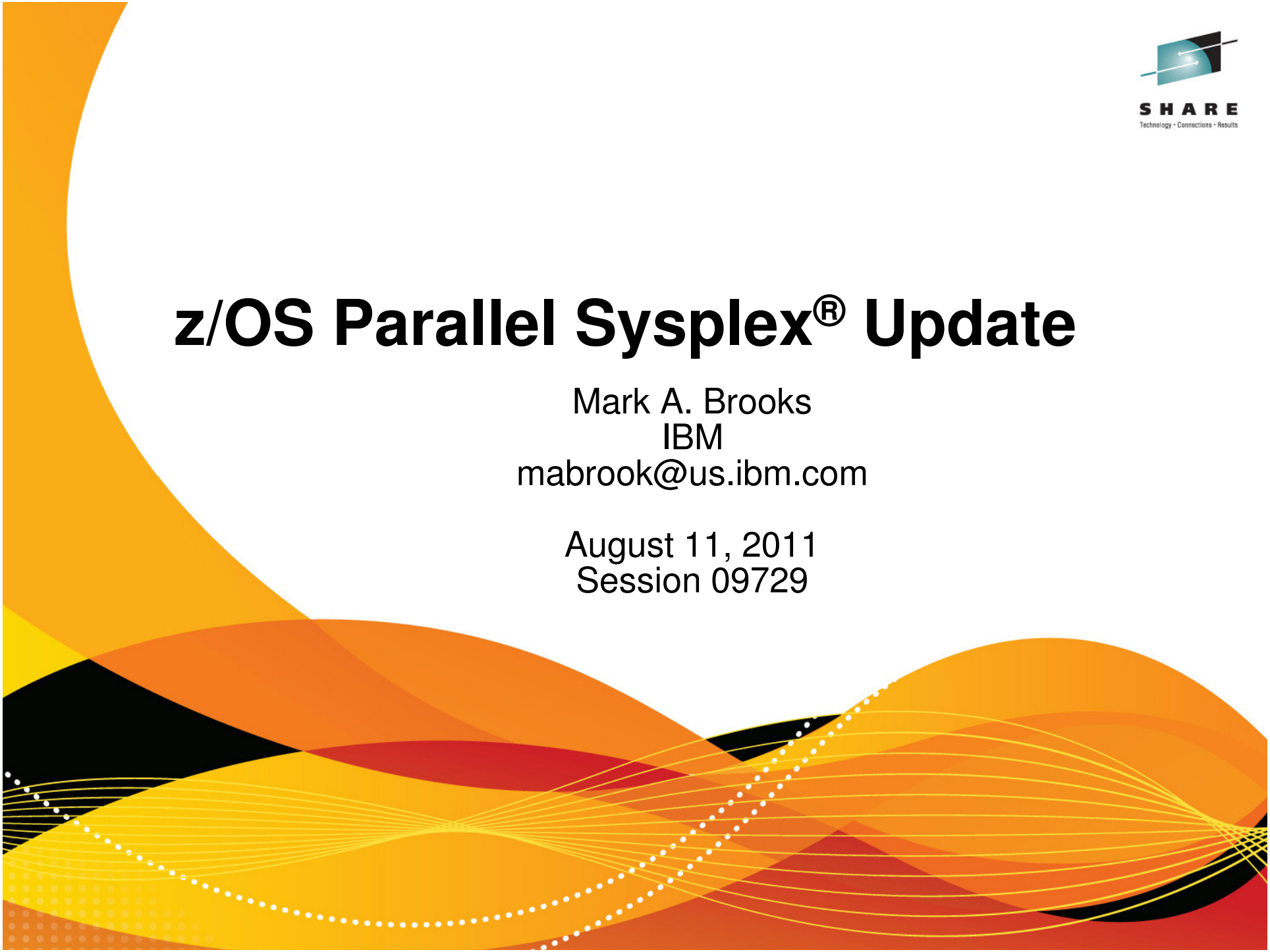


z/OS Parallel Sysplex® Update

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

August 11, 2011
Session 09729

Trademarks

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

IBM®	MQSeries®	S/390®	z9®
ibm.com®	MVS™	Service Request Manager®	z10™
CICS®	OS/390®	Sysplex Timer®	z/Architecture®
CICSPlex®	Parallel Sysplex®	System z®	zEnterprise™
DB2®	Processor Resource/Systems Manager™	System z9®	z/OS®
eServer™	PR/SM™	System z10®	z/VM®
ESCON®	RACF®	Tivoli®	z/VSE®
FICON®	Redbooks®	VTAM®	zSeries®
HyperSwap®	Resource Measurement Facility™	WebSphere®	
IMS™	RETAIN®		
IMS/ESA®	GDPS®		
	Geographically Dispersed Parallel Sysplex™		

The following are trademarks or registered trademarks of other companies.

IBM, z/OS, Predictive Failure Analysis, DB2, Parallel Sysplex, Tivoli, RACF, System z, WebSphere, Language Environment, zSeries, CICS, System x, AIX, BladeCenter and PartnerWorld are registered trademarks of IBM Corporation in the United States, other countries, or both.
DFSMShsm, z9, DFSMSrmm, DFSMSdfp, DFSMSdss, DFSMS, DFS, DFSORT, IMS, and RMF are trademarks of IBM Corporation in the United States, other countries, or both.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United Sta/tes, 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.
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.

* 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.

Statements of Direction

- **z196 announcement, July 22, 2010**
- **z114 announcement, July 12, 2011**
- **The z196 and z114 will be the last System z servers to:**
  - **Offer ordering of ESCON channels**
  - **Offer ordering of ISC-3**
  - **Support dial-up modem**
- Implications
  - If using CTC devices for XCF signalling paths, need to migrate to FICON from ESCON
  - Migrate from ISC-3 coupling links to infiniband
  - Migrate to alternatives for dial-up time services

Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

CFCC Level 17 – Constraint Relief

- Up to 2047 structures per CF image (prior limit 1023)
  - Many data sharing groups, SAP, CICS, merging plexes
  - New version of CFRM CDS needed to define more than 1024 structures
- Supports up to 255 connectors for all structure types
  - Cache structures already support 255 connectors
  - z/OS imposes smaller limits for lock structures (247) and serialized list structures (127)
    - Will require exploiter changes as well (none yet!)
- Prerequisites
  - z/OS V1.10 or later with PTF for OA32807
  - z/VM V5.4 for guest virtual coupling

CFCC Level 17 - Serviceability

- CF Diagnostics
  - Non-disruptive dumping
  - Coordinated dump capture
    - Gathers z/OS, CF, and link diagnostics at same time
    - Use DUPLEXCFDIAG to enable

Also available for z10
CFCC Level 16 (need MCLs)
z/OS APAR OA33723

- Prerequisites
  - z/OS V1.12
  - z/OS 1.10 or 1.11 with PTFs for OA31387

CFCC Level 17 - Migration

- 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
  - Improperly sized structures can lead to outages !
- Note that minimum CFCC image size is now 512MB

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

Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

Glossary for System z Coupling

Acronym	Full name	Comments
AID	Adapter identification	HCA fanout has AID instead of a PCHID
CIB	Coupling using InfiniBand	CHPID type z196, z10, System z9
HCA	Host Channel Adapter	Path for communication
MBA	Memory Bus Adapter	Path for communication
PSIFB	Parallel Sysplex using InfiniBand	InfiniBand Coupling Links
12x IFB	12x InfinBand	12 lanes of fiber in each direction
1x IFB	1x InfiniBand	Long Reach - one pair of fiber
12x IFB3	12x InfiniBand3	Improved service times of 12x IFB on HCA3-O

Glossary for System z Coupling …

Type	System z10	System z196
HCA1-O fanout	NA	NA
HCA2-O fanout	Optical - Coupling 12x InfiniBand	Optical - Coupling 12x InfiniBand
HCA2-O LR fanout	Optical - Coupling 1x InfiniBand	Optical - Coupling 1x InfiniBand
HCA3-O fanout	NA	Optical - Coupling 12x InfiniBand
HCA3-O LR fanout	NA	Optical - Coupling 1x InfiniBand
MBA fanout	Copper - Coupling (ICB-4)	N/A

Coupling link choices

Short Distance

- IC (Internal Coupling Channel)
  - Microcode - no external connection
  - Only between partitions on same processor
- ICB-3 and ICB-4 (Integrated Cluster Bus)
  - Copper cable plugs close to memory bus
  - 10 meter max length

- 12x IFB, 12x IFB3
  - 150 meter max distance optical cabling

Extended Distance

- ISC-3 (InterSystem Channel)
  - Fiber optics
  - I/O Adapter card
  - 10km and longer distances with qualified DWDM solutions
- 1x IFB
  - Fiber optics – uses same cabling as ISC
  - 10km and longer distances with qualified DWDM solutions

InfiniBand (IFB) Overview

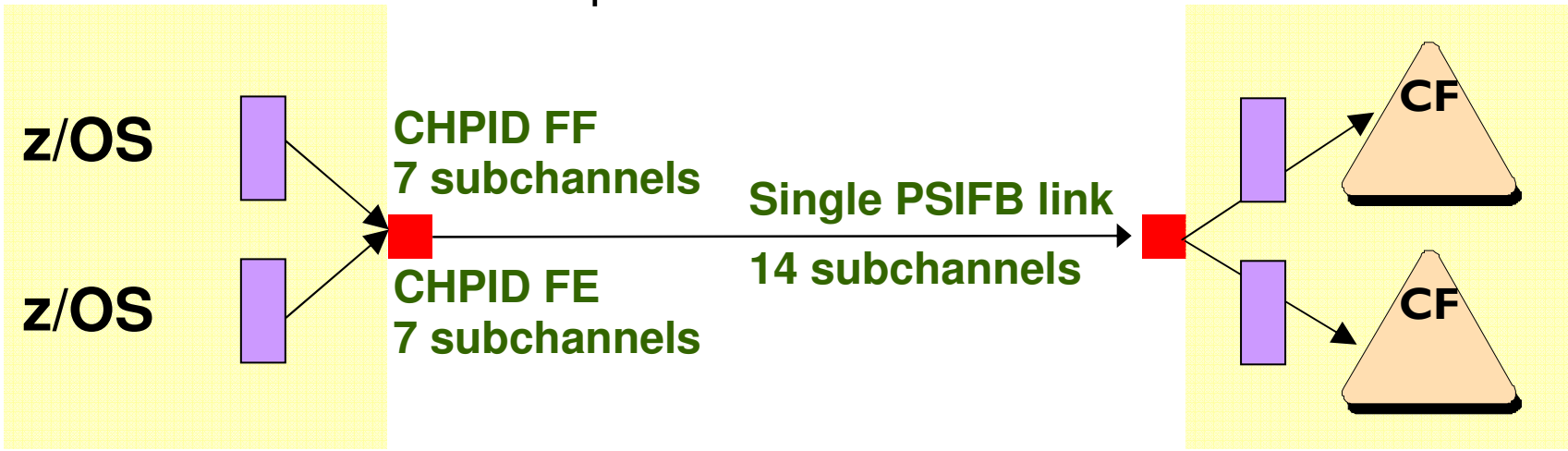
- **Physical lane**
  - Link based on a two-fiber 2.5 Gbps bidirectional connection for an optical (fiber cable) implementation
  - Grouped as either 12 physical lanes (12x) or 1 physical lane (1x)
- **Link speeds**
  - Single data rate (SDR) delivering 2.5 Gbps per physical lane
  - Double data rate (DDR) delivering 6.0 Gbps per physical lane
- **Host Channel Adapter (HCA)**
  - Physical devices that create and receive packets of information
- **CHPID type (CIB) for both 12x IFB and 1x IFB**
  - 7 subchannels per CHPID (12x)
  - 32 for HCA2-O LR and HCA3-O LR (1x)
- IFB available on z196, z114, z10, z9
  - HCA3-O exclusive to z196 and z114

New Infiniband Support
July 12, 2011 Announce

- Improved service times with 12x IFB links
  - New 12x IFB3 protocol, applies when:
  - HCA3-O to HCA3-O connectivity, <= 4 CHPIDs/port
  - Designed to improve link service time up to 40%
- Improved physical connectivity with 1x IFB links
  - HCA3-O LR has 4 ports instead of 2
  - Helps with migration from ISC-3 links
- Up to 32 subchannels per CHPID with 1x IFB links
  - HCA3-O LR or HCA2-O LR
  - Helps with bandwidth at longer distances

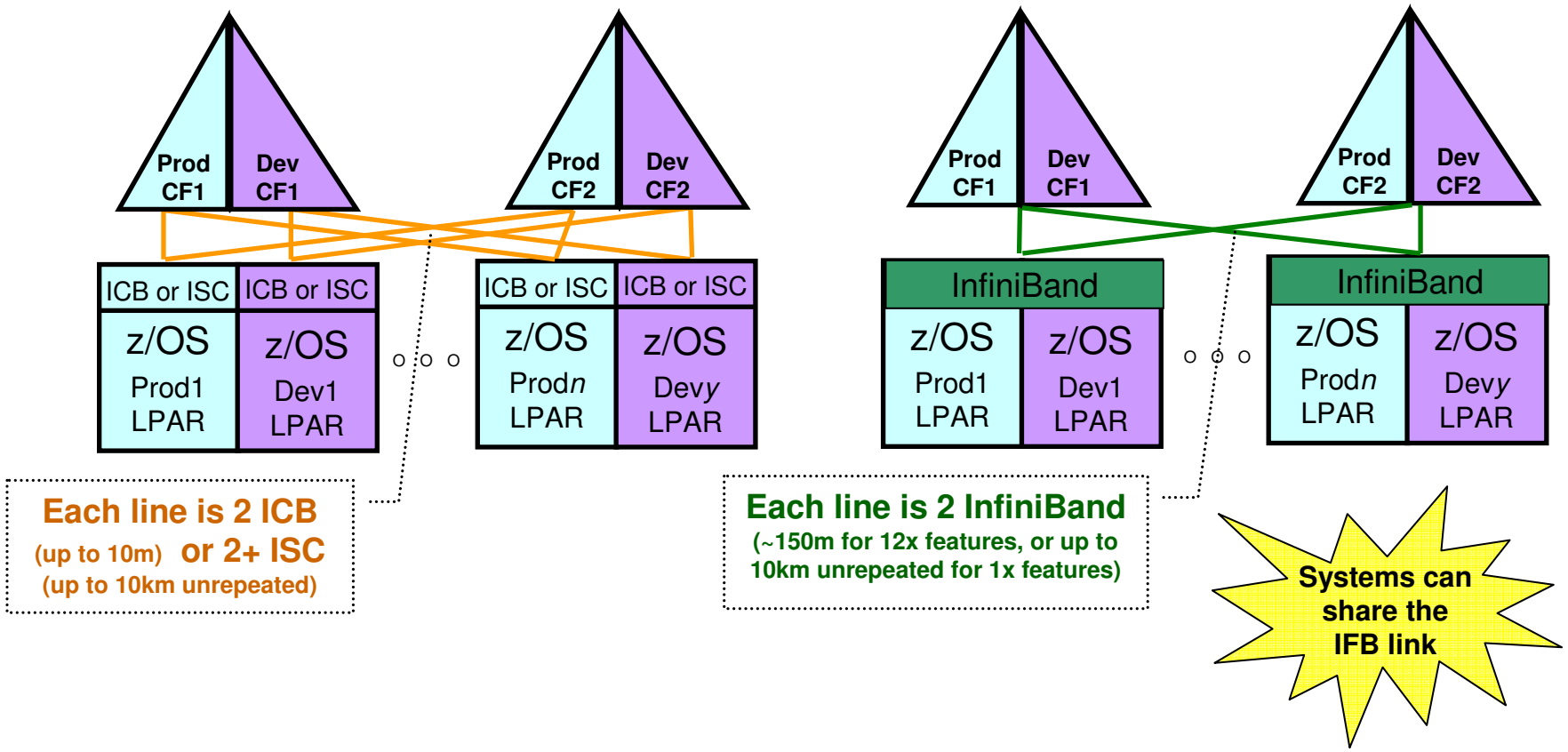
IFB Supports Multiple Channel Paths per Physical Link

- Up to 16 CHPIDs across the ports of a single InfiniBand coupling HCA
  - Allows more subchannels per physical link
- Can connect to multiple CF LPARs



- MIF uses same address, 7 subchannels per CHPID
- With HCA2-O LR or HCA3-O LR, 32 subchannels per CHPID

Lower Cost Coupling Infrastructure – consolidating coupling links



Consolidating links with IFB

- Pure Capacity
  - 1 12x IFB replaces 1 ICB-4
  - 1 12x IFB replaces 4 ISC-3s

Be sure to maintain redundancy !

- Eliminating subchannel and path delays
  - Often >2 ICB-4s configured not for capacity but for extra subchannels/paths to eliminate delays
  - 2 12x IFB links with multiple CHPIDs can replace >2 ICB-4s in this case
- Multiple sysplexes sharing hardware
  - Production, development, test sysplexes may share hardware
    - Previously each required own ICB-4 or ISC-3 links
  - 2 12x or 1x IFB links with multiple CHPIDs can replace >2 ICB-4s or ISC-3s in this case
- Multiple CHPID recommendations
  - Max 16 per HCA (2 ports per HCA)
    - Use up to all 16 for lightly loaded connectivity
    - Limit to max of 8 per HCA for heavy loads

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	z9 BC	z9 EC	z10 BC	z10 EC	z114	z196
z9 BC ISC3	14%	15%	17%	19%	18%	23%
z9 BC 12x IFB	NA	NA	13%	14%	13%	16%
z9 BC ICB4	9%	10%	11%	12%	NA	NA
z9 EC ISC3	13%	14%	16%	18%	17%	22%
z9 EC 12x IFB	NA	NA	13%	14%	13%	16%
z9 EC ICB4	8%	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	8%	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	10%	12%
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%

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

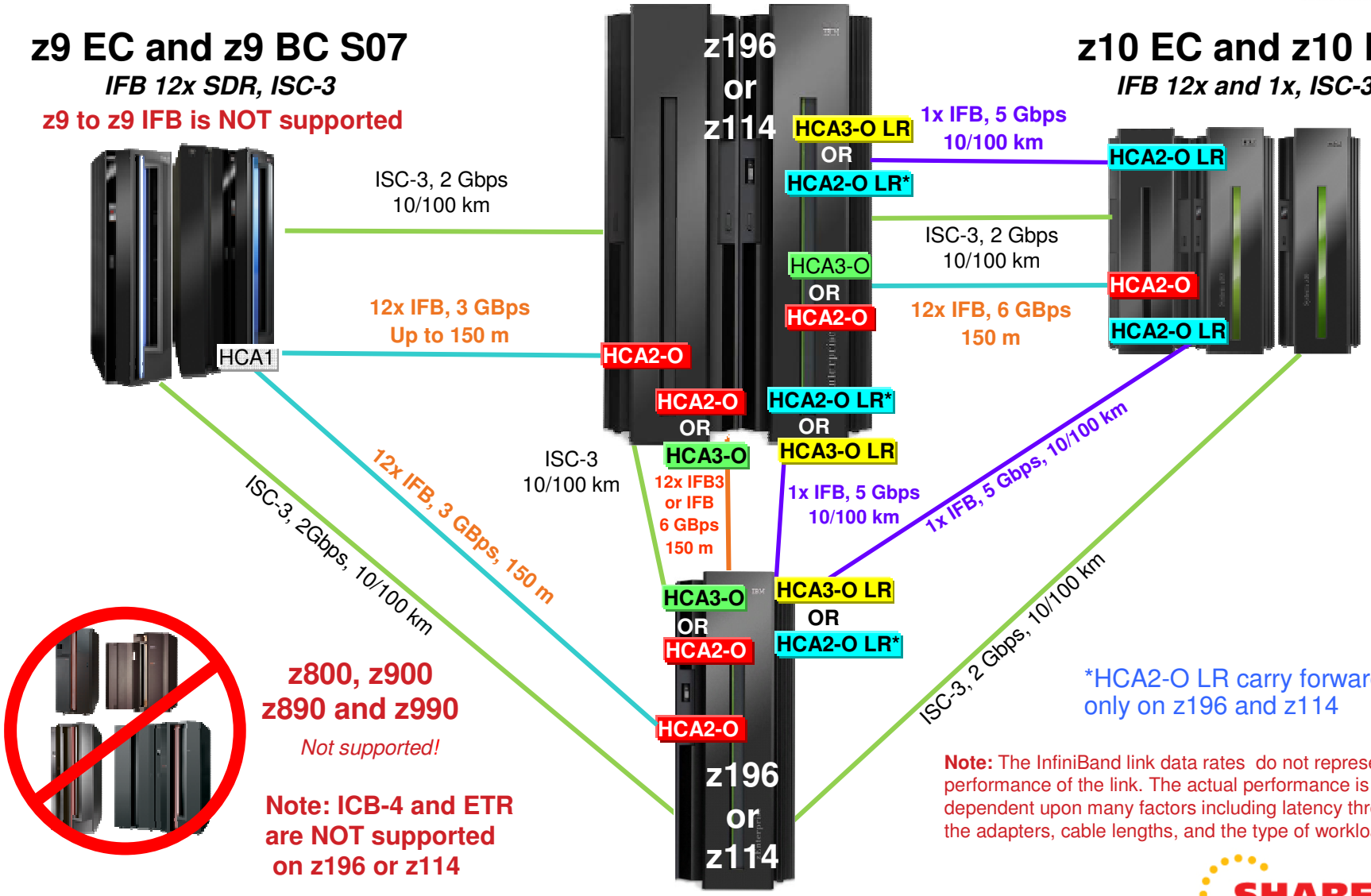
Maximum Coupling Links and CHPIDs

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*	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*	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 (4)	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	16 S08 - 12	32	16	16	48	64	64
z9 BC	N/A	N/A	N/A	12	32	16	16	48	64	64

1. z196 & z114 do not have an inherent maximum external link limit. The effective limit depends on the HCA fanout slots available and combined 12x IFB and 1x IFB limit of 16 HCA features
   z196 M49, M66 or M80 supports max 96 extended distance links (48 1x IFB and 48 ISC-3) plus 8 12x IFB links
   z196 M32 supports max 96 extended distance links (48 1x IFB and 48 ISC-3) plus 4 12x IFB links*
   z196 M15 supports max 72 extended distance links (24 1x IFB and 48 ISC-3) with no 12x IFB links*
2. z114 M10 supports max 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*
4. ICB-4 not supported on Model E64. 32 ICB-4 links with RPQ on z10 EC

* Uses all available fanout slots. Allows no other I/O or coupling

z114 and z196 GA2 Parallel Sysplex Coupling Connectivity



IFB Resources For Making the Transition

- Parallel Sysplex Website: http://www.ibm.com/systems/z/advantages/pso/ifb.html
- Redbooks and Whitepapers
  - "Coupling Configuration Options" whitepaper
  - "Getting Started with InfiniBand on System z10 and System z9" SG24-7539
  - "IBM System z Connectivity Handbook" SG24-5444
- Tools
  - STG Lab Services – Specialized studies available for complex situations.
    - Send a note to stgls@us.ibm.com.
  - **zCP3000** (Performance Analysis and Capacity Planning) – Includes reports of CF and CP utilization given a change of coupling link types
    - Contact your IBM representative to use this tool

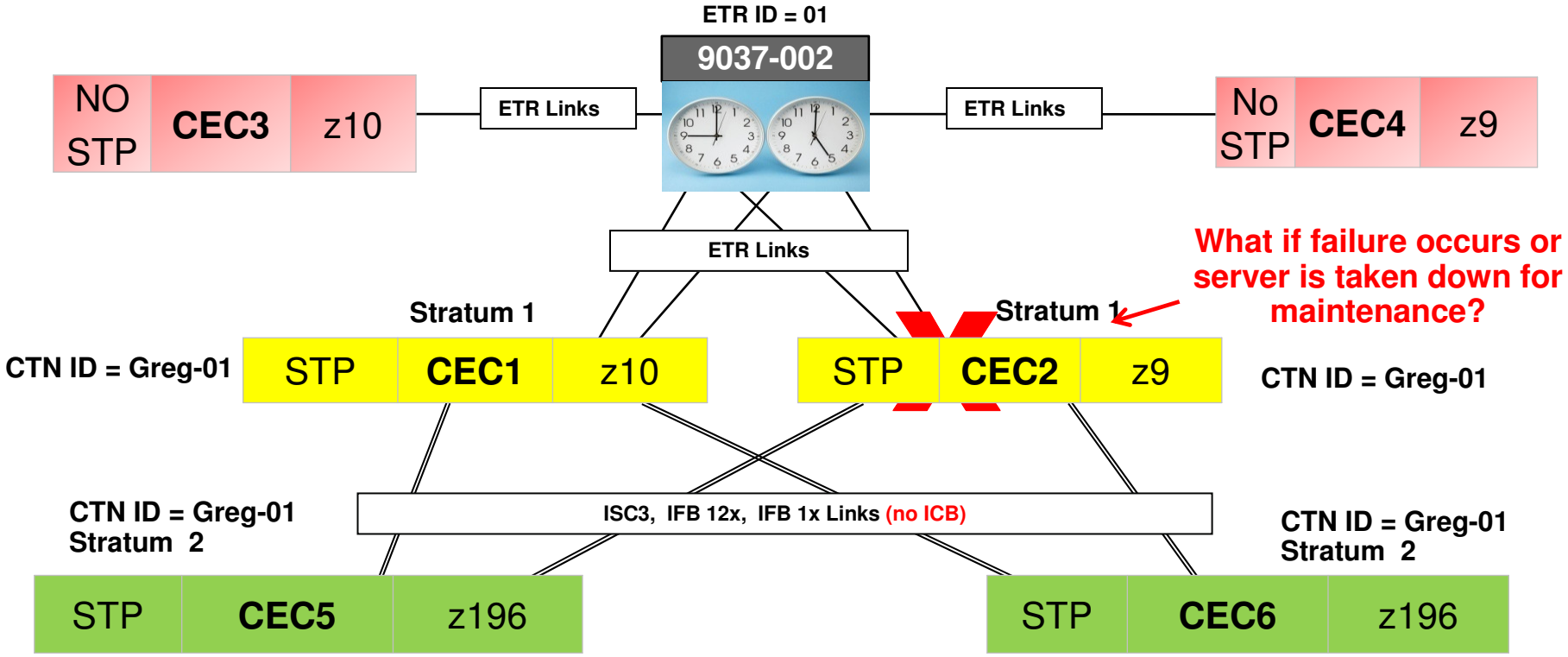
Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

Glossary for Server Time Protocol (STP)

Acronym	Full name	Comments
Arbiter	Arbiter	Server assigned by the customer to provide additional means for the Backup Time Server to determine whether it should take over as the Current Time Server.
BTS	Backup Time Server	Server assigned by the customer to take over as the Current Time Server (stratum 1 server) because of a planned or unplanned reconfiguration.
CST	Coordinated Server Time	The Coordinated Server Time in a CTN represents the time for the CTN. CST is determined at each server in the CTN.
CTN	Coordinated Timing Network	A network that contains a collection of servers that are time synchronized to CST.
CTN ID	Coordinated Timing Network Identifier	Identifier that is used to indicate whether the server has been configured to be part of a CTN and, if so, identifies that CTN.
CTS	Current Time Server	A server that is currently the clock source for an STP-only CTN.
PTS	Preferred Time Server	The server assigned by the customer to be the preferred stratum 1 server in an STP-only CTN.

No Support for ETR with z196 and z114
Migrate to STP



- It is possible to have a z196 server as a Stratum 2 server in a Mixed CTN as long as there are at least two z10s or z9s attached to the Sysplex Timer operating as Stratum 1 servers
- Two Stratum 1 servers are highly recommended to provide redundancy and avoid a single point of failure
- Suitable for a customer planning to migrate to an STP-only CTN.
- Neither a z196 nor z114 can be in the same Mixed CTN as a z990 or z890 (n-2)

STP Recovery Enhancement

- Reliable unambiguous “going away” signal allows CTS in an STP-only CTN to notify BTS of its demise
- The BTS can then safely take over as the CTS
  - Dependencies on OLS and CAR removed in a 2 server CTN
  - Dependency on BTS>Arbiter communication removed in CTNs with 3 or more servers
  - BTS can also use GOSIG to take over as CTS for CTNs with 3 or more servers without communicating with Arbiter
- Hardware Requirements
  - STP-only CTN
  - z196 or z114
  - HCA3-O or HCA3-O LR connecting BTS and CTS

Improved Time Synchronization for zBX componentry

- NTP clients on blades in zBX can synchronize to the NTP server in the Support Element (SE)
- The SE's battery operated clock can now maintain time accuracy within 100 milliseconds of the NTP server
- And so components in zBX can maintain time accuracy within 100 milliseconds of the NTP server

Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

z/OS V1R13 - Summary

- D XCF,SYSPLEX – Revised output
- CF Structure Placement – more explanation
- SETXCF MODIFY - Disable structure alter processing
- ARM – New timeout parameter for application cleanup
- XCF – New API for message passing

- SDSF – Sysplex wide data gathering without MQ
- Runtime Diagnostics – Detects more contention
- zFS – Direct access to shared files throughout sysplex

z/OS V1R13 - DISPLAY XCF,SYSPLEX

- D XCF,SYSPLEX command is a popular command used to display the systems in the sysplex
- But, prior to z/OS V1R13:
  - Output not as helpful for problem diagnosis as it could be
  - Much useful system and sysplex status information is kept by XCF, but not externalized in one central place
- So z/OS V1R13 enhances the output
  - You can still get the same output (perhaps with new msg #)
  - And you can get more details than before

z/OS V1R13 – D XCF,SYSPLEX,ALL

	z/OS 1.12
D XCF,S,ALL	IXC335I 12:55:00 DISPLAY XCF FRAME LAST F E SYS=SY1 SYSPLEX PLEX1 SYSTEM TYPE SERIAL LPAR STATUS TIME SYSTEM STATUS SY1 4381 9F30 N/A 04/22/2011 12:55:00 ACTIVE TM=SIMETR SY2 4381 9F30 N/A 04/22/2011 12:54:56 ACTIVE TM=SIMETR SY3 4381 9F30 N/A 04/22/2011 12:54:56 ACTIVE TM=SIMETR SYSTEM STATUS DETECTION PARTITIONING PROTOCOL CONNECTION EXCEPTIONS: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE SSD PROTOCOL
	z/OS 1.13
D XCF,S,ALL	IXC337I 12.29.36 DISPLAY XCF FRAME LAST F E SYS=SY1 SYSPLEX PLEX1 MODE: MULTISYSTEM-CAPABLE SYSTEM SY1 STATUS: ACTIVE TIMING: SIMETR NETID: 0F STATUS TIME: 05/04/2011 12:29:36.000218 JOIN TIME: 05/04/2011 10:31:08.072275 SYSTEM NUMBER: 01000001 SYSTEM IDENTIFIER: AC257038 01000001 SYSTEM TYPE: 4381 SERIAL: 9F30 LPAR: N/A NODE DESCRIPTOR: SIMDEV.IBM.PK.D13ID31 PARTITION: 00 CPCID: 00 RELEASE: z/OS 01.13.00 SYSTEM STATUS DETECTION PARTITIONING PROTOCOL CONNECTION EXCEPTIONS: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE SSD PROTOCOL

z/OS V1R13 – CF Structure Placement

- Why did it put my structure in that CF ?
  - A dark art, often a mystery to the observer
- Existing messages updated to help explain
  - IXL015I: Initial/rebuild structure allocation
    - Also has “CONNECTIVITY=” insert
  - IXC347I: Reallocate/Reallocate test results
  - IXC574I: Reallocate processing, system managed rebuild processing, or duplexing feasibility

z/OS V1R13 – CF Structure Placement …

```
IXL015I STRUCTURE ALLOCATION INFORMATION FOR
STRUCTURE THRLST01, CONNECTOR NAME THRLST0101000001,
CONNECTIVITY=SYSPLEX
 CFNAME        ALLOCATION STATUS/FAILURE REASON
 --------      ----------------------------------------
 LF01          ALLOCATION NOT PERMITTED
               COUPLING FACILITY IS IN MAINTENANCE MODE
 A             STRUCTURE ALLOCATED CC007B00
 TESTCF        PREFERRED CF ALREADY SELECTED CC007B00
               PREFERRED CF HIGHER IN PREFLIST
 LF02          PREFERRED CF ALREADY SELECTED CC007300
               EXCLLIST REQUIREMENT FULLY MET BY PREFERRED CF
 SUPERSES      NO CONNECTIVITY 98007800
```

CF Structure Alter Processing

- CF Structure Alter processing is used to dynamically reconfigure storage in the CF and its structures to meet the needs of the exploiting applications
  - Size of structures can be changed
  - Objects within structures can be reapportioned
- Alter processing can be initiated by the system, the application, or the operator
- There have been occasional instances, either due to extreme duress or error, where alter processing has contributed to performance problems
- Want an easy way to inhibit alter processing ….

z/OS V1R13 – Enable/Disable Start Alter Processing

- SETXCF MODIFY,STRNAME=pattern,ALTER=DISABLED
- SETXCF MODIFY,STRNAME=pattern,ALTER=ENABLED
  - STRNAME=strname
  - STRNAME=strprfx*
  - STRNAME=ALL | STRNAME=*
- D XCF,STRUCTURE, ALTER={ENABLED|DISABLED}
- Only systems with support will honor ALTER=DISABLED indicator in the active policy
  - So you may not get the desired behavior until the function is rolled around the sysplex
  - But fall back is trivial since downlevel code ignores it
- APAR OA34579 for z/OS V1R10 and up

Automatic Restart Management (ARM)

- If you have an active ARM policy, then:
  - After 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

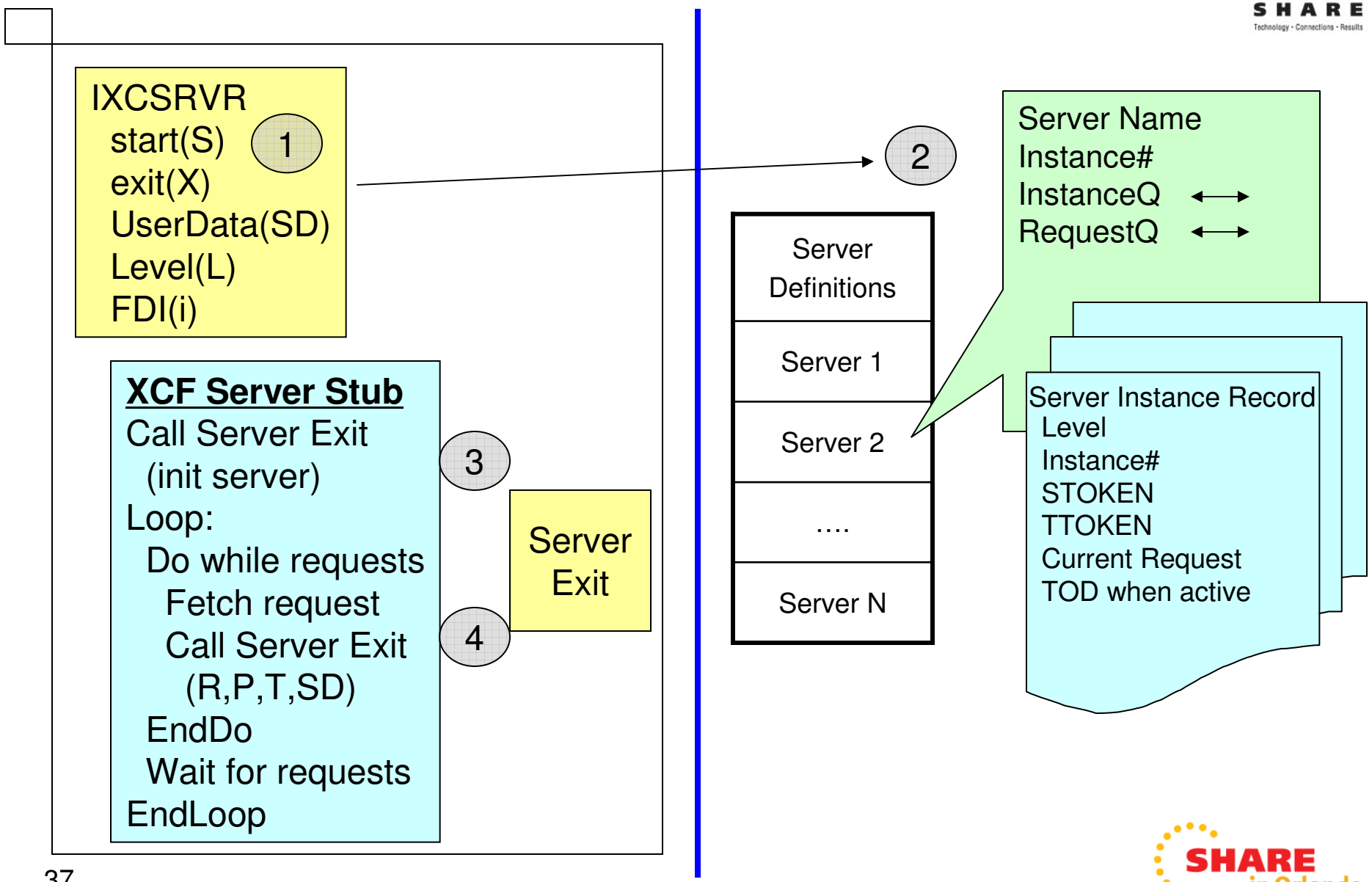
z/OS V1R13 – New ARM Parameter

- CLEANUP_TIMEOUT
  - New parameter for the ARM policy specifies how long ARM should wait for survivors to cleanup for a failed system
  - Specified in seconds, 120..86400 (2 min to 24 hours)
- If parameter not specified
  - Defaults to 300 seconds (5 minutes, not 2)
  - Code 120 if you want to preserve old behavior
- If greater than 120:
  - Issues message IXC815I after two minutes to indicate that restart is being delayed
  - If the timeout expires, issues message IXC815I to indicate restart processing is continuing despite incomplete cleanup
- Available for z/OS V1R10 and up with APAR OA35357

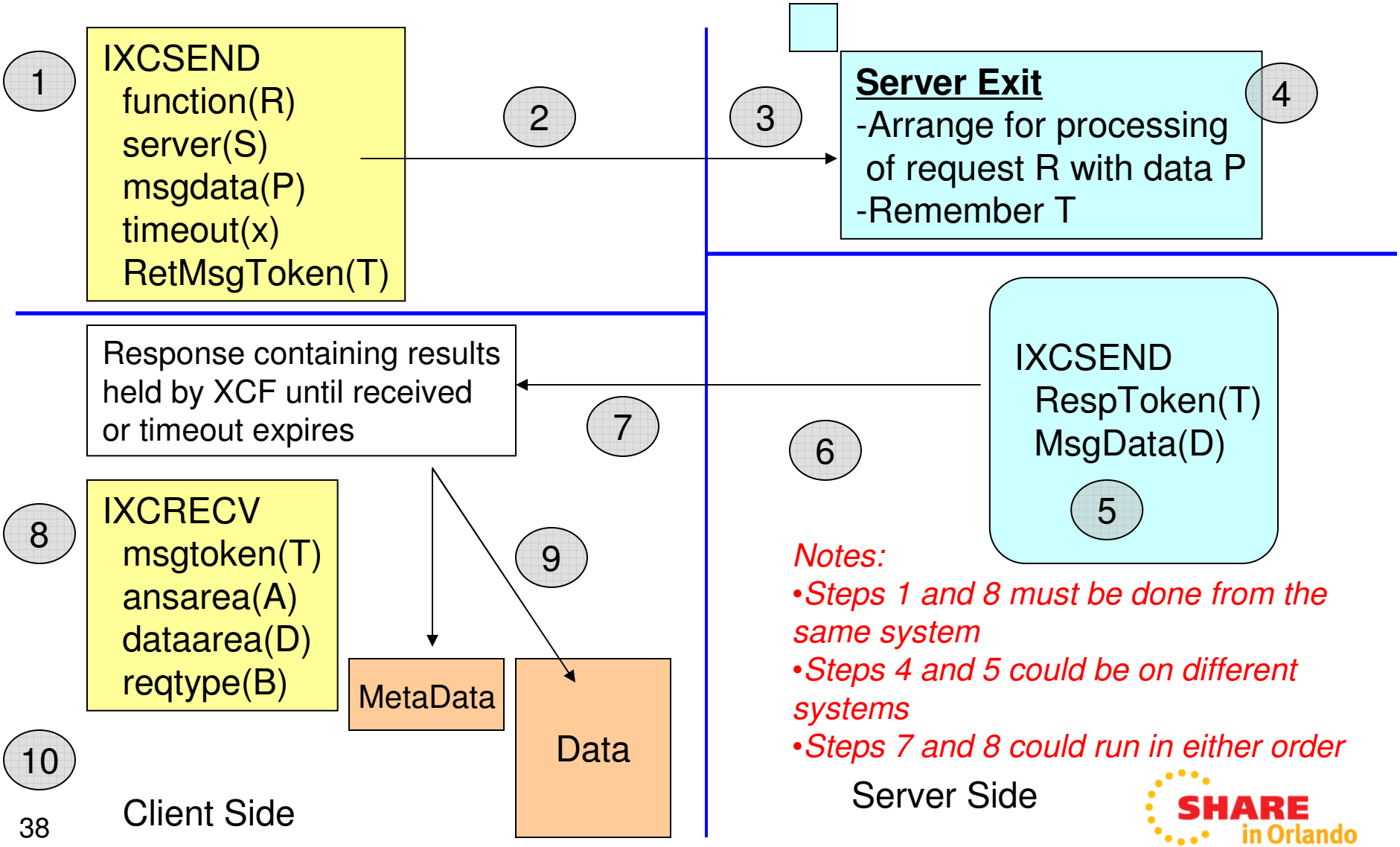
z/OS V1R13 – New XCF API for Message Passing

- **XCF Client/Server Interfaces**
  - Allows authorized programs to send and receive signals within a sysplex without joining an XCF Group
  - XCF does communication and failure handling
  - Simplifies development, reduces complexity, implementation and support costs by eliminating some of the XCF exploitation costs
  - Servers run in task mode

XCF Client/Server – Server Task



XCF Client/Server - Send/Receive



Notes:

- *Steps 1 and 8 must be done from the same system*
- *Steps 4 and 5 could be on different systems*
- *Steps 7 and 8 could run in either order*

z/OS V1R13 - SDSF

- SDSF provides sysplex view of panels:
  - Health checks; processes; enclaves; JES2 resources
- Data gathered on each system using the SDSF server
- Consolidated on client for display so user can see data from all systems
- Previously used MQ series to send and receive requests
  - Requires configuration and TCP/IP, instance of MQ queue manager on each system
- z/OS V1R13 implementation uses XCF Client/Server
  - No additional configuration requirements

Session 09720: z/OS 1.13 SDSF Update
Thu 4:30

z/OS V1R13 – 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 in z/OS V1R12
  - S HZR
- Starts at IPL in z/OS V1R13
  - F HZR,ANALYZE command initiates report

Session 9867: z/OS Diagnostics Extensions:
Runtime Diagnostics and Base Diagnostic Aids
Wed 4:30

z/OS V1R13 – Runtime Diagnostics …

Does what you might do manually today:

- Review critical messages in the log
- Analyze contention
  - GRS ENQ
  - GRS Latches
  - z/OS UNIX file system latches
- Examine address spaces with high CPU usage
- Look for an address space that might be in a loop
- Evaluate local lock conditions
- Perform additional analysis based on what is found
  - For example, if XES reports a connector as unresponsive, RTD will investigate the appropriate address space

z/OS V1R13 - zFS

- Full read/write capability from anywhere in the sysplex for shared file systems
  - Better performance for systems that are not zFS owner
  - Reduced overhead on the owner system
- Expected to improve performance of applications that use zFS services
  - z/OS UNIX System Services
  - WebSphere® Application Server

Session 09739: Significant Enhancements in z/OS V1R13 zFS
Wed 3:00

Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

z/OS V1R12 Summary

- **Critical Members**
- **CFSTRHANGTIME**
- **REALLOCATE**
- Support for CFLEVEL 17
- Health Checks
- Auto Reply
- Run Time Diagnostics
- XCF Programming Interfaces

Due to time restrictions, only the topics in bold will be discussed.
Slides for the remaining topics are included in the Appendix

z/OS V1R12 - 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 OR it should be removed to avoid sympathy sickness

z/OS V1R12 - 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
  - SFM parameter MEMSTALLTIME determines “long enough”
- GRS is a “system critical member”
  - XCF will remove a system from the sysplex if GRS on that system becomes “impaired”

z/OS V1R12 - 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 V1R12 - 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 V1R12 - Critical Members …

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

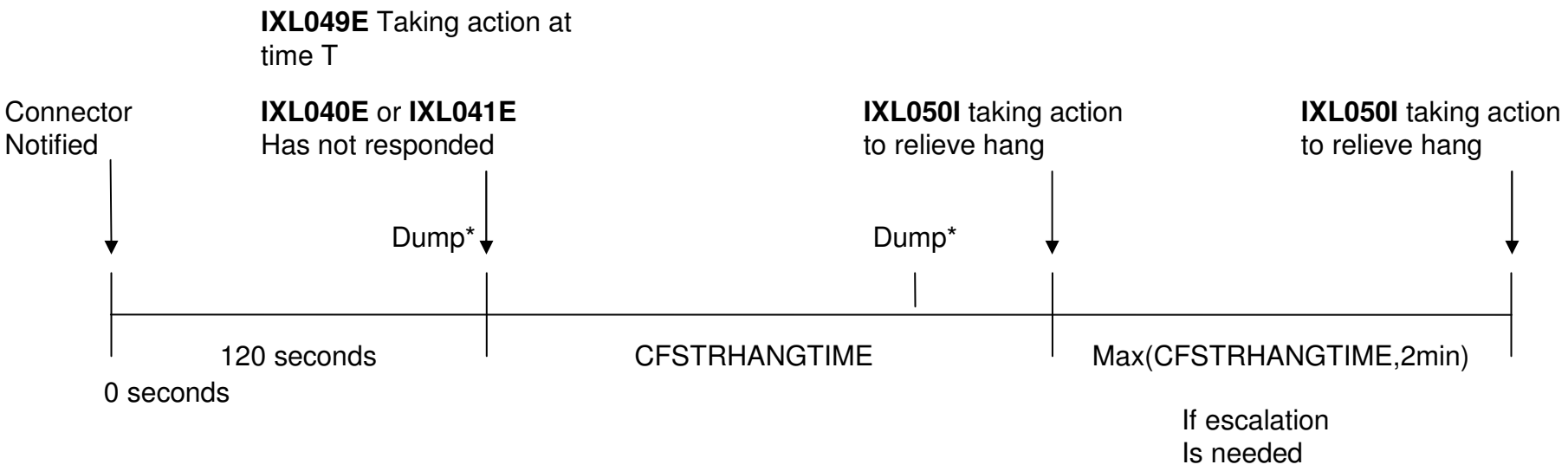
XES Connector Hang Detection

- 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
  - Impact: sympathy sickness, delays, outages
- Need a way to resolve this automatically …

z/OS 1VR12 – 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 V1R12 – CFSTRHANGTIME Processing



Dump* = Base release, dump is taken either when hang is announced or just prior to termination.
With OA34440, dump taken only when hang is announced

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 V1R12 – CFSTRHANGTIME …

- Coexistence
  - Toleration APAR OA30880 for z/OS V1R10 and z/OS V1R11 makes reporting of the CFSTRHANGTIME keyword with IXCMIAPU utility possible on those releases.
  - However the capability to take action to resolve the problem is not rolled back to previous releases

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 V1R12 - 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

z/OS V1R12 – REALLOCATE …

Caveats for 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

z/OS V1R12 – REALLOCATE …

Caveats for REPORT option

- 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

Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

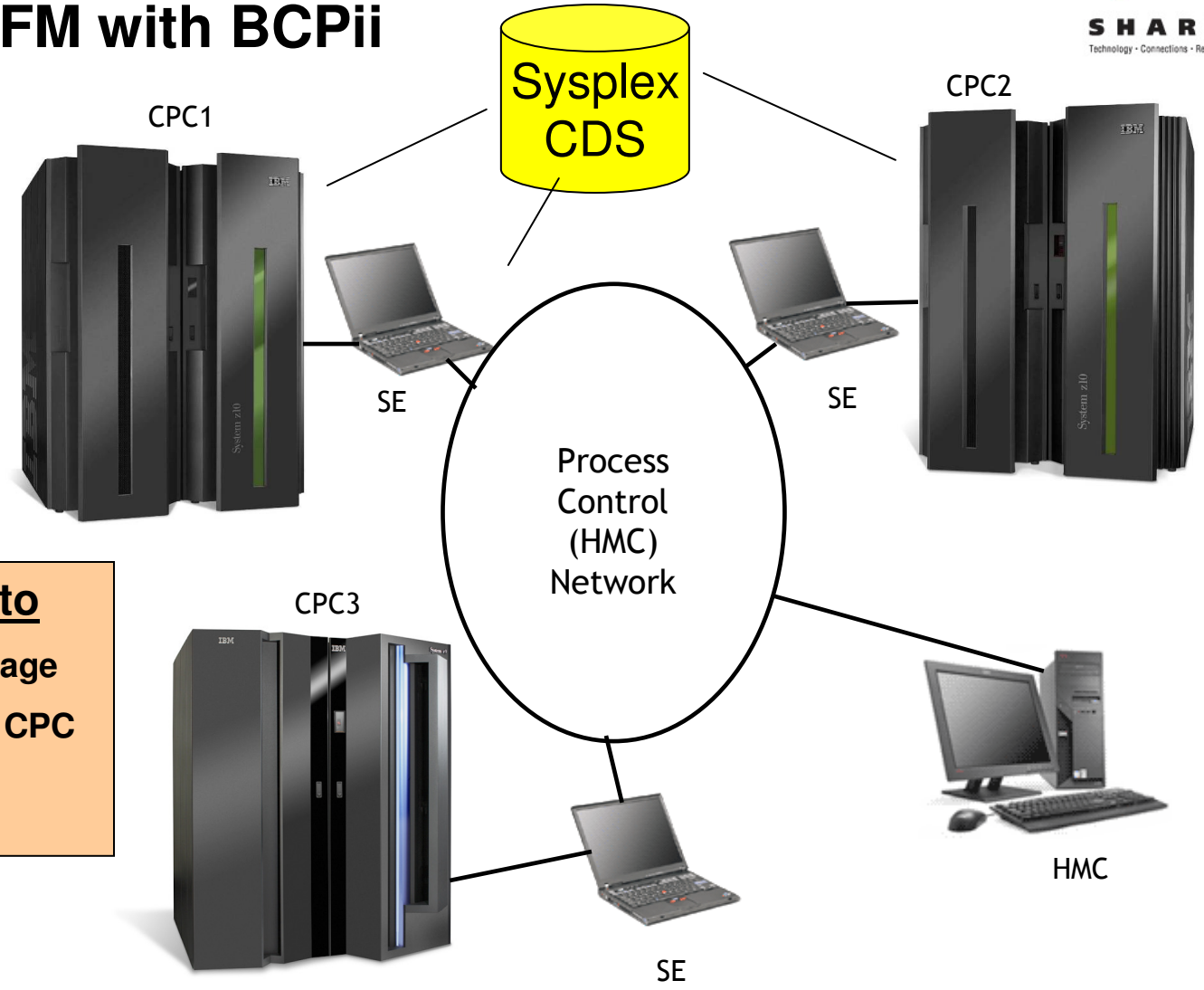
z/OS V1R11 - Summary

- **SFM with BCPii**
- System Default Action
- XCF FDI Consistency

z/OS V1R11 - 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

z/OS V1R11 - SFM with BCPii



z/OS Images

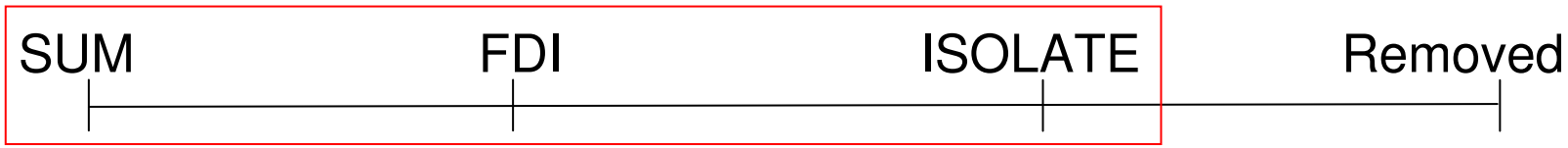
- Oops
- EPO

XCF uses BCPii to

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

z/OS V1R11 - 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



z/OS V1R11 - 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 or TRUSTED attribute
  - z10 GA2, or z196, or z114 (all with appropriate MCL's)
  - 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

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

Agenda

- Hardware Updates
  - CFCC Level 17
  - InfiniBand (IFB) coupling links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V1R13
  - z/OS V1R12
  - z/OS V1R11
- Summary

Highlights

- **CFLEVEL 17 for z196 and z114**
  - More structures and nondisruptive dumping
- **Infiniband links for**
  - Bandwidth
    High performance links at 150 meters
    Fewer physical links
    Additional connectivity with HCA3-O LR (four ports)
- **Automatic resolution of sympathy sickness**
  - **SFM with BCPii for better availability**
  - CFSTRHANGTIME, Critical Members
- **CF Structure Management**
  - REALLOCATE test and report
  - Dynamically Disable/Enable structure alter if needed

Sysplex-related Redbooks of potential interest

- System z Parallel Sysplex Best Practices, SG24-7817
- Considerations for Multi-Site Sysplex Data Sharing, SG24-7263
- Server Time Protocol Planning Guide, SG24-7280
- Server Time Protocol Implementation Guide, SG24-7281
- Server Time Protocol Recovery Guide, SG24-7380
- System z Parallel Sysplex Performance, SG24-7654

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

- Available at www.redbooks.ibm.com

Other Sources of Information

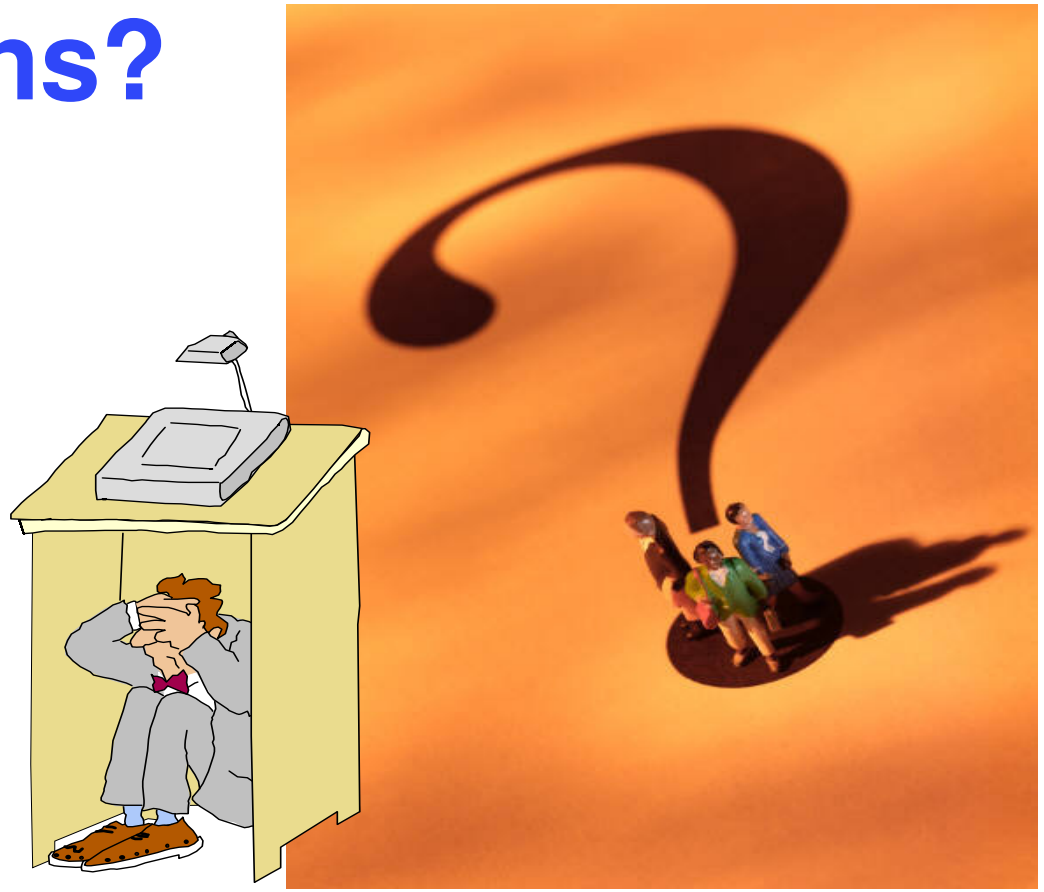
- *MVS Setting Up a Sysplex (SA22-7625)*
- *MVS Initialization and Tuning (SA22-7591)*
- *MVS Systems Commands (SA22-7627)*
- *MVS Diagnosis: Tools and Service Aids (GA22-7589)*
- *z/OS V1R13.0 Migration (GA22-7499)*
- *z/OS V1R13.0 Planning for Installation (GA22-7504)*
- *z/OS MVS Programming: Callable Services for High Level Languages (SA22-7613)*
  - Documents BCPii Setup and Installation and BCPii APIs
- *Migration to the IBM zEnterprise System for z/OS V1R7 through z/OS V1R12 (SA23-2269)*

Parallel Sysplex Web Site

http://www.ibm.com/systems/z/advantages/pso/index.html

Questions?

- Questions?



Appendix – z/OS V1R12

- Support for CFLEVEL 17
- Health Checks
- Auto Reply
- XCF Programming Interfaces

z/OS 1.12 – Support for CFLEVEL 17

- 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 – Support for CFLEVEL 17 …

- A new version of the CFRM CDS is needed 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 – Support for CFLEVEL 17 …

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

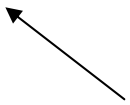
- Coherent Parallel-Sysplex Data Collection Protocol
  - Exploited for duplexed requests
  - Triggering event will result in non-disruptive dump from both CFs, dumps from all connected z/OS images, and capture of relevant link diagnostics within a short period
  - Prerequisites:
    - Installation must ENABLE the XCF function DUPLEXCFDIAG
    - z/OS 1.12
    - z/OS 1.10 or 1.11 with OA31392 (IOS) and OA31387 (XES)
  - Note that full functionality requires that:
    - z/OS image initiating the CF request reside on a z196
    - CF that “spreads the word” reside on a z196

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



Initially complained if more than 300 (5 minutes). APAR OA34439 changed it to 900 (15 minutes) to allow more time for operator intervention and more time for all rebuilds to complete after losing connectivity to a CF

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 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 - 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

Appendix – z/OS V1R11

- System Default Action
- XCF FDI Consistency

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)

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)

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

z/OS 1.11 - XCF FDI Consistency