC
  - 10km and longer distances with qualified WDM solutions
- PSIFB (12x IFB)
  - 150 meter max distance optical cabling
  - Supports multiple CHPIDs per physical link
  - Multiple CF partitions can share physical link
- ICB (Integrated Cluster Bus)
  - Copper cable plugs close to memory bus
  - 10 meter max length
  - Not available on z196
- IC (Internal Coupling Channel)
  - Microcode - no external connection
  - Only between partitions on same processor



Relative Performance
Based on avg data xfer size

Coupling Technology versus Host Processor Speed

Host effect with primary application involved in data sharing

Chart below is based on 9 CF ops/Mi - may be scaled linearly for other rates

Host CF	z890	z990	z9 BC	z9 EC	z10 BC	z10 EC	z196
z890 ISC	13%	15%	16%	17%	19%	21%	NA
z890 ICB	9%	10%	10%	11%	12%	13%	NA
z990 ISC	13%	14%	14%	15%	17%	19%	NA
z990 ICB	9%	9%	9%	10%	12%	13%	NA
z9 BC ISC	12%	13%	14%	15%	17%	19%	23%
z9 BC PSIFB 12X	NA	NA	NA	NA	13%	14%	16%
z9 BC ICB	8%	9%	9%	10%	11%	12%	NA
z9 EC ISC	12%	13%	13%	14%	16%	18%	22%
z9 EC PSIFB 12X	NA	NA	NA	NA	13%	14%	16%
z9 EC ICB	8%	8%	8%	9%	10%	11%	NA
z10 BC ISC	12%	13%	13%	14%	16%	18%	22%
z10 BC PSIFB 12X	NA	NA	11%	12%	13%	14%	15%
z10 BC ICB	8%	8%	8%	9%	10%	11%	NA
z10 EC ISC	11%	12%	12%	13%	15%	17%	22%
z10 EC PSIFB 12X	NA	NA	10%	11%	12%	13%	15%
z10 EC ICB	7%	7%	7%	8%	9%	10%	NA
z196 ISC	NA	NA	11%	12%	14%	16%	21%
z196 PSIFB 12X	NA	NA	9%	10%	11%	12%	14%

With z/OS 1.2 and above, synch->asynch conversion caps values in table at about 18%

PSIFB 1X links would fall approximately halfway between PSIFB 12X and ISC links

IC links scale with speed of host technology and would provide an 8% effect in each case

Maximum CF Links

Server	IC	IFB	ICB-4	ICB-3	ICB	ISC-3	Max # Links
z800	32	-	-	5 6 (0CF)	-	24	26 + 32
z900-100 CF	32	-	-	16	16	32 **42** w/RPQ	64
z900	32	-	-	16	8 **16** w/RPQ	32	64
z890	32	-	8	16	-	48	64
z990	32	-	16	16	8	48	64
z9 EC	32	16	16	16	-	48 Peer	64
z9 BC	32	12	16	16	-	48 Peer	64
z10 EC	32	32	16	-	-	48 Peer	64 32 IFB + ICB-4
z10 BC	32	12	12	-	-	48 Peer	64 56 External 12 IFB + ICB-4
z196	32	32	-	-	-	48 Peer	80

PSIFB Configurations Supported

z/OS CF	z9	z10	z196
z9	No	Yes	Yes
z10	Yes	Yes	Yes
z196	Yes	Yes	Yes

Distance Considerations

Distance	IC	ICB-4	12x IFB	ISC-3 1x IFB
Within server	Yes	n/a	n/a	n/a
<10 m		Yes	Yes	Yes
10 m – 150 m			Yes	Yes
150 m – 100+ km				Yes

For more information

- **“Coupling Technology Overview and Planning - What’s the Right Stuff for Me?”**
  - Thursday August 5 8:00-9:00
- **“PSIFB (Infiniband) Coupling Links Overview and User Experience”**
  - Thursday August 5 9:30-10:30

- “IBM System z Connectivity Handbook” (SG24-5444)
- “Getting Started with Infiniband on System z10 and System z9” (SG24-7539)
  - Available at www.redbooks.ibm.com

- http://www.ibm.com/systems/z/advantages/pso/whitepaper.html
  - CF Configuration Options White Paper

Agenda

- Hardware Updates
  - CFCC Level 17
  - CFCC Level 16
  - Parallel Sysplex InfiniBand Links
- z/OS Updates
  - **Sysplex Failure Management**
  - z/OS V1R12
- Summary

Sysplex Failure Management (SFM) Subtopics

MEMSTALLTIME	z/OS 1.8
SSUMLIMIT	z/OS 1.9
SFM and AutoIPL	z/OS 1.10
SFM with BCPii	z/OS 1.11
System Default Action	z/OS 1.11
XCF FDI Consistency	z/OS 1.11
Critical Members	z/OS 1.12
CFSTRHANGTIME	z/OS 1.12

SFM is the subcomponent within XCF that deals with the detection and resolution of sympathy sickness conditions that can arise when a system or sysplex application is unresponsive

Sysplex Failure Management – z/OS 1.8 MEMSTALLTIME

- XCF detects and surfaces inter-system signalling sympathy sickness caused by stalled group member(s)
- SFM policy MEMSTALLTIME specification determines how long XCF should wait before taking action to resolve the problem
- After expiration, the stalled member is terminated
  - For GRS, XCF, or Consoles, implies system termination
- Provides a backstop that can take automatic action in case your automation or manual procedures fail to resolve the issue

Sysplex Failure Management - z/OS 1.9 SSUMLIMIT

- Systems in sysplex monitor each other for signs of life:
  - Status updates in sysplex couple data set
  - XCF signal transfers
- A system could be sending signals but not updating status
- The SFM Policy SSUMLIMIT specification determines how long a system is allowed to persist in this state
- When the SSUMLIMIT interval expires, the system will be partitioned from the sysplex
  - Without SSUMLIMIT, SFM will not take action because the monitored system has signs of life. But something *is* wrong.
  - Allows the installation to “bound” the amount of time that a sick system might impact the remainder of the sysplex
  - Not too aggressive, perhaps 15 minutes

Sysplex Failure Management - z/OS 1.10 Auto-IPL

- Auto-IPL support provides an automated way to recover from system wait-states without operator intervention
  - Can optionally take stand alone dump
  - Can re-IPL the system
- z/OS determines the set of wait-states to which Auto-IPL applies and what actions are applicable
  - No action, SADUMP, SADUMP + IPL, IPL
- The Auto-IPL policy determines the recovery process

Faster Recovery

Sysplex Failure Management - z/OS 1.10 Auto-IPL and VARY XCF

- New XCF support allows operator to indicate Auto-IPL actions to be performed after the system is removed from the sysplex
  - VARY XCF,sysname,OFFLINE,SADMP
  - VARY XCF,sysname,OFFLINE,REIPL
  - VARY XCF,sysname,OFFLINE,SADMP,REIPL
- Designated system must have an Auto-IPL policy that permits the indicated action(s)

Sysplex Failure Management
Auto-IPL and SFM

- z/OS 1.10 (with APARs and appropriate server LIC's)
  - Re-IPL can cause fencing to fail, which interferes with isolation and removal of system by SFM
    - And so may require manual intervention
  - So Auto-IPL will delay its action in an attempt to give SFM time to isolate (fence) the system
- z/OS 1.11
  - If SFM is able to exploit BCPii to detect failed systems and perform system reset, Auto-IPL has cases where it need not allow time for fencing to occur
  - Otherwise, as above for z/OS 1.10

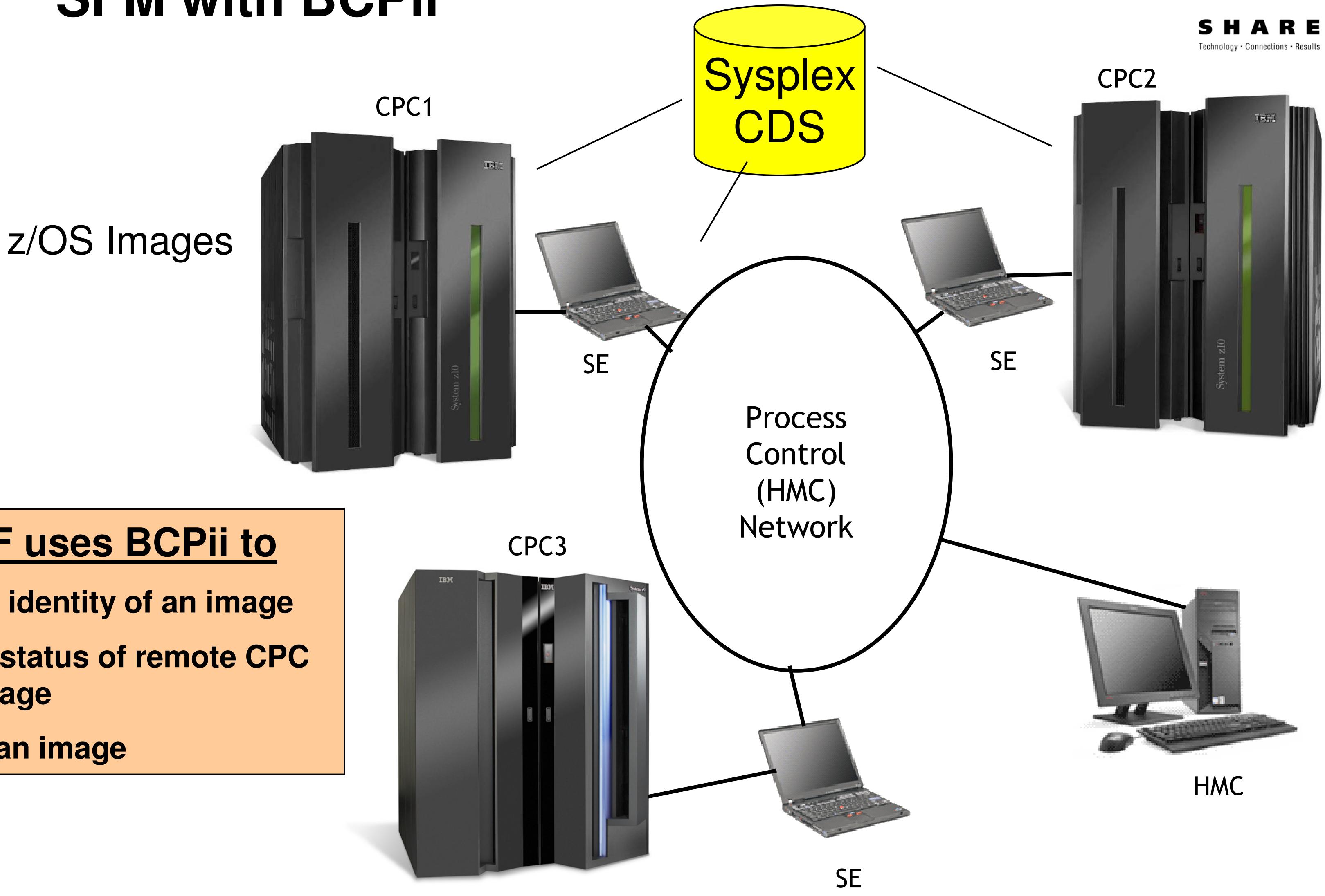
Sysplex Failure Management
Auto-IPL and GDPS

- GDPS automation is intended to be the sole manager of IPLs and re-IPLs of z/OS images in a GDPS environment
- To avoid conflicts, Auto-IPL should be not be configured for use on any system being managed as part of a GDPS environment

Sysplex Failure Management – z/OS 1.11 SFM with BCPii

- Expedient removal of unresponsive or failed systems is essential to high availability in sysplex
- XCF exploits new BCPii services to:
  - Detect failed systems
  - Reset systems
- Benefits:
  - Improved availability by reducing duration of sympathy sickness
  - Eliminate manual intervention in more cases
  - Potentially prevent human error that can cause data corruption

Sysplex Failure Management – z/OS 1.11 SFM with BCPii



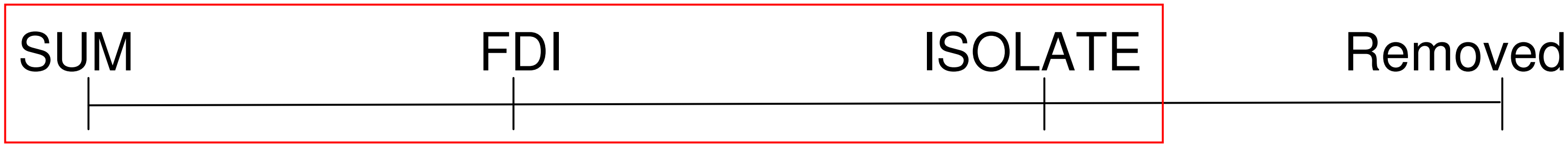
XCF uses BCPii to

- Obtain identity of an image
- Query status of remote CPC and image
- Reset an image

Sysplex Failure Management – z/OS 1.11 SFM with BCPii

- With BCPii, XCF can know that system is dead, and:
  - Bypass the Failure Detection Interval (FDI)
  - Bypass the Indeterminate Status Interval (ISI)
  - Bypass the cleanup interval
  - Reset the system even if fencing fails
  - Avoid IXC102A, IXC402D and IXC409D manual intervention
  - Validate “down” to help avoid corruption of shared data

Often saves 2-3 minutes



Helps improve availability

Sysplex Failure Management – z/OS 1.11
SFM with BCPii

- SFM will automatically exploit BCPii and as soon as the required configuration is established:
  - Pairs of systems running z/OS 1.11 or later
  - BCPii configured, installed, and available
  - XCF has security authorization to access BCPii defined FACILITY class resources
  - z10 GA2 with appropriate MCL's, or z196
  - New version of sysplex CDS is primary in sysplex
    - Toleration APAR OA26037 for z/OS 1.9 and 1.10
    - Does NOT allow systems to use new SSD function or protocols

For more information

- **BCPii for Dummies: Start to finish installation, setup and usage**
  - Tuesday, August 3, 2010: 4:30 PM-5:30 PM

- **“Sysplex Partitioning Using BCPii”**
  - Session 2251 (proceedings from March SHARE in Seattle)
- **“BCPii: Secure z/OS Interface to Your HMC and SE”**
  - Session 2227 (proceedings from March SHARE in Seattle)

Enabling SFM to use BCPii will have a big impact on availability. Make it happen !

Sysplex Failure Management – z/OS 1.11 System Default Action

- SFM Policy lets you define how XCF is to respond to a Status Update Missing condition
- Each system “publishes” in the sysplex couple data set the action that is to be applied by its peers
- The system “default action” is published if:
  - The policy does not specify an action for it
  - There is no SFM policy active

- Prior to z/OS 1.11, the “default action” was PROMPT
- With z/OS 1.11, the system default action is ISOLATETIME(0)

Sysplex Failure Management – z/OS 1.11 System Default Action

- The resulting behavior for system “default action” depends on who is monitoring who:
  - z/OS 1.11 will isolate a peer z/OS 1.11
  - z/OS 1.11 will PROMPT for lower level peer
  - Lower level system will PROMPT for z/OS 1.11

- D XCF,C shows what the system *expects*
  - *But it may not get that in a mixed sysplex*

- Note: z/OS 1.11 may fence even if action is PROMPT
  - Lower level releases performed fencing only when the system was taking automatic action to remove the system (ISOLATETIME)

Sysplex Failure Management – z/OS 1.11
XCF FDI Consistency

- Enforces consistency between the system Failure Detection Interval (FDI) and the excessive spin parameters
- Allows system to perform full range of spin recovery actions before it gets removed from the sysplex
- Avoids false removal of system for a recoverable situation

Helps prevent false SFM removals

Sysplex Failure Management – z/OS 1.11 XCF FDI Consistency

```
IXC357I  15.12.46  DISPLAY XCF                        E   SYS=D13ID71
SYSTEM D13ID71 DATA
   INTERVAL    OPNOTIFY       MAXMSG     CLEANUP       RETRY    CLASSLEN
        165         170         3000          60          10         956

   SSUM ACTION   SSUM INTERVAL   SSUM LIMIT      WEIGHT  MEMSTALLTIME
        PROMPT             165          N/A         N/A           N/A

   PARMLIB USER INTERVAL:      60
   DERIVED SPIN INTERVAL:     165
   SETXCF  USER OPNOTIFY: +     5
< - - - snip - - - >
OPTIONAL FUNCTION STATUS:
  FUNCTION NAME                      STATUS      DEFAULT
  DUPLEXCF16                         ENABLED     DISABLED
  SYSSTATDETECT                      ENABLED     ENABLED
  USERINTERVAL                       DISABLED    DISABLED
```

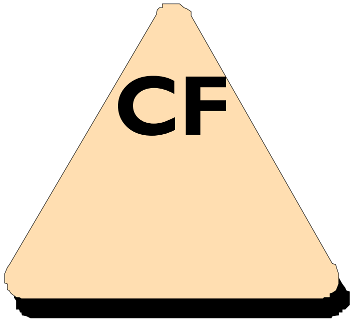
Effective Values

User FDI
Spin FDI
User OpNotify
- Absolute
- Relative

Switch

Agenda

- Hardware Updates
  - CFCC Level 17
  - CFCC Level 16
  - Parallel Sysplex InfiniBand Links
- z/OS Updates
  - Sysplex Failure Management
  - **z/OS V1R12**
- Summary



z/OS 1.12

- **Critical Members**
- **CFSTRHANGTIME**
- **REALLOCATE**
- Large Structure Support
- Non-disruptive CF dumping
- Health Checks
- Auto Reply
- **Run Time Diagnostics**
- XCF Programming Interfaces

z/OS 1.12 - Critical Members

- A system may appear to be healthy with respect to XCF system status monitoring, namely:
  - Updating status in the sysplex CDS
  - Sending signals
- But is the system actually performing useful work?

- There may be critical functions that are non-operational
- Which in effect makes the system unusable, and perhaps induces sympathy sickness elsewhere in the sysplex

- Action should be taken to restore the system to normal operation

z/OS 1.12 - Critical Members …

- Member Impairment
  - A member is **confirmed** to be impaired when its status exit indicates “status missing”
  - A member is **deemed** to be impaired if it is stalled with no signs of activity
- XCF now surfaces impairment for <u>all</u> members

z/OS 1.12 - Critical Members …

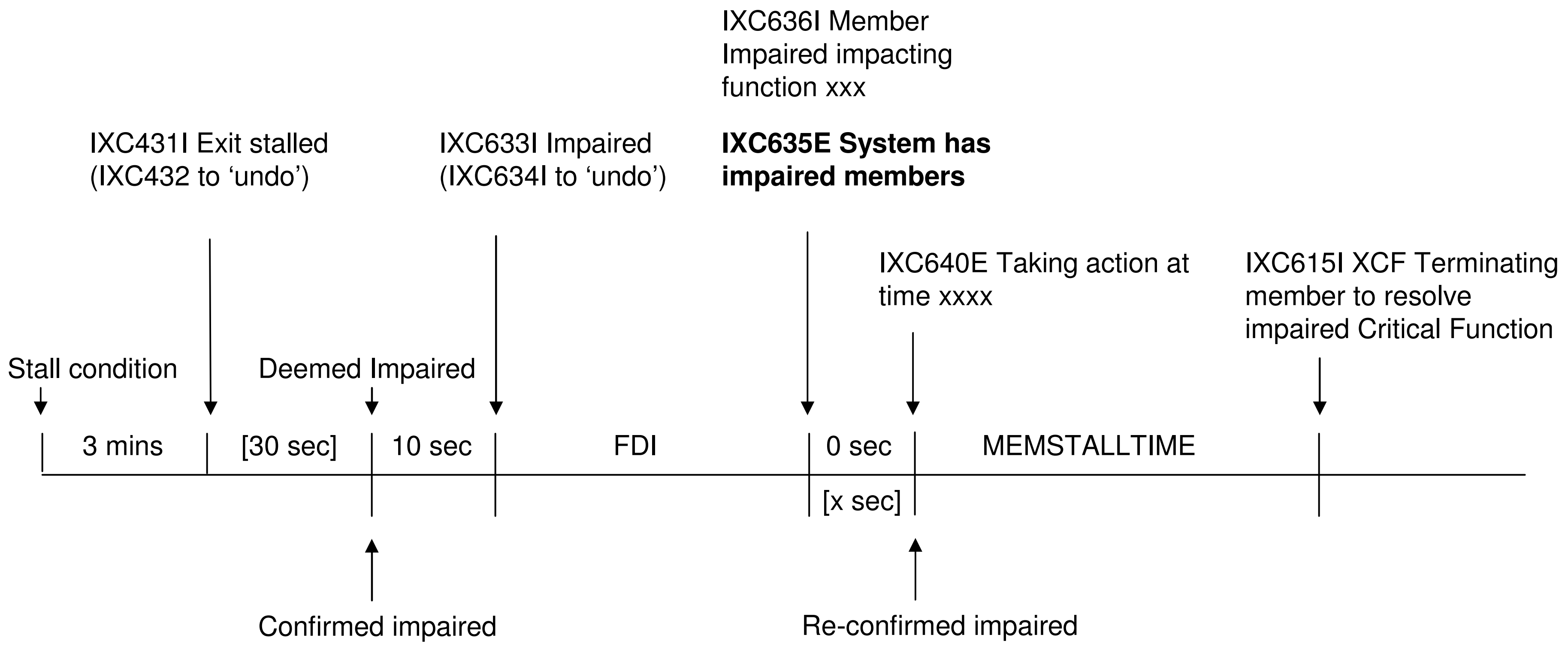
- A Critical Member is a member of an XCF group that Identifies itself as "critical" when joining its group
- If a critical member is impaired for long enough, XCF will eventually terminate the member
  - Per the member's specification: task, space, or system
  - MEMSTALLTIME determines "long enough"

- GRS is a "system critical member"

z/OS 1.12 - Critical Members …

- New Messages
  - IXC633I “member is impaired”
  - IXC634I “member no longer impaired”
  - **IXC635E “system has impaired members”**
  - IXC636I “impaired member impacting function”
- Changed Messages
  - IXC431I “member stalled” (includes status exit)
  - IXC640E “going to take action”
  - IXC615I “terminating to relieve impairment”
  - IXC333I “display member details”
  - IXC101I, IXC105I, IXC220W “system partitioned”

z/OS 1.12 - Critical Members …



Stall condition

IXC431I Exit stalled
(IXC432 to 'undo')

Deemed Impaired

IXC633I Impaired
(IXC634I to 'undo')

IXC636I Member
Impaired impacting
function xxx

IXC635E System has
impaired members

IXC640E Taking action at
time xxxx

IXC615I XCF Terminating
member to resolve
impaired Critical Function

3 mins

[30 sec]

10 sec

FDI

0 sec

[x sec]

MEMSTALLTIME

Confirmed impaired

Re-confirmed impaired

Stall condition = At least 1 exit stalled for 30 seconds or work item on head of queue for 30 seconds

Deemed impaired = IXC431I issued and for last 30 seconds, either all scheduled <u>user</u> exits stalled or no <u>user</u> exits scheduled

z/OS 1.12 - Critical Members …

- **Coexistence considerations**
  - Toleration APAR OA31619 for systems running z/OS V1R10 and z/OS V1R11 should be installed before IPLing z/OS V1R12
  - The APAR allows the down level systems to understand the new sysplex partitioning reason that is used when z/OS V1R12 system removes itself from the sysplex because a system critical component was impaired
  - If the APAR is not installed, the content of the IXC101I and IXC105I messages will be incorrect

z/OS 1.12 - Critical Members …

- **Potential migration action**
  - Evaluate, perhaps change MEMSTALLTIME parameter

z/OS 1.12 - CFSTRHANGTIME

- Connectors to CF structures need to participate in various processes and respond to relevant events
- XES monitors the connectors to ensure that they are responding in a timely fashion
- If not, XES issues messages (IXL040E, IXL041E) to report the unresponsive connector
- Users of the structure may hang until the offending connector responds or is terminated

z/OS 1.12 – CFSTRHANGTIME …

- CFSTRHANGTIME
  - A new SFM Policy specification
  - Indicates how long the system should allow a structure hang condition to persist before taking corrective action(s) to remedy the situation
- Corrective actions may include:
  - Stopping rebuild
  - Forcing the user to disconnect
  - Terminating the connector task, address space, or system

z/OS 1.12 – CFSTRHANGTIME …

New Messages

IXL049E HANG RESOLUTION ACTION FOR CONNECTOR NAME: conname
TO STRUCTURE strname, JOBNAME: jobname, ASID: asid:
actiontext

IXL050I CONNECTOR NAME: conname TO STRUCTURE strname,
JOBNAME: jobname, ASID: asid
HAS NOT PROVIDED A REQUIRED RESPONSE AFTER noresponsetime SECONDS.
TERMINATING termtarget TO RELIEVE THE HANG.

z/OS 1.12 – CFSTRHANGTIME …

- Initiates diagnostic dump as appropriate

Background - REALLOCATE

- SETXCF START,REALLOCATE
  - Well-received, widely exploited for CF structure management
  - For example, to apply “pure” CF maintenance:
    - SETXCF START,MAINTMODE,CFNAME=cfname
    - SETXCF START,REALLOCATE to move structures out of CF
    - Perform CF maintenance
    - SETXCF STOP,MAINTMODE,CFNAME=cfname
    - SETXCF START,REALLOCATE to restore structures to CF

Background - REALLOCATE

But…

- Difficult to tell what it did
  - Long-running process
  - Messages scattered all over syslog
  - Difficult to find and deal with any issues that arose

- And people want to know in advance what it will do

z/OS 1.12 - REALLOCATE

- DISPLAY XCF,REALLOCATE,option

- TEST option
  - Provides detailed information regarding what REALLOCATE would do if it were to be issued
  - Explains why an action, if any, would be taken
- REPORT option
  - Provides detailed information about what the most recent REALLOCATE command actually did do
  - Explains what happened, but not why

REALLOCATE TEST Example (part 1)

```
D XCF,REALLOC,TEST
IXC347I  10.31.05  DISPLAY XCF

COUPLING FACILITY STRUCTURE ANALYSIS PERFORMED FOR REALLOCATE TEST.
----------------------------------------------------------------------
STRUCTURE(S) WITH AN ERROR/EXCEPTION CONDITION

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

NONE
----------------------------------------------------------------------
```

results from a simulated REALLOCATE

REALLOCATE TEST Example (part 2)

```
STRUCTURE(S) REALLOCATED SUCCESSFULLY

STRNAME: BIGONE                                    INDEX: 38
   SIMPLEX STRUCTURE ALLOCATED IN CF(S) NAMED: LF02
   CFNAME     STATUS/FAILURE REASON
   --------   --------------------------------------------------------
   LF01       PREFERRED CF 1
                                   INFO110: 00000003 AC007800 00010011
   LF02       PREFERRED CF ALREADY SELECTED
                                   INFO110: 00000003 AC007800 00020011

   1 REALLOCATE STEP(S): REBUILD
----------------------------------------------------------------------
STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF(S)

...
----------------------------------------------------------------------
```

REALLOCATE TEST Example (part 3)

```
COUPLING FACILITY STRUCTURE ANALYSIS OUTPUT FOR REALLOCATE TEST

CFNAME: LF01
   COUPLING FACILITY      :   SIMDEV.IBM.EN.ND0100000000
                              PARTITION: 00   CPCID: 00
   CONNECTED SYSTEM(S):
   SY1      SY2      SY3

   ACTIVE STRUCTURE(S):
   BIGONE                 CACHE01(OLD)           CACHE02(OLD)
   CACHE12                CACHE128               CACHE16
   CACHE256               CACHE32                CACHE64
…
```

This is like message IXC362I from DISPLAY XCF,CF,CFNAME=ALL and shows approximately what that message would look like AFTER performing the REALLOCATE.

REALLOCATE TEST Example (part 4)

```
REALLOCATE TEST RESULTED IN THE FOLLOWING:
         1  STRUCTURE(S) REALLOCATED - SIMPLEX
         0  STRUCTURE(S) REALLOCATED - DUPLEXED
         0  STRUCTURE(S) POLICY CHANGE MADE - SIMPLEX
         0  STRUCTURE(S) POLICY CHANGE MADE - DUPLEXED
         6  STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF - SIMPLEX
         2  STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF - DUPLEXED
         0  STRUCTURE(S) NOT PROCESSED
       102  STRUCTURE(S) NOT ALLOCATED
        17  STRUCTURE(S) NOT DEFINED
  --------
       128  TOTAL

         0  STRUCTURE(S) WITH AN ERROR/EXCEPTION CONDITION
```

z/OS 1.12 – REALLOCATE …

Caveats for REALLOCATE TEST option

- Actual REALLOCATE could have different results
  - Environment could change
  - For structures processed via user-managed rebuild, the user could make “unexpected” changes
  - Capabilities of systems where REALLOCATE runs differ from the system where TEST ran
    - For example, connectivity to coupling facilties
- TEST cannot be done:
  - While a real REALLOCATE (or POPCF) is in progress
  - If there are no active allocated structures in the sysplex

REALLOCATE REPORT Example (part 1)

```
D XCF,REALLOC,REPORT
IXC347I  10.37.45  DISPLAY XCF

THE REALLOCATE PROCESS STARTED ON 08/07/2009 AT 10:31:23.98.
THE REALLOCATE PROCESS ENDED ON 08/07/2009 AT 10:36:09.81.
----------------------------------------------------------------------
STRUCTURE(S) WITH AN ERROR/EXCEPTION CONDITION

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

NONE
----------------------------------------------------------------------
```

results from a real REALLOCATE

dividing line at END of each section

REALLOCATE REPORT Example (part 2)

```
STRUCTURE(S) REALLOCATED SUCCESSFULLY

STRNAME: CACHE01                                   INDEX: 2
   3 REALLOCATE STEP(S): KEEP=OLD, REBUILD, DUPLEX
   COMPLETED ON SYSTEM SY1 ON 08/07/2009 AT 10:31:40.01.

STRNAME: CACHE02                                   INDEX: 6
   3 REALLOCATE STEP(S): KEEP=OLD, REBUILD, DUPLEX
   COMPLETED ON SYSTEM SY1 ON 08/07/2009 AT 10:31:53.03.
----------------------------------------------------------------------
STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF(S)

STRNAME: IXC2                                      INDEX: 22
EVALUATED ON SYSTEM SY1 ON 08/07/2009 AT 10:32:32.60.
----------------------------------------------------------------------
```

REALLOCATE processes structures in INDEX order

REALLOCATE REPORT Example (part 3)

```
REALLOCATE PROCESSING RESULTED IN THE FOLLOWING:
        91  STRUCTURE(S) REALLOCATED - SIMPLEX
         8  STRUCTURE(S) REALLOCATED - DUPLEXED
         0  STRUCTURE(S) POLICY CHANGE MADE - SIMPLEX
         0  STRUCTURE(S) POLICY CHANGE MADE - DUPLEXED
         1  STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF - SIMPLEX
         0  STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF - DUPLEXED
         0  STRUCTURE(S) NOT PROCESSED
        12  STRUCTURE(S) NOT ALLOCATED
        16  STRUCTURE(S) NOT DEFINED
  --------
       128  TOTAL

         0  STRUCTURE(S) WITH AN ERROR/EXCEPTION CONDITION

         0  STRUCTURE(S) MISSING PREVIOUS REALLOCATE DATA
```

z/OS 1.12 – REALLOCATE …

Caveats for REPORToption

- Can be done during or after a real REALLOCATE (but not before a real REALLOCATE is started)
- A REPORT is internally initiated by XCF if a REALLOCATE completes with exceptions

z/OS 1.12 – REALLOCATE …

- Software Dependencies
  - Available with z/OS V1.12
  - Coexistence apar OA29236 is required on V1R10 and V1R11

OA29236 is PE
Need OA33533
(even for z/OS 1.12)

z/OS 1.12 – Large Structure Support

- Large CF Structures
  - Increased CF structure size supported by z/OS to 1TB
  - Usability enhancements for structure size specifications
    - CFRM policy sizes
    - Display output
- More CF Structures can be defined
  - New z/OS limit is 2048 (CF limit is 2047)
- More Structure Connectors (CF limit is 255)
  - Lock structure – new limit is 247
  - Serialized list – new limit is 127
  - Unserialized list – new limit is 255

z/OS 1.12 – Large Structure Support

- A new version of the CFRM CDS is needed in order to define more than 1024 structures in a CFRM policy
- May need to roll updated software around the sysplex for any exploiter that wants to request more than 32 connectors to list and lock structures
  - Not aware of any at this point (so really just positioning for future growth)

z/OS 1.12 – Non-disruptive CF dumps

- z/OS requests non-disruptive CF dumps as appropriate

- Coherent Parallel-Sysplex Data Collection Protocol
  - Exploited for duplexed requests
  - Triggering event will result in non-disruptive dump from both CFs, dumps from both z/OS images, and capture of relevant link diagnostics within a short period
  - Prerequisites:
    - Installation must ENABLE the XCF function DUPLEXCFDIAG
    - z/OS 1.12

z/OS 1.12 Health Checks

- XCF_CF_PROCESSORS
  - Ensure CF CPU’s configured for optimal performance
- XCF_CF_MEMORY_UTILIZATION
  - Ensure CF storage is below threshold value
- XCF_CF_STR_POLICYSIZE
  - Ensure structure SIZE and INITSIZE values are reasonable

z/OS 1.12 Health Checks …

- XCF_CDS_MAXSYSTEM
  - Ensure function CDS supports at least as many systems as the sysplex CDS
- XCF_CFRM_MSGBASED
  - Ensure CFRM is using desired protocols
- XCF_SFM_CFSTRHANGTIME
  - Ensure SFM policy using desired CFSTRHANGTIME specification

z/OS 1.12 Auto-Reply

- Fast, accurate, knowledgeable responses can be critical
- Delays in responding to WTOR's can impact the sysplex

- Parmlib member defines a reply value and a time delay for a WTOR. The system issues the reply if the WTOR has been outstanding longer than the delay

- Very simple automation
- **Can be used during NIP !**

z/OS 1.12 Auto-Reply

- For example:

  IXC289D REPLY U TO USE THE DATA SETS LAST USED FOR *typename* OR C TO USE THE COUPLE DATA SETS SPECIFIED IN COUPLExx

- The message occurs when, for example, the couple data sets specified in the COUPLExx parmlib member do not match the ones in use by the sysplex (as might happen when the couple data sets are changed dynamically via SETXCF commands to add a new alternate or switch to a new primary)
- Most likely always reply “U”

z/OS 1.12 – Runtime Diagnostics

- Allows installation to quickly analyze a system experiencing “sick but not dead” symptoms
- Looks for evidence of “soft failures”
- Reduces the skill level needed when examining z/OS for “unknown” problems where the system seems “sick”
- Provides timely, comprehensive analysis at a critical time period with suggestions on how to proceed

- Runs as a started task

z/OS 1.12 – Runtime Diagnostics …

- Does what you might do manually today:
  - Review critical messages in the log
  - Analyze contention
  - Examine address spaces with high CPU usage
  - Look for an address space that might be in a loop
  - Evaluate local lock conditions
- Additional analysis based on what it finds
  - For example, if XES reports connector as unresponsive, RTD will investigate the appropriate address space

z/OS 1.12 – Runtime Diagnostics …

For more information:

- z/OS V1R12 Problem Management (G325-2564)

XCF Programming Interfaces

- IXCMSGOX
  - 64 bit storage for sending messages
  - Duplicate message toleration
  - Message attributes: Recovery, Critical
- IXCMSGIX
  - 64 bit storage for receiving messages
- IXCJOIN
  - Recovery Manager
  - Critical Member
  - Termination level

Agenda

- Hardware Updates
  - CFCC Level 17
  - CFCC Level 16
  - Parallel Sysplex InfiniBand Links
- z/OS Updates
  - Sysplex Failure Management
  - z/OS V1R12
- **Summary**

Highlights

- **CFLEVEL 17 for z196**
- **Infiniband links for**
  - High performance links for z196
  - Bandwidth
  - Fewer physical links
  - High performance links at 150 meters
- **SFM with BCPii for better availability**
- **z/OS 1.12**
  - Sympathy sickness resolution for better availability
  - REALLOCATE test and report for CF structure management

Recent sysplex-related Redbooks

- System z Parallel Sysplex Best Practices, SG24-7817
- System z Parallel Sysplex Performance, SG24-7654
- Considerations for Multi-Site Sysplex Data Sharing, SG24-7263
- Sysplex Recovery Considerations in an STP Environment, SG24-7670

- Exploiting the IBM Health Checker for z/OS Infrastructure, REDP-4590

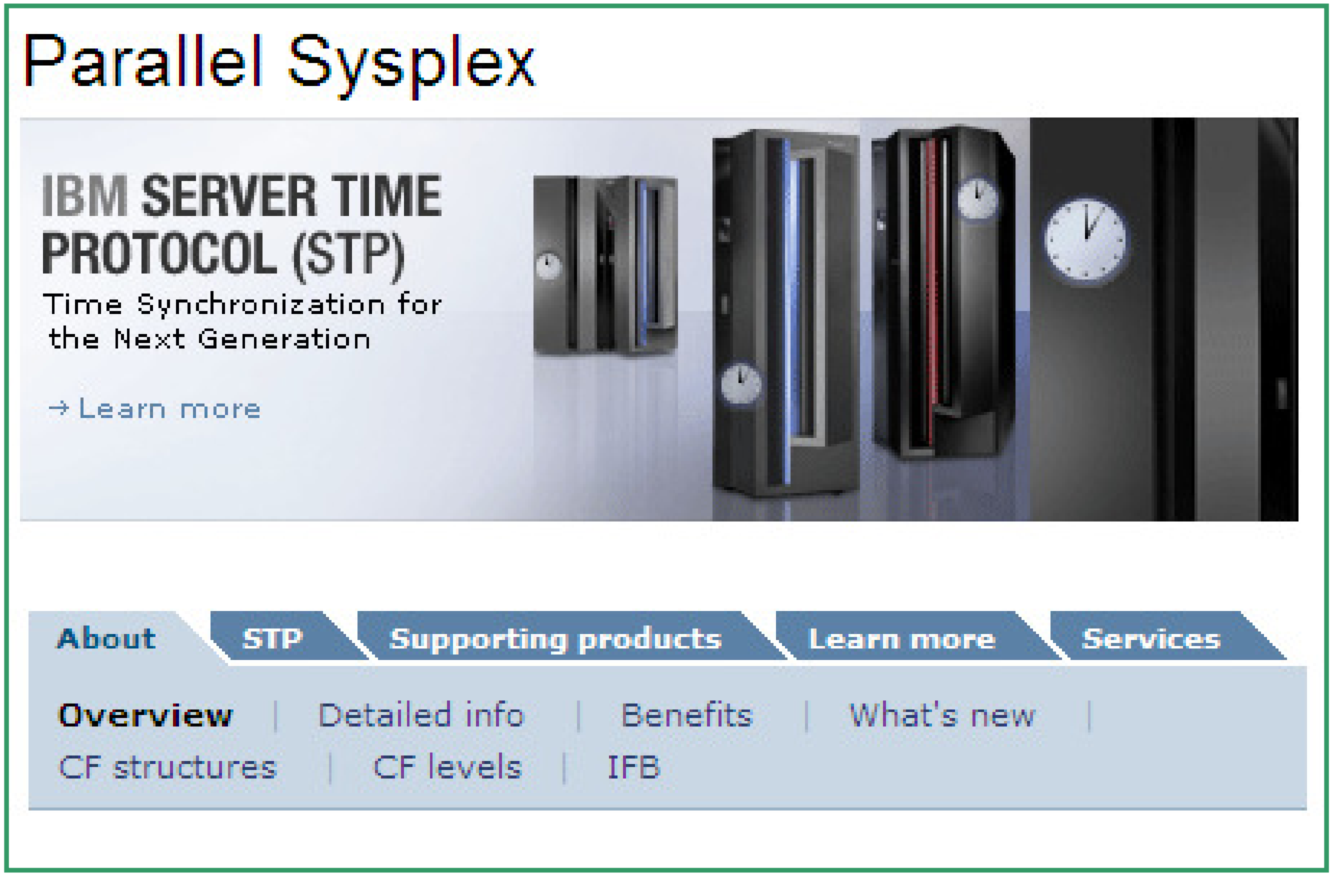
- Available at www.redbooks.ibm.com

Other Sources of Information

- *MVS Setting Up a Sysplex (SA22-7625)*
- *MVS Initialization and Tuning (SA22-7591)*
- *MVS Systems Commands (SA22-7627)*
- *MVS Diagnosis: Tools and Service Aids (GA22-7589)*
- *z/OS V1R12.0 Planning for Installation (GA22-7504)*
- z/OS MVS Programming: Callable Services for High Level Languages (SA22-7613)
  - Documents BCPii Setup and Installation and BCPii APIs

Parallel Sysplex Web Site

- www.ibm.com/systems/z/pso



Parallel Sysplex

IBM SERVER TIME PROTOCOL (STP)

Time Synchronization for the Next Generation

→ Learn more

About
STP
Supporting products
Learn more
Services
Overview | Detailed info | Benefits | What's new |
CF structures | CF levels | IFB

Questions?

- Questions?

Appendix

Additional topics and details of potential interest
These slides will not be presented during the session

Appendix

- D XCF,C shows sync/async conversion thresholds
- SETXCF FORCE,PENDSTR,CFNAME=xxx
- Large Number of Subchannels Support

DISPLAY XCF,COUPLE – z/OS 1.11 Sync/Async Conversion Thresholds

```
IXC357I  16.03.35  DISPLAY XCF 963
SYSTEM SYSB DATA
   INTERVAL    OPNOTIFY      MAXMSG     CLEANUP       RETRY   CLASSLEN
        105         108        2500          15          10        956

   SSUM ACTION  SSUM INTERVAL  SSUM LIMIT     WEIGHT  MEMSTALLTIME
       ISOLATE              0          60         80           180

   PARMLIB USER INTERVAL:      85
   DERIVED SPIN INTERVAL:     105
   DEFAULT USER OPNOTIFY: +     3

   MAX SUPPORTED CFLEVEL: 16

   MAX SUPPORTED SYSTEM-MANAGED PROCESS LEVEL: 16

   SIMPLEX SYNC/ASYNC THRESHOLD:                     34
   DUPLEX SYNC/ASYNC THRESHOLD:                      37
   SIMPLEX LOCK SYNC/ASYNC THRESHOLD:                34
   DUPLEX LOCK SYNC/ASYNC THRESHOLD:                 45
```

Available with APAR OA28603 at z/OS 1.8 and up

Structure “Pending Deallocation” The Problem

- Sysplex loses connectivity to CF
- Structures rebuilt or failover to duplexed copy
- XES checkpoints old structure instances so it can remember to deallocate them when connectivity is restored to CF
- These structure instances show up as “pending deallocation” on DISPLAY XCF output
- They cause confusion (and consume space in the policy)
  - Particularly when the CF is not coming back

Structure "Pending Deallocation" Solution

- New command option available in z/OS 1.11 or later
- SETXCF FORCE,PNDSTR,CFNAME=cfname
  - System will remove "pending deallocation" structure entries from the CFRM Active policy
  - The CF should not be connected to any system in the sysplex
- Intended for use when, either:
  - CF will remain inaccessible for an extended period
    - Such as forever
  - CF will be brought back online with all structures removed

Structure "Pending De-Allocation" Coexistence

- PNDSTR keyword only recognized as of z/OS 1.11
- Lower releases will not recognize the command
- But if a z/OS 1.11 system deletes the records, the lower releases will correctly:
  - Deal with the deleted checkpoint records
  - Display the correct information

Large Number of Subchannels Support
The problem

- Number and type of CF Links determine the number of message subchannels
- Which determines the number of concurrent operations that can be sent to the CF
- As number of concurrent operations increases, tendency for z/OS to convert synchronous requests to asynchronous processing increases
  - Especially if CF can’t keep up
- Which increases the tendency of z/OS to have long queues of message subchannel operations that require asynchronous completion processing

Large Number of Subchannels Support
The problem …

- Long queues can elongate service times for asynchronous requests
- Processing of CF Link Timeouts and CF Failures requires additional recovery processing time for each CF Request
  - And this time was elongating service times for other requests on the queue
- MVS Abends for Spin Loop Timeouts can occur

Large Number of Subchannels Support Solution

- Redesign the algorithm for processing completion of asynchronous CF Requests
  - Processing of CF Link Timeouts and CF Failures no longer affects service times of unrelated operations on the queue
  - Eliminate MVS Abends that are encountered for Spin Loop Timeouts while processing long queues
- Update the D CF and D M=CHP Command Output to compress the larger number of message subchannels/devices output to as few pages as necessary

Large Number of Subchannels Support Applicability

- Available in z/OS 1.11 and up
- Available at z/OS 1.7 and up with APAR OA26033

- Likely most beneficial to:
  - Installations that define close to the maximum number of CF links and CF message subchannels
  - Installations that need to increase number of concurrent operations (more message subchannels)
    - As might be done with PSIFB Links