SHARE in Orlando 2015

SHARE Educate · Network · Influence

z/OS Parallel Sysplex Update

Session 17443   12 August 2015

Mark A Brooks
mabrook@us.ibm.com
z/OS Sysplex Development
IBM Poughkeepsie, NY

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides education, professional networking and industry influence.

Copyright (c) 2015 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

CELEBRATING 60 YEARS OF SHARE Influencing IT Since 1955

Trademarks

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

IBM®
ibm.com®
CICS®
CICSPlex®
DB2®
eServer™
ESCON®
FICON®
HyperSwap®
IMS™
IMS/ESA®

MQSeries®
MVS™
OS/390®
Parallel Sysplex®
Processor Resource/Systems Manager™
PR/SM™
RACF®
Redbooks®
Resource Measurement Facility™
RETAIN®
GDPS®
Geographically Dispersed Parallel Sysplex™

S/390®
Service Request Manager®
Sysplex Timer®
System z®
System z9®
System z10®
System/390®
Tivoli®
VTAM®
WebSphere®

z9®
z10™
z/Architecture®
zEnterprise™
z/OS®
z/VM®
z/VSE®
zSeries®

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.

Please note:

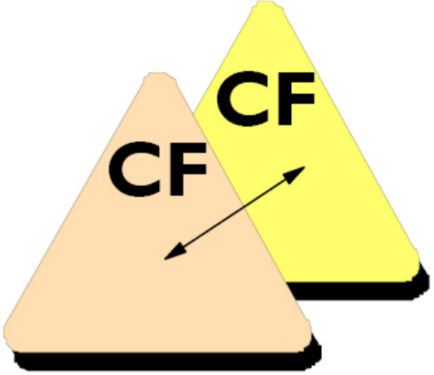
There are far more slides in this handout than I will present during the session. In an hour, I only have time to talk about the highlights. I include the additional slides as reference material in order to:

- Provide details that I may fail to verbalize during the session
- Leave you with a narrative that will (hopefully) stand on its own when you review the material in a few weeks and have long forgotten my verbal explanation
- Highlight features and changes that you may have missed because you are not upgrading to each new processor and z/OS release as they are made available, or you are not able to attend the technical conference that covered the material.

© 2015 IBM Corporation

Agenda

- Hardware Updates
  - CFCC Level 20
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2
  - z/OS V2R1
  - z/OS V1R13
- Summary



© 2015 IBM Corporation

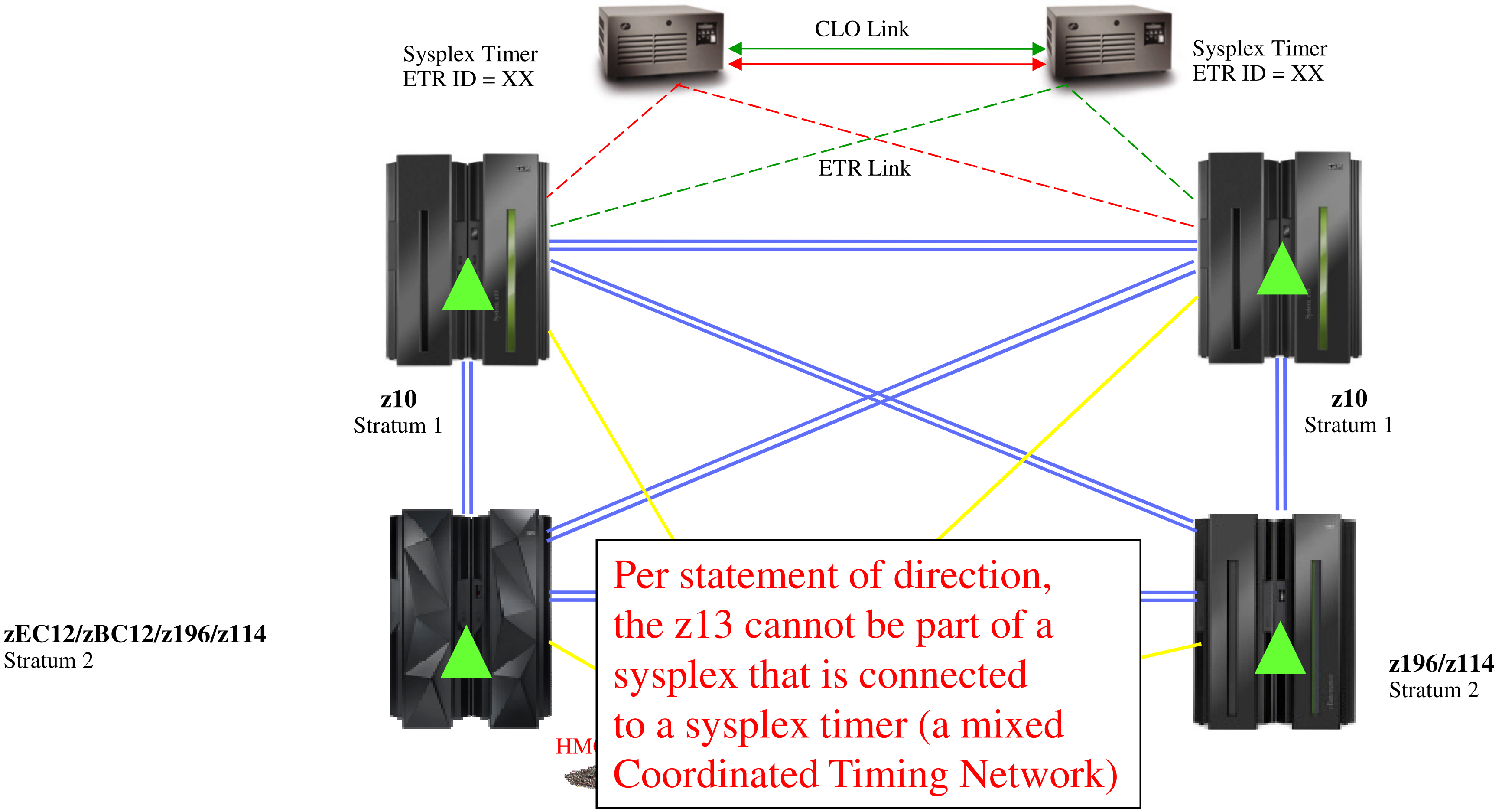
IBM z13^{TM} from a sysplex perspective

- Up to 141 Coupling Facility processors
  - Any given CF LPAR can have at most 16 logical CP’s (no change)
- 256 Coupling CHPIDs per CEC (up from 128)
  - At most 128 can be configured for given CF LPAR
- CTN must be STP-Only (not mixed)
- Does not support older links
  - ISC3
  - HCAO-2 Infiniband

So z10 and prior generation processors cannot participate in a sysplex with z13

© 2015 IBM Corporation

No Support for ETR with zEC12/zBC12/z196/z114 – Use Mixed CTN



Per statement of direction, the z13 cannot be part of a sysplex that is connected to a sysplex timer (a mixed Coordinated Timing Network)

Statement of Direction: **Removal of support for connections to an STP Mixed CTN**
The zEC12 and zBC12 are the last System z servers to support connections to an STP Mixed CTN. After the zEC12 and the zBC12, servers that require time synchronization (e.g., those in a sysplex) will require Server Time Protocol (STP) and all servers in that network must be configured in STP-only mode.

© 2015 IBM Corporation

Statement of Direction

Per statements of direction, the z13 cannot use ISC-3 links and older IFB links. That is, z10 and older machines cannot participate in a parallel sysplex that includes a z10 or any earlier generation processor.

- Removal of ISC-3 support on system z

  The zEC12 and zBC12 are planned to be the last System z servers to offer support of the InterSystem Channel-3 (ISC-3) for Parallel Sysplex environments at extended distances. ISC-3 will not be supported on future System z servers as carry forward on an upgrade. Previously we announced that the IBM zEnterprise 196 (z196) and IBM zEnterprise 114 (z114) servers were the last to offer ordering of ISC-3. Enterprises should continue migrating from ISC-3 features (#0217, #0218, #0219) to 12x InfiniBand (#0171 - HCA3-O fanout) or 1x InfiniBand (#0170 - HCA3-O LR fanout) coupling links.

- Removal of support for the HCA2-O fanouts for 12x IFB and 1x IFB InfiniBand coupling links

  The zEC12 and zBC12 are planned to be the last System z servers to support the following features as carry forward on an upgrade: HCA2-O fanout for 12x IFB coupling links (#0163) and HCA2-O LR fanout for 1x IFB coupling links (#0168). Enterprises should continue migrating to the HCA3-O fanout for 12x IFB (#0171) and the HCA3-O LR fanout for 1x IFB (#0170).

© 2015 IBM Corporation

CFLEVEL 20

- IBM z13^{TM}

- Requirements for CFLEVEL 20
  - –z/OS V2.2
  - –z/OS V1.13 and V2.1 need APARs:
    - IOS OA44287; XES OA44440; HCD OA47336; RMF OA44502

- CFLEVEL 20 provides support for
  - New ICA SR links
  - Cache structure detach performance improvements

© 2015 IBM Corporation

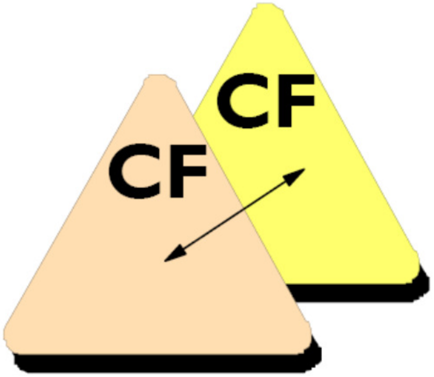
Large Cache Structure Detach Processing Improvements

- Prior to CFCC LEVEL=20, detach processing required a scan of every active entry and registration block in the structure
  - The larger the structure, the longer the scan takes to complete
  - Exploiter had to wait for scan to complete before deemed to be detached
  - Attachment "slot" cannot be reused until detach completes
  - Which could delay (re)attach by exploiter
- With CFCC LEVEL=20, the detach completes "instantly"
  - Attachment "slot" immediately released; exploiter can immediately (re)attach
  - CF still scans structure to clean up artifacts associated with the attachment that is no longer valid
  - The CF can distinguish between artifacts associated with the new instance of the attachment versus those associated with the old instance

© 2015 IBM Corporation

Agenda

- Hardware Updates
  - CFCC Level 20
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2 preview
  - z/OS V2R1
  - z/OS V1R13
- Summary



© 2015 IBM Corporation

zEC12 and zBC12 from Sysplex Perspective

- Neither machine supports ESCON
  - If using CTC devices for XCF signalling paths, must be FICON CTC
- zEC12 supports
  - Up to 101 ICF *prior limit:* 16
  - 64 1x IFB HCA3-O LR links *prior limit:* 48
    - Facilitates migration from ISC-3 links

© 2015 IBM Corporation

CFLEVEL 19

- IBM zEnterprise™ EC12 GA2 (zEC12) and BC12 (zBC12)

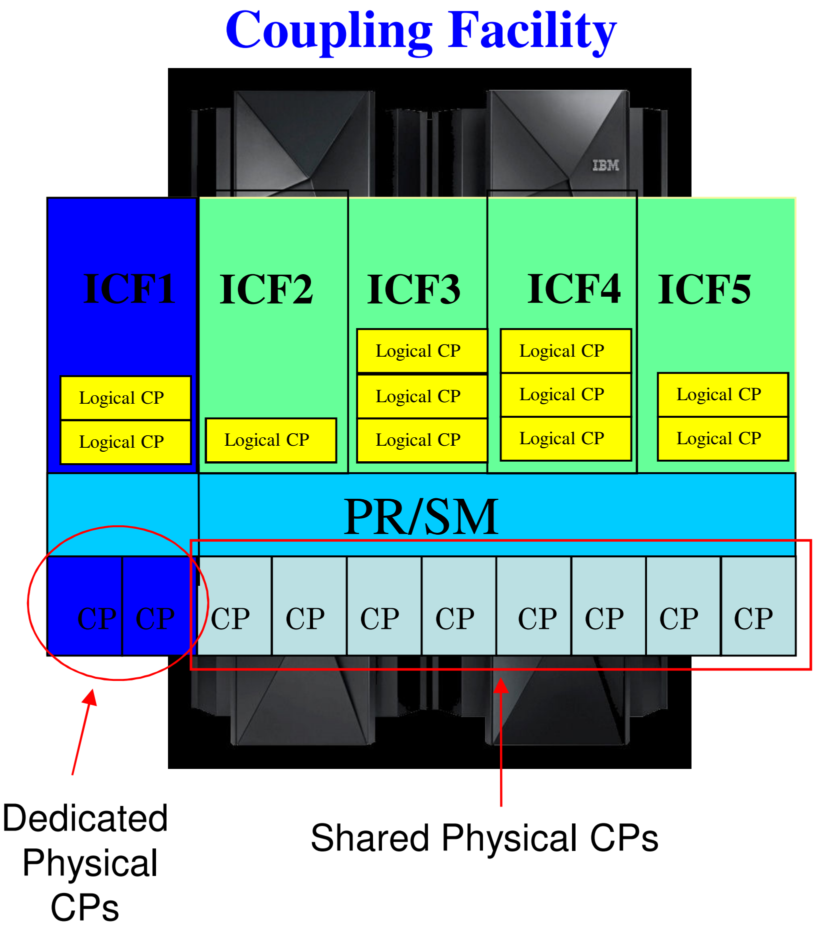
- Requirements for CFLEVEL 19
  - z/OS V2.1 and up
  - z/OS V1.10 and up need toleration APAR OA42372
    - Addresses anomalies with object counts for SM duplexed structures
    - Should install APAR everywhere before CFLEVEL 19 introduced
  - z/VM V6.3 or later with PTFs for guest exploitation

- CFLEVEL 19 provides support for
  - Thin Interrupts to reduce parallel sysplex costs
  - Flash Memory to enhance resiliency

© 2015 IBM Corporation

Review: Dedicated vs Shared CPs for Coupling Facility

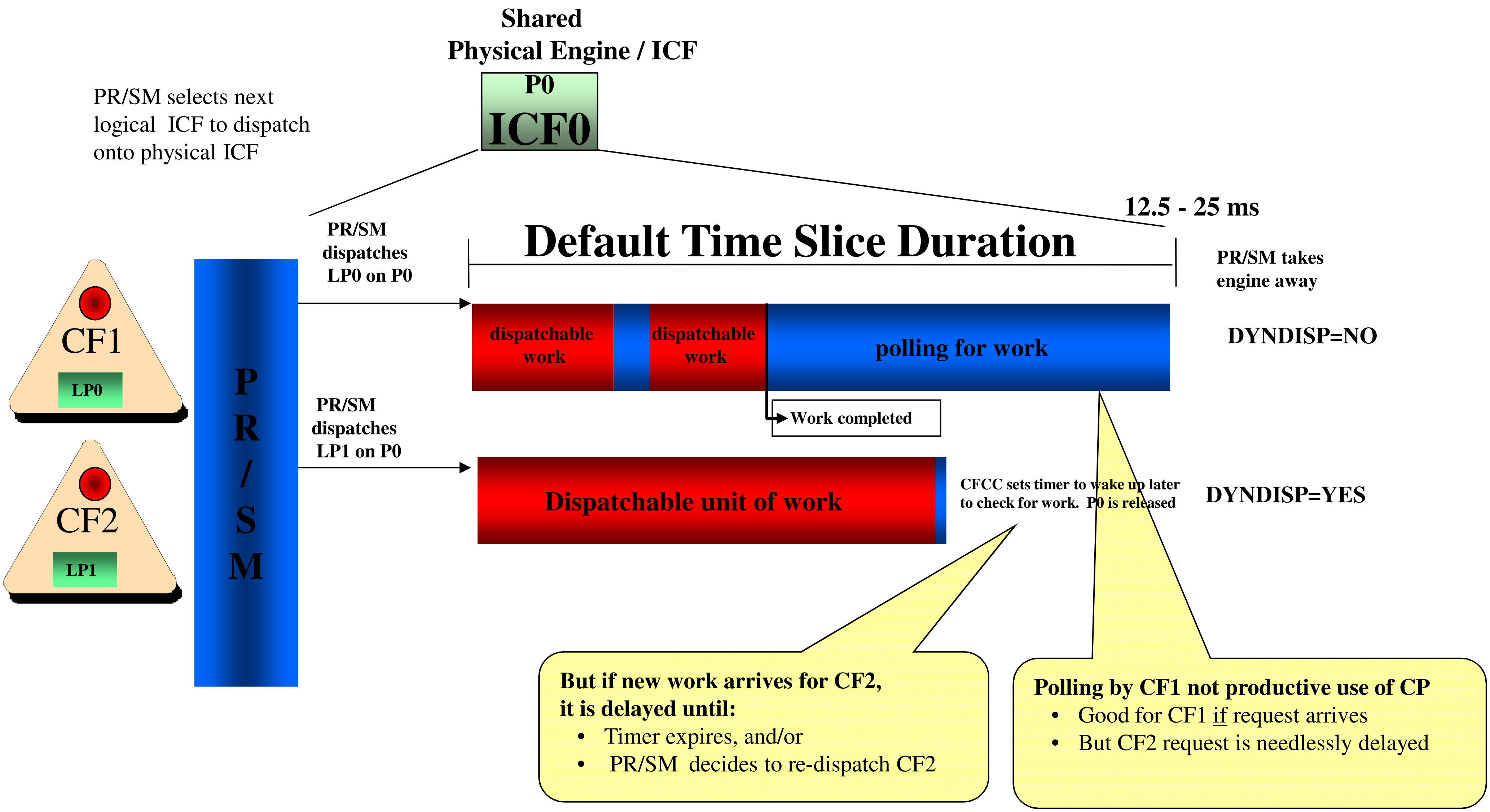
- CFCC uses a polling model
  - Provides consistently good services when dedicated engine assigned to CF
  - But “hurts” when sharing engines since needlessly consumes cycles looking for work when there is none
- With shared engines, CF service times are unpredictable and highly variable
  - PR/SM determines which logical partition should be dispatched on a physical engine based on LPAR weights and relative share used
    - Moderately complex configuration and tuning tasks to get reasonable behavior
  - Request with service time of a few microseconds when processed by CF with dedicated engines might have service time of tens of milliseconds when processed by CF with shared engines
    - z/OS heuristics likely converts sync to async



Let's see what's going on ....

© 2015 IBM Corporation

Review: Dynamic CF Dispatch Options



Shared
Physical Engine / ICF
P0
ICF0
PR/SM selects next logical ICF to dispatch onto physical ICF
12.5 - 25 ms
Default Time Slice Duration
PR/SM takes engine away
PR/SM dispatches LP0 on P0
CF1
LP0
CF2
LP1
P R / S M
dispatchable work
dispatchable work
polling for work
DYNDISP=NO
Work completed
PR/SM dispatches LP1 on P0
Dispatchable unit of work
CFCC sets timer to wake up later to check for work. P0 is released
DYNDISP=YES

But if new work arrives for CF2, it is delayed until:

- Timer expires, and/or
- PR/SM decides to re-dispatch CF2

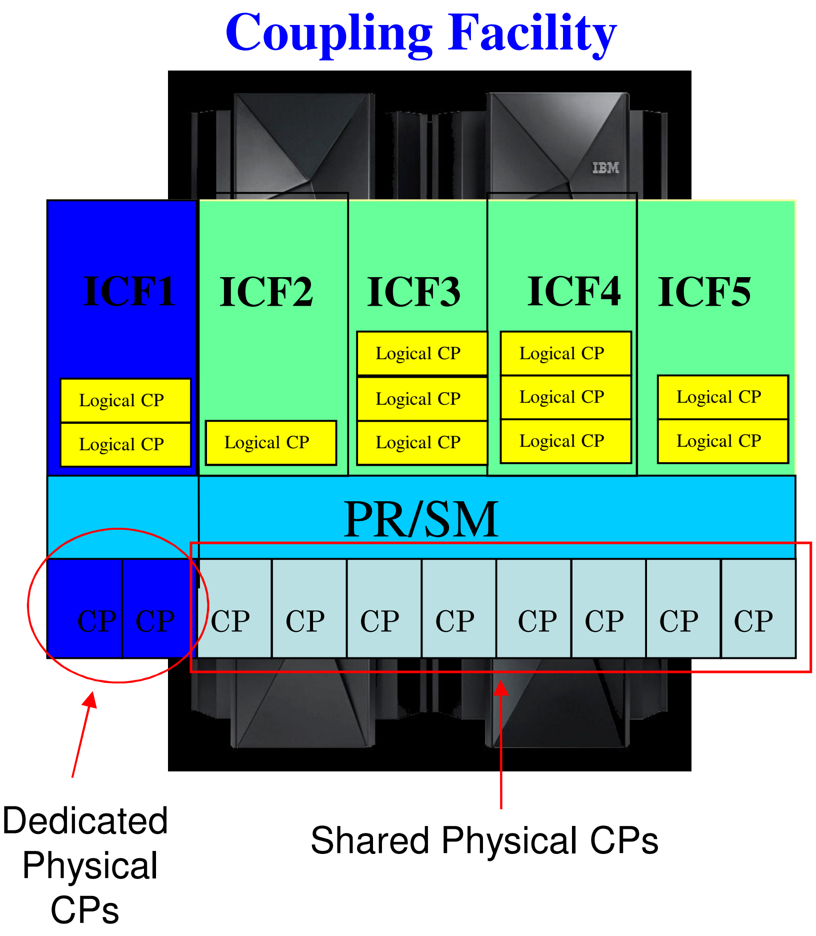
Polling by CF1 not productive use of CP

- Good for CF1 if request arrives
- But CF2 request is needlessly delayed

© 2015 IBM Corporation

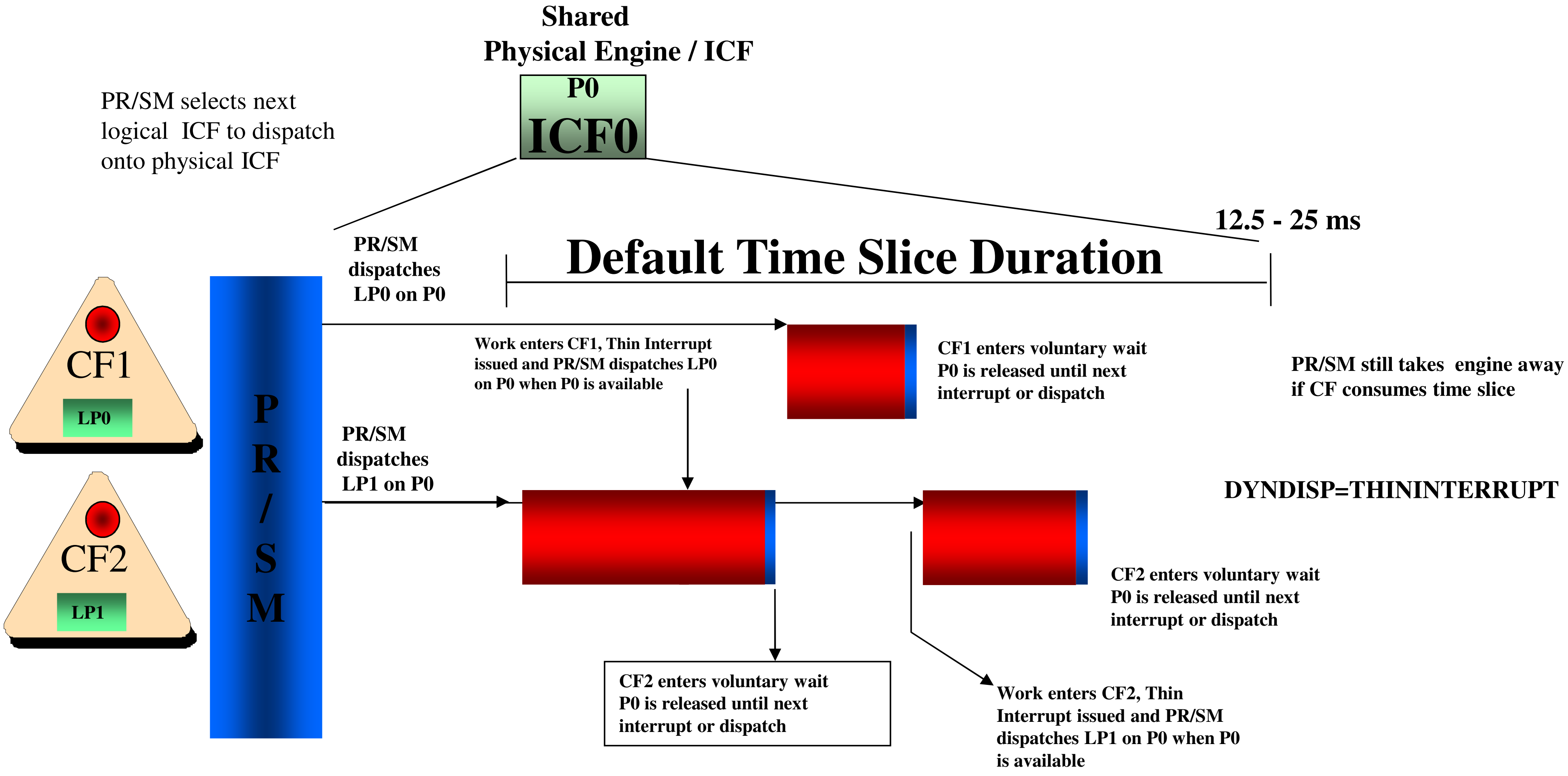
Review: Dynamic CF Dispatch and Shared CF CPs

- DYNDISP=NO
  - CF runs polling model until PR/SM takes engine
  - Excellent service time while running
  - Poor use of shared engine if no work
- DYNDISP=YES
  - When dispatched, CF runs polling loop as usual
  - Processes whatever work it finds
  - If runs out of work, sets timer and releases the engine
    - Duration of timer heuristically adjusted
    - Tends to be shorter if finding work
    - Tends to be longer if not finding work
  - More effective use of shared engines, but
  - New latencies can increase CF request service time
- Ideally, we want PR/SM to dispatch the CF when work arrives ...



© 2015 IBM Corporation

New Dynamic CF Dispatch Option – Thin Interrupts



PR/SM selects next logical ICF to dispatch onto physical ICF

PR/SM dispatches LP0 on P0

Work enters CF1, Thin Interrupt issued and PR/SM dispatches LP0 on P0 when P0 is available

CF1 enters voluntary wait P0 is released until next interrupt or dispatch

PR/SM still takes engine away if CF consumes time slice

PR/SM dispatches LP1 on P0

DYNDISP=THININTERRUPT

CF2 enters voluntary wait P0 is released until next interrupt or dispatch

CF2 enters voluntary wait P0 is released until next interrupt or dispatch

Work enters CF2, Thin Interrupt issued and PR/SM dispatches LP1 on P0 when P0 is available

© 2015 IBM Corporation

Coupling Thin Interrupts

Goal: Expedite the dispatching of the partition when work arrives

- Thin interrupt driven for:
  - Arrival of new request from z/OS
  - Arrival of duplexing signal from peer CF
  - Back end async completion of duplexing signal sent to peer CF
- Enables timely dispatch of shared processor when work arrives
  - Reduces latency of waiting for timer pop or normal time slice
  - Immediate dispatch if physical CP available at time of interrupt
- Once the CF image gets dispatched, the traditional polling mechanism is used to locate and process the work
- CF will give up control when work is exhausted (or when LPAR kicks it off the shared processor)

© 2015 IBM Corporation

Benefits of Using Thin Interrupts for Shared Engine CF

- You should get:
  - Faster and more consistent CF service times
  - More effective utilization of physical processor
- Which might allow you to reduce costs:
  - Use shared-engine CF in a broader range of configurations
  - Smaller pool of physical engines to support the workload
- Simplification
  - More similar to use of shared engines with z/OS images

Dedicated engines still provide best service times for CF requests

© 2015 IBM Corporation

CFLEVEL 19 - DYNDISP Comparisons

CF Polling	Dynamic CF Dispatching	Coupling Thin Interrupts
DYNDISP=NO	**DYNDISP=YES**	**DYNDISP=THININTERRUPT**
LPAR Time slicing	CF time-based algorithm for CF engine sharing	CF releases shared engine if no work left to be done
- CF does not "play nice" with other shared images sharing the processor - CF controls processor long after work is exhausted	- CF does own time slicing - More effective engine sharing than polling - Blind to presence or absence of work to do - Relies on timer or LPAR time slice to check for work	- Event-Driven Dispatching - Most effective use of shared engines across multiple CF images - CF relies on generation of thin interrupt to dispatch processor when new work arrives

© 2015 IBM Corporation

References

- **White Paper: Coupling Thin Interrupts and Coupling Facility Performance in Shared Process Environments**
  - www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102400
- White Paper: Coupling Facility Configuration Options
  - public.dhe.ibm.com/common/ssi/ecm/en/zsw01971usen/ZSW01971USEN.PDF

- Processor Resource/Systems Manager Planning Guide (SB10-7156)
  - Available on Resource Link

- www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/tips0237.html
  - Written in 2003, so no mention of thin interrupts
  - But nicely explains considerations for DYNDISP=ON or OFF

© 2015 IBM Corporation

CFLEVEL 19 - “CF Flash”

aka “Storage Class Memory (SCM)”

- Initially targeted to MQ shared queue structures
- Provides emergency capacity to handle MQ shared queue buildups during abnormal situations
  - Transient mismatch between producers and consumers of messages on a shared queue can lead to long queues
    - Regulatory requirement to store 8 hours of message traffic for 24 hours and then work through the backlog in 3 hours
- Requirements:
  - CFCC CFLEVEL 20, or 19 with appropriate MCL
  - z/OS V2R2
  - z/OS V2R1 or z/OS V1R13 (both with appropriate PTFs)
  - A new level of MQ is not required

APARs
OA40747 enables
OA45920 fix rebuild
OA45746 fix 0C4
OA45224 + OA45421
better accounting
for flash faults

© 2015 IBM Corporation

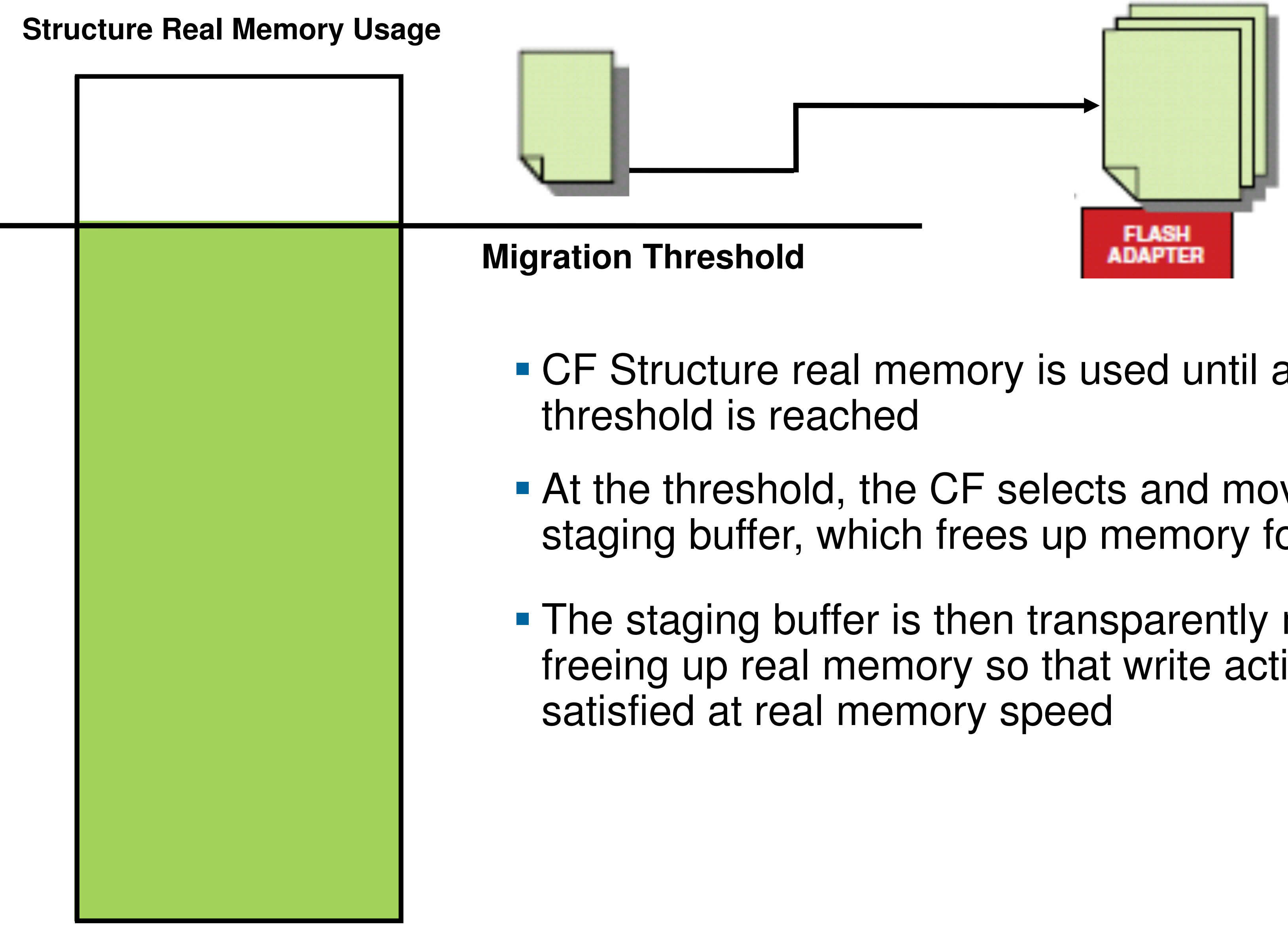
CF Flash – For Capacity not Performance

- Currently, the CF is a pure “real memory” system
  - All structures allocated and backed entirely by real memory in the CF image
  - No paging, no virtual storage, no disk I/O at all
- Adding relatively slower Flash memory to a CF structure, therefore, cannot speed anything up
  - So CF Flash exploitation is not a performance enhancement item
  - Indeed, a request that needs to access a structure object residing in flash memory will be rejected by the CF
- CF Flash truly intended for temporary “abnormal” capacity issues
  - AutoAlter or application-specific offload mechanisms preferable to use of flash for normal operation

So how will this work ...

© 2015 IBM Corporation

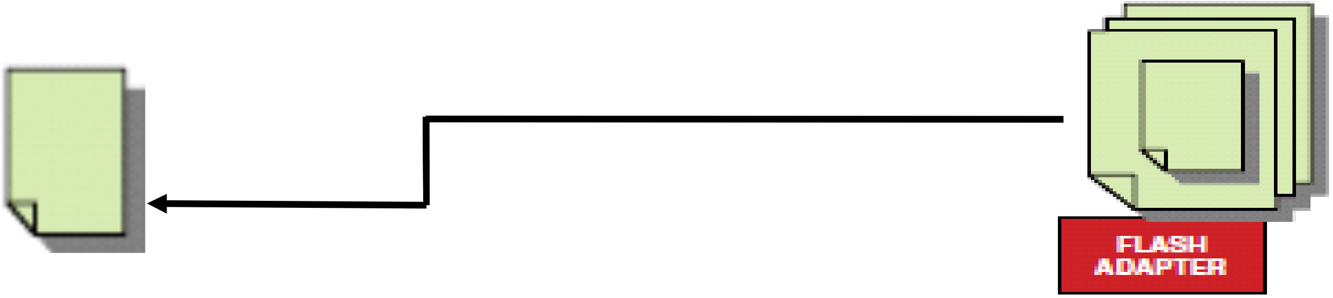
CF Flash: Migrating Objects to Flash Memory



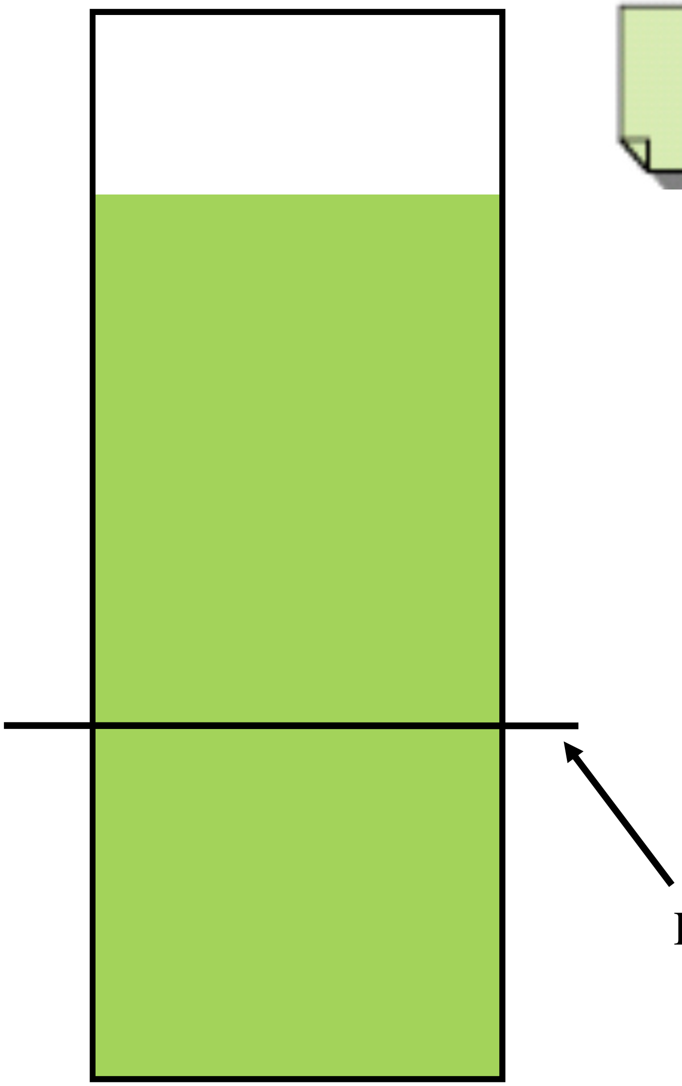
- CF Structure real memory is used until a migration threshold is reached
- At the threshold, the CF selects and moves objects to a staging buffer, which frees up memory for more ‘puts’
- The staging buffer is then transparently moved to flash, freeing up real memory so that write activity continues to be satisfied at real memory speed

© 2015 IBM Corporation

CF Flash: Prefetch of Objects from Flash Memory



- When the structure’s real memory utilization goes below a prefetch threshold, or when certain other list-specific criteria are reached, the CF starts to retrieve objects from flash memory back into a staging buffer in real memory
- The structure is then repopulated from the staging buffer so the objects can be retrieved on “get” references at real memory speeds



© 2015 IBM Corporation

CF Flash Considerations - Migration

- If migration to flash keeps up:
  - Structure real memory never fills
  - Write activity continues to be satisfied at real memory speeds
- If migration does not keep up
  - Structure real memory may fill up
  - Causing writes to fail due to lack of available real memory
  - Writes automatically redriven by z/OS
  - But will not succeed until migration moves enough objects to flash
    - Can still get full conditions if run out of flash (or “augmented space”)

© 2015 IBM Corporation

CF Flash Considerations - PreFetch

- If prefetching keeps up:
  - –Structure objects never reside in flash when referenced by application
  - –References continue to be satisfied at real memory speeds
- If prefetching does not keep up, could have a “flash fault”:
  - –Object might reside in flash when referenced by application
  - –Request is rejected, but CF initiates retrieval of object from flash
    - CF never waits for object to be retrieved from flash
  - –z/OS automatically redrives the request (possibly more than once)

- Random references to objects can also experience flash faults
  - –Likely restricts practical exploitation of flash to structures with particular “well behaved” access patterns

© 2015 IBM Corporation

CF Flash – Tailoring Migration/Prefetch to Structure Usage

- The CF migration/prefetch algorithm will be successful if:
  - Real memory available for every put
  - Target object resides in real memory for every get
- To be successful, the CF needs to select and move the correct set of objects to/from flash in time
  - Implies need to understand access patterns of the application
- New specification in CFRM policy defines the algorithm to be used
  - Only one choice today, and it is specifically tailored for MQ usage
  - I do not believe this algorithm will work for applications other than MQ
    - But vendors will need to make determination and provide guidance

© 2015 IBM Corporation

CF Flash – Configuration

- Flash Memory Assignment
  - Flash memory exists on a flash card in the CPC
  - Assigned to a CF partition via hardware definition panels
    - Just like it is for z/OS partitions
- CFRM Policy
  - Indicates maximum allotment of flash memory to be used for given structure
  - If permitted, increases size requirements of that structure
    - Even if flash is never used by the structure
    - CF requires additional control objects to manage things
    - CFSIZER updated accordingly
- CF may need additional memory as well ...

© 2015 IBM Corporation

CF Flash – Configuration Considerations

- Flash memory is not pre-allocated
  - In contrast to normal structure allocation
  - Acquired on an as needed basis
  - Returned to “free pool” when no longer needed
- CFRM Policy can over commit flash memory in the aggregate
- CF may need to acquire “augmented space” from real memory as flash memory is acquired for structure
  - Has implications for memory requirements of CF partition
  - CFSIZER indicates maximum amount of augmented space that would be needed for specified amount of flash memory
- New conditions where structure deemed to be full

© 2015 IBM Corporation

CF Flash – Migration and Coexistence Considerations

- All systems in sysplex must have appropriate z/OS support before CF flash will be exploited
- Down-level systems cannot connect to, rebuild, alter, display, or dump a structure that is capable of using flash

- Alter processing disabled for structure while flash in use
- Duplexed structures
  - Once duplexed, a structure can begin using flash
  - XES will not (normally) permit duplexing to begin for a structure already using flash

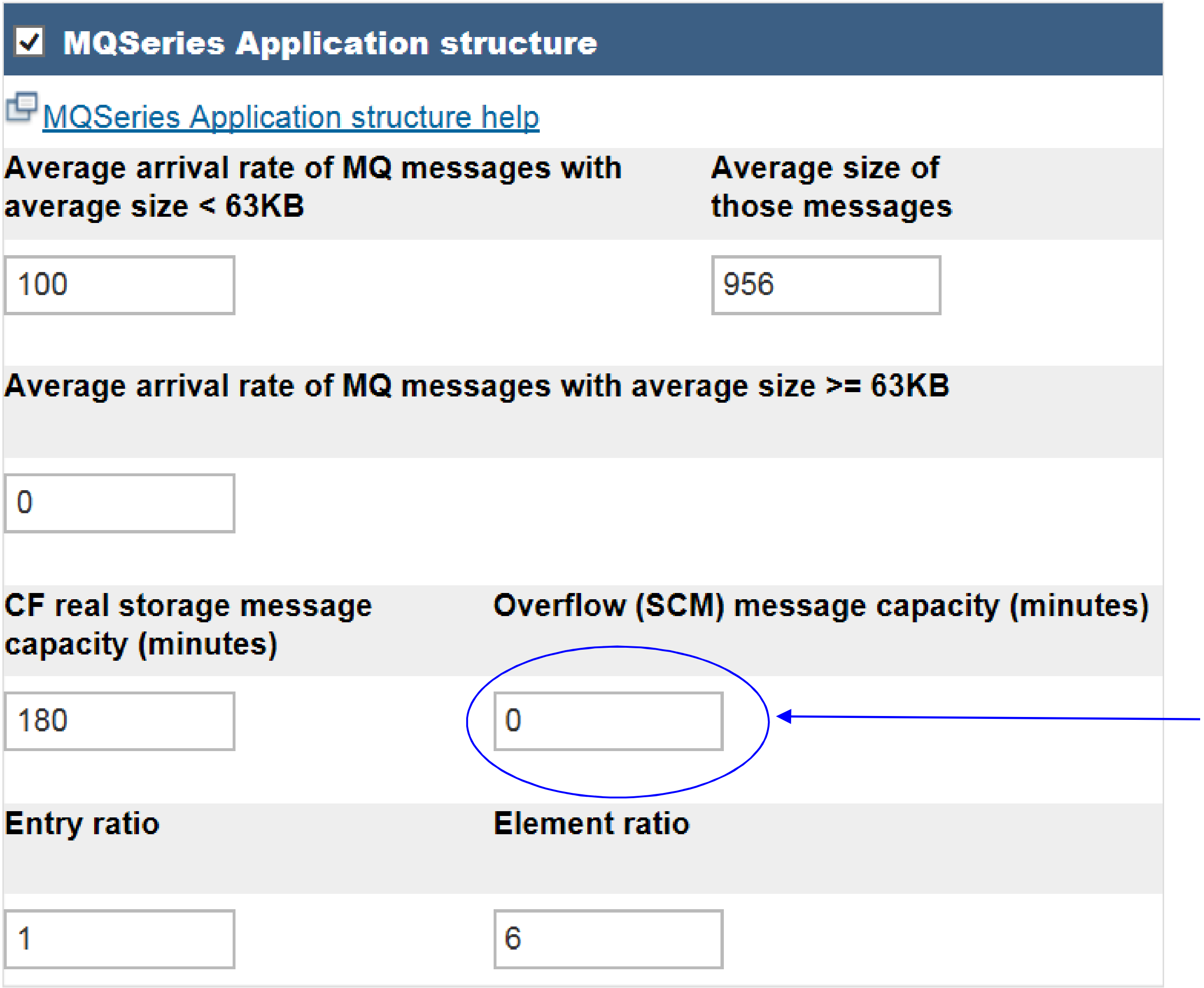
© 2015 IBM Corporation

CF Sizing Updates

- Newest version of the SIZER utility
  - www.ibm.com/systems/support/resources/sizer.zip
  - Supports SCM
  - Supports output to a file (instead of just the console)
- CFSIZER
  - www.ibm.com/systems/support/z/cfsizer
  - Supports SCM
  - Produces explanatory messages when appropriate

© 2015 IBM Corporation

New CFSIZER input panel for MQ



☑ **MQSeries Application structure**

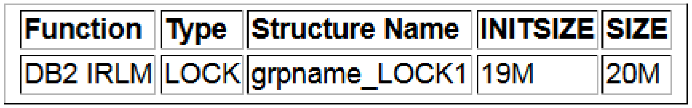
MQSeries Application structure help

Average arrival rate of MQ messages with average size < 63KB	**Average size of those messages**
100	956
Average arrival rate of MQ messages with average size >= 63KB	
0	
CF real storage message capacity (minutes)	**Overflow (SCM) message capacity (minutes)**
180	0
Entry ratio	**Element ratio**
1	6

Nonzero value triggers SCM calculations

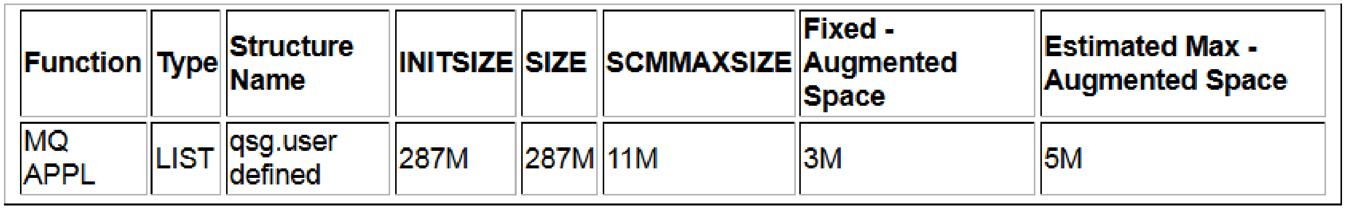
© 2015 IBM Corporation

CFSIZER – Explanatory Messages



Function	Type	Structure Name	INITSIZE	SIZE
DB2 IRLM	LOCK	grpname_LOCK1	19M	20M

Lock table entry count=2097152. Specify this count as your IRLMPROC LTE value to ensure that the structure is allocated with sufficient record table entries (RTEs).



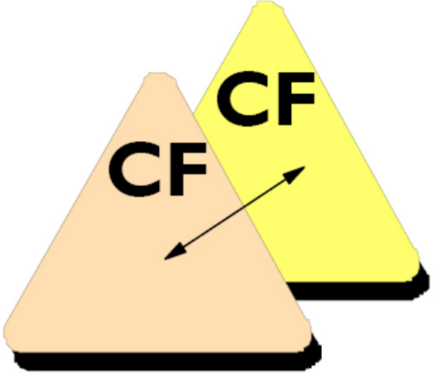
Function	Type	Structure Name	INITSIZE	SIZE	SCMMAXSIZE	Fixed - Augmented Space	Estimated Max - Augmented Space
MQ APPL	LIST	qsg.user defined	287M	287M	11M	3M	5M

Augmented space is not included in the structure sizes. It is additional CF storage required to exploit storage-class memory.

© 2015 IBM Corporation

Agenda

- Hardware Updates
  - CFCC Level 20
  - CFCC Level 19
  - **CFCC Level 18**
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2
  - z/OS V2R1
  - z/OS V1R13
- Summary



© 2015 IBM Corporation

CFLEVEL 18

- IBM zEnterprise^{TM} EC12 (zEC12 GA1)
  - Available September, 2012
- Serviceability and Performance Enhancements
- Requirements
  - z/OS V2R2
  - z/OS V2R1, or z/OS V1R13 and V1R12 with PTFs
  - z/VM 5.4 or later with PTFs for guest exploitation

© 2015 IBM Corporation

CFLEVEL 18 Overview of Enhancements

- **CF Cache Write Around**
- Internal CFCC Changes for Cache Structures
- Delete Name Extensions for Cache Structures
- Register Attach Validation
- RAS Enhancements
- Enhanced RMF Channel Path Reporting

*Due to time restrictions, I will only discuss topics in **bold**.*
See slides in uploaded presentation for details.

© 2015 IBM Corporation

CFLEVEL 18 - Performance Enhancements Cache Write Around

- Enhancements to the IXLCACHE macro interface and CFCC allow exploiters to optionally request that writes to CF Cache be suppressed if:
  - The data is not currently stored in the CF Cache structure, and
  - Only the local cache has registered interest
- Can intelligently decide which entries should be written to the cache and which should just be “written around” directly to disk
  - Helps preserve application “working set”
  - Suppressing writes reduces work that CF must perform
- Requires application exploitation, and:
  - APAR OA40966 at z/OS V1R12 and up
  - z/VM 5.4 with PTFs for guest exploitation
- Also note: Roll back to CFCC Release 17 (MCL12)

APAR OA37550 provided write around. OA40966 drags it and fixes some unrelated issues.

© 2015 IBM Corporation

CFCC Level 18 - Performance Enhancements
Cache Write Around …

- IBM DB2 11 for z/OS exploits cache write around for batch update/insert processing
  - Conditionally writes to group buffer pool (GBP)
  - Helps avoid over running cache structures with directory entries and changed data that are not part of the normal working set
  - Avoids thrashing the cache through LRU processing
  - Avoids castout processing backlogs and delays
- Intended to improve DB2 batch performance
  - We saw 50% improvement in some of our tests**
- Online transactions may encounter less delay during large concurrent batch updates

*** These were not formal performance measurements. Your results may vary..*

© 2015 IBM Corporation

CFCC Level 18 - Performance Enhancements
Internal CFCC Changes for Cache Structures

- Elapsed time improvements when dynamically altering cache structure
  - Entry / Element ratio
  - Size
- CF Storage Class and Castout Class contention avoidance
  - Changes the way serialization is performed on individual storage class and castout class queues
  - Reduces storage class and castout class latch contention.
- Throughput enhancements for parallel cache castout processing.

© 2015 IBM Corporation

CFLEVEL 18 – Resiliency and Performance
Delete Name Extensions for Cache Structures

- Halt on Changed
  - Allows exploiter to redrive cast out processing when changed data is unexpectedly encountered
  - Helps avoid the accidental deletion of directory entries which might lead to data corruption
- Optional suppression of cross invalidate signals
  - Helps improve delete name performance, particularly at distance
  - But local vector will not reflect validity of locally cached data
- Requires
  - OA38419 at z/OS V1R12 and up
  - Exploitation by application (DB2 APAR PM67544)
- Also note
  - Rolled back to CFCC Release 17

© 2015 IBM Corporation

CFLEVEL18 – Resiliency Register Attach Validation

- Verification of local cache controls for a Coupling Facility cache structure connector.
  - Performed when connection registers interest in data
  - System gathers diagnostics if discrepancies detected
  - System proactively takes steps to mitigate problem before data corruption can occur

XES Guilty. Get APAR OA42519

- Also note
  - Rolled back to CFCC Release 17 (MCL12)
  - Requires z/OS APAR OA37550 or z/OS 2.1
    - But install OA40966 (to avoid unrelated issues)

© 2015 IBM Corporation

CFLEVEL 18 – RAS Enhancements

- Background structure deallocation
  - XES task freed to perform other work and requests instead of redriving the deallocate command in the foreground

- This change also allows for better structure dumps
  - Extended dumping can be performed when structure damage is detected allowing for the capturing of more content for error analysis
  - In the past, foreground deallocation could cause structure dumps to be truncated

© 2015 IBM Corporation

CFCC Level 18 – RAS Enhancements

- Enhanced CFCC tracing support
  - Significantly enhanced trace points, especially in troublesome areas
    - Latching (CP and suspend),
    - Locate queue and suspend queue management/dispatching,
    - Duplexing protocols (especially suppression and clear-off processing),
    - Sublist notification,
    - Alter/ECR,
    - Castout processing
    - RCC cursors, etc.
  - Quantity gathered
    - Trace buffer size increase
  - Trace buffer granularity –
    - Special trace buffers for specific types of traces (e.g. Alter/ECR)
  - Controls
    - Default/detail/exception levels of tracing, activated via OPERMSG commands

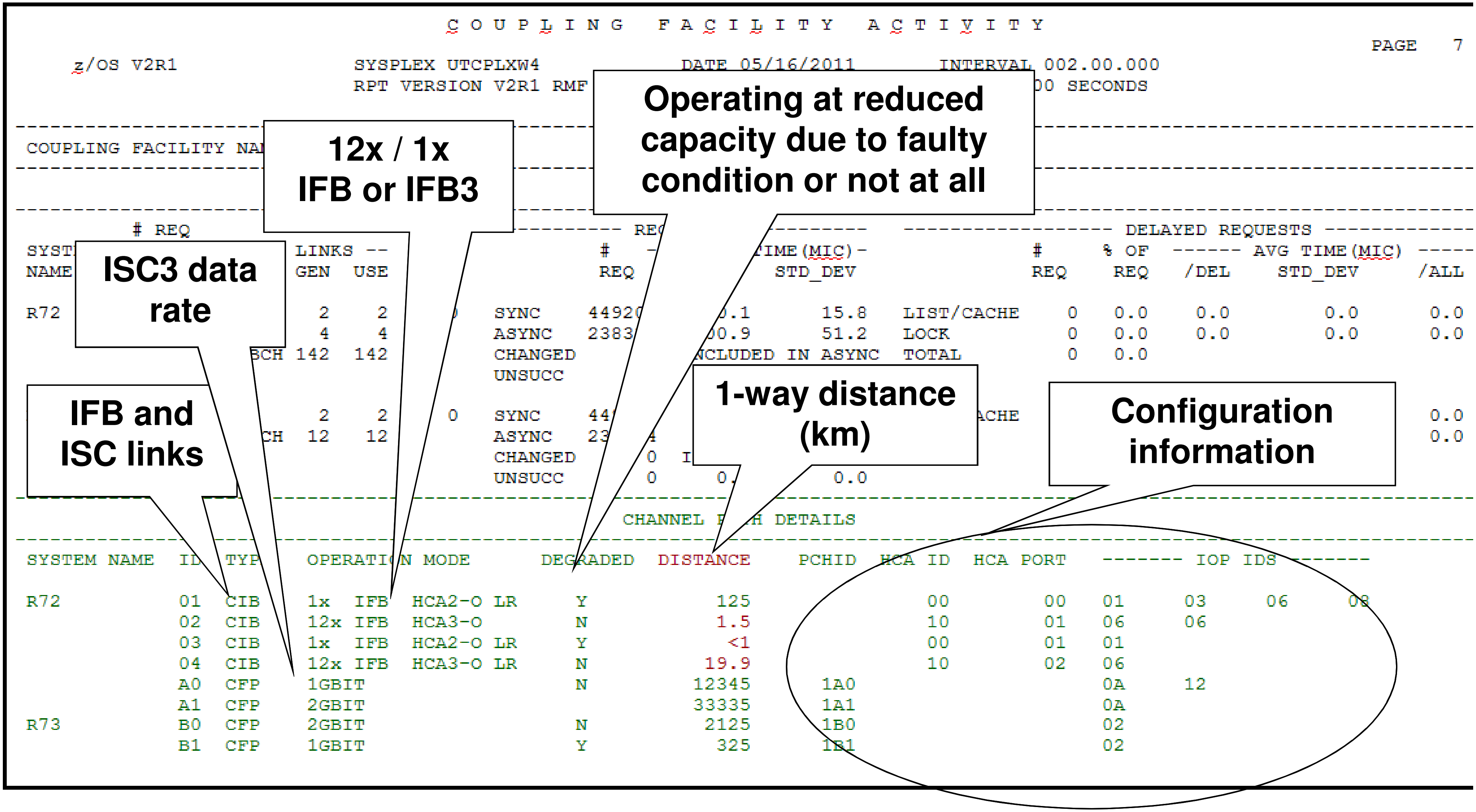
© 2015 IBM Corporation

CFCC Level 18 – RMF Channel Path Details

- Provides enhanced reporting of channel path characteristics for Parallel Sysplex Coupling Facility CIB or CFP links
- Helps understand link performance, response times and coupling overheads
  - Channel path ID
  - Channel path operation mode
  - Channel path degraded status
  - Channel path distance
  - Accessible I/O processors
  - Channel path type acronym
  - Physical channel path ID
  - Host channel adapter ID
  - Host channel adapter port number
- New Channel Path Details section
  - RMF Coupling Facility Postprocessor Report
  - RMF Monitor III CFSYS Report
  - XML report
- New display commands
- APAR OA38312 for support on z/OS V1R12 and up
- APAR OA37826 for RMF support

© 2015 IBM Corporation

CFCC Level 18 – RMF Channel Path Details



```
                                C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                  PAGE    7
    z/OS V2R1                SYSPLEX UTCPLXW4              DATE 05/16/2011         INTERVAL 002.00.000
                             RPT VERSION V2R1 RMF                                  00 SECONDS

---------------------------------------------------------------------------------------------------------
 COUPLING FACILITY NA
---------------------------------------------------------------------------------------------------------

---------------------------------------------------------------------------------------------------------
          # REQ                                    ----------- RE            ---------   ------------------ DELAYED REQUESTS -------------
 SYST                   LINKS --                         #     -        TIME(MIC)-           #      % OF    ------ AVG TIME(MIC)  -----
 NAME                   GEN   USE                        REQ                STD_DEV           REQ     REQ     /DEL       STD_DEV       /ALL

 R72                      2     2        0   SYNC      44920       0.1        15.8   LIST/CACHE    0     0.0     0.0          0.0        0.0
                          4     4            ASYNC     2383       00.9        51.2   LOCK          0     0.0     0.0          0.0        0.0
                   BCH  142   142            CHANGED             NCLUDED IN ASYNC    TOTAL         0     0.0
                                             UNSUCC

                          2     2        0   SYNC      44                                  ACHE                                           0.0
                    CH   12    12            ASYNC     23        4                                                                        0.0
                                             CHANGED             0    I
                                             UNSUCC              0      0.            0.0
---------------------------------------------------------------------------------------------------------
                                               CHANNEL P  H DETAILS
---------------------------------------------------------------------------------------------------------
 SYSTEM NAME   ID  TYP    OPERATION MODE        DEGRADED  DISTANCE    PCHID  HCA ID  HCA PORT   ------- IOP IDS -------

 R72           01  CIB    1x  IFB  HCA2-O LR       Y          125               00        00    01    03    06    08
               02  CIB    12x IFB  HCA3-O          N          1.5               10        01    06    06
               03  CIB    1x  IFB  HCA2-O LR       Y           <1               00        01    01
               04  CIB    12x IFB  HCA3-O LR       N         19.9               10        02    06
               A0  CFP    1GBIT                    N        12345      1A0                      0A    12
               A1  CFP    2GBIT                             33335      1A1                      0A
 R73           B0  CFP    2GBIT                    N         2125      1B0                      02
               B1  CFP    1GBIT                    Y          325      1B1                      02
```

12x / 1x
IFB or IFB3

Operating at reduced capacity due to faulty condition or not at all

ISC3 data rate

IFB and ISC links

1-way distance (km)

Configuration information

© 2015 IBM Corporation

Messages – IEE174I (D M=CHP)

Configuration information

```
COUPLING FACILITY  type.mfg.plant.sequence
                   PARTITION: partition side  CPCID: cpcid
                   CONTROL UNIT ID: cuid
NAMED cfname

PATH            PHYSICAL                LOGICAL   CHANNEL TYPE       AID  PORT
chpid[/pchid]  phystatus               logstatus chtype [pathmode] [aid  port]

COUPLING FACILITY SUBCHANNEL STATUS
TOTAL:  totdev  IN USE:  usedev   NOT USING: nusedev   NOT USABLE: unusedev
 [NOT] OPERATIONAL DEVICES / SUBCHANNELS:
     dev / subch     dev / subch     dev / subch     dev / subch
```

May now indicate:
ONLINE-DEGRADED

H
F
(ISC3 data rate)

1X-IFB
12X-IFB
12X-IFB3

© 2015 IBM Corporation

Messages – IXL150I (DISPLAY CF output)

```
IXL150I hh.mm.ss DISPLAY CF
COUPLING FACILITY type.mfg.plant.sequence
                  PARTITION: partition side  CPCID: cpcid
                  LP NAME: lparname   CPC NAME: cpcname
                  CONTROL UNIT ID: cuid
NAMED cfname
. . .
    DYNAMIC CF DISPATCHING: ON|OFF]
    COUPLING FACILITY IS standalonestate
. . .
PATH           PHYSICAL              LOGICAL   CHANNEL TYPE       AID  PORT
chpid[/pchid]  phystatus             logstatus chtype [pathmode]  aid  port
. . .
REMOTELY CONNECTED COUPLING FACILITIES
       CFNAME              COUPLING FACILITY
       ---------           --------------------------
       rfcfname            rftype.rfmfg.rfplant.rfsequence
                           PARTITION: partition rfside CPCID: rfcpcid
                           CHPIDS ON cfname CONNECTED TO REMOTE FACILITY
                           RECEIVER: CHPID    TYPE
                                     rfchpid  rfchtype [rfpmode]
                           SENDER:   CHPID    TYPE
                                     rfschpid rfschtype [rfspmode]
                   [* = PATH OPERATING AT REDUCED CAPACITY]
```

© 2015 IBM Corporation

Whenever you migrate to a new CFLEVEL

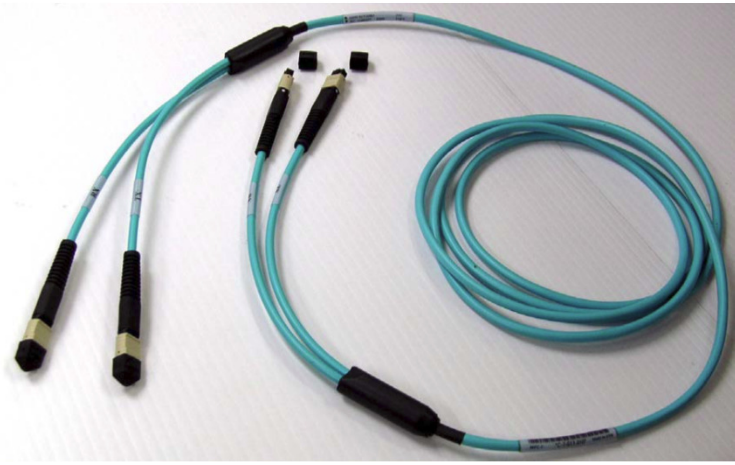
- In general, get to most current LIC levels
- Use CFSIZER to check/update structure sizes:
  - CF structure sizes may increase when migrating to newer level from earlier levels due to additional CFCC controls
  - Improperly sized structures can lead to outages !
- Minimum CFCC image size is 512MB as of CFLEVEL 17

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

© 2015 IBM Corporation

Agenda

- Hardware Updates
  - CFCC Level 20
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2
  - z/OS V2R1
  - z/OS V1R13
- Summary



© 2015 IBM Corporation

New Coupling Links for z13

- Integrated Coupling Adapter Short Range (ICA SR)
  - New CHPID type CS5
  - Uses PCIe Gen3 technology
- Only for z13 to z13 connectivity
  - At most 16 ICA SR features and 32 ICA SR ports supported by z13
  - ICA SR feature resides in PCIe I/O fanout slot in CPC drawer
  - A drawer supports up to 10 ICA SR features and up to 20 ICA SR ports
- Supports distances up to 150 meters
  - Can use OM3 fiber up to 100 meters
  - Must use OM4 fiber to get to 150 meters
- Up to 4 CHPIDs per physical link, with 7 Subchannels per CHPID
- Can be used as timer links

Greater coupling connectivity for single node relative to prior processor generations

© 2015 IBM Corporation

The IBM Integrated Coupling Adapter (ICA SR), introduced on the IBM z13 platform, is a two-port, short distance coupling fanout that utilizes a new coupling channel type: CS5. The ICA SR utilizes PCIe Gen3 technology, with x16 lanes that are bifurcated into x8 lanes for coupling. The ICA SR is designed to drive distances up to 150 m and support a link data rate of 8 GBps. It is also designed to support up to 4 CHPIDs per port and 7 subchannels (devices) per CHPID. The maximum number of ICA SR fanout features is limited to 16 per system.

The ICA SR fanout resides in the PCIe I/O fanout slot on the IBM z13 CPC drawer, which supports 10 PCIe I/O slots. Up to 10 ICA SR fanouts and up to 20 ICA SR ports are supported on an IBM z13 CPC drawer, enabling greater connectivity for short distance coupling on a single processor node compared to prior generations.

The ICA SR can only be used for coupling connectivity between IBM z13 servers, and the ICA SR can only connect to another ICA SR. IBM recommends that you order ICA SR (#0172) on the IBM z13 processors used in a Parallel Sysplex **to help ensure coupling connectivity with future processor generations.**

The ICA SR fanout requires new cabling. For distances up to 100 m, clients can choose the OM3 fiber type. For distances up to 150 m, clients must choose the OM4 fiber type. Refer to IBM z Systems Planning for Fiber Optic Links (FICON/FCP, Coupling Links, and Open System Adapters), GA23-1407, and to IBM z Systems Maintenance for Fiber Optic Links (FICON/FCP, Coupling Links, and Open System Adapters), SY27-7694, which can be found in the Library section of Resource Link®
at
http://www.ibm.com/servers/resourcelink/svc03100.nsf?OpenDatabase
Refer to the Software requirements section.

© 2015 IBM Corporation

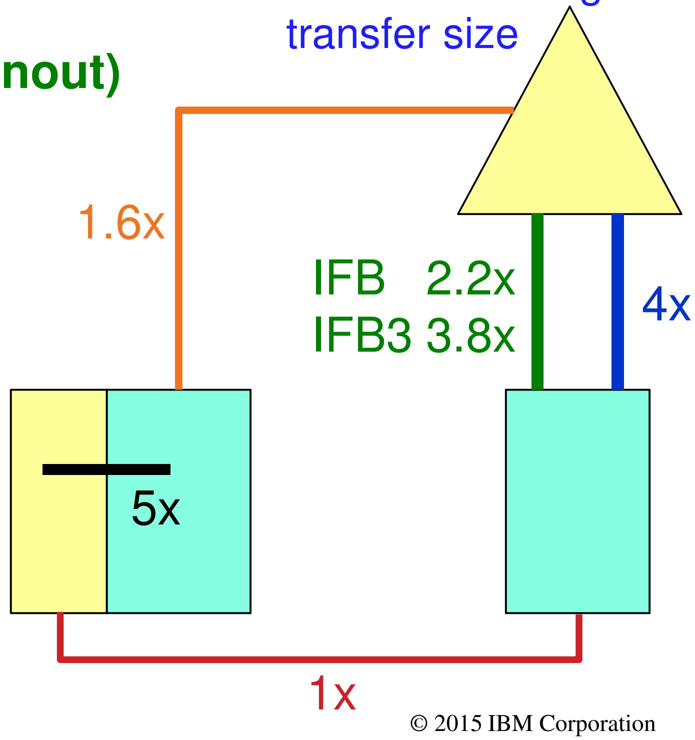
Coupling Link Choices - Overview

Not supported by z13

- **ISC (Inter-System Channel / HCA2-C / CPC-to-IO Fanout)**
  - Fiber optics
  - I/O Adapter card
  - 10km, 20km support with RPQ 8P2197 as carry forward only, and longer distances with qualified DWDM solutions
- **PSIFB (1x IFB / HCA2-O LR or HCA3-O LR / CPC-to-CPC Fanout)**
  - Fiber optics – uses same cabling as ISC
  - 10km and longer distances with qualified WDM solutions
  - Supports multiple CHPIDs per physical link
  - Multiple CF partitions can share physical link
- **PSIFB (12x IFB / HCA2-O or HCA3-O / CPC-to-CPC Fanout)**
  - 150 meter max distance optical cabling
  - Supports multiple CHPIDs per physical link
  - Multiple CF partitions can share physical link
- **ICA SR (Integrated Coupling Adapter)**
  - Connects to PCIe fanout
  - 150 meter max distance
  - Supports up to 4 CHPIDs per physical link
- **IC (Internal Coupling Channel / Internal CPC)**
  - Microcode, no external connection
  - Only between partitions on same processor

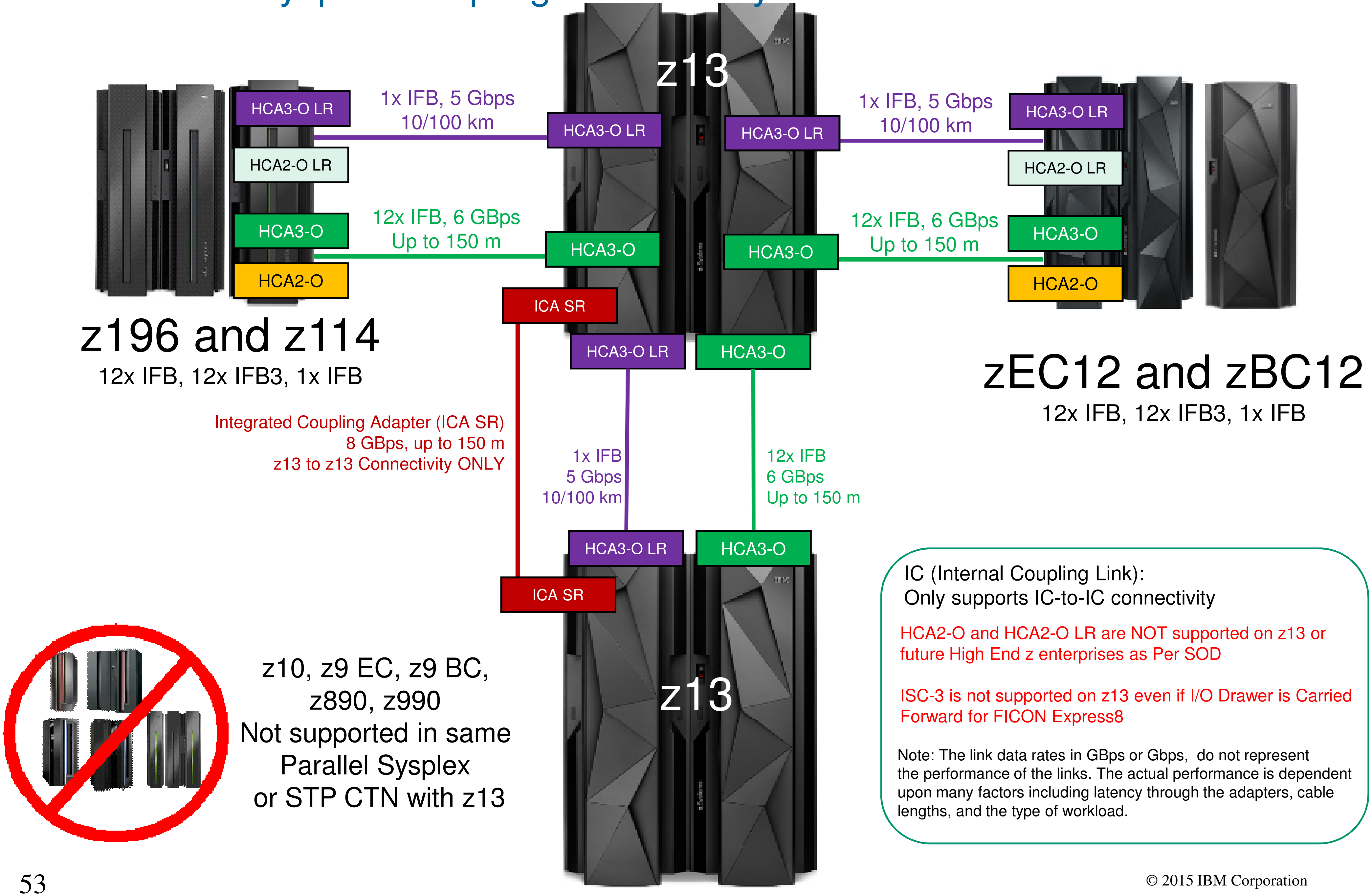
Only for z13 to z13

Relative performance based on average data transfer size



© 2015 IBM Corporation

z13 Parallel Sysplex Coupling Connectivity



Integrated Coupling Adapter (ICA SR)
8 GBps, up to 150 m
z13 to z13 Connectivity ONLY

z10, z9 EC, z9 BC,
z890, z990
Not supported in same
Parallel Sysplex
or STP CTN with z13

IC (Internal Coupling Link):
Only supports IC-to-IC connectivity

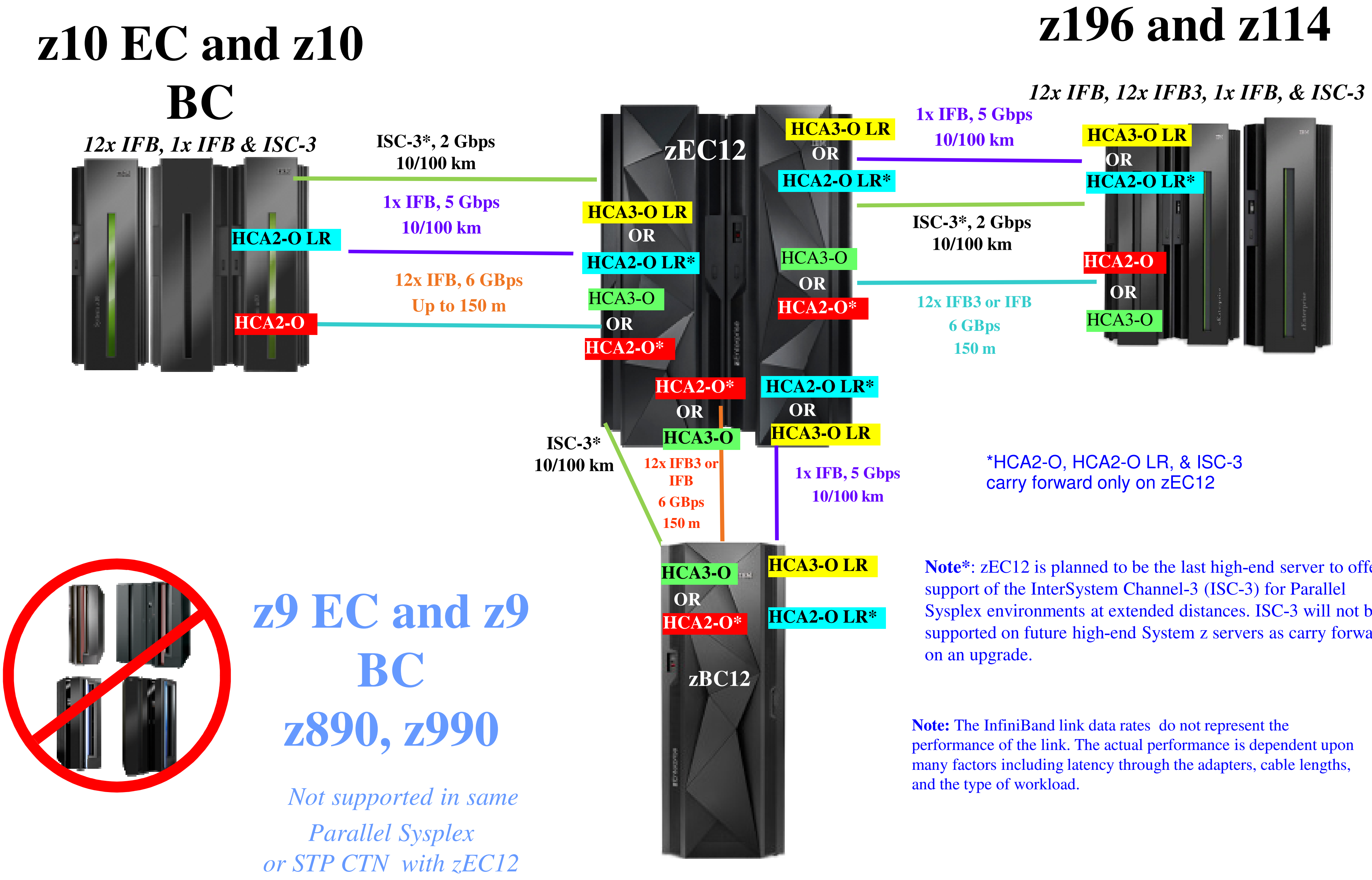
HCA2-O and HCA2-O LR are NOT supported on z13 or future High End z enterprises as Per SOD

ISC-3 is not supported on z13 even if I/O Drawer is Carried Forward for FICON Express8

Note: The link data rates in GBps or Gbps, do not represent the performance of the links. The actual performance is dependent upon many factors including latency through the adapters, cable lengths, and the type of workload.

© 2015 IBM Corporation

zEC12/zBC12 Parallel Sysplex Coupling Connectivity



z10 EC and z10 BC

12x IFB, 1x IFB & ISC-3

HCA2-O LR

HCA2-O

z196 and z114

12x IFB, 12x IFB3, 1x IFB, & ISC-3

HCA3-O LR OR HCA2-O LR*

HCA2-O OR HCA3-O

zEC12

HCA3-O LR OR HCA2-O LR*

HCA3-O OR HCA2-O*

HCA3-O LR OR HCA2-O LR*

HCA3-O OR HCA2-O*

HCA2-O* OR HCA3-O

HCA2-O LR* OR HCA3-O LR

ISC-3*, 2 Gbps 10/100 km

1x IFB, 5 Gbps 10/100 km

12x IFB, 6 GBps Up to 150 m

1x IFB, 5 Gbps 10/100 km

ISC-3*, 2 Gbps 10/100 km

12x IFB3 or IFB 6 GBps 150 m

ISC-3* 10/100 km

12x IFB3 or IFB 6 GBps 150 m

1x IFB, 5 Gbps 10/100 km

HCA3-O OR HCA2-O*

HCA3-O LR

HCA2-O LR*

zBC12

z9 EC and z9 BC z890, z990

Not supported in same Parallel Sysplex or STP CTN with zEC12

*HCA2-O, HCA2-O LR, & ISC-3 carry forward only on zEC12

Note*: zEC12 is planned to be the last high-end server to offer support of the InterSystem Channel-3 (ISC-3) for Parallel Sysplex environments at extended distances. ISC-3 will not be supported on future high-end System z servers as carry forward on an upgrade.

Note: The InfiniBand link data rates do not represent the performance of the link. The actual performance is dependent upon many factors including latency through the adapters, cable lengths, and the type of workload.

© 2015 IBM Corporation

IFB and ICA SR Link Configuration Advantages vs ISC3 links

- Capacity perspective
  - 1 1x IFB = 1 ISC3
  - 1 ICA SR = 1 12x IFB3 = 4 ISC3s
- Eliminating subchannel and path delays
  - Multiple ISC3 links are sometimes configured not for capacity, but to help eliminate delays due to subchannel busy and path busy conditions
  - In lieu of adding links beyond the two needed for redundancy, you could take advantage of:
    - Multiple CHPID support with ICA and IFB links
    - 32 subchannel support for 1x IFB
- Multiple sysplex sharing hardware
  - With ISC3 links, each sysplex needs its own links
  - Multiple CHPID support with ICA and IFB links can often be used in lieu of configuring separate links for each sysplex

© 2015 IBM Corporation

Multiple CHPID recommendations for ICA and IFB links

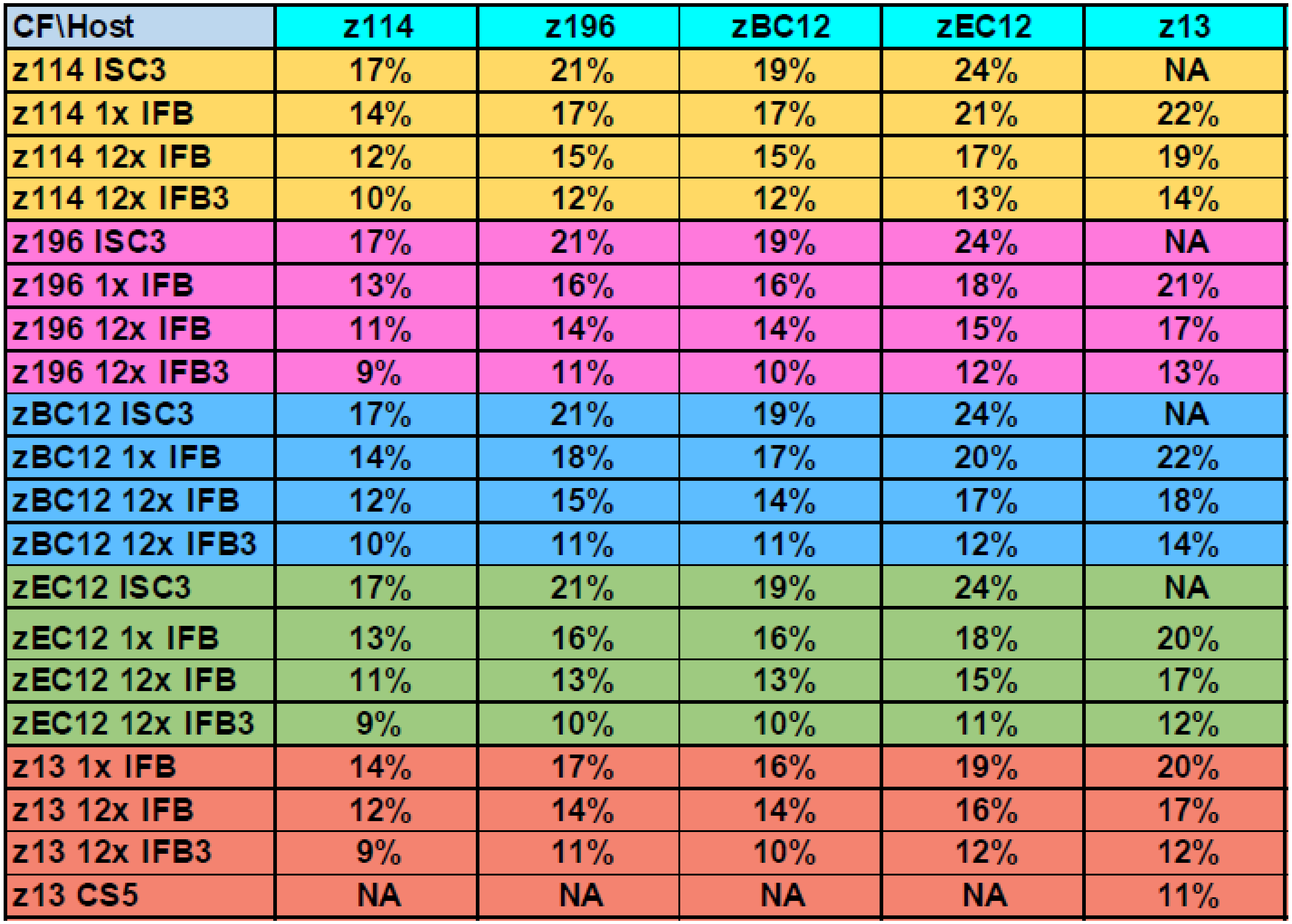
- Two links with two CHPIDs often sufficient for many installations
  - Provides redundancy for availability
  - Provides 28 subchannels, which is generally sufficient to keep percentage of delayed requests below the 10% guideline
- May define up to a max of 4 CHPIDs per link
  - Additional CHPIDs can provide additional connectivity or help reduce busy conditions for heavy loads
  - 4 is practical limit for IFB, and actual limit for ICA

© 2015 IBM Corporation

Coupling Technology vs Host Processor Speed

Host effect with primary application involved in data sharing

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



CF\Host	z114	z196	zBC12	zEC12	z13
z114 ISC3	17%	21%	19%	24%	NA
z114 1x IFB	14%	17%	17%	21%	22%
z114 12x IFB	12%	15%	15%	17%	19%
z114 12x IFB3	10%	12%	12%	13%	14%
z196 ISC3	17%	21%	19%	24%	NA
z196 1x IFB	13%	16%	16%	18%	21%
z196 12x IFB	11%	14%	14%	15%	17%
z196 12x IFB3	9%	11%	10%	12%	13%
zBC12 ISC3	17%	21%	19%	24%	NA
zBC12 1x IFB	14%	18%	17%	20%	22%
zBC12 12x IFB	12%	15%	14%	17%	18%
zBC12 12x IFB3	10%	11%	11%	12%	14%
zEC12 ISC3	17%	21%	19%	24%	NA
zEC12 1x IFB	13%	16%	16%	18%	20%
zEC12 12x IFB	11%	13%	13%	15%	17%
zEC12 12x IFB3	9%	10%	10%	11%	12%
z13 1x IFB	14%	17%	16%	19%	20%
z13 12x IFB	12%	14%	14%	16%	17%
z13 12x IFB3	9%	11%	10%	12%	12%
z13 CS5	NA	NA	NA	NA	11%

With z/OS V1.2 and above, synch-> asynch conversion caps values in the table at about 18%
IC links scale with the speed of the host technology and would provide an 8% effect in each case

© 2015 IBM Corporation

For More Information

- “IBM z Systems Connectivity Handbook” (SG24-5444)
  - Updated April 2015
- “Implementing and Managing InfiniBand Coupling Links on System z” (SG24-7539)
  - Available at www.redbooks.ibm.com
- www.ibm.com/systems/z/advantages/pso/whitepaper.html
  - CF Configuration Options White Paper

© 2015 IBM Corporation

Agenda

- Hardware Updates
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R1
  - z/OS V1R13
  - z/OS V1R12
- Summary

© 2015 IBM Corporation

Glossary for System z 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.
	Going Away Signal	A reliable unambiguous signal to indicate that the CPC is about to enter a check stopped state.
PTS	Preferred Time Server	The server assigned by the customer to be the preferred stratum 1 server in an STP-only CTN.

© 2015 IBM Corporation

z13 STP Enhancements

- Enable STP communications via the IBM Integrated Coupling Adapter (ICA SR)
- HMC Panels enhanced

Initialized Time Panel

- Lists time zone and leap second offset
- Indicates if the system time was set

Can quickly check fields during CTN configuration

Set Date and Time Panel

- Encourages use of External Time Source to set CTN time

Time Zone panel

- Confirmation messages when setting STP time zone on Current Time Server
- Lists scheduled switch times for leap seconds and time zone/daylight savings time on Timing Network Tab

Also added support for view-only STP panels

© 2015 IBM Corporation

New messages for significant STP events (with z/OS V2R2)

New z/OS console messages regarding events that can affect sysplex timing:

```
IEA396E ARBITER ASSISTED RECOVERY IS DISABLED FOR THE CTN. CODE = hh
IEA397I ARBITER ASSISTED RECOVERY IS ENABLED FOR THE CTN.
IEA398I STP ROLE SERVER ATTACHMENT STATE CHANGE. STATE = cccccccc
     REASON = h
```

FULL
PARTIAL
DEGRADED

You might want automation actions to be driven from these events.

© 2015 IBM Corporation

Server Time Protocol Enhancements

- Improved SE Time Accuracy – zEC12 GA2 and zBC12
  - Optionally, the SE can be configured to connect to an external time source periodically to maintain highly accurate time that can be used, if required, to initialize CTN time during POR.
- Broadband Security Improvements for STP
  - Authenticates NTP servers when accessed by the HMC client through a firewall
  - Authenticates NTP clients when the HMC is acting as an NTP server
  - Provides symmetric key (NTP V3-V4) and autokey (NTP V4) authentication (Autokey is not supported if Network Address Translation is used)
  - This is the highest level of NTP security available

© 2015 IBM Corporation

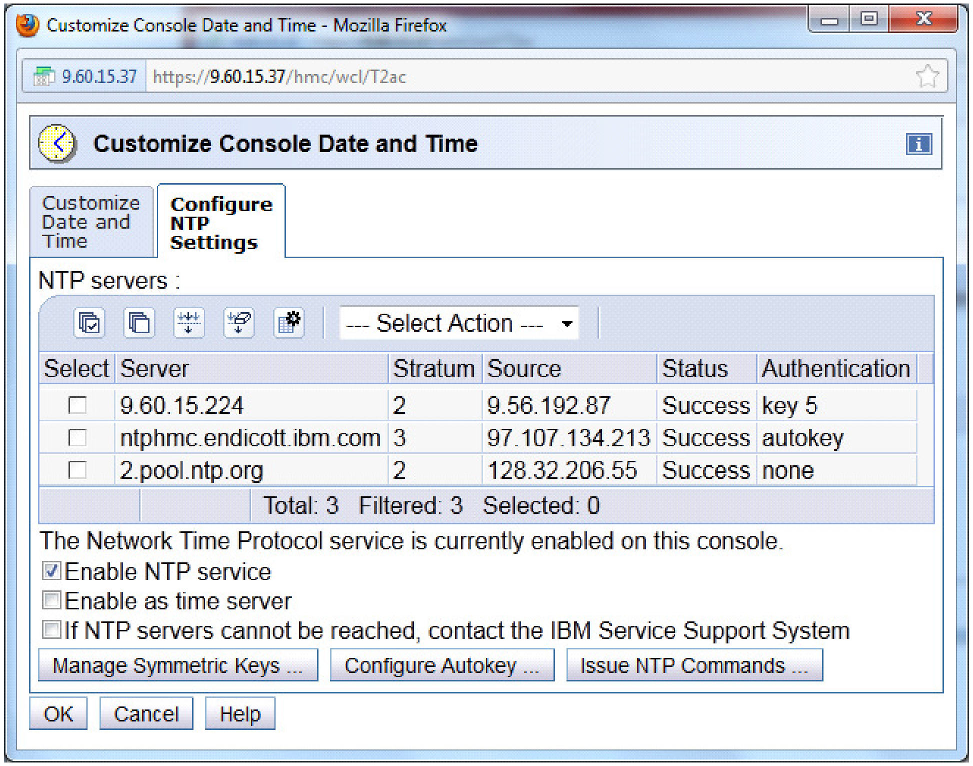
zEC12/zBC12 Server Time Protocol Enhancements

- Improved NTP Commands panel on HMC/SE
  - Shows command response details
- Telephone modem dial out to an STP time source is no longer supported
  - All STP dial functions are still supported by broadband connectivity
  - zEC12 HMC LIC no longer supports dial modems
    (Fulfills the Statement of Direction in Letter 111-167, dated October 12, 2011)

© 2015 IBM Corporation

NTP Broadband Authentication Support for zEC12/zBC12

- Highest level of NTP security available
- Panels accept and generate key information to be configured into HMC NTP configuration.
- Autokey authentication not available with network address translating (NAT) firewall. Symmetric key still supports NAT.
- Autokey availability based on MCP level. First supported at zEC12 MCP level.


Customize Console Date and Time - Mozilla Firefox
9.60.15.37 https://9.60.15.37/hmc/wcl/T2ac
Customize Console Date and Time
Customize Date and Time
Configure NTP Settings
--- Select Action ---
Manage Symmetric Keys ...
Configure Autokey ...
Issue NTP Commands ...
OK
Cancel
Help

NTP servers :

Select	Server	Stratum	Source	Status	Authentication
□	9.60.15.224	2	9.56.192.87	Success	key 5
□	ntphmc.endicott.ibm.com	3	97.107.134.213	Success	autokey
□	2.pool.ntp.org	2	128.32.206.55	Success	none

Total: 3 Filtered: 3 Selected: 0

The Network Time Protocol service is currently enabled on this console.

☑ Enable NTP service

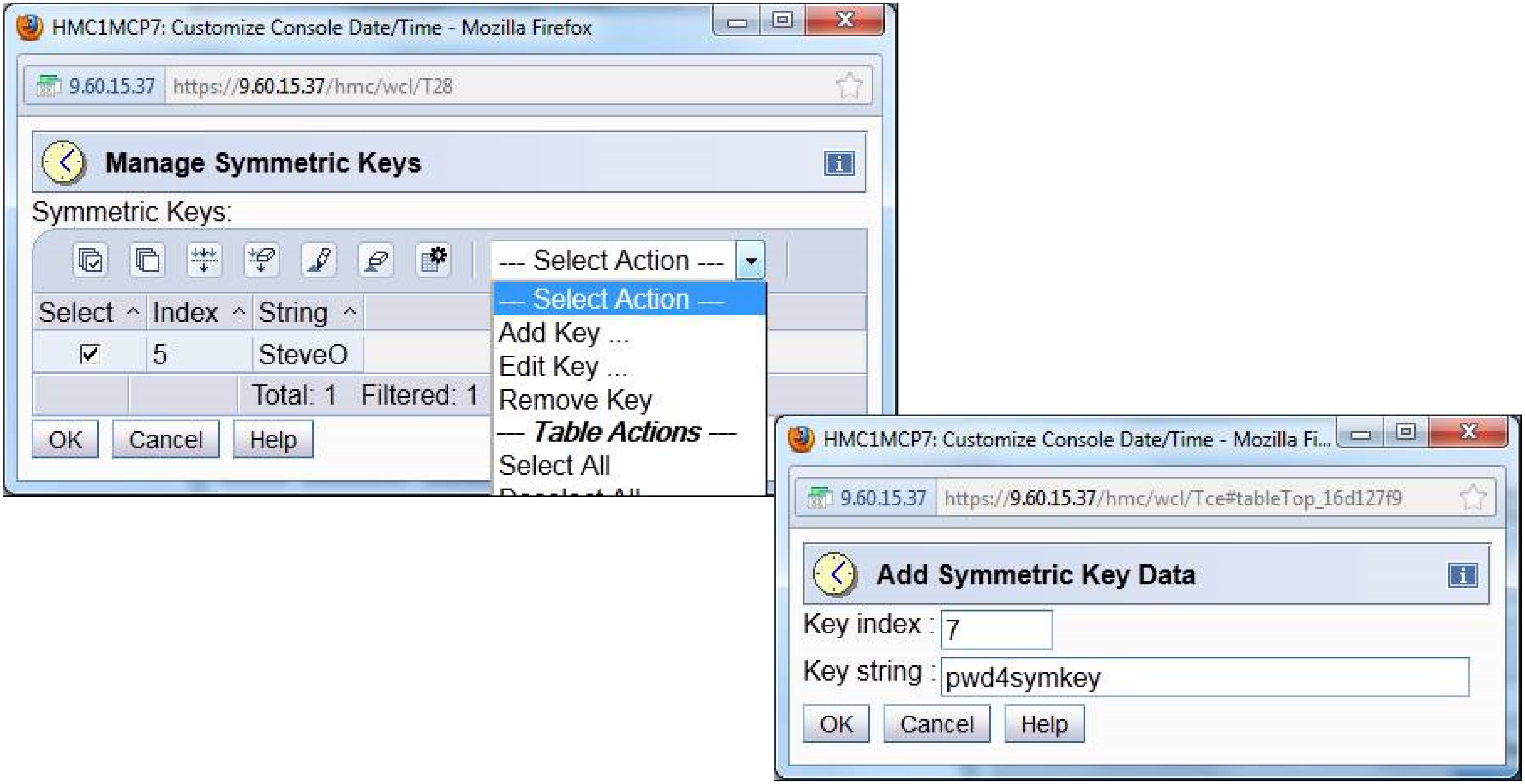
□ Enable as time server

□ If NTP servers cannot be reached, contact the IBM Service Support System

© 2015 IBM Corporation

NTP Authentication - Symmetric Key

- Symmetric key encryption uses the same key for both encryption and decryption. Users exchanging data keep this key to themselves. Message encrypted with a secret key can be decrypted only with the same secret key.


HMC1MCP7: Customize Console Date/Time - Mozilla Firefox
9.60.15.37 https://9.60.15.37/hmc/wcl/T28
Manage Symmetric Keys
--- Select Action ---
— Select Action —
Add Key ...
Edit Key ...
Remove Key
— Table Actions —
Select All
Deselect All
OK
Cancel
Help
HMC1MCP7: Customize Console Date/Time - Mozilla Fi...
9.60.15.37 https://9.60.15.37/hmc/wcl/Tce#tableTop_16d127f9
Add Symmetric Key Data
OK
Cancel
Help

Symmetric Keys:

Select	Index	String
☑	5	SteveO

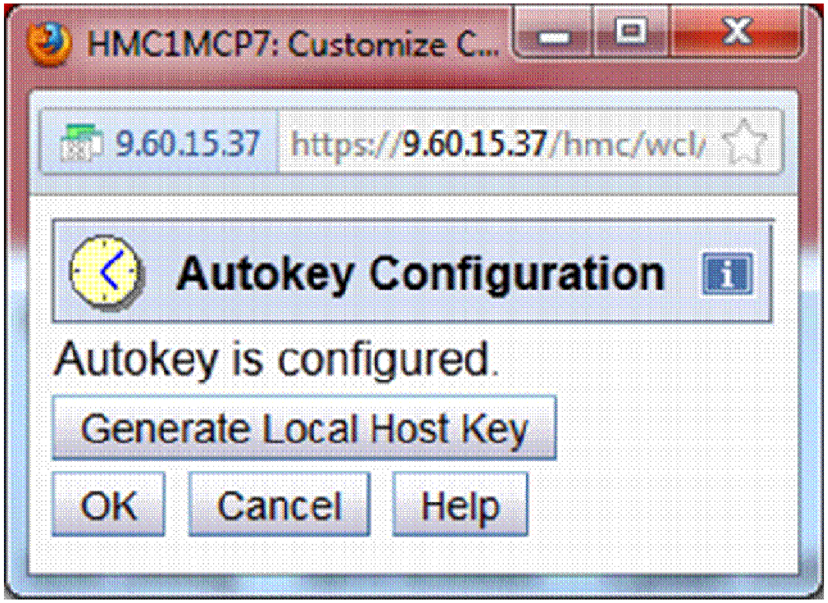
Total: 1 Filtered: 1

Key index : 7

Key string : pwd4symkey

© 2015 IBM Corporation

NTP Authentication - AutoKey



HMC1MCP7: Customize C...

9.60.15.37 https://9.60.15.37/hmc/wcl/

Autokey Configuration

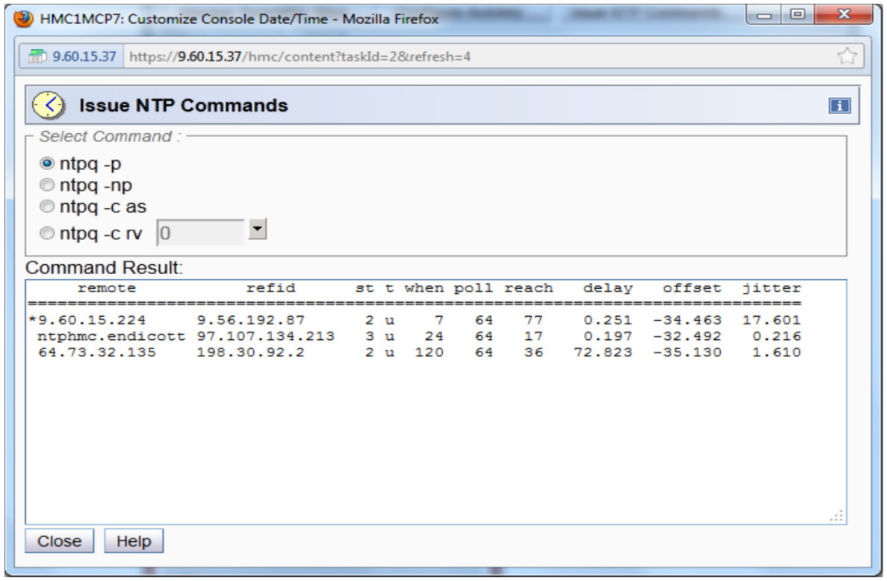
Autokey is configured.

- An autokey cipher (also known as the autoclave cipher) is a cipher which incorporates the message (the plaintext) into the key
- http://www.eecis.udel.edu/~mills/ntp/html/autokey.html

© 2015 IBM Corporation

NTP Authentication – Issue NTP Commands Panel

- One customer complaint in regard to the HMC NTP server panels, as they stand today, has been the status of the connection to target NTP servers.
- With the addition of NTP authentication, this display will aid in the determination of failures during configuration.

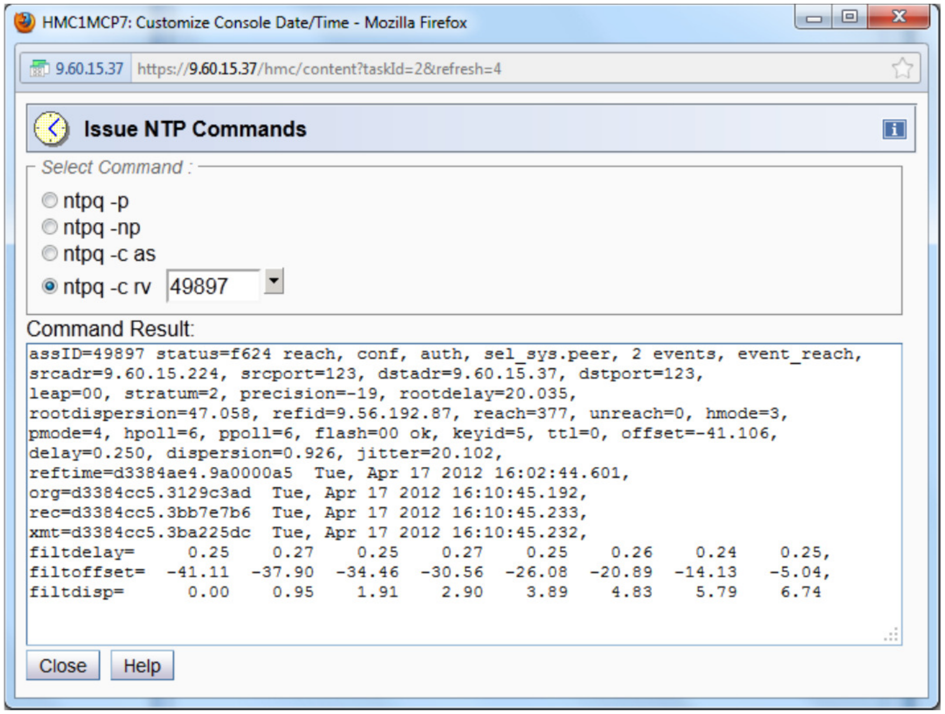

HMC1MCP7: Customize Console Date/Time - Mozilla Firefox
9.60.15.37 https://9.60.15.37/hmc/content?taskId=28&refresh=4
Issue NTP Commands
Close
Help

Select Command :

● ntpq -p
○ ntpq -np
○ ntpq -c as
○ ntpq -c rv 0

Command Result:

```
     remote           refid      st t when poll reach   delay   offset  jitter
==============================================================================
*9.60.15.224     9.56.192.87      2 u    7   64   77    0.251  -34.463  17.601
 ntphmc.endicott 97.107.134.213   3 u   24   64   17    0.197  -32.492   0.216
 64.73.32.135    198.30.92.2      2 u  120   64   36   72.823  -35.130   1.610
```


HMC1MCP7: Customize Console Date/Time - Mozilla Firefox
9.60.15.37 https://9.60.15.37/hmc/content?taskId=28&refresh=4
Issue NTP Commands
Close
Help

Select Command :

○ ntpq -p
○ ntpq -np
○ ntpq -c as
● ntpq -c rv 49897

Command Result:

```
assID=49897 status=f624 reach, conf, auth, sel_sys.peer, 2 events, event_reach,
srcadr=9.60.15.224, srcport=123, dstadr=9.60.15.37, dstport=123,
leap=00, stratum=2, precision=-19, rootdelay=20.035,
rootdispersion=47.058, refid=9.56.192.87, reach=377, unreach=0, hmode=3,
pmode=4, hpoll=6, ppoll=6, flash=00 ok, keyid=5, ttl=0, offset=-41.106,
delay=0.250, dispersion=0.926, jitter=20.102,
reftime=d3384ae4.9a0000a5  Tue, Apr 17 2012 16:02:44.601,
org=d3384cc5.3129c3ad  Tue, Apr 17 2012 16:10:45.192,
rec=d3384cc5.3bb7e7b6  Tue, Apr 17 2012 16:10:45.233,
xmt=d3384cc5.3ba225dc  Tue, Apr 17 2012 16:10:45.232,
filtdelay=     0.25    0.27    0.25    0.27    0.25    0.26    0.24    0.25,
filtoffset=  -41.11  -37.90  -34.46  -30.56  -26.08  -20.89  -14.13   -5.04,
filtdisp=      0.00    0.95    1.91    2.90    3.89    4.83    5.79    6.74
```

The explanation for the ntpq commands are all located on the following link:
http://www.eecis.udel.edu/~mills/ntp/html/ntpq.html

© 2015 IBM Corporation

STP References for Additional Information

- **Redbooks**
  - Server Time Protocol Planning Guide, SG24-7280
    - http://www.redbooks.ibm.com/redpieces/abstracts/sg247280.html
  - Server Time Protocol Implementation Guide, SG24-7281
    - http://www.redbooks.ibm.com/redpieces/abstracts/sg247281.html
- **TechDocs**
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102019 (avoid outages)
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102037 (recovery voting)
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105103 (restore STP config after power on reset)
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102081 (STP and leap seconds)
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS2398 (STP overview)
- **Education**
  - Introduction to Server Time Protocol (STP)
    - Available on Resource Link at General Availability (GA)
    - www.ibm.com/servers/resourcelink/hom03010.nsf?OpenDatabase
- **STP Web site**
  - www.ibm.com/systems/z/pso/stp.html
- **Systems Assurance**
  - The IBM team is required to complete a Systems Assurance Review (SAPR Guide, SA06-012) and to complete the Systems Assurance Confirmation Form via Resource Link
  - http://w3.ibm.com/support/assure/assur30i.nsf/WebIndex/SA779
- **For further information on NTP and the NTP Public services project, refer to the Web sites:**
  - http://www.ntp.org
  - http://support.ntp.org
  - http://www.faqs.org/rfcs/rfc1305.html

© 2015 IBM Corporation

Agenda

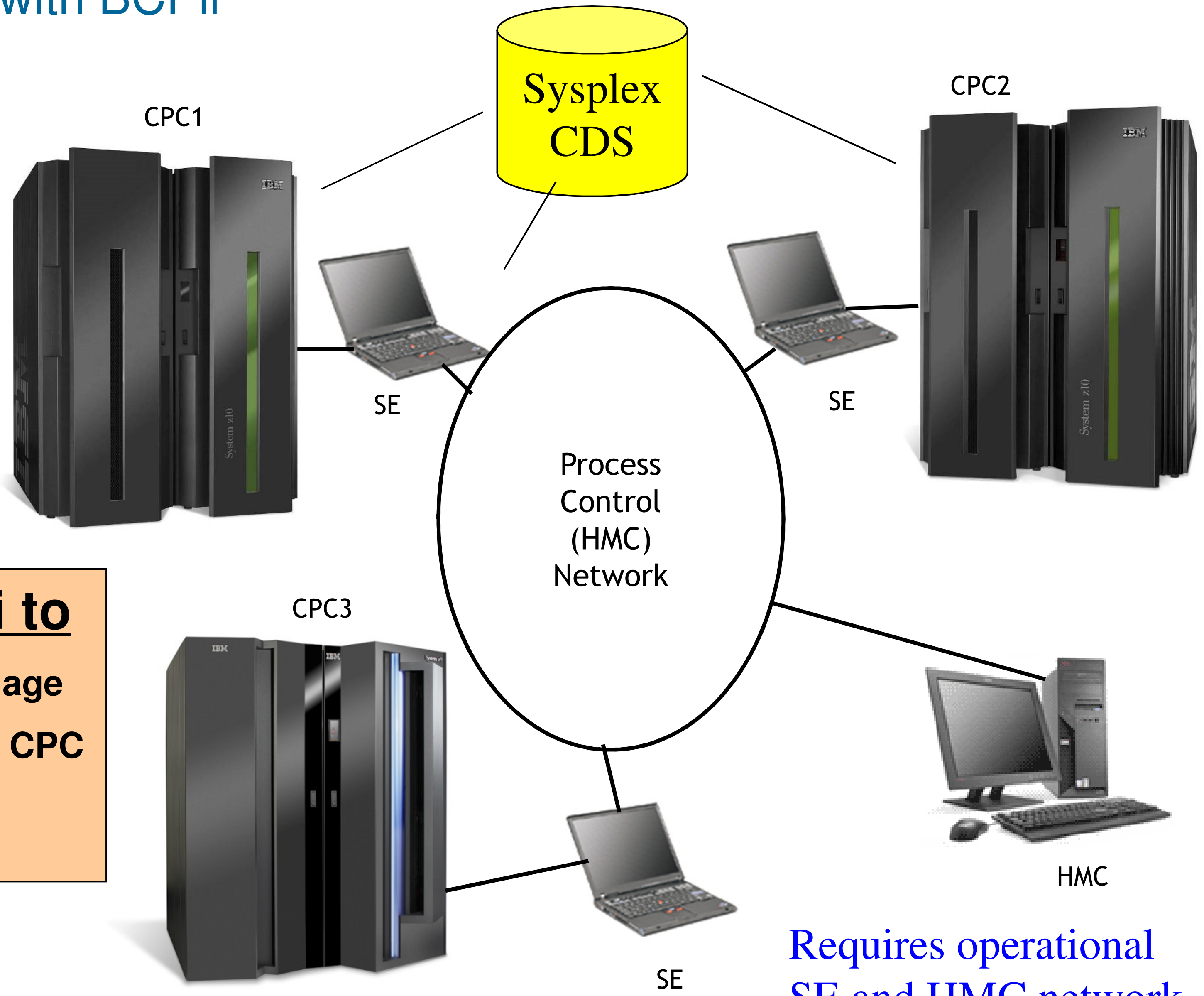
- Hardware Updates
  - CFCC Level 20
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2
  - z/OS V2R1
  - z/OS V1R13
- Summary

© 2015 IBM Corporation

z/OS V1R11 - SFM with BCPii

z/OS Images

(not VM guests)



XCF uses BCPii to

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

Requires operational SE and HMC network

© 2015 IBM Corporation

z/OS V1R11 - SFM with BCPii

- Expedient removal of unresponsive or failed systems is essential to high availability in sysplex
- XCF exploits BCPii services to:
  - Detect failed systems
  - Reset systems
- **Benefits**:
  - Improved availability by reducing duration of sympathy sickness
    - No waiting for FDI to expire
  - Eliminate manual intervention in more cases
    - Avoid IXC102A, IXC402D, IXC409D
  - Potentially prevent human error that can cause data corruption
    - Validate “down”

© 2015 IBM Corporation

z/OS V1R11 - SFM with BCPii

- SFM will automatically exploit BCPii 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, z196, z114, zEC12, or zBC12, z13
  - **New version of sysplex CDS is primary in sysplex**
    - Formatted with: ITEM NAME(SSTATDET) NUMBER(1)

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

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

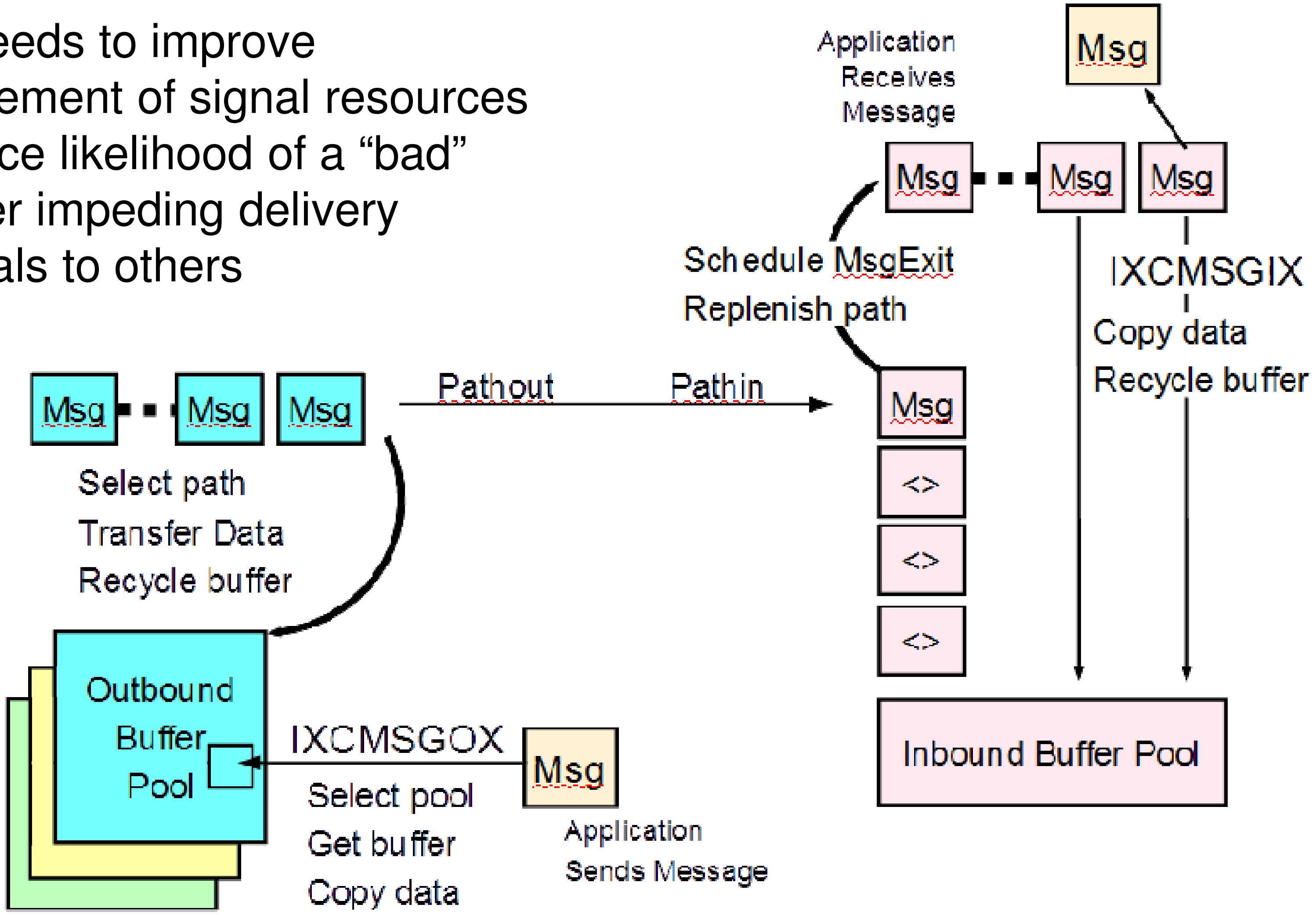
CF Gain Ownership protection

IXCQUERY for CDS and policies

© 2015 IBM Corporation

XCF Signal Service

XCF needs to improve management of signal resources to reduce likelihood of a “bad” member impeding delivery of signals to others



© 2015 IBM Corporation

Message Isolation

Goal: Prevent a target member from consuming so much signal resource that it impedes the delivery of signals to other members

- Primary issue is consumption of inbound signal buffers
- Secondary issue is consumption of common storage

Once a signal is accepted by XCF, it must be delivered

- Receiving system cannot reject or discard incoming signals
- The sending system must be the one to refuse messages targeted to a “bad” member

The inbound side identifies “bad” members and “isolates” them

- Tells sending systems to stop accepting signals for the bad member
- (and when to resume sending)

© 2015 IBM Corporation

Consequences of being Message Isolated

- XCF either delays or rejects signals targets to an isolated member
- By default, the reason is “no buffer”
  - “*no buffer for you*”. Signals sent to non-isolated members are still accepted.
  - Exploiters can optionally request unique reason code of “isolated”

- Delayed signals are held by sending system until the target member becomes “not isolated” or the message completes (timeout, cancel, …)
- If XCF accepts a signal, but the target member is isolated before the signal can be queued to a signal path, the signal is held until the member becomes “not isolated”, terminates, or message completes

© 2015 IBM Corporation

Consequences of being “Message Isolated” …

- So impacted senders may see more rejects, delays, or timeouts
  - But this is the intended behavior
  - If a member is not able to participate effectively, there will likely be impacts to its peers
- Goal is to limit scope of impact to the offending application (group)
  - In the past, all group members had potential for impact due to no buffer conditions or transfer delays
  - Signals for unrelated applications (groups) should flow freely
- Note! There can still be impact to other applications
  - Others may depend on the services of the offending application
  - There may be (are) scenarios that could defeat the XCF algorithms

© 2015 IBM Corporation

Terminology

Message isolation

The process by which XCF monitoring identifies a member that is not processing its signals in a timely manner and then arranges to have sending systems reject or delay messages targeted to that member

Isolated member

A target member whose signals are subject to being rejected or delayed by XCF due to message isolation. We say the member is “message isolated”.

Impacted member

A sending member whose signals are being rejected or delayed because the target member is message isolated

© 2015 IBM Corporation

Terminology …

Isolation window

The period of time during which a target member is message isolated

Impact window

The period of time during which a sending member was impacted by message isolation of a given target member.

- Member impact window – an impact window for a given sending member
- System impact window – an impact window for a given sending system is the union of the set of member impact windows for all the members on that system

Note:

Isolation windows do not necessarily induce impact windows.

Impact windows can span isolation windows.

© 2015 IBM Corporation

New XCF Messages

- IXC638I – documents isolation window for given member
- IXC6371 – documents impact window for given member

  By default, these messages are issued to hardcopy log.

  Issued at start and end of a window.

  Periodically reissued if window persists long enough.
- IXC645E – alerts operator to existence of isolated members
- IXC440E – alerts operator to existence of impacted members

  These are issued as highlighted messages.

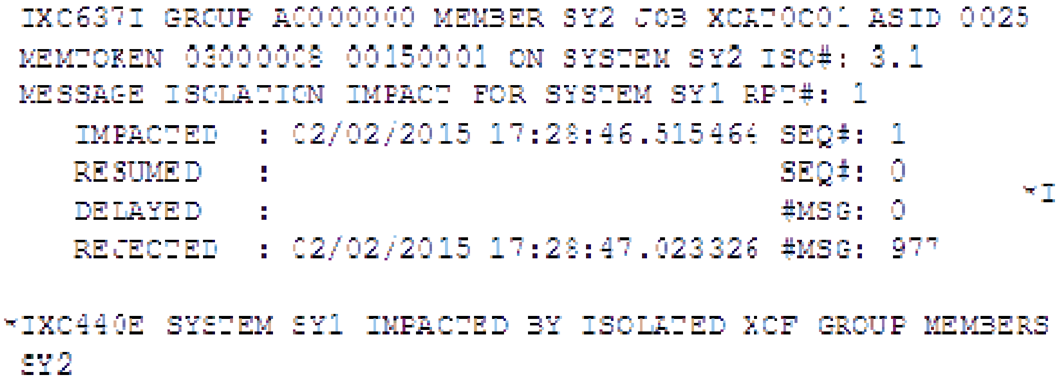
  Persist until no members on the issuing system qualify.

  Investigate with D XCF,G and/or review of IXC637I, IXC638I.

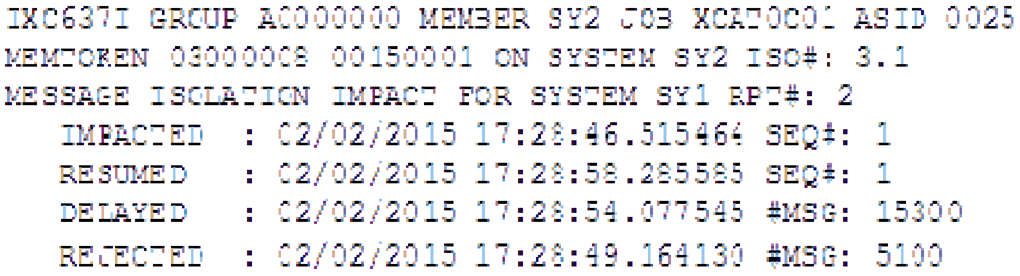
© 2015 IBM Corporation

XCF Messages show isolation and impact

Systems with Impacted Member

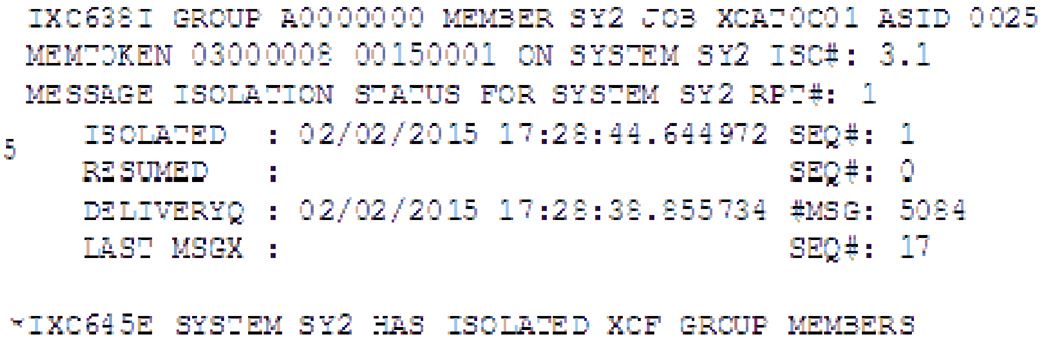


```
IXC637I GROUP A0000000 MEMBER SY2 JOB XCAT0C01 ASID 0025
MEMTOKEN 03000008 00150001 ON SYSTEM SY2 ISO#: 3.1
MESSAGE ISOLATION IMPACT FOR SYSTEM SY1 RPT#: 1
   IMPACTED  : 02/02/2015 17:28:46.515464 SEQ#: 1
   RESUMED   :                            SEQ#: 0
   DELAYED   :                            #MSG: 0
   REJECTED  : 02/02/2015 17:28:47.023326 #MSG: 977

*IXC440E SYSTEM SY1 IMPACTED BY ISOLATED XCF GROUP MEMBERS ON SYSTEM
 SY2
```

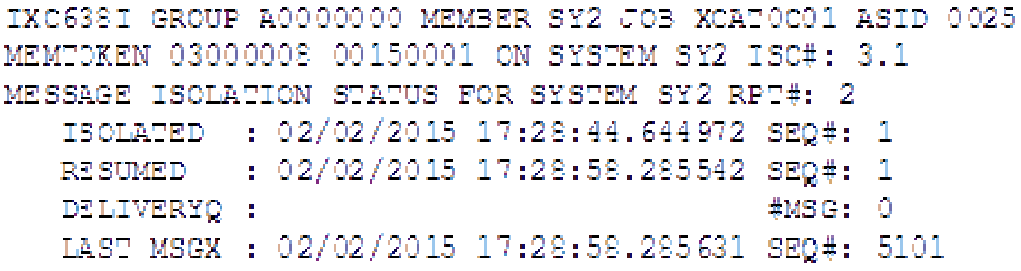


```
IXC637I GROUP A0000000 MEMBER SY2 JOB XCAT0C01 ASID 0025
MEMTOKEN 03000008 00150001 ON SYSTEM SY2 ISO#: 3.1
MESSAGE ISOLATION IMPACT FOR SYSTEM SY1 RPT#: 2
   IMPACTED  : 02/02/2015 17:28:46.515464 SEQ#: 1
   RESUMED   : 02/02/2015 17:28:58.285585 SEQ#: 1
   DELAYED   : 02/02/2015 17:28:54.077545 #MSG: 15300
   REJECTED  : 02/02/2015 17:28:49.164130 #MSG: 5100
```

System with Isolated Member



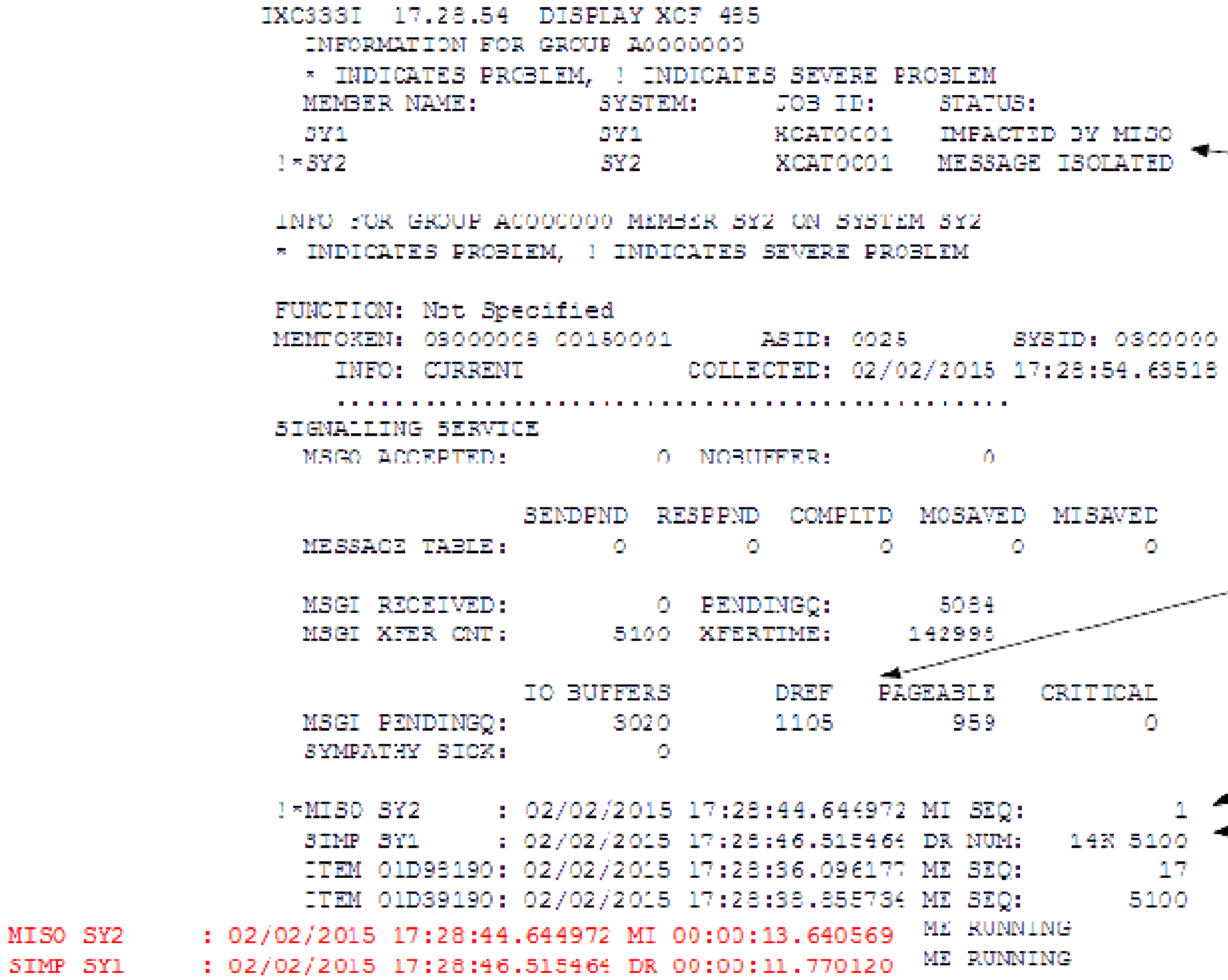
```
IXC638I GROUP A0000000 MEMBER SY2 JOB XCAT0C01 ASID 0025
MEMTOKEN 03000008 00150001 ON SYSTEM SY2 ISO#: 3.1
MESSAGE ISOLATION STATUS FOR SYSTEM SY2 RPT#: 1
   ISOLATED  : 02/02/2015 17:28:44.644972 SEQ#: 1
   RESUMED   :                            SEQ#: 0
   DELIVERYQ : 02/02/2015 17:28:38.855734 #MSG: 5084
   LAST MSGX :                            SEQ#: 17

*IXC645E SYSTEM SY2 HAS ISOLATED XCF GROUP MEMBERS
```



```
IXC638I GROUP A0000000 MEMBER SY2 JOB XCAT0C01 ASID 0025
MEMTOKEN 03000008 00150001 ON SYSTEM SY2 ISO#: 3.1
MESSAGE ISOLATION STATUS FOR SYSTEM SY2 RPT#: 2
   ISOLATED  : 02/02/2015 17:28:44.644972 SEQ#: 1
   RESUMED   : 02/02/2015 17:28:58.285542 SEQ#: 1
   DELIVERYQ :                            #MSG: 0
   LAST MSGX : 02/02/2015 17:28:58.285631 SEQ#: 5101
```

© 2015 IBM Corporation

D XCF,GROUP,grpname,memname



```
IXC333I  17.28.54  DISPLAY XCF 485
   INFORMATION FOR GROUP A0000000
   * INDICATES PROBLEM, ! INDICATES SEVERE PROBLEM
   MEMBER NAME:          SYSTEM:     JOB ID:    STATUS:
   SY1                   SY1         XCATOC01   IMPACTED BY MISO
 !*SY2                   SY2         XCATOC01   MESSAGE ISOLATED

 INFO FOR GROUP A0000000 MEMBER SY2 ON SYSTEM SY2
 * INDICATES PROBLEM, ! INDICATES SEVERE PROBLEM

 FUNCTION: Not Specified
 MEMTOKEN: 03000008 00150001      ASID: 0025       SYSID: 0300000
     INFO: CURRENT            COLLECTED: 02/02/2015 17:28:54.63518
     ...............................................
 SIGNALLING SERVICE
   MSGO ACCEPTED:          0  NOBUFFER:          0

                   SENDPND  RESPPND  COMPLTD  MOSAVED  MISAVED
   MESSAGE TABLE:        0        0        0        0        0

   MSGI RECEIVED:          0  PENDINGQ:       5084
   MSGI XFER CNT:       5100  XFERTIME:     142995

                   IO BUFFERS       DREF   PAGEABLE   CRITICAL
   MSGI PENDINGQ:        3020       1105        959          0
   SYMPATHY SICK:           0

 !*MISO SY2     : 02/02/2015 17:28:44.644972 MI SEQ:          1
   SIMP SY1     : 02/02/2015 17:28:46.515464 DR NUM:   14K 5100
   ITEM 01D98190: 02/02/2015 17:28:36.096177 ME SEQ:         17
   ITEM 01D39190: 02/02/2015 17:28:38.855734 ME SEQ:       5100
MISO SY2     : 02/02/2015 17:28:44.644972 MI 00:00:13.640569 ME RUNNING
SIMP SY1     : 02/02/2015 17:28:46.515464 DR 00:00:11.770120 ME RUNNING
```

New status inserts indicate when an active member is isolated or impacted

Migration has moved signals to DREF and pageable

SY2 Isolated

SY1 Impacted
>14K delayed
5100 rejects

When windows close, can see duration of isolation and impact

© 2015 IBM Corporation

Programming Interfaces

- IXCJOIN
  - New MSGISO keyword to request “isolated” instead of “no buffer”
- IXCMSGO, IXCMSGOX, IXCYCON
  - New “isolated” reason code
- Various query services will indicate when member is isolated or impacted
  - IXCMG, IXCYAMDA
  - IXCQUERY, IXCYQUAA
  - (these are what DISPLAY XCF,GROUP uses)

© 2015 IBM Corporation

Coexistence, Migration, Exploitation

- New behaviors only apply when sender and target reside on a system running z/OS V2R2
  - Does apply to local message traffic, though we seldom see issues there
  - So not likely to see any new behavior until there are at least two systems running z/OS V2R2 in the sysplex
- Simply IPLing system with z/OS V2R2 activates the new behavior
  - Can be disabled via new XCF FUNCTIONS switch MSGISO
- When communicating with down level system, the old behaviors apply (and so derive no benefit)
  - Members on down level system will not be isolated
  - Signals from down level sender to isolated target member will be sent
- Down level systems do not require any compatibility support

© 2015 IBM Corporation

Exploiter messages regarding “no buffer” may be inaccurate !

- On z/OS V2R2 systems, XCF might now selectively indicate “no buffer” for signals targeted to an isolated member
- Some XCF exploiters issue messages to complain when their msgout request is rejected for a “no buffer” condition
  - In the past, you might then go look at your MAXMSG specifications
  - But with z/OS V2R2, those exploiter messages might be the result of the target member being “message isolated”
  - So with z/OS V2R2, you must first look to see whether message isolation might apply
- XCF query services (and therefore measurement products such as RMF) only indicate “no buffer” for true MAXMSG constraints

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

CF Gain Ownership protection

IXCQUERY for CDS and policies

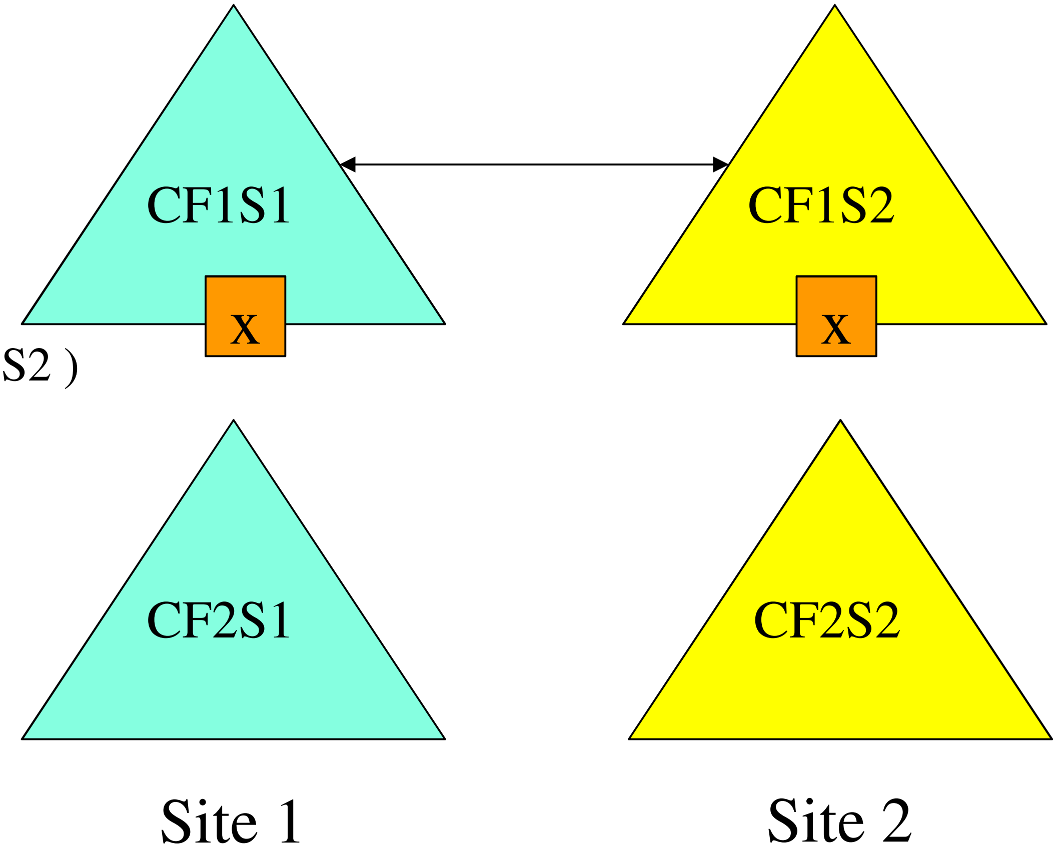
© 2015 IBM Corporation

PREFLIST might not achieve desired placement for duplexed structures

Want structures duplexed across sites for DR or availability

CFRM Policy

```
CF NAME(CF1S1) SITE(SITE1) …
CF NAME(CF2S1) SITE(SITE1) …
CF NAME(CF1S2) SITE(SITE2)
CF NAME(CF2S2) SITE(SITE2)

STRUCTURE STRNAME(x)
  DUPLEX(ENABLED)
  PREFLIST( CF1S1, CF1S2, CF2S1, CF2S2 )
```



But what if CF1S2 is down for maintenance?

Structure likely duplexed across the site 1 CF’s.

Fail to achieve desired cross site availability.

© 2015 IBM Corporation

New CFRM Policy Specifications

For a structure that is eligible for duplexing, can now specify a site preference

DUPLEX(ALLOWED,*site_preference*)

DUPLEX(ENABLED,*site_preference*)

Where *site_preference* is one of:

- ANYSITE – default, ignore site (same as today)
- CROSSSITE – prefer host CF's be in different sites
- SAMESITE – prefer host CF's be in the same site
- SAMESITEONLY – host CF's must reside in the same site

© 2015 IBM Corporation

You may want to revise PREFLIST when specifying site preference

Objective

- Prefer structure be duplexed across sites for availability
- Want primary instance in site 1 for locality to workload

Current structure definition might be:

ENFORCEORDER(NO)
DUPLEX(ENABLED)
PREFLIST(CF1SITE1,**CF1SITE2,CF2SITE1**,CF2SITE2)

List first pair of CF's in site 1, site 2 order to encourage CFRM to allocated primary in site 1 and secondary in site 2. But if CF1SITE1 not available, primary instance is placed in in site 2 CF – fail to get desired locality.

New structure definition might be:

ENFORCEORDER(NO)
DUPLEX(ENABLED**,CROSSSITE**)
PREFLIST(CF1SITE1,**CF2SITE1,CF1SITE2,**CF2SITE2)

List both site 1 CF's first to maintain locality for primary. CROSSSITE will get site 2 CF's used for secondary instance.

© 2015 IBM Corporation

CFRM Policy Sanity Checks – Message IXC754I

If you specify DUPLEX keyword

THE PREFLIST MUST CONTAIN TWO OR MORE FACILITIES.

If you specify CROSSSITE:

THE PREFLIST MUST CONTAIN TWO OR MORE FACILITIES WITH DIFFERENT SITE VALUES.

If you specify SAMESITE or SAMESITEONLY:

THE PREFLIST MUST CONTAIN TWO OR MORE FACILITIES WITH THE SAME SITE VALUE

© 2015 IBM Corporation

Health check XCF_CF_STR_PREFLIST revised to account for site preference

Most preferred CF depends on DUPLEX SITE preference:

- ANYSITE (specified or defaulted)
  - Primary in 1st CF in PREFLIST
  - Secondary in 2nd CF in PREFLIST
- SAMESITE or SAMESITEONLY
  - Primary in 1st CF with SITE specified
  - Secondary in next CF with same SITE
- CROSSSITE
  - Primary in 1st CF with SITE specified
  - Secondary in next CF with other SITE

New message IXCH0226I when DUPLEX SITE preference used

New message IXCH0227I if DUPLEX SITE preference not met

Messages IXCH0202I and IXCH0206E also affected

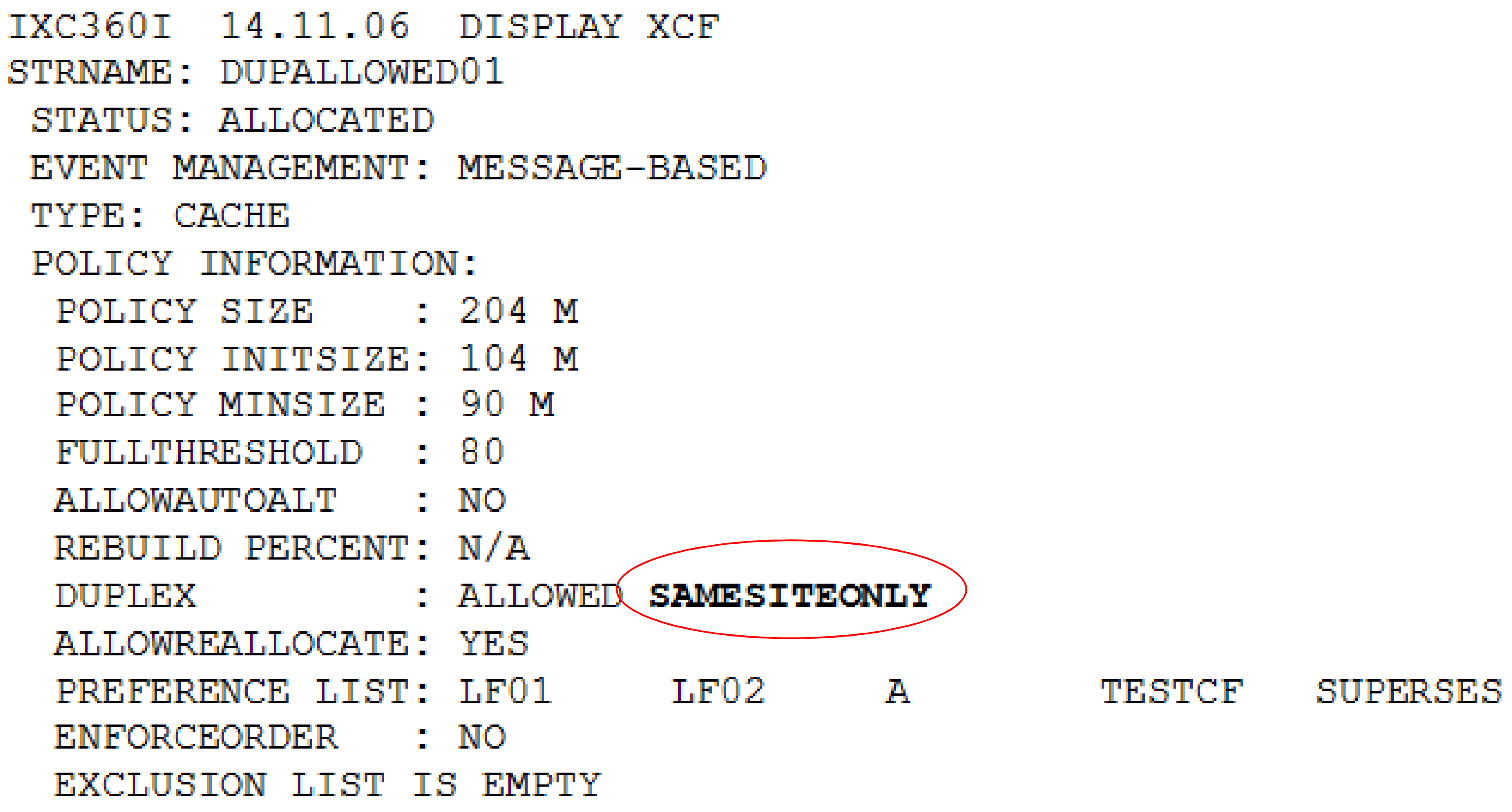
© 2015 IBM Corporation

Should get better adherence to PREFLIST for duplexed structures

- Allocation for duplexed structures was reworked in z/OS V2R2 to account for fact that certain attributes are “composite” when duplexed:
    - CFLEVEL - effective level is minimum of two
    - Volatility - satisfied as long as one is non-volatile
    - Failure isolation - satisfied by duplexed failure isolation
    - Exclusion list - generally satisfied by duplexing
- Previously, XCF would override PREFLIST order to meet attribute requirements as if structure would NOT be duplexed
    - Has generated concern when duplex instances swapped

© 2015 IBM Corporation

DISPLAY XCF,STR,STRNAME=x shows site preference

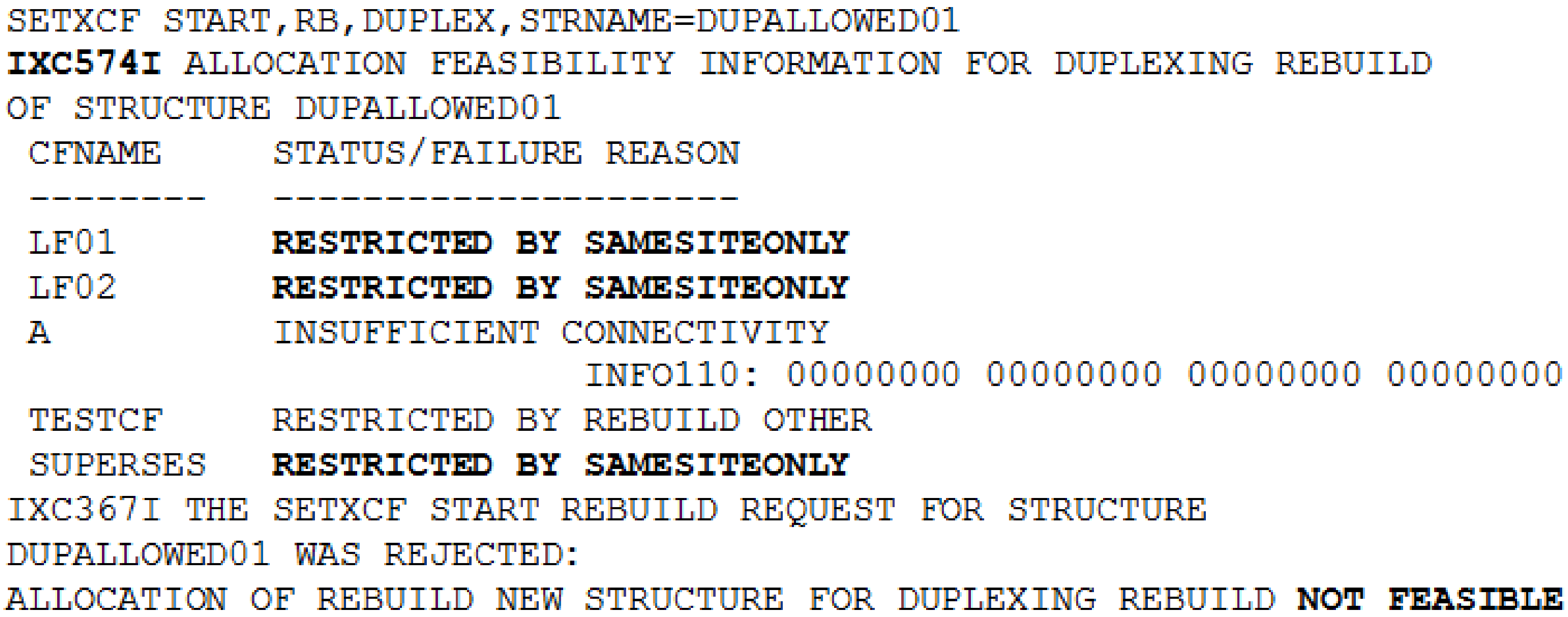


```
IXC360I  14.11.06  DISPLAY XCF
STRNAME: DUPALLOWED01
 STATUS: ALLOCATED
 EVENT MANAGEMENT: MESSAGE-BASED
 TYPE: CACHE
 POLICY INFORMATION:
  POLICY SIZE    : 204 M
  POLICY INITSIZE: 104 M
  POLICY MINSIZE : 90 M
  FULLTHRESHOLD  : 80
  ALLOWAUTOALT   : NO
  REBUILD PERCENT: N/A
  DUPLEX         : ALLOWED SAMESITEONLY
  ALLOWREALLOCATE: YES
  PREFERENCE LIST: LF01     LF02     A        TESTCF   SUPERSES
  ENFORCEORDER   : NO
  EXCLUSION LIST IS EMPTY
```

© 2015 IBM Corporation

SAMESITEONLY might exclude CF from consideration

Messages IXC347I, IXC574I, and IXL015I may indicate RESTRICTED BY SAMESITEONLY



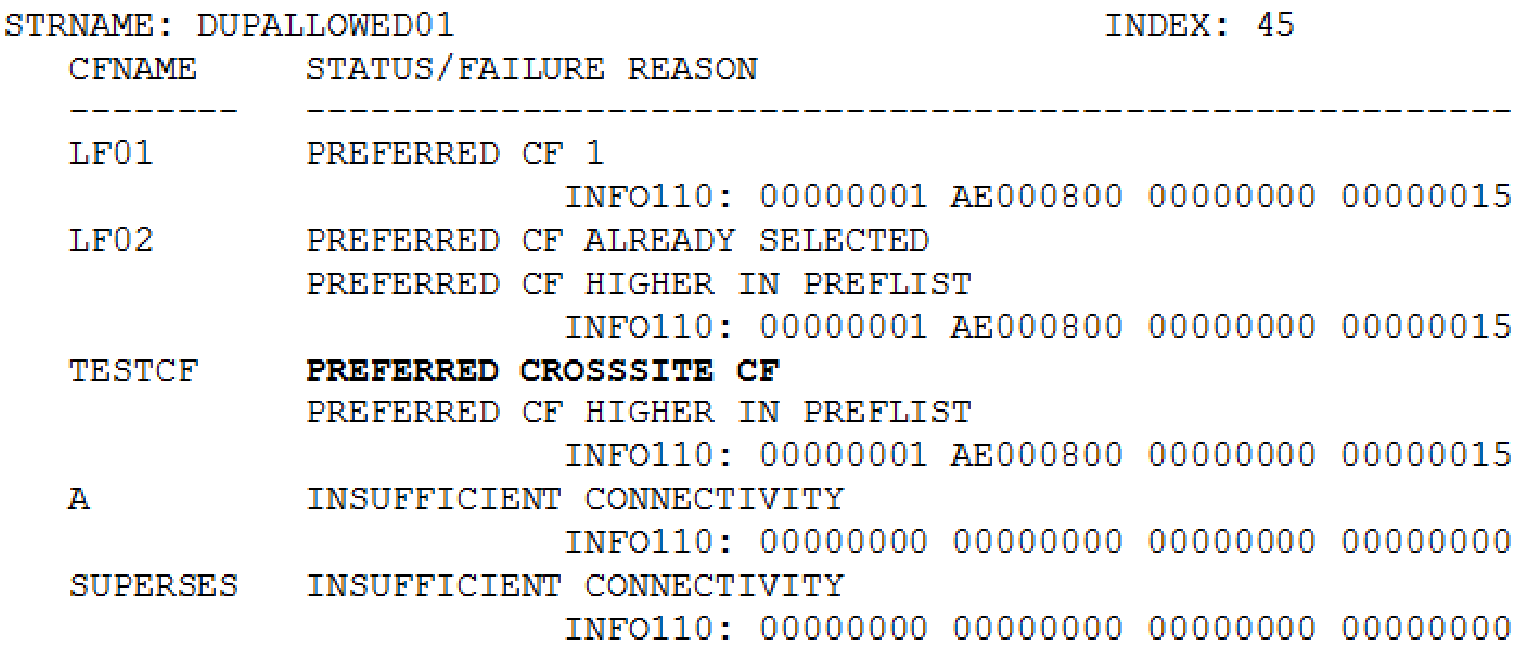
```
SETXCF START,RB,DUPLEX,STRNAME=DUPALLOWED01
IXC574I ALLOCATION FEASIBILITY INFORMATION FOR DUPLEXING REBUILD
OF STRUCTURE DUPALLOWED01
 CFNAME     STATUS/FAILURE REASON
 --------   ---------------------
 LF01       RESTRICTED BY SAMESITEONLY
 LF02       RESTRICTED BY SAMESITEONLY
 A          INSUFFICIENT CONNECTIVITY
                          INFO110: 00000000 00000000 00000000 00000000
 TESTCF     RESTRICTED BY REBUILD OTHER
 SUPERSES   RESTRICTED BY SAMESITEONLY
IXC367I THE SETXCF START REBUILD REQUEST FOR STRUCTURE
DUPALLOWED01 WAS REJECTED:
ALLOCATION OF REBUILD NEW STRUCTURE FOR DUPLEXING REBUILD NOT FEASIBLE
```

© 2015 IBM Corporation

REALLOCATE accounts for DUPLEX site preference

- Secondary CF tagged as one of:
  - CF 2
  - SAMESITEONLY CF
  - SAMESITE CF
  - CROSSSITE CF
- Affects messages IXC347I and IXC574I

The CF chosen for the secondary structure instance might not be same as the second choice for the primary structure instance.

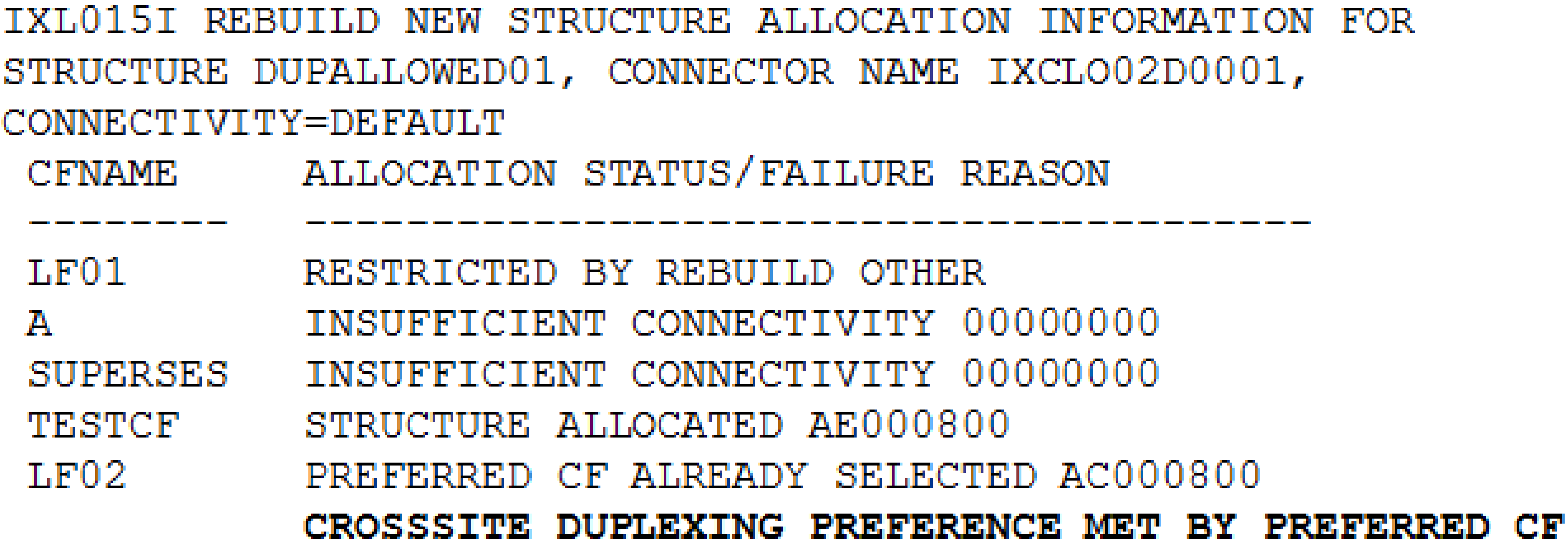


```
STRNAME: DUPALLOWED01                                    INDEX: 45
   CFNAME     STATUS/FAILURE REASON
   --------   ----------------------------------------------------
   LF01       PREFERRED CF 1
                          INFO110: 00000001 AE000800 00000000 00000015
   LF02       PREFERRED CF ALREADY SELECTED
              PREFERRED CF HIGHER IN PREFLIST
                          INFO110: 00000001 AE000800 00000000 00000015
   TESTCF     PREFERRED CROSSSITE CF
              PREFERRED CF HIGHER IN PREFLIST
                          INFO110: 00000001 AE000800 00000000 00000015
   A          INSUFFICIENT CONNECTIVITY
                          INFO110: 00000000 00000000 00000000 00000000
   SUPERSES   INSUFFICIENT CONNECTIVITY
                          INFO110: 00000000 00000000 00000000 00000000
```

© 2015 IBM Corporation

New reasons to choose one CF over another

Affects messages IXC347I, IXC574I, and IXL015I



```
IXL015I REBUILD NEW STRUCTURE ALLOCATION INFORMATION FOR
STRUCTURE DUPALLOWED01, CONNECTOR NAME IXCLO02D0001,
CONNECTIVITY=DEFAULT
 CFNAME     ALLOCATION STATUS/FAILURE REASON
 --------   ----------------------------------------
 LF01       RESTRICTED BY REBUILD OTHER
 A          INSUFFICIENT CONNECTIVITY 00000000
 SUPERSES   INSUFFICIENT CONNECTIVITY 00000000
 TESTCF     STRUCTURE ALLOCATED AE000800
 LF02       PREFERRED CF ALREADY SELECTED AC000800
            CROSSSITE DUPLEXING PREFERENCE MET BY PREFERRED CF
```

© 2015 IBM Corporation

Wait until all systems in sysplex are running z/OS V2R2

Do not specify site preferences in your CFRM policy until all systems in the sysplex are running z/OS V2R2

- In general, ANY system in the sysplex might perform CF structure allocation. Downlevel systems will only use PREFLIST.
- Downlevel systems will not DISPLAY new policy options
- Downlevel XCF_CF_STR_PREFLIST health check will not consider new policy options

© 2015 IBM Corporation

When all systems in sysplex are running z/OS V2R2

- Consider specifying site preference for duplexed structures as appropriate.
- Activate CFRM policy that specifies the new keywords
  - Run IXCMIAPU with updated policies
  - Use SETXCF START,POL command to start updated policy

    Note that site preference changes will NOT be “PENDING”
- Use DISPLAY XCF,REALLOCATE,TEST to check for changes
- Use SETXCF START,REALLOCATE to reconfigure structures

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

CF Gain Ownership protection

IXCQUERY for CDS and policies

© 2015 IBM Corporation

Health Based Routing

- Today, middle-ware servers perform self-assessments and report their “health” to WLM
- Through WLM, these health scores are available for use by others
  - For example, Sysplex Distributor uses this information (and more) to distribute incoming TCP connections within the sysplex to meet service goals (load balancing, availability, …)
- Problems
  - Self-assessment issues
  - “Storm Drain”
- Assessments by an external agent could help provide a more complete view of the health of the server …

© 2015 IBM Corporation

Health Based Routing in z/OS V2R2 incorporates XCF and XES health assessments

- Historically, XCF and XES have monitored exploiter use of their services and externalized “stall” conditions via various messages
  - IXC431I – member stalled
  - IXL040E, IXL041E – connector not responding
- The monitors are extended in z/OS V2R2 to report to WLM, as applicable, the XCF and XES view of health of an address space
  - XCF assesses health of the XCF group members in the address space
  - XES assesses health of the CF structure connectors in the address space
  - These assessments are independently reported to WLM
- WLM merges health scores from all reporting sources to determine an overall health score for an address space

© 2015 IBM Corporation

Caveats

- We assume there is a relationship between the ability of an address space to perform its intended function and its ability to process its XCF and XES work
- Seems plausible
  - Presumably would not bother to join an XCF group or connect to a structure if those services were not needed to accomplish its intended function
  - Therefore a failure to process XCF and XES work in a timely fashion likely suggests that the intended function is experiencing delay
- But we do not really know if XCF and XES services are part of the critical path for the address space’s intended function

© 2015 IBM Corporation

XCF and XES Address Space Health Monitoring

- XCF Monitoring
  - Only monitors health of address spaces that have members with “problems”
    - So there has to be a “problem” to trigger reporting of health of address space
    - Stops monitoring health of address space when all members in the space are deemed healthy
  - Periodically assesses health of each member in the address space
  - Reports to WLM the average health score of all members in address space
- XES Monitoring
  - Monitors health of every address space that has an active connector
  - Periodically assesses health of each connector in the address space
  - Reports to WLM the average health score of all connectors in address space

Health score is an integer value in range 0..100 representing
the degree (percentage) to which address space is deemed healthy

© 2015 IBM Corporation

XCF Member Health Scoring

- Member is unhealthy if:
  - Has stalled user exit (signal, group, status)
  - Has stalled work queue (signal, group)
  - User status exit confirms member is status update missing
  - Is message isolated
- Factors considered
  - Proportion of stalled exits relative to “good” exits
  - Time since recent activity
  - Depth of work queues
  - Duration of stall conditions

“stall” = 30 seconds

XCF reduces score more aggressively as depth of work queues increases

XCF may report “not healthy” to WLM long before the IXC431I “member stalled” message is issued (if ever)

© 2015 IBM Corporation

XES Connector Health Scoring

- Healthy if connector has responded to all outstanding events that require a response (or there are no such events)
- Unhealthy if fails to respond in timely fashion to an event that requires a response
  - XES will have issued IXL040E or IXL041E message
    - So for first two minutes, deemed to be healthy
  - XES computes a score for each outstanding event
    - Score for a particular event is percentage of CFSTRHANGTIME that remains for the event
  - Health score is minimum score for all outstanding events

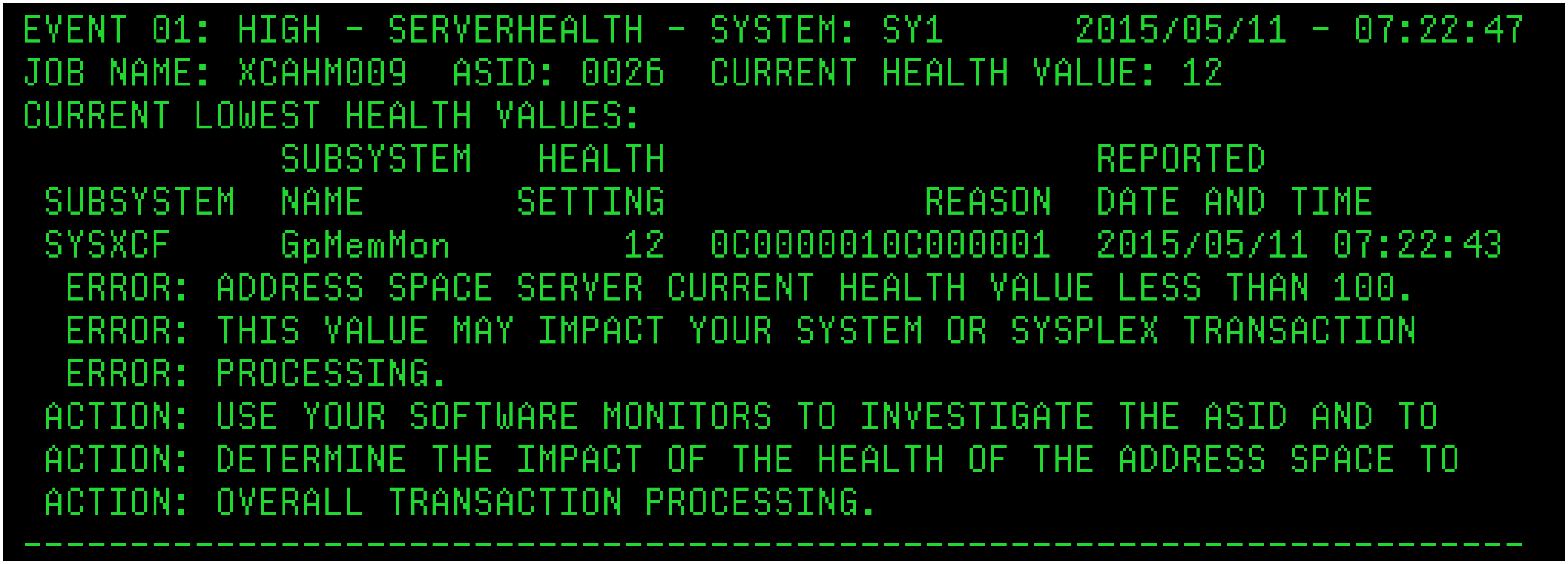
30 minutes
if “NO”

XES monitor runs under an SRB in the connector address space.
That SRB must get dispatched for XES to be able to report health score.

© 2015 IBM Corporation

Viewing Health Scores

- IWM4QHLT
  - WLM macro used to query health of address space(s)
- z/OS Runtime Diagnostics (RTD)
  - Lists any address space whose health score falls below 100%
  - F HZR,ANALYZE command initiates report



```
EVENT 01: HIGH - SERVERHEALTH - SYSTEM: SY1      2015/05/11 - 07:22:47
JOB NAME: XCAHM009  ASID: 0026  CURRENT HEALTH VALUE: 12
CURRENT LOWEST HEALTH VALUES:
            SUBSYSTEM   HEALTH                    REPORTED
 SUBSYSTEM  NAME       SETTING           REASON  DATE AND TIME
 SYSXCF     GpMemMon        12  0C0000010C000001  2015/05/11 07:22:43
  ERROR: ADDRESS SPACE SERVER CURRENT HEALTH VALUE LESS THAN 100.
  ERROR: THIS VALUE MAY IMPACT YOUR SYSTEM OR SYSPLEX TRANSACTION
  ERROR: PROCESSING.
 ACTION: USE YOUR SOFTWARE MONITORS TO INVESTIGATE THE ASID AND TO
 ACTION: DETERMINE THE IMPACT OF THE HEALTH OF THE ADDRESS SPACE TO
 ACTION: OVERALL TRANSACTION PROCESSING.
----------------------------------------------------------------------
```

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

CF Gain Ownership protection

IXCQUERY for CDS and policies

© 2015 IBM Corporation

Simultaneous Multithreading (SMT) Exploitation

- Available on z13
- SMT is designed to improve both core capacity and single thread performance
- When in SMT mode, up to two execution threads can be active per core
- Threads can dynamically share the caches, TLBs and execution resources of each IFL and zIIP core.
- PR/SM dispatches cores, z/OS can dispatch threads

- Implications for synchronous CF requests …

© 2015 IBM Corporation

SMT and Synchronous CF Requests

- Advantageous to allow the core to process other work while waiting for a synchronous CF request to complete
- Synchronous CF request processing:
  - Send request to CF
  - Test for request completion, looping until done
- Revised completion loop so that when running on z13, the core can process another thread (when operating in SMT mode)
- However:
  - z/OS SMT exploitation restricted to zIIP
  - Relatively small amount of workload issues CF requests while running on zIIP

© 2015 IBM Corporation

Sync/Async Heuristics

- XES Heuristic Algorithm
  - Measures synchronous service times for each kind of CF operation
  - Determines whether the opportunity cost of running the operation synchronously exceeds the overhead of running the request asynchronously
  - If so, the request will be processed asynchronously
- Helps optimize use of CPU resources so as to maximize the amount of work performed
- With SMT, the opportunity costs are different …

© 2015 IBM Corporation

SMT and Sync/Async Heuristics

- Synchronous service time more or less independent of SMT mode
  - Transmit time for command and results, plus time to process command in CF
- But the opportunity cost of a synchronous CF request depends on SMT mode
  - Single thread: Cost is equivalent to (and expressed as) CPU time
  - Multithread: Cost is something less than elapsed time since the core can perform other work while waiting for CF request to complete
- Heuristic decision must account for the lower cost of synchronous operation when the core is running in multithread mode
- Requests more likely to run synchronous since the opportunity cost is less

© 2015 IBM Corporation

Requirements for z/OS exploitation of SMT when processing CF request

- z13 configured with zIIP
- Software
  - z/OS V2R2
  - z/OS V2R1 with PTFs
    - OA44439 – XES
    - OA43366 – Supervisor
    - OA43622 – WLM/SRM
    - OA44101 – RMF
    - OA44624 – Unix System Services
- Need workload that:
  - Runs on zIIP
  - Issues synchronous CF requests

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

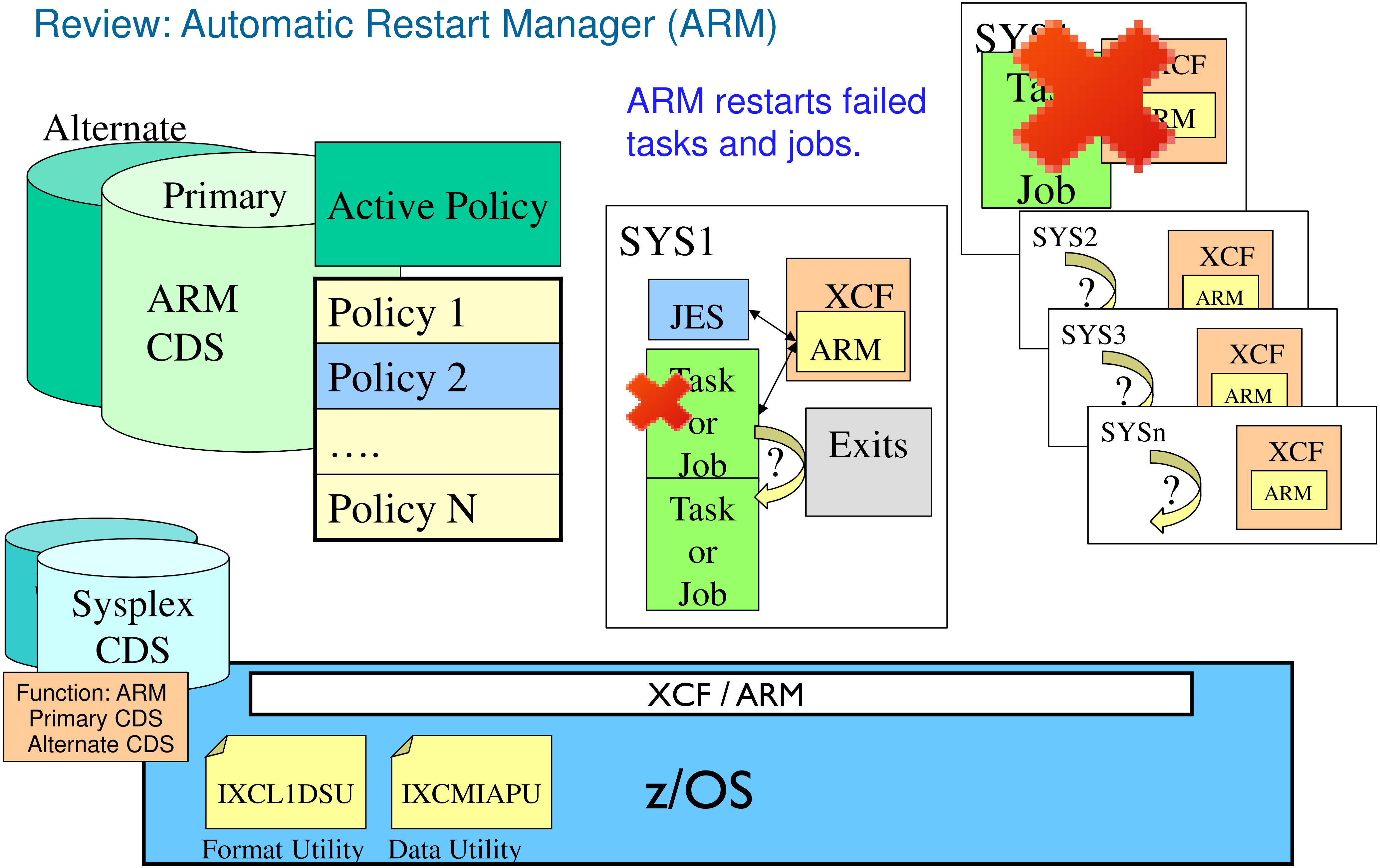
CF Gain Ownership protection

IXCQUERY for CDS and policies

© 2015 IBM Corporation

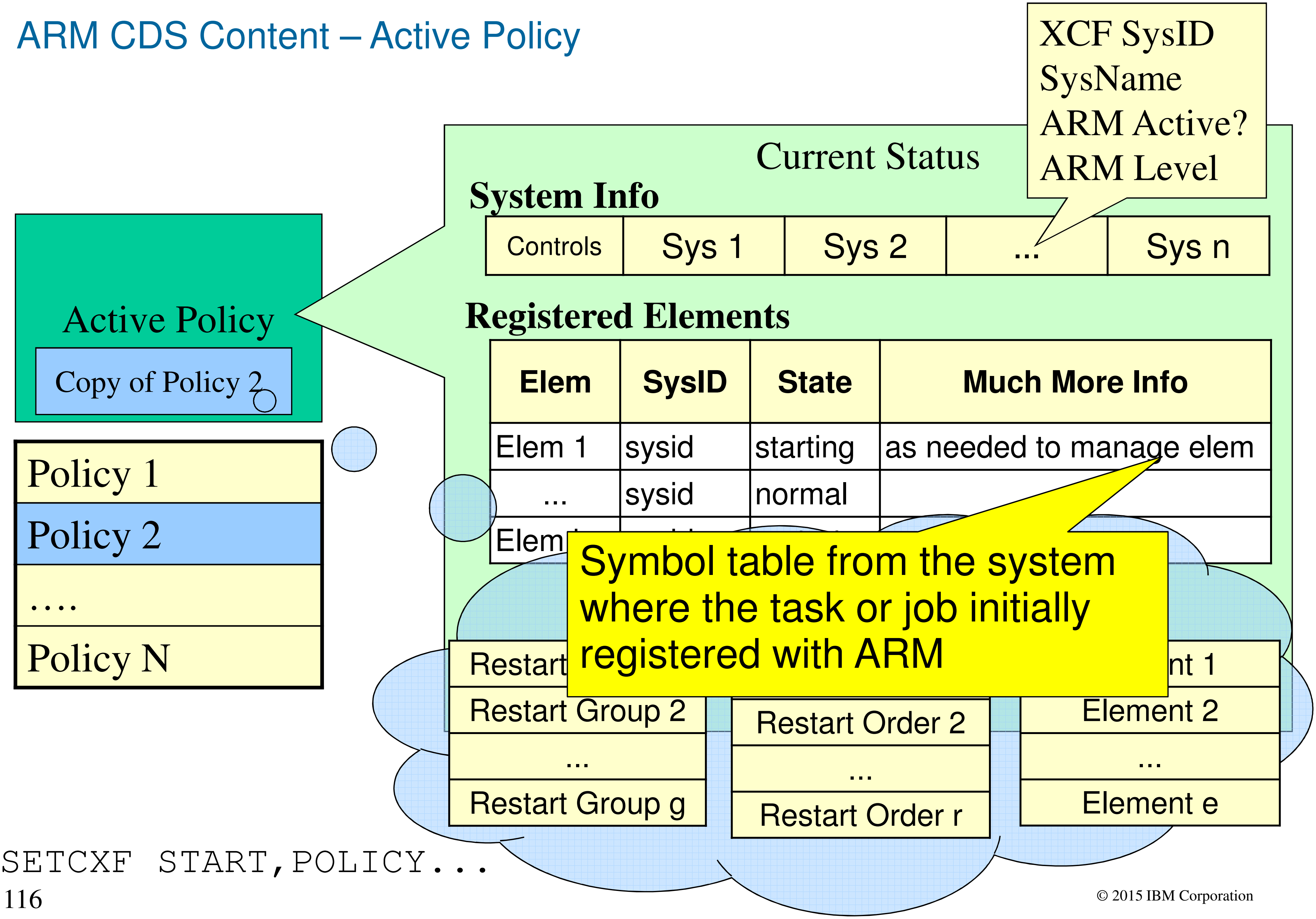
Review: Automatic Restart Manager (ARM)

ARM restarts failed tasks and jobs.



© 2015 IBM Corporation

ARM CDS Content – Active Policy



System Info

Controls	Sys 1	Sys 2	...	Sys n

Registered Elements

Elem	SysID	State	Much More Info
Elem 1	sysid	starting	as needed to manage elem
...	sysid	normal	
Elem			

Symbol table from the system where the task or job initially registered with ARM

```
SETCXF START,POLICY...
```

© 2015 IBM Corporation

ARM and System Symbols

- z/OS V2R2 supports system symbols with longer names and values
- Automatic Restart Manager (ARM) is impacted
  - ARM stores system symbol tables in the ARM Couple Data Set (CDS)
  - The CDS needs larger records to accommodate the longer symbols
- If your existing sysplex has an ARM CDS, then before you IPL the very first z/OS V2R2 system into the sysplex, you should:
  - Format new ARM CDS's
  - Bring them into service
- If you fail to do so, the z/OS V2R2 system will not be ARM capable and will not be able to participate in automatic restart management

© 2015 IBM Corporation

z/OS V2R2 Migration Action for ARM

The new ARM CDS is fully compatible with systems running older z/OS releases

Reformat your primary and alternate ARM couple data sets to the new level using your current (old) system, following these steps:

1. Use a STEPLIB DD statement to access your z/OS V2R2 system's SYS1.MIGLIB data set.
2. Use the z/OS V2R2 IXCL1DSU utility located in your z/OS V2R2's SYS1.MIGLIB to format new ARM couple data sets with support for the z/OS V2R2 system symbol table
   - Primary, alternate, etc. as needed
3. Issue the SETXCF COUPLE,ACOUPLE system command to enable the new couple data set as the alternate couple data set, and then use the SETXCF COUPLE,PSWITCH system command to switch the alternate couple data set to the primary couple data set.
4. Issue SETXCF COUPLE,ACOUPLE system command to enable the sysplex to use the second newly formatted ARM CDS as the alternate CDS
5. Update your COUPLExx parmlib member to point to the new ARM couple data sets so that they will be used during the IPL.

© 2015 IBM Corporation

APAR OA46977

- New function APAR enables DISPLAY XCF,COUPLE,TYPE=ARM to indicate whether the ARM CDS supports the z/OS V2R2 symbol tables
- Output message IXC358I has insert to describe which symbol table is supported by the ARM CDS:

  HBB5520 SYMBOL TABLE SUPPORT
  HBB7707 SYMBOL TABLE SUPPORT
  HBB77A0 SYMBOL TABLE SUPPORT ← New
- Likely helpful when making the migration
- Also means display output produced by down-level systems will be consistent with output of z/OS V2R2 systems

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

CF Gain Ownership protection

IXCQUERY for CDS and policies

© 2015 IBM Corporation

CF Gain Ownership

- Problem Statement / Need Addressed
  - Configuration errors can cause multiple sysplexes to attempt to use the same coupling facility (CF).
  - The operator must decide whether a sysplex should use the CF or not – with little provided information.
  - Incorrectly deciding to use the CF can cause a sysplex outage.
- Solution
  - Provide new COUPLExx CFRMTAKEOVERCF keyword to control whether the operator is prompted or XCF rejects use of a CF that may be in use by another sysplex. New default is NO.
- Benefit / Value
  - Avoid operator errors by forcing the installation to reactivate a CF in order to pass it from one sysplex to another

© 2015 IBM Corporation

Background: CF Authorities

- CF authority (sysplex name and TOD) storage
  - Authority stored in the CF before used by a sysplex
  - XCF also stores same authority in CFRM couple data set (CDS)
  - Authority in CF cleared when removing CF from sysplex (if possible)
  - Authority in CF (non-zero authority) generally indicates CF is in use
- CF authorities in CFRM CDS
  - When sysplex is initialized authorities in CFRM CDS are **old**
  - While running, authorities stored in CFRM CDS are **current**
- Gain ownership (storing a new authority in a CF)
  - Only happens when the CF is in the CFRM active policy
  - No questions asked if CF has no authority (zero authority)
  - No questions asked if CF authority matches a **current** authority in the CFRM CDS (for example total connectivity loss/re-gain)

© 2015 IBM Corporation

CFRMOWNEDPROMPT(YES)

- When sysplex is re-IPLed, scrub the sysplex name portion of all the authority values found in the CFRM policy
- Later when looking to gain ownership of a CF:
  - If CF authority is zero, take ownership of the CF
  - If CF authority is nonzero, and:
    - Matches CFRM policy, take ownership of CF
    - Matches CFRM policy except for the zero sysplex name, prompt the operator
    - Differs from CFRM policy then:
      - Prompt operator if CFRMTAKEOVERCF is PROMPT
      - Reject use of CF if CFRMTAKEOVERCF is NO

© 2015 IBM Corporation

CFRMOWNEDPROMPT(NO)

- When sysplex is re-IPLed, the authority values found in the CFRM policy are not changed
- Later when looking to gain ownership of a CF:
  - If CF authority is zero, take ownership of the CF
  - If CF authority is nonzero, and:
    - Matches CFRM policy, take ownership of CF
    - Differs from CFRM policy then:
      - Prompt operator if CFRMTAKEOVERCF is PROMPT
      - Reject use of CF if CFRMTAKEOVERCF is NO

© 2015 IBM Corporation

Gaining Ownership of a CF (storing new authority value in CF)

- If CF authority matches the **old** authority in the CFRM CDS (for example, sysplex-wide IPL with same CFRM CDS), depends on COUPLExx
  - CFRMOWNEDCFPROMPT(NO) – default
    - Sysplex will gain ownership
  - CFRMOWNEDCFPROMPT(YES)
    - Operator will be prompted and need to decide
- Other CF authority mismatches (for example, sysplex-wide IPL with a new CFRM CDS)
  - CFRMTAKEOVERCF(NO) – new default
    - XCF will reject use of the CF
  - CFRMTAKEOVERCF(PROMPT) – old behavior
    - Operator will be prompted and need to decide

© 2015 IBM Corporation

Authority Mismatches

- Why would an **old** authority in the CFRM CDS match the CF?
  - Sysplex-wide IPL
    - Fine to use CF
  - IPL another sysplex (different sysplex CDS) using a copy (from mirroring or otherwise) of the CFRM CDS
    - Bad if sysplex using current copy of CFRM CDS still using CF
- Why would an authority in the CFRM CDS NOT match the CF at all?
  - Sysplex-wide IPL with a different CFRM CDS
    - Fine to use CF
  - IPL another sysplex (different sysplex CDS) using an old copy (from mirroring or otherwise) of the CFRM CDS
    - Bad if sysplex using current copy of CFRM CDS still using CF
  - IPL another sysplex (different sysplex CDS) with a CFRM policy that has a CF that does not belong to it
    - Bad

© 2015 IBM Corporation

Usage & Invocation

- SYS1.PARMLIB(COUPLExx)

```
COUPLE
      CFRMTAKEOVERCF(NO)
      …
```

- Statements for the “safest” configuration

```
CFRMOWNEDCFPROMPT(YES)
CFRMTAKEOVERCF(NO)               z/OS V2R2 new default behavior
```

- Statements with most automatic gain ownership of CF (susceptible to more configuration/operator errors)

```
CFRMOWNEDCFPROMPT(NO)       default
CFRMTAKEOVERCF(PROMPT)      pre-z/OS V2R2 behavior
```

© 2015 IBM Corporation

Message changes

- Use of CF may be rejected for CFRMTAKEOVERCF(NO)

```
IXC518I SYSTEM S1 NOT USING
           COUPLING FACILITY SIMDEV.IBM.EN.ND0100000000
                             PARTITION: 00      CPCID: 00
                             LP NAME: N/A       CPC NAME: TINK6
           NAMED LF01
           AUTHORITY DATA: PLEX1 01/30/2015 08:08:55.195784
          REASON: TAKEOVER PROHIBITED.
          REASON FLAG: 13340009.
```

- Messages IXC500I, IXC517I, and IXC518I enhanced with additional information when available
  - Partition and CPC name
  - Additional authority data TOD resolution
    - Authority data is new for IXC517I and IXC518I

© 2015 IBM Corporation

Message changes …

- Additional authority data in messages IXC500I, IXC518I, and IXC362I (result of DISPLAY XCF,CF,CFNAME=) when applicable

```
IXC362I  18.13.16  DISPLAY XCF
   CFNAME: LF02
     COUPLING FACILITY     :   SIMDEV.IBM.EN.ND0200000000
                               PARTITION: 00   CPCID: 00
     SITE                  :   SITE1
     POLICY DUMP SPACE SIZE:   1 M
     ACTUAL DUMP SPACE SIZE:   1 M
     STORAGE INCREMENT SIZE:   1 M

     AUTHORITY DATA        :   PLEX1 01/30/2015 18:06:13.281887
     CFRM AUTHORITY        :   PLEX1 01/30/2015 17:53:35.128473
```

- AUTHORITY DATA is from the CF
- CFRM AUTHORITY is from the CFRM CDS (not provided if none stored)

Message changes …

- Enhanced explanations in IXL150I (result of DISPLAY CF)

```
IXL150I  18.18.51  DISPLAY CF
   COUPLING FACILITY SIMDEV.IBM.EN.ND0100000000
                     PARTITION: 00  CPCID: 00
                     LP NAME: N/A   CPC NAME: TINK6
                     CONTROL UNIT ID: 0001
   NAMED LF01

    NOT IN USE BY SYSTEM
```

- NOT IN USE BY SYSTEM (Hint: see IXC518I)
- NOT CONNECTED TO SYSTEM (Hint: check paths)
- NOT IN THE CFRM ACTIVE POLICY (Hint: check CFRM policy)

Migration & Coexistence Considerations

- Migration action
  - Update COUPLExx to specify **CFRMTAKEOVERCF(PROMPT)** if new CFRMTAKEOVERCF(NO) behavior is not desired
- Coexistence consideration when using **CFRMOWNEDCFPROMPT(YES)**
  - A downlevel system clears CF authorities in the CFRM CDS when initializing the sysplex (sysplex-wide IPL)
  - This may cause a z/OS V2R2 system using CFRMTAKEOVERCF(NO) to reject use of a CF when the desired behavior might have been to PROMPT
  - To avoid the strange behavior, do one of the following
    - Create a new COUPLExx for z/OS V2R2 with CFRMTAKEOVERCF(PROMPT)
    - Update COUPLExx to use CFRMOWNEDCFPROMPT(NO)
- No toleration/coexistence APARs/PTFs

© 2015 IBM Corporation

Installation

- Nothing needed if new CFRMTAKEOVERCF(NO) behavior is acceptable
- Update COUPLExx PARMLIB member to specify CFRMTAKEOVERCF(PROMPT) if new CFRMTAKEOVERCF(NO) behavior is not desired
- May still need a copy of the COUPLExx without the CFRMTAKEOVERCF keyword for use by downlevel systems that do not support the keyword

© 2015 IBM Corporation

Health Check Updated

- XCF_CF_CONNECTIVITY health check
  - New message IXCH0459E for specific case where CF is connected and in the CFRM active policy, but not in use by the sysplex
    - Hint: See IXC518I

© 2015 IBM Corporation

z/OS V2R2 Summary

XCF message isolation

CFRM policy site preferences

Health based routing

SMT exploitation

Need new ARM CDS

CF Gain Ownership protection

IXCQUERY for CDS and policies

© 2015 IBM Corporation

IXCQUERY Support for CDS and Policies

- Problem Statement / Need Addressed / User Stories
  - Automation products used to manage System and Sysplex resources must use DISPLAY XCF commands and/or submit policy batch jobs along with parsing/screen scraping the output to obtain Couple Dataset and Policy information
- Solution
  - Provide an authorized programming interface through the IXCQUERY service to return Couple Dataset and Policy information
- Benefit / Value
  - The IXCQUERY interface is both more efficient and simpler

© 2015 IBM Corporation

z/OS V2R1 - Summary

APAR OA41203	Spikes in CF structure requests
Serial Rebuild	**Better MTTR**
Thin Interrupts	**Better async service time**
Sync/Async Thresholds	Response time vs CPU
XCF Note Pads	Simpler API for List Structures
D XCF STR,STATUS	New filters
XCF Signal Throughput	100K/sec
Couple Data Set Accessibility Verification	Avoid outages: OA38311
Cache Vector Corruption Detection	Avoid data corruption: OA42519

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

© 2015 IBM Corporation

APAR OA41203

- New function APAR
  - Available since January 2014
  - We recently marked it HIPER
- Optionally provides an alert when the system has a significant number of delayed requests targeted to a given CF Structure
  - Controlled by new XCF FUNCTIONS switch CFSTRQMON
    - DISABLED by default
  - IXL053E “requests for cfname are delayed”
  - IXL054I “requests for cfname not delayed”
  - IXL055I “requests for strname are delayed” (along with some context)
  - IXL056I “requests for strname not delayed”

© 2015 IBM Corporation

APAR OA41203 ...

- Also reworked XES selection algorithm for queued CF requests
  - In effect as soon as the APAR is installed
  - Independent of the the CFSTRQMON switch
- Old: FIFO
  - A spike in requests for one structure induces delay for unrelated structures
  - Delay is a function of the number of predecessor requests
- New: Fair Queue
  - For a given structure, FIFO
  - But round robin among the structures
  - Delay is a function of the number of structures with queued requests

© 2015 IBM Corporation

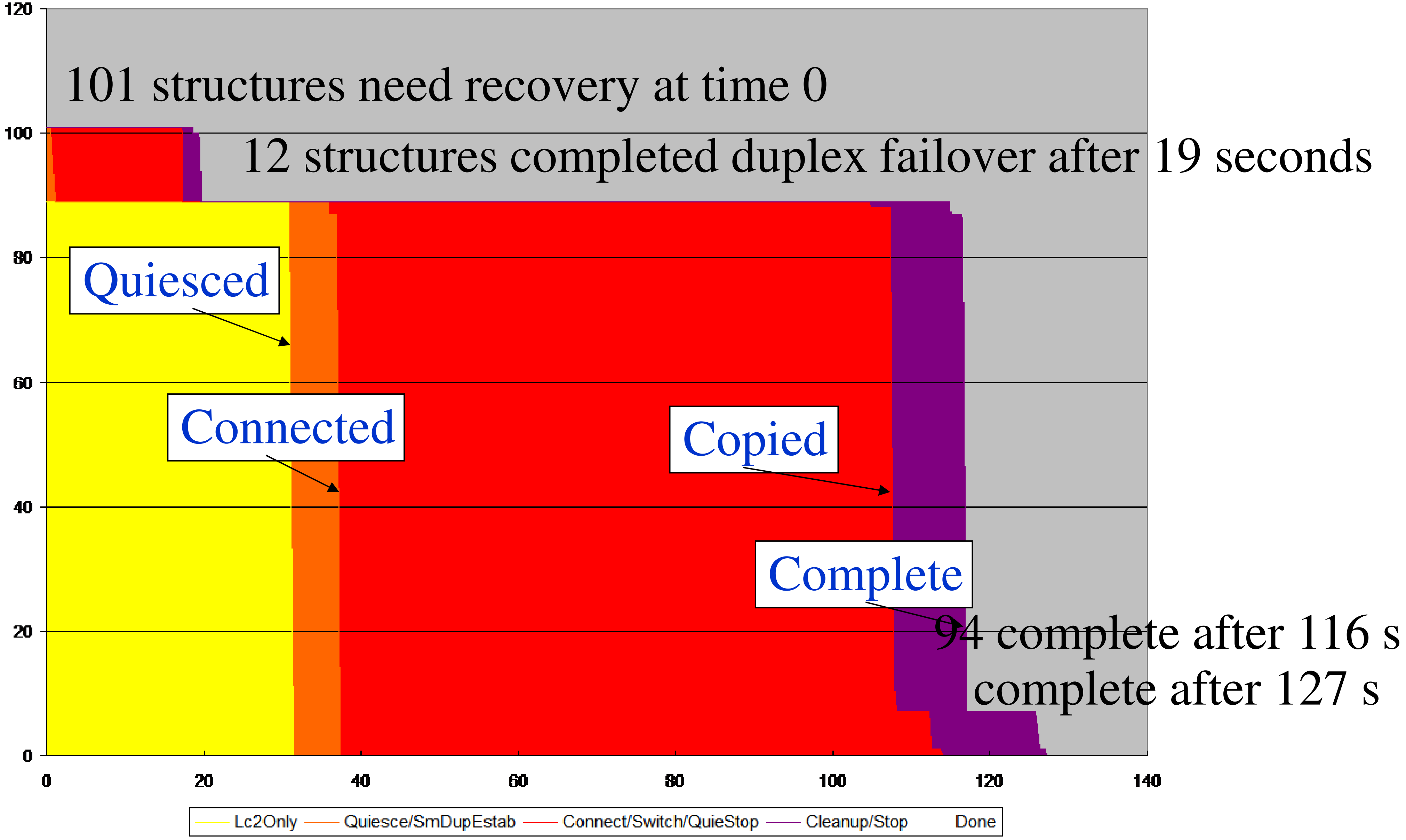
z/OS V2R1 – Serial Rebuild

- When a system loses connectivity to a coupling facility the sysplex tries to recover the structures by:
  - Failing over to the accessible copy of a duplexed structure
  - Rebuilding the structure in some other coupling facility
- During the recovery, the structure is unavailable for use by the workload, so applications tend to hang for the duration
- A coupling facility could have dozens, perhaps hundreds of structures to recover
- Today, there is one massive burst of recovery processing launched in parallel for all of the impacted structures …

© 2015 IBM Corporation

Parallel Rebuild - Test Results
Policy Based Event Processing

Not formal performance test measurements



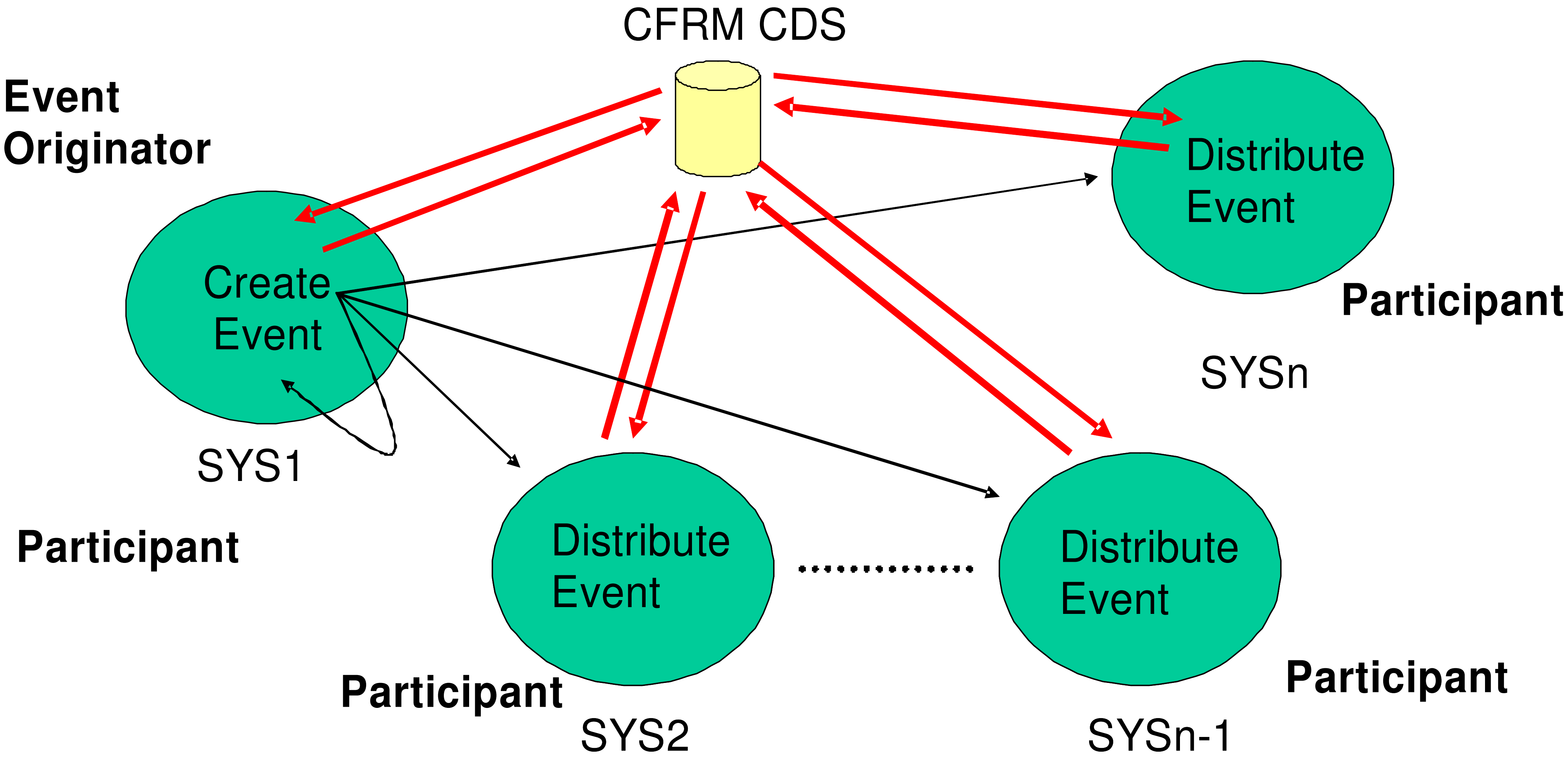
© 2015 IBM Corporation

Policy Based Event Processing

→ CDS I/O

→ GAT

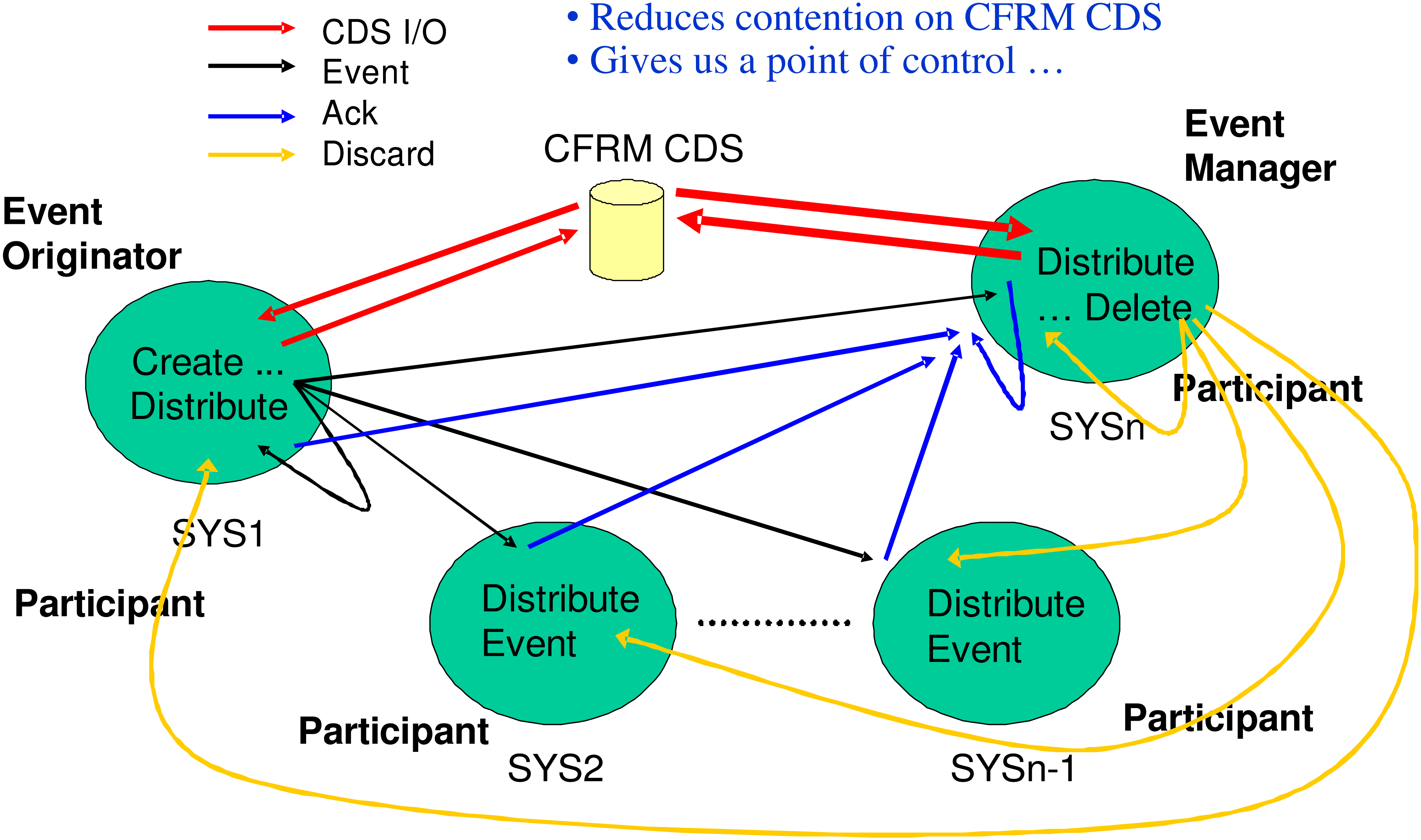
- Lots of contention on CFRM CDS
- Systems work independently



© 2015 IBM Corporation

Message Based Event Processing

- Reduces contention on CFRM CDS
- Gives us a point of control …



© 2015 IBM Corporation

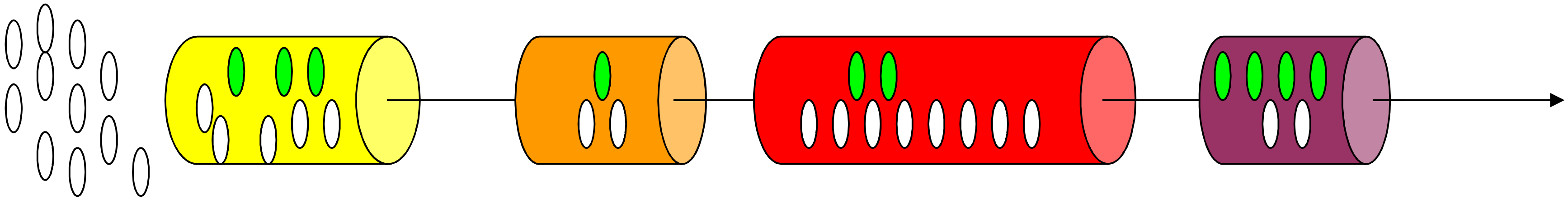
Serial Rebuild Improves Availability by Reducing MTTR

- Recovery is actually faster when done with less parallelism since there is less resource contention in several areas:
  - CFRM CDS
  - Coupling Facility
  - Participating systems
- Faster recovery means
  - Structures will be inaccessible for shorter periods, so
  - Applications are down for shorter periods
- Finishing recovery of the “important” structures sooner can reduce the business impact of the failure

© 2015 IBM Corporation

Serial Rebuild – New Behavior

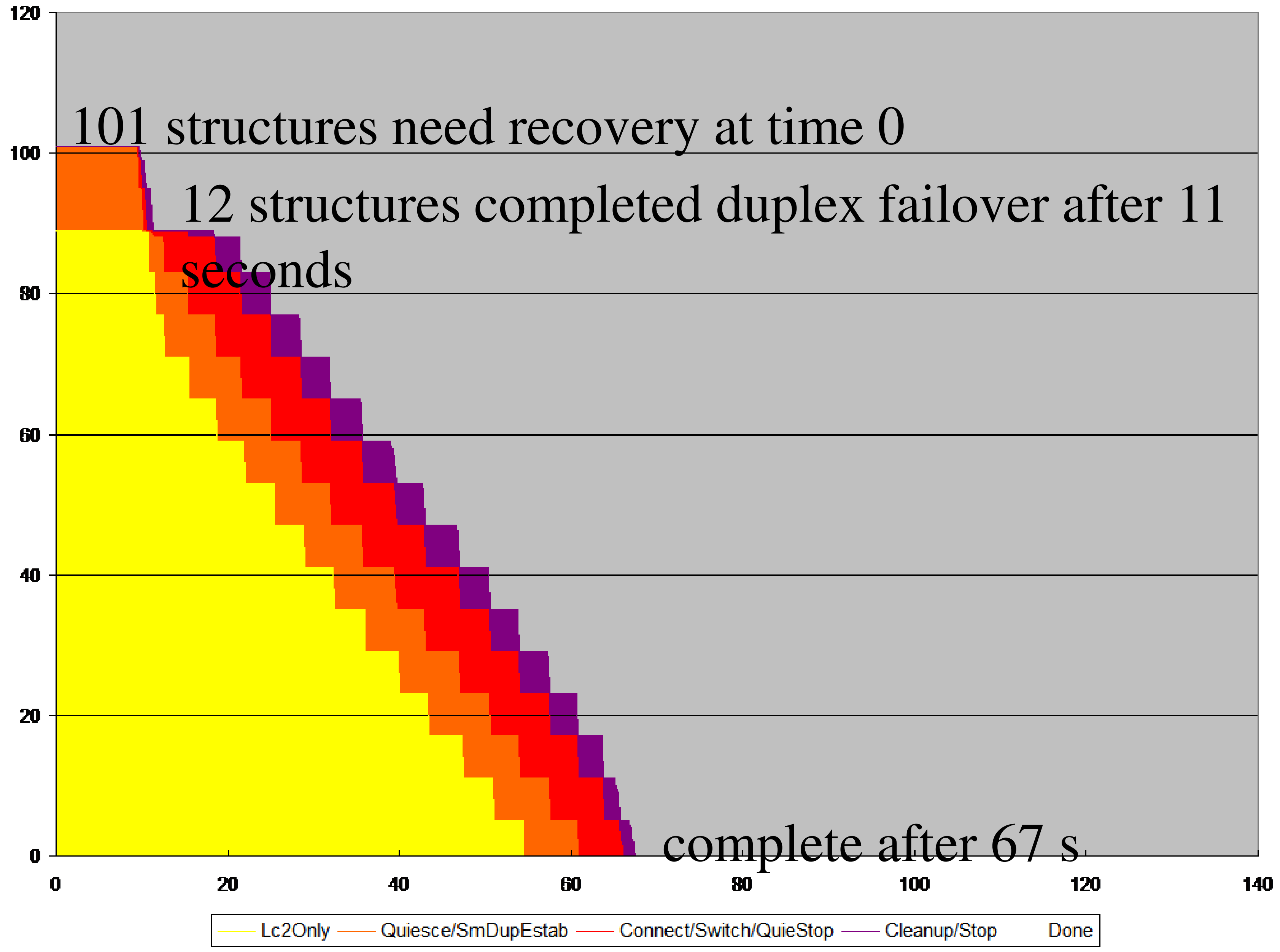
- Issue message IXC568I “starting recovery”
- Sort structures according to recovery criteria
- Prime the pipe by initiating work to do rebuild or duplex failover for a batch of highest priority structures
- Do until done:
  - As work higher in the pipe completes a phase, push finite amount of completed work lower in the pipe ahead to the next phase
  - If no progress is being made in a phase, pull in finite amount of work from a lower phase
  - Use priority to determine what work to move to next phase
- Issue message IXC568I “finished recovery”



© 2015 IBM Corporation

Serial Rebuild - Test Results

Not formal performance test measurements



© 2015 IBM Corporation

Serial Rebuild – Influencing Priority of Rebuild when LossConn

- You have some input as to what order structures will be selected for processing during LossConn Recovery
  - Presumably this would relate to the order in which you want your business applications to be restored to service
- Optional RECPRTY(value) specification in CFRM Policy
  - Decimal value in the range 1 to 4, default is 3
  - Low numbers imply rebuild sooner, high numbers later
  - Takes effect immediately when policy activated
- Order is determined by:
  - RECPRTY specification
  - “Distance” from completion
  - Lock structures
  - Other structures

© 2015 IBM Corporation

Serial Rebuild – Concerns About Rebuild Order

- There may be unknown dependencies or relationships between the structures
  - The rebuild of one structure might not be able to complete until the rebuild of a second structure has completed if …
  - The rebuild process for the first structure calls a service that needs access to the second structure
- What if the rebuild priorities are reversed?
- Serial Rebuild will not deadlock
  - Pulls in more work if not enough progress being made
  - So all structures will eventually complete
  - But “eventually” might be longer than necessary

© 2015 IBM Corporation

Serial Rebuild – Exploitation

- All systems in the sysplex must be z/OS V2R1 or later
- Primary CFRM CDS must be formatted for MSGBASED
- The new XCF CFLCRMGMT switch must be ENABLED on all systems in the sysplex
  - COUPLExx PARMLIB member at IPL, or
  - Dynamically using SETXCF
- Consider use of RECPRTY keyword in CFRM policy
  - Looks like ISGLOCK should be a “1”

© 2015 IBM Corporation

Another Undesirable Old Behavior

- Scenario:
  - –CF is reset. Looks like LossConn to all the systems.
  - –Duplex structures fail over and are available in simplex mode
  - –Remaining structures in massive wave of rebuilds
  - –CF reboots and connectivity is restored
  - –CFRM immediately tries to re-duplex the structures
  - –Re-duplexing effort bogged down in massive wave of rebuilds
- Well this is rather annoying
  - –Duplex structures quickly restored to service after initial failure
  - –And now unavailable for duration of the non-duplex rebuilds

© 2015 IBM Corporation

Sysplex-wide CFRM Processing is Now Prioritized

- Most important
  - LossConn Recovery (disconnect, rebuild, or duplex failover)
  - REALLOCATE/POPULATECF structure evaluation/action
  - Policy-initiated STOP CF structure duplexing for policy change
  - Policy-initiated START CF structure duplexing for DUPLEX(ENABLED)
- Least important

- Work of lesser importance is deferred if there is more important work …

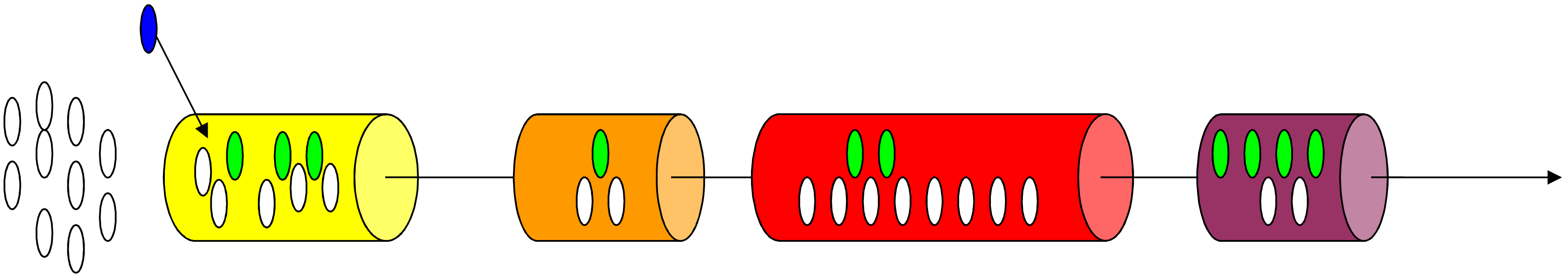
New Behavior – Serial Duplexing

- Re-duplexing effort is deferred until after the LossConn recovery is completed
  - The duplex structures are quickly restored to service as they fail over to simplex mode as part of the LossConn recovery
  - Let the other structures complete their recovery before launching a new duplexing effort
- When launched, CFRM will re-duplex the structures:
  - Sequentially, one at a time
  - In a system determined order
    - Recovery priority is not used

© 2015 IBM Corporation

What About Other Rebuild Requests

- During LossConn Recovery, the rebuild processing is being carefully managed
- An externally initiated rebuild request could arrive
  - SETXCF
  - Application initiated
- The new rebuild request is immediately initiated, but is otherwise managed along with all the others



© 2015 IBM Corporation

Recent results from more customer like test environment

- Roughly 180 structures
- “Bouncing the CF”
  - Restored to service about 2 minutes into the test

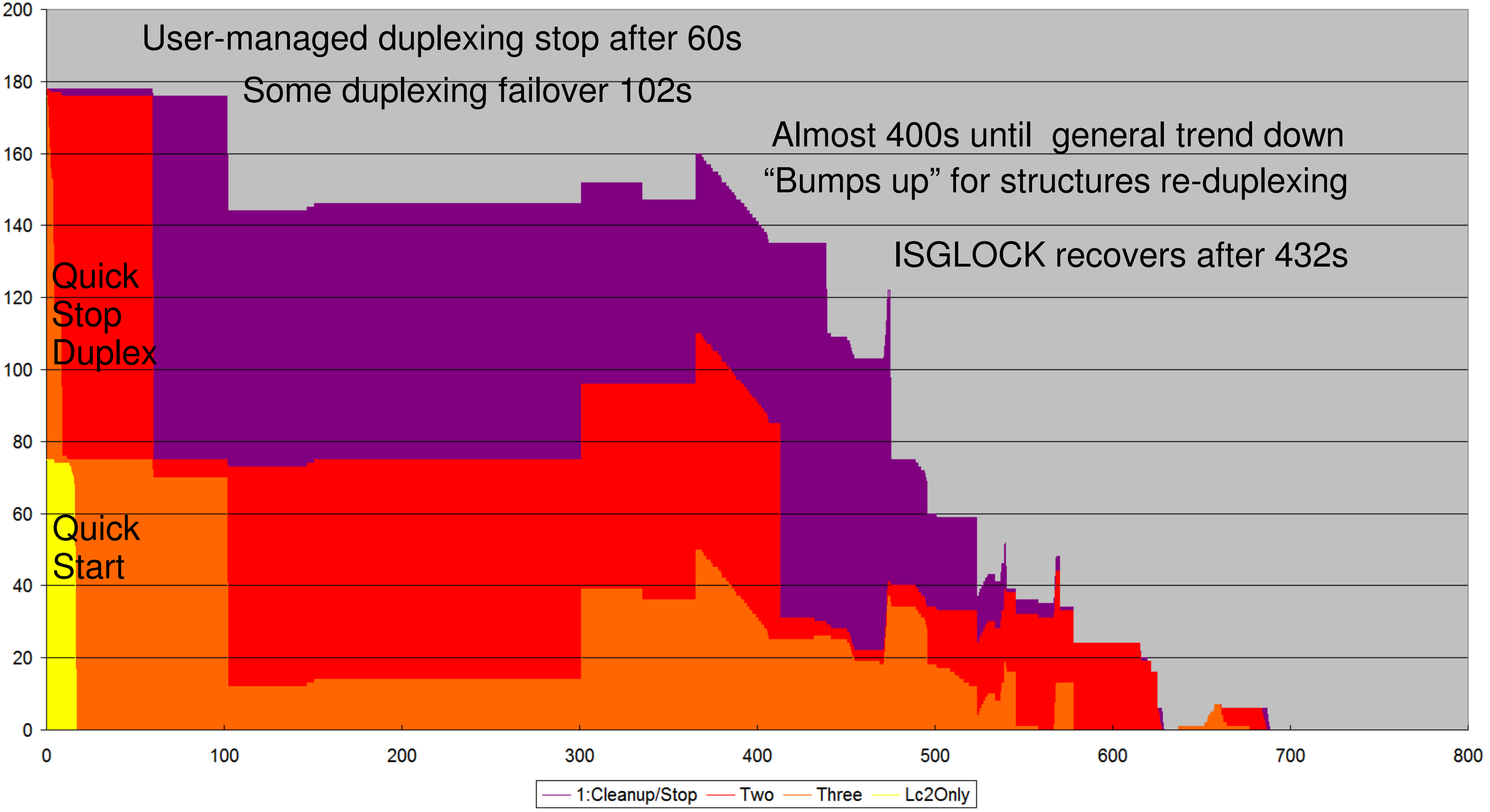
- Still in the investigation phase
- Working to get repeatable results
  - Eliminating “interference”
  - Such as launching of health checks

- But looks promising ...

© 2015 IBM Corporation

CFLCRMGMT DISABLED

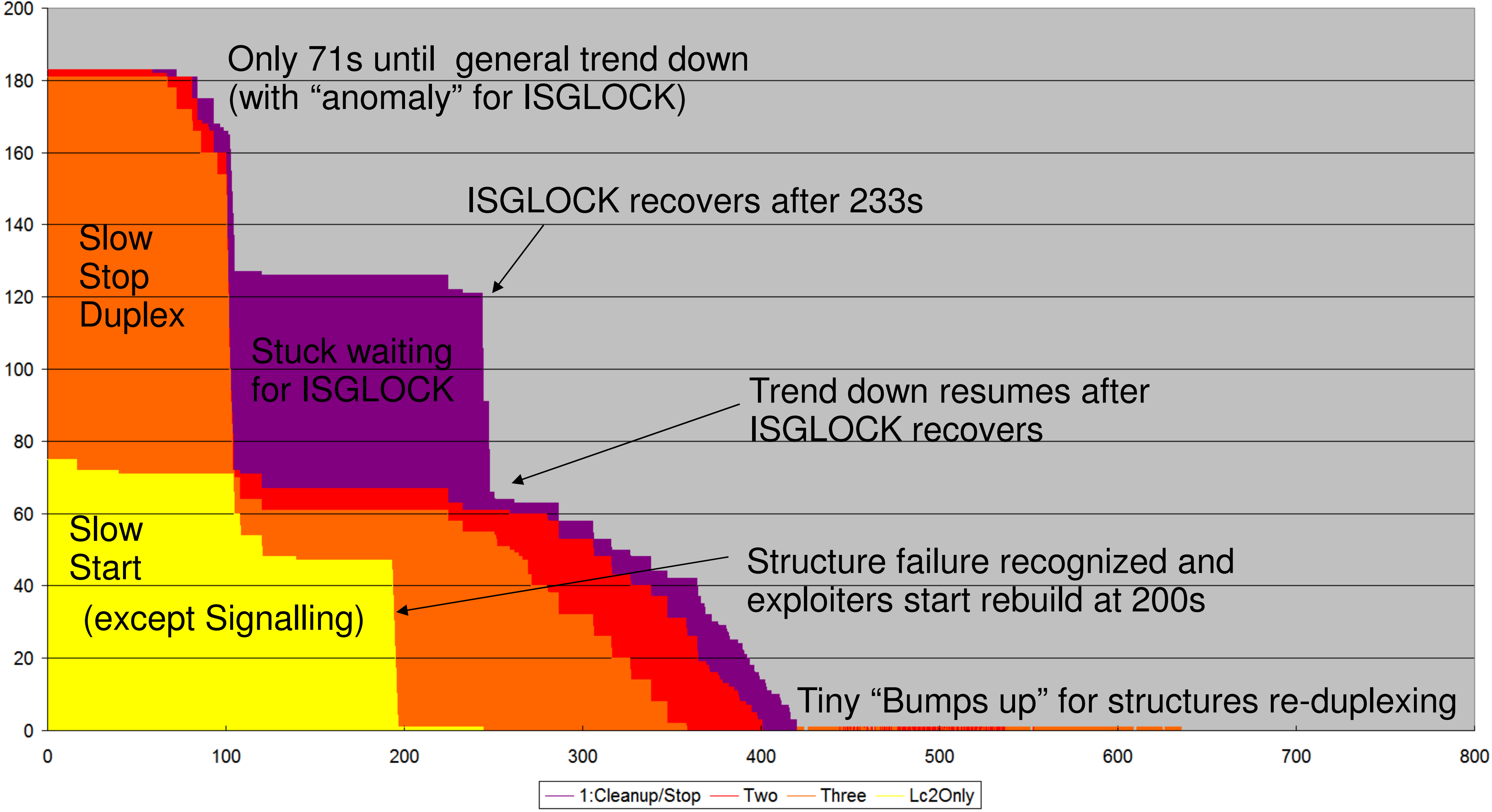
Not formal performance test measurements



© 2015 IBM Corporation

Serial Rebuild (CFLCRMGMT ENABLED)

Not formal performance test measurements



Only 71s until general trend down (with “anomaly” for ISGLOCK)

ISGLOCK recovers after 233s

Trend down resumes after ISGLOCK recovers

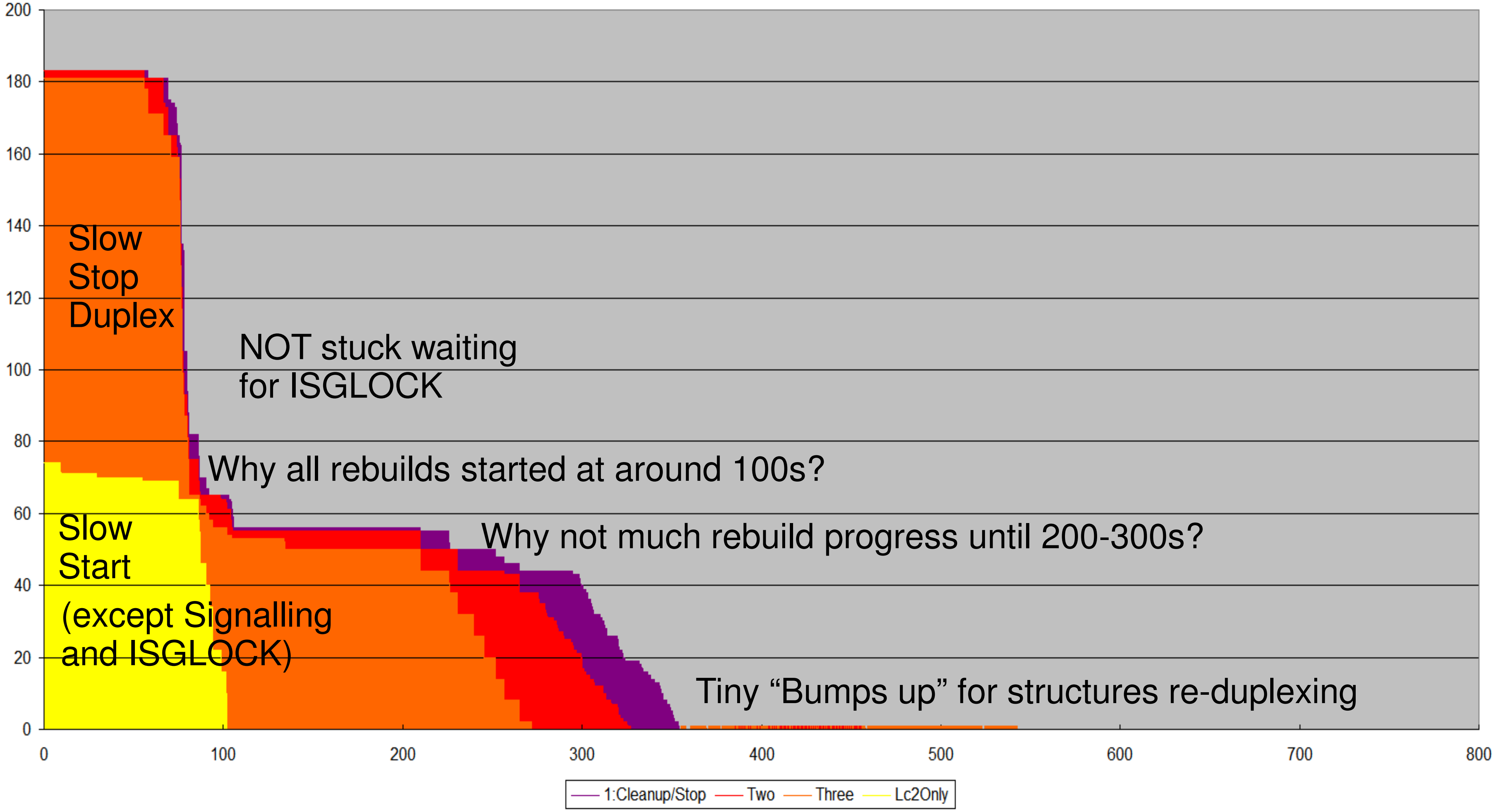
Structure failure recognized and exploiters start rebuild at 200s

Tiny “Bumps up” for structures re-duplexing

© 2015 IBM Corporation

Serial Rebuild ISGLOCK RECPRTY(1)

Not formal performance test measurements



Slow Stop Duplex

NOT stuck waiting for ISGLOCK

Why all rebuilds started at around 100s?

Why not much rebuild progress until 200-300s?

Slow Start

(except Signalling and ISGLOCK)

Tiny “Bumps up” for structures re-duplexing

© 2015 IBM Corporation

Comparing these particular runs

Not formal performance test measurements

CFLCRMGMT	DISABLED	ENABLED	ENABLED w/ISGLOCK RECPRTY(1)
Structures with duplexing complete end point	103	108	109
Structures with rebuild process complete end point	75	75	74
Average time to complete rebuild to recover	481 seconds	337 seconds	254 seconds
Average time to complete duplexing failover to recover	327 seconds	174 seconds	78 seconds
Average time to duplex	79 seconds	2 seconds	2 seconds
Average structure quiesce time	435 seconds	335 seconds	150 seconds
Elapsed time until last rebuild completed	over 10 minutes	about 7 minutes	about 6 minutes
Elapsed time until last duplex established	about 11 minutes	about 11 minutes	about 9 minutes

© 2015 IBM Corporation

Comparing these particular runs ...

- Fewer complaint messages from XES monitors regarding “unresponsive connectors”
- Hundreds of messages reduced to just a few

© 2015 IBM Corporation

z/OS V2R1 - Thin Interrupts for z/OS

- Not to be confused with thin interrupts for the CF
  - Same idea, same technology
  - But exploitation completely independent of one another
- For z/OS, helps reduce latencies arising from:
  - Waiting for PR/SM to dispatch CP
  - Waiting for timer to drive dispatcher
- Which helps improve:
  - Service time for asynchronous requests
  - Responsiveness to list transition notifications
- So relevant workload should have:
  - Shorter elapsed time
  - More throughput

Should help any exploiter with significant async request activity or list notifications

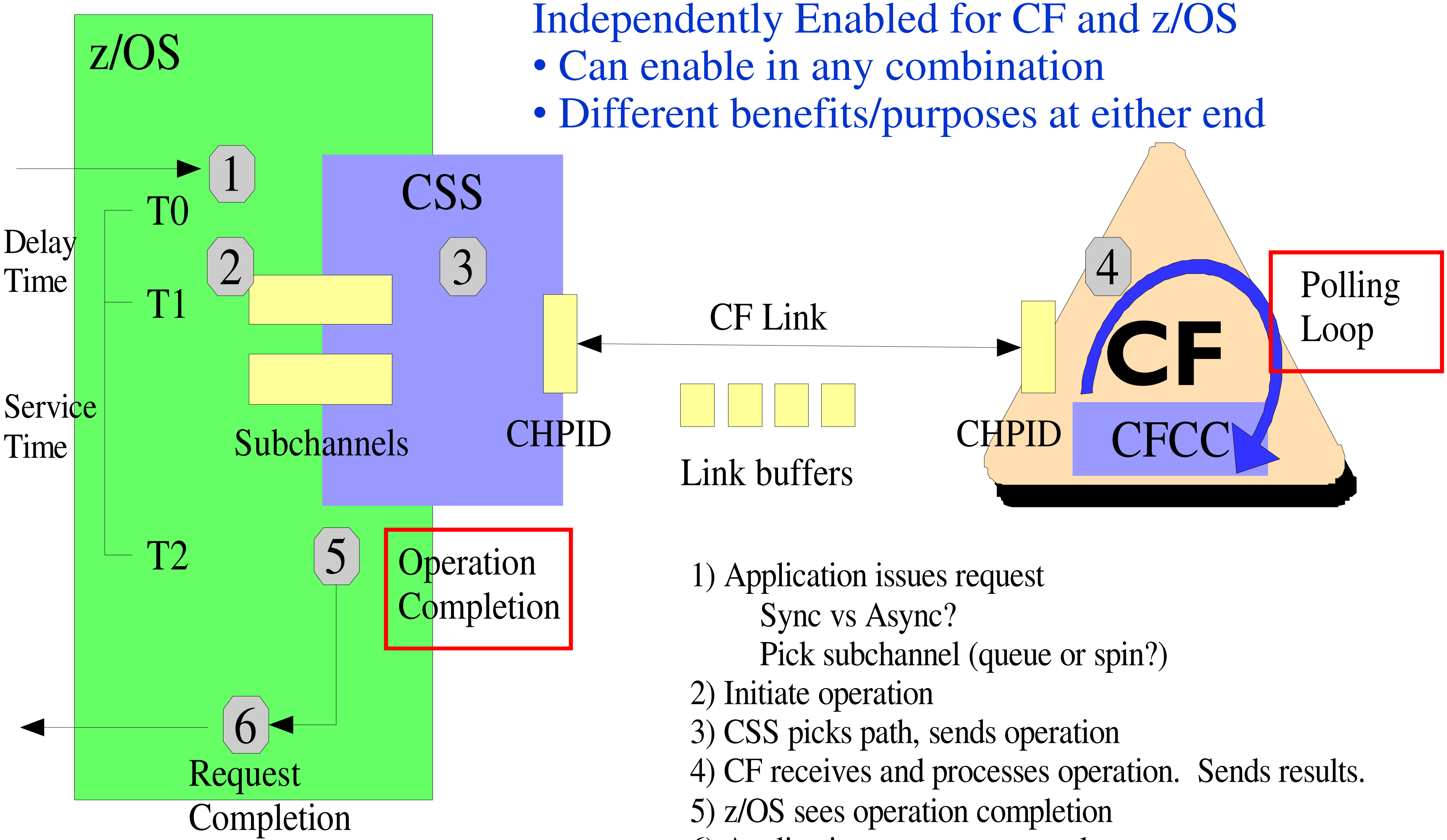
- MQ
- XCF signaling
- IMS SMQ
- Heuristic sync to async conversions

© 2015 IBM Corporation

z/OS V2R1 – Thin Interrupts

Independently Enabled for CF and z/OS

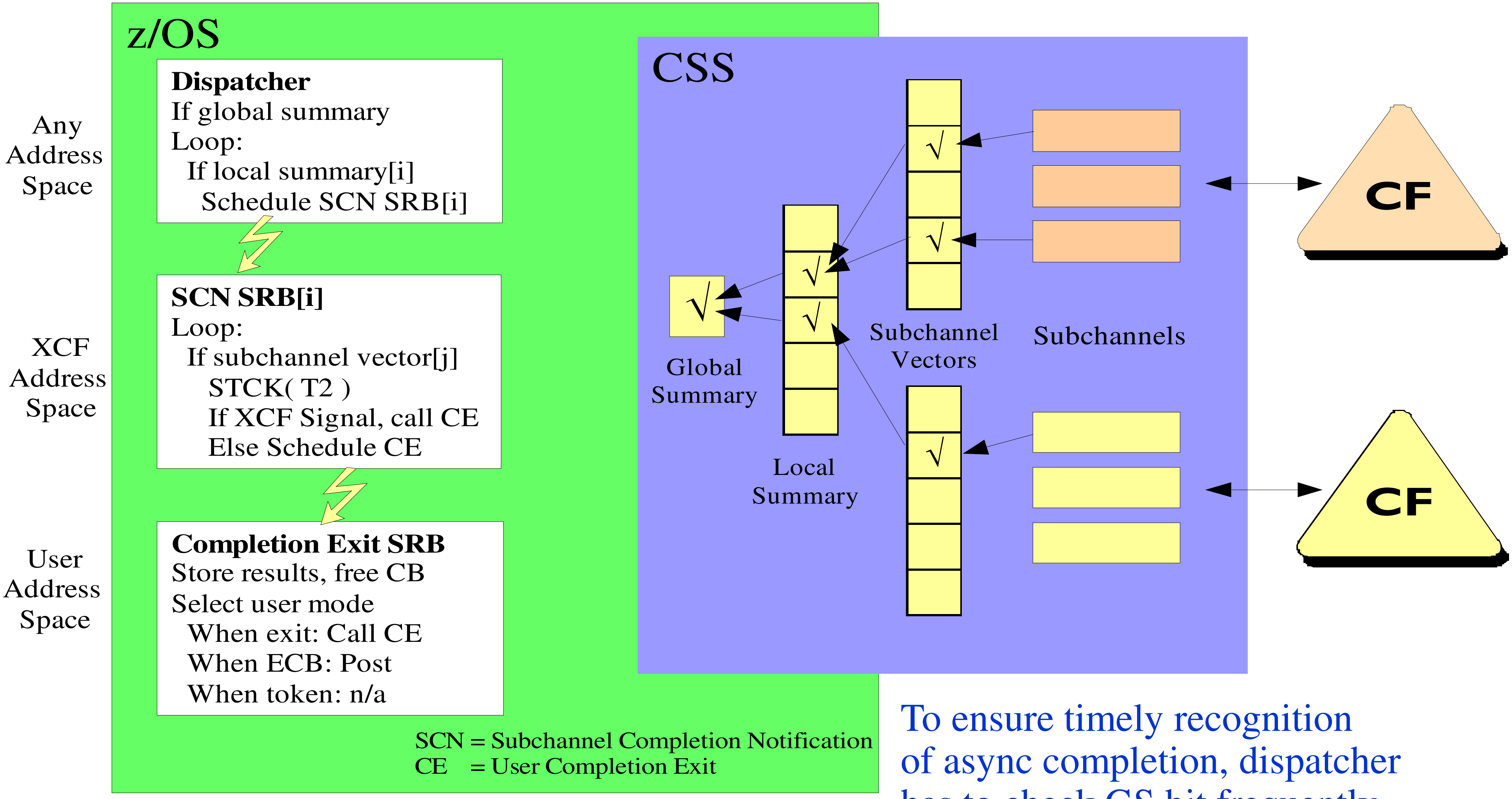
- Can enable in any combination
- Different benefits/purposes at either end



1) Application issues request
   Sync vs Async?
   Pick subchannel (queue or spin?)
2) Initiate operation
3) CSS picks path, sends operation
4) CF receives and processes operation. Sends results.
5) z/OS sees operation completion
6) Application gets request results

© 2015 IBM Corporation

Asynchronous Operation Completion - Today



```
Dispatcher
If global summary
Loop:
  If local summary[i]
    Schedule SCN SRB[i]
```

```
SCN SRB[i]
Loop:
  If subchannel vector[j]
     STCK( T2 )
     If XCF Signal, call CE
     Else Schedule CE
```

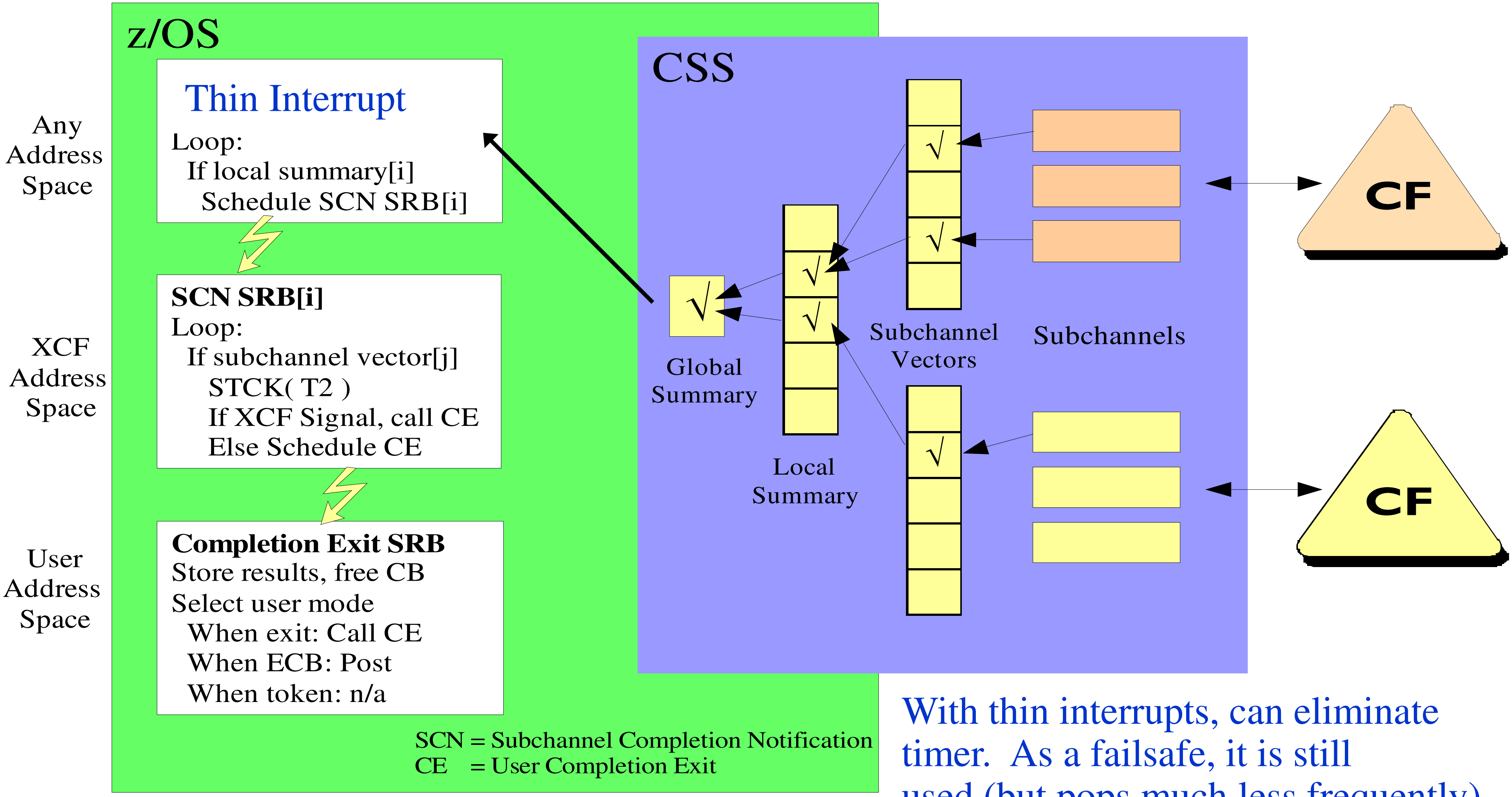
```
Completion Exit SRB
Store results, free CB
Select user mode
  When exit: Call CE
  When ECB: Post
  When token: n/a
```

SCN = Subchannel Completion Notification
CE = User Completion Exit

To ensure timely recognition of async completion, dispatcher has to check GS bit frequently

© 2015 IBM Corporation

Asynchronous Operation Completion - With Thin Interrupt



Thin Interrupt

```
Loop:
  If local summary[i]
    Schedule SCN SRB[i]
```

SCN SRB[i]

```
Loop:
  If subchannel vector[j]
     STCK( T2 )
     If XCF Signal, call CE
     Else Schedule CE
```

Completion Exit SRB

```
Store results, free CB
Select user mode
  When exit: Call CE
  When ECB: Post
  When token: n/a
```

SCN = Subchannel Completion Notification
CE = User Completion Exit

With thin interrupts, can eliminate timer. As a failsafe, it is still used (but pops much less frequently).

© 2015 IBM Corporation

Thin Interrupts for z/OS - Configuration

- Software Requirements (for z/OS exploitation)
  - z/OS V2R1 or later, or
  - z/OS V1R12 and V1R13 with the following service installed:
    - APAR OA38734 (XES)
    - APAR OA37186 (Supervisor)
    - APAR OA38781 (IOS)
    - APAR OA42682 (RMF)
  - By default, z/OS will exploit thin interrupts if hardware supports them
    - If not wanted, disable XCF FUNCTIONS switch COUPLINGTHININT
- Hardware Requirements
  - zEC12 GA2, BC12, or z13 for z/OS coupling thin interrupts

© 2015 IBM Corporation

Thin Interrupts – Switch

- XCF FUNCTIONS switch to enable or disable use of thin interrupts on the z/OS side
  - Does not change the CF behavior at all
  - Default is for COUPLINGTHININT to be ENABLED
- If enabled (and hardware supports it)
  - CSS is told to drive thin interrupts when asynchronous operation completes
  - Timer for dispatcher to check global summary bit fires occasionally
- If disabled
  - CSS is told to not generate thin interrupts (if hardware supports it)
  - Timer for dispatcher to check global summary bit fires frequently

© 2015 IBM Corporation

Thin Interrupts – Messages

- D XCF,C
  - Reports COUPLETHININT switch setting
- D XCF,CF
  - CF DYNDISP setting – add "thin interrupts"
  - Are thin interrupts supported and/or enabled on the CF CEC
- New Messages
  - IXL163I – XES could not enable/disable thin interrupts
  - IXL164I – enabled thin interrupts
- Health Check updated
  - Should have thin interrupts if using shared CPs
  - Configuration data includes thin interrupt information

© 2015 IBM Corporation

z/OS V2R1 – Sync/Async Thresholds

- XES Heuristic Algorithm
  - Measures synchronous service times for each kind of CF operation
  - Determines whether the opportunity cost of running the operation synchronously exceeds the overhead of running the request asynchronously
  - If so, the request will be processed asynchronously
- Helps optimize use of CPU resources so as to maximize the amount of work performed
- At the expense of increasing the elapsed time of the request

© 2015 IBM Corporation

Some Prefer a Different Tradeoff

- Some customers would rather sacrifice CPU efficiency in order to maintain shorter service times
  - –The longer service times impact their business objectives
  - –They can tolerate getting less total work done
- To accommodate this need, you can now set the conversion thresholds used by the heuristic algorithm to tailor the tradeoff between CPU cost and service time.
  - –Also available on z/OS V1R13 and V1R12 via APAR OA41661
  - –On z/OS V1R12, OA41661 went PE since it broke “SFM with BCPii” so you'll need OA43435 to fix that

© 2015 IBM Corporation

Setting Conversion Threshold

- COUPLExx parmlib member
  - On new SYNCASYNC statement specify: keyword(value)
- SETXCF MODIFY,SYNCASYNC,keyword=value
- “keyword” is one of the following thresholds
  - SIMPLEX - for simplex list and cache requests
  - DUPLEX - for duplexed list and cache requests
  - LOCKSIMPLEX - for simplex lock requests
  - LOCKDUPLEX - for duplexed lock requests
- “value” can be:
  - Numeric value in range 1 to 10000 (microseconds)
  - DEFAULT – to use the system determined threshold value

© 2015 IBM Corporation

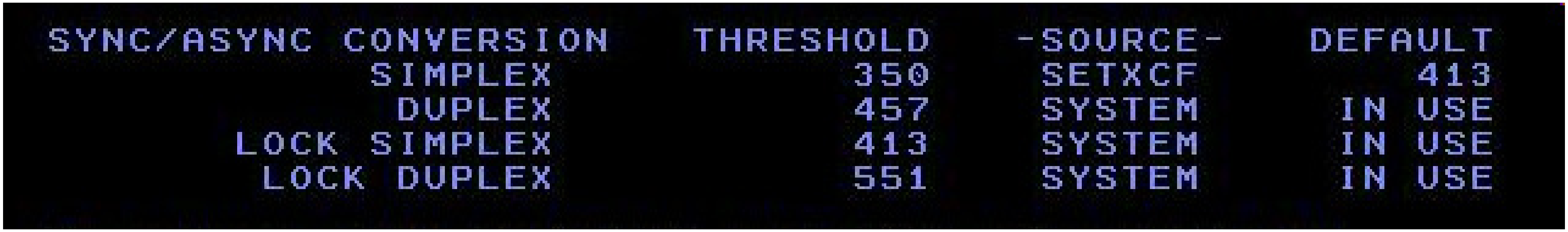
Some Cautions

- The default threshold value is dynamically computed to account for factors that would affect the opportunity cost
  - –Speed of the processor
  - –Number of CPs
- When you set the threshold
  - –It is a fixed value
  - –Never adjusted for dynamic changes that might occur
- Playing with these knobs could significantly impact your workload.
  - – Be careful.

© 2015 IBM Corporation

What Are My Threshold Values?

- D XCF,COUPLE



```
SYNC/ASYNC CONVERSION    THRESHOLD    -SOURCE-    DEFAULT
          SIMPLEX              350     SETXCF         413
           DUPLEX              457     SYSTEM      IN USE
     LOCK SIMPLEX              413     SYSTEM      IN USE
      LOCK DUPLEX              551     SYSTEM      IN USE
```

Which threshold

Current value being used

How was it set?

Current system default if not already in use

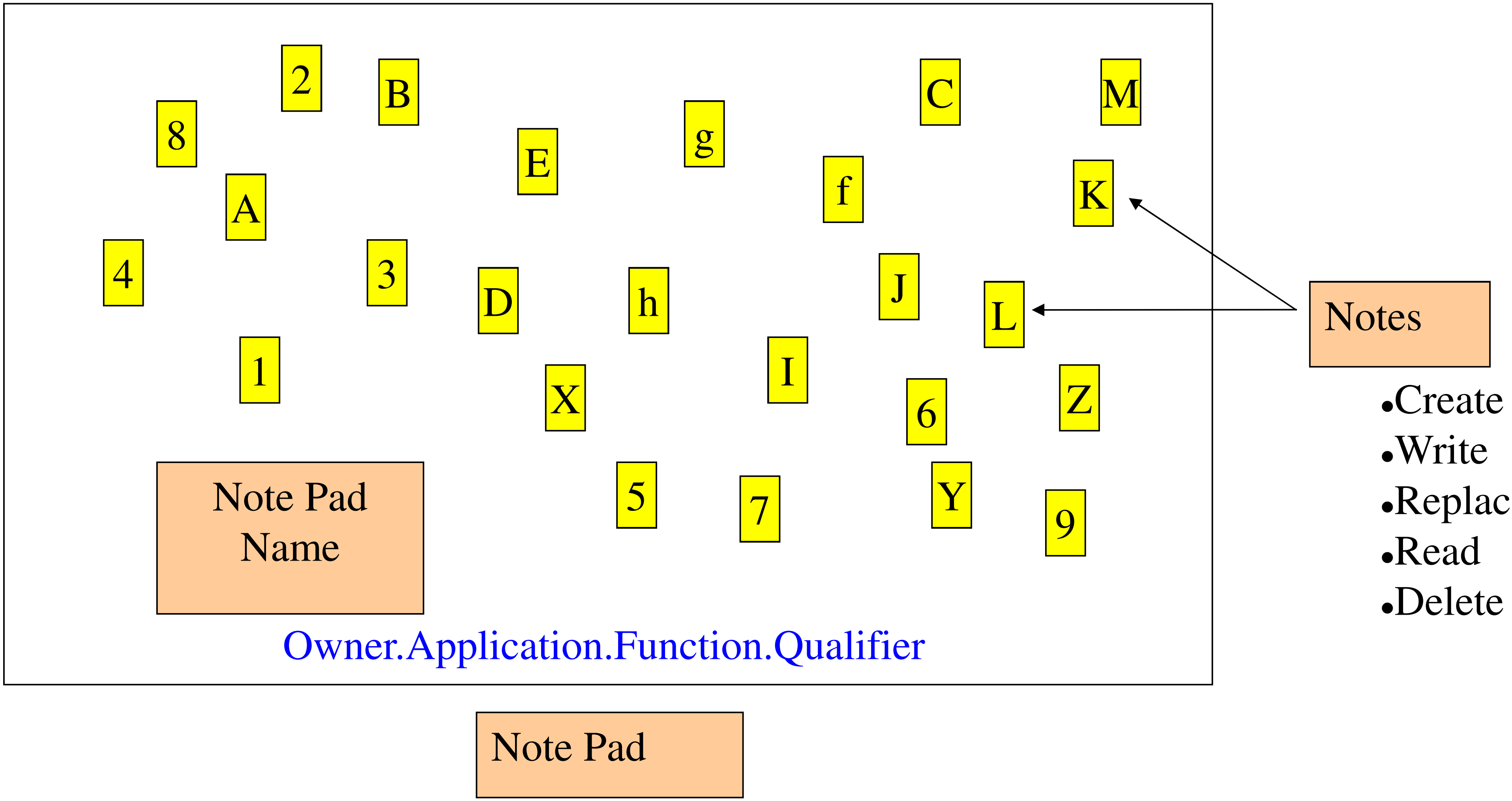
© 2015 IBM Corporation

z/OS V2R1 – XCF Note Pad Service (IXCNOTE macro)

- Programs can read and write notes (list entries) in an XCF note pad (CF list structure)
  - Supports unauthorized callers
  - One or more note pads can reside in the same list structure
  - Each note pad can contain finite number of 1K notes
- Useful for applications that can exploit the “note pad” model
  - High performance access to (state) data from any system
  - Not useful for message passing or work flow since no notification
  - Does not expose full functionality of list structure
- XCF connects to CF structure and deals with various XES exits and protocols
- Simplifies development, reduces complexity, decreases implementation and support costs by masking most of the traditional CF exploitation overhead

© 2015 IBM Corporation

Abstract View of an XCF Note Pad



Notes

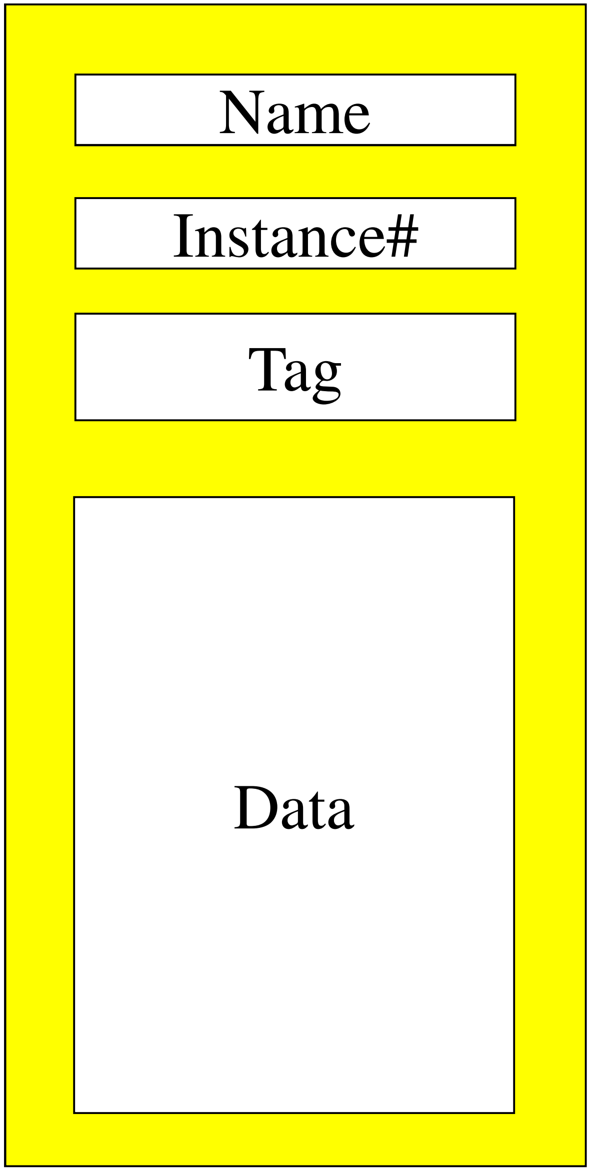
- Create
- Write
- Replace
- Read
- Delete

Note Pad

- Create
- Query
- Delete

© 2015 IBM Corporation

Note in a Note Pad



Name

Instance#

Tag

Data

8 byte user note name

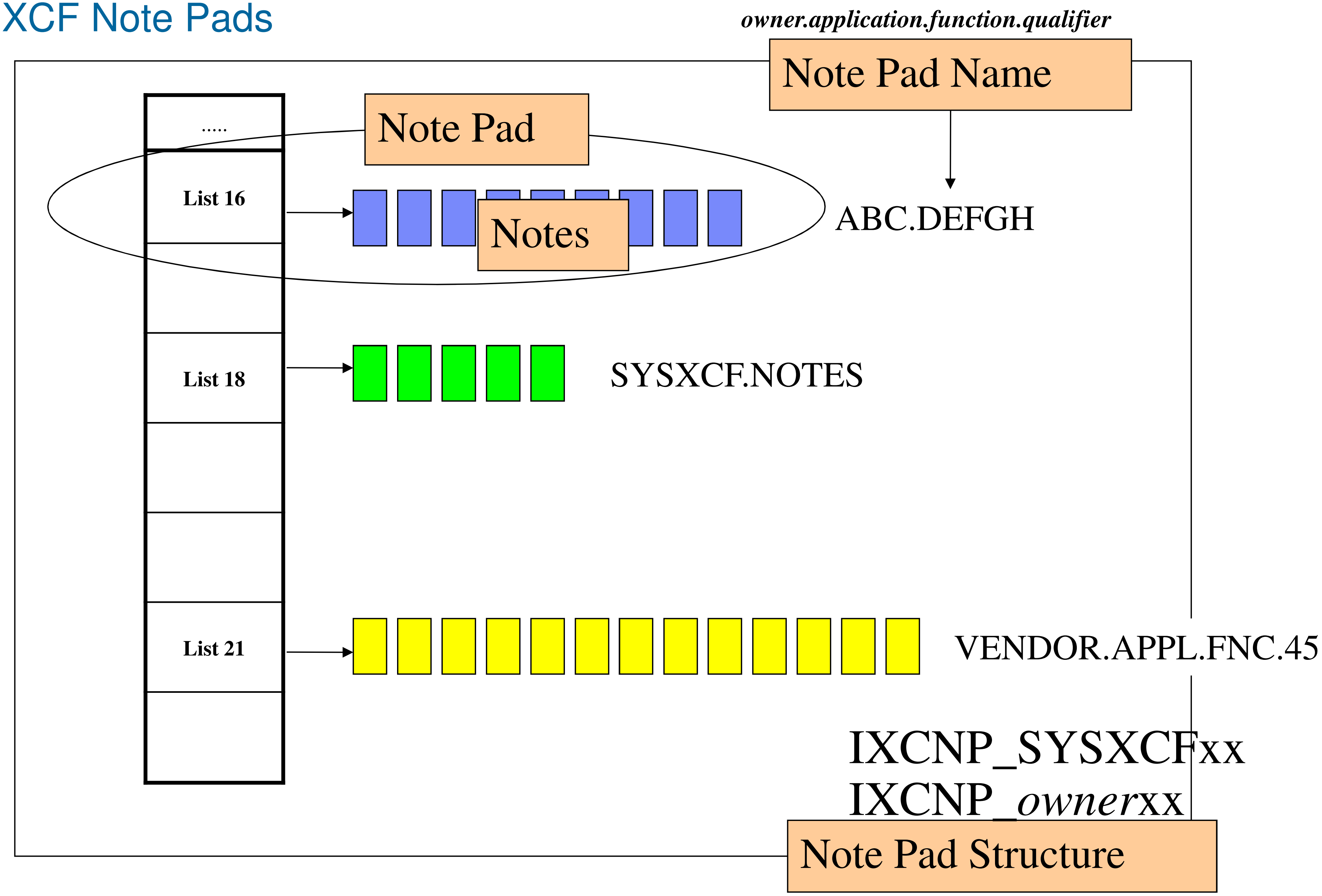
8 byte XCF instance number

16 bytes of user metadata

1024 bytes of user data
(or none)

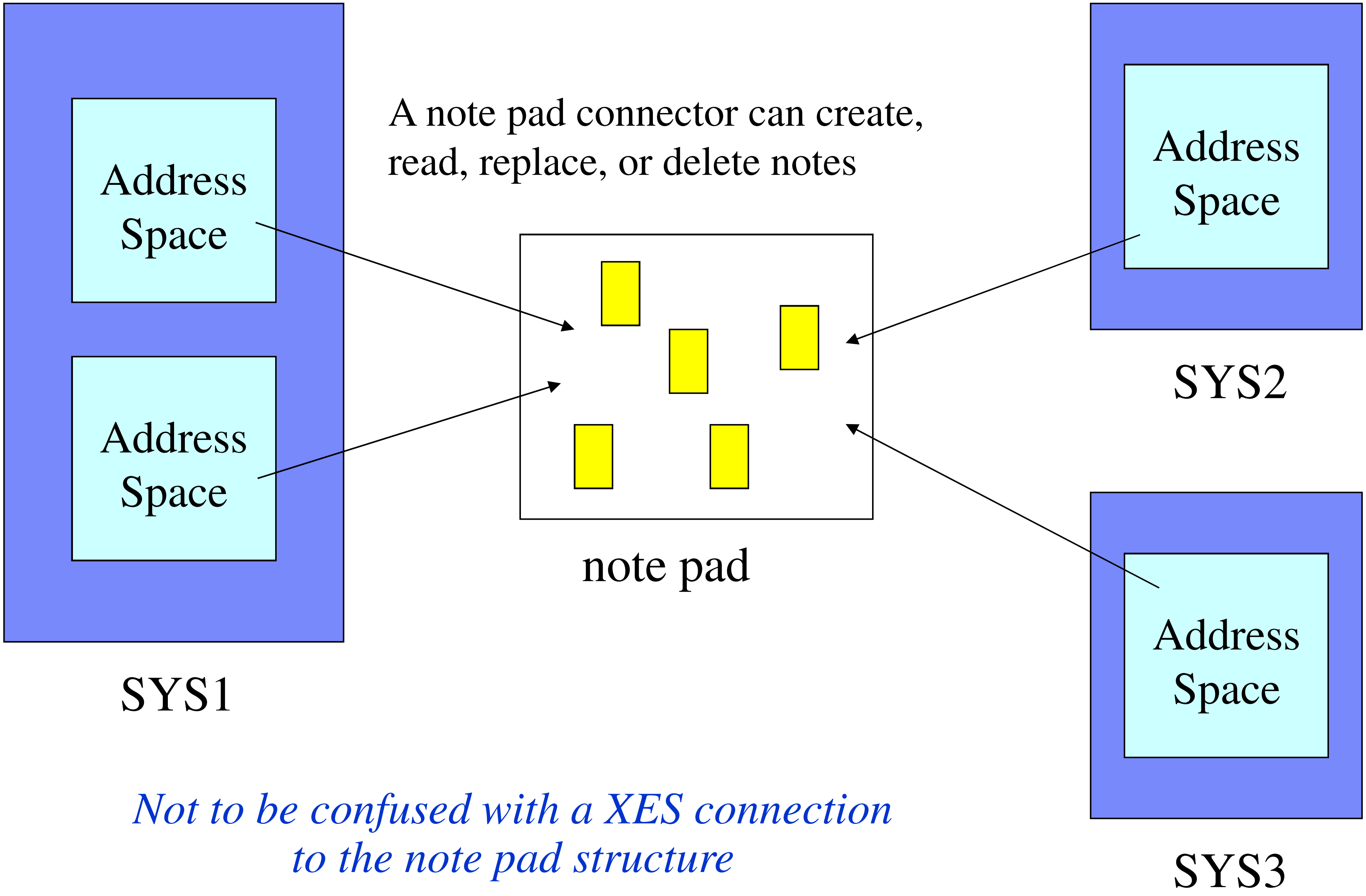
© 2015 IBM Corporation

XCF Note Pads



© 2015 IBM Corporation

Connections to a Note Pad

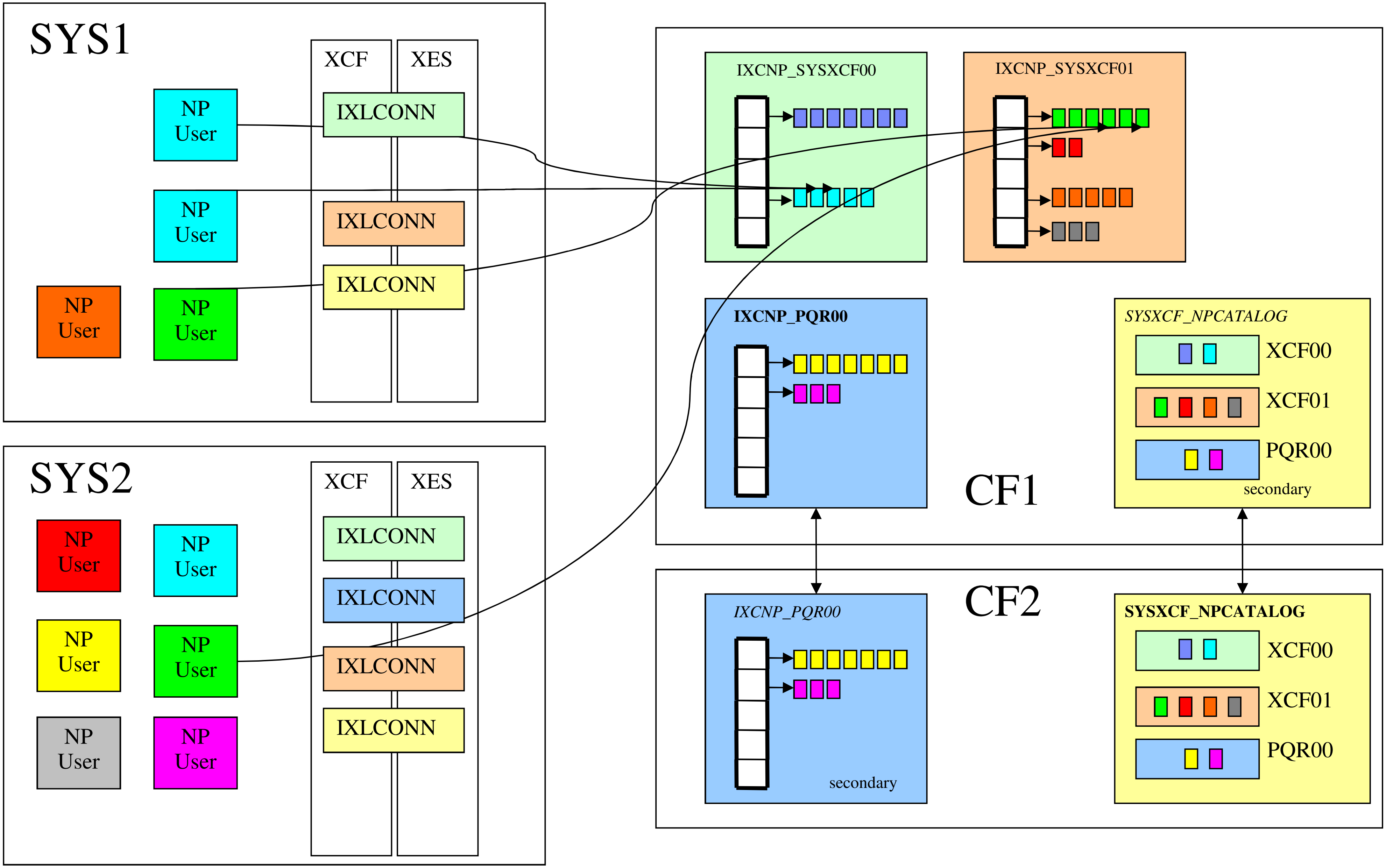


A note pad connector can create, read, replace, or delete notes

Not to be confused with a XES connection to the note pad structure

© 2015 IBM Corporation

XCF Note Pads in the Sysplex



© 2015 IBM Corporation

XCF Note Pad - System Programmer Perspective

- **Requirements**
  - z/OS V2R1, or
  - z/OS 1.13 with OA38450
  - CFLEVEL 9 or later
- **Note Pad Catalog**
  - Size
  - Duplex
- **Note Pad Structure(s)**
  - Names
  - Size
  - Simplex or duplex ?
- **Security**
  - Note pads
  - Structures
- **Management**
  - D XCF,NP
  - Messages
  - Delete Utility
  - Delete Structures
  - Measurement
- Diagnostics
  - XCF CTRACE options

© 2015 IBM Corporation

D XCF,STR – New Status Filters

```
D XCF,STR,STAT=?
IXC352I DISPLAY XCF SYNTAX ERROR, COULD NOT RECOGNIZE:
?.  ONE OF THE FOLLOWING WAS EXPECTED:
  ( ALLOCATED NOTALLOCATED REBUILD
  STRDUMP DEALLOCPENDING POLICYCHANGE LARGERCFRMDS
  FPCONN NOCONN ALTER INCLEANUP
  DUPREBUILD DUPMISMATCH LOSSCONN RBPROC
  RBPEND DUPENAB DUPALLOW
```

- DUPMISMATCH
  - Allocated but DUPLEXED state does not match policy – start or stop duplexing pending
- LOSSCONN
  - A connector has lost connectivity to the structure
- RBPROC
  - Structure in rebuild processing (other than duplex established)
- RBPEND
  - POPCF or REALLOCATE evaluation pending
- DUPENAB/DUPALLOW
  - Structure with policy DUPLEX specification of ENABLED or ALLOWED, respectively

© 2015 IBM Corporation

D XCF,STR – Use of New Filters

- Did z/OS duplex all my DUPLEX(ENABLED) structures?
  - D XCF,STR,STAT=DUPENAB
  - If not, maybe delayed for more important work, rebuild processing, or stop duplex
    - D XCF,STR,STAT=(DUPMISMATCH,RBPROC,RBPEND)

      ```
      CF LOSSCONN RECOVERY MANAGEMENT IS IN PROGRESS.
      THE REALLOCATE PROCESS IS IN PROGRESS.
      POPULATECF REBUILD PENDING
      REBUILD IN PROGRESS
      ```

- Did z/OS resolve all duplexing mismatches?
  - D XCF,STR,STAT=DUPMISMATCH
  - If not, maybe delayed for more important work or rebuild processing
    - D XCF,STR,STAT=(RBPROC,RBPEND)

      ```
      CF LOSSCONN RECOVERY MANAGEMENT IS IN PROGRESS.
      THE REALLOCATE PROCESS IS IN PROGRESS.
      POPULATECF REBUILD PENDING
      REBUILD IN PROGRESS
      ```

D XCF,STR – Use of New Filters …

- Did REALLOCATE (or POPCF) complete?
  - D XCF,STR,STAT=RBPEND

    ```
    THE REALLOCATE PROCESS IS IN PROGRESS.
    POPULATECF REBUILD PENDING
    POPULATECF REBUILD IN PROGRESS
    ```

  - If not, maybe delayed for more important work

    ```
    CF LOSSCONN RECOVERY MANAGEMENT IS IN PROGRESS.
    ```

- Did CF LOSSCONN RECOVERY complete?
  - D XCF,STR,STAT=LOSSCONN,STRNM=*

    ```
    CF LOSSCONN RECOVERY MANAGEMENT IS IN PROGRESS.
    ```

© 2015 IBM Corporation

XCF Signal Throughput Improvement

- XCF manages message exit SRBs to provide for:
  - –Responsive message delivery without …
  - –Starving tasks in the target member address space
    - Many members drop messages off to tasks for processing
- To do so, XCF controls
  - –Number of SRBs
  - –Number of signals an SRB can process
  - –Frequency with which SRBs are scheduled
- We have adjusted these controls to permit delivery of signals in the neighborhood of 100,000 signals/second (per member)
  - –Roughly 4X improvement over prior releases
  - –Assuming the target member can keep up

© 2015 IBM Corporation

Couple Data Set Accessibility Verification

Problem

- A loss of DASD power at one of two sysplex sites can cause loss of all couple data sets (CDSes) and a sysplex-wide outage.

Solution

- XCF processing to remove a CDS will now attempt to verify the accessibility of the remaining CDS. If XCF can determine that both CDSes of a given type have been lost simultaneously, it will refrain from sending the signals that can trigger the sysplex outage.
- So when a subset of sysplex systems have lost both CDSes of a given type, only those systems will be required to remove both CDSes from service. The remaining systems may lose only one CDS or neither.

Also available with APAR OA38311 at z/OS V1R12 and V1R13

© 2015 IBM Corporation

Cache Vector Corruption Detection

- Had an issue in the field which caused buffer invalidation (XI) signals to be missed by DB2 for a Group Buffer Pool
- Unable to determine source of problem
  - –DB2? XES? Links? CFCC?
- So we added support in the CF to:
  - –Detect possible occurrences
  - –Gather timely diagnostic data
  - –Fail in a way that avoids data corruption
- And it seems to have worked ….

© 2015 IBM Corporation

Cache Vector Corruption Detection …

- XES is a guilty party
  - –Timing window where cleanup of vector used to manage XI signals was not properly handled
- APAR OA42519

© 2015 IBM Corporation

Agenda

- Hardware Updates
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2
  - z/OS V2R1
  - z/OS V1R13
- Summary

© 2015 IBM Corporation

z/OS V1R13 - Summary

▪ D XCF,SYSPLEX	– Revised output
▪ CF Structure Placement	– more explanation
▪ ARM	– New timeout for application cleanup
▪ New API for XCF signalling	– IXCSRVR,IXCSEND,IXCRECV
▪ SETXCF MODIFY	– Disable structure alter processing
▪ SDSF	– Sysplex wide data gather without MQ
▪ Runtime Diagnostics	– Detects more contention
▪ zFS	– Sysplex wide direct access to shared files

Due to time restrictions, I can only summarize.
Slides with more information on each topic are included in the Appendix

Agenda

- Hardware Updates
  - CFCC Level 20
  - CFCC Level 19
  - CFCC Level 18
  - Parallel Sysplex Coupling Links
  - Server Time Protocol (STP)
- Software Updates
  - z/OS V2R2
  - z/OS V2R1
  - z/OS V1R13
- Summary

© 2015 IBM Corporation

Highlights

- z13
- ICA SR coupling links
- STP
- Thin Interrupts:
  - On CF for configuration flexibility
  - On z/OS for better async CF service time
- XCF message isolation
- CFRM site preferences for duplexed structures
- CFRMTAKEOVERCF to avoid sysplex outages
- Serial Rebuild for better MTTR

© 2015 IBM Corporation

Please complete your session evaluation

z/OS Parallel Sysplex Update
Session 17443

z/OS Publications

- *MVS Setting Up a Sysplex*
- *MVS Initialization and Tuning*
- *MVS Systems Commands*
- *MVS Diagnosis: Tools and Service Aids*
- *z/OS V2R2 Migration*
- *z/OS V2R2 Planning for Installation*
- *z/OS MVS Programming: Callable Services for High Level Languages*
  - Documents BCPii Setup and Installation and BCPii APIs

© 2015 IBM Corporation

Sysplex-related Redbooks

- 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

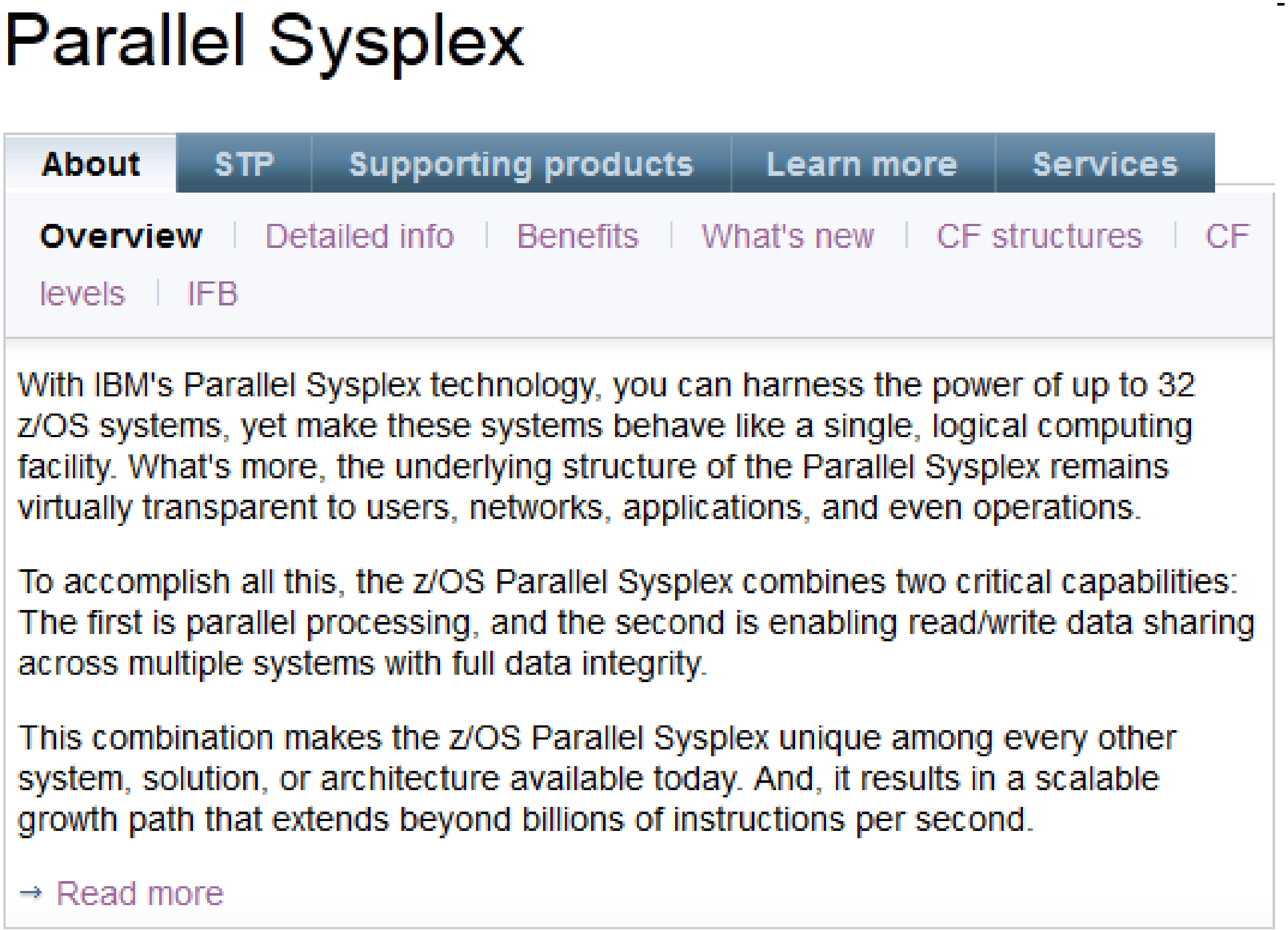
- Exploiting the IBM Health Checker for z/OS Infrastructure, REDP-4590

- Available at www.redbooks.ibm.com

© 2015 IBM Corporation

Parallel Sysplex Web Site

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



With IBM's Parallel Sysplex technology, you can harness the power of up to 32 z/OS systems, yet make these systems behave like a single, logical computing facility. What's more, the underlying structure of the Parallel Sysplex remains virtually transparent to users, networks, applications, and even operations.

To accomplish all this, the z/OS Parallel Sysplex combines two critical capabilities: The first is parallel processing, and the second is enabling read/write data sharing across multiple systems with full data integrity.

This combination makes the z/OS Parallel Sysplex unique among every other system, solution, or architecture available today. And, it results in a scalable growth path that extends beyond billions of instructions per second.

Appendix

Sysplex Highlights from z/OS V1R13

© 2015 IBM Corporation

Appendix – z/OS V1R13

- D XCF,SYSPLEX – Revised output
- CF Structure Placement – more explanation
- ARM – New timeout parameter for application cleanup

- New XCF Client/Server API for sending signals
- SETXCF MODIFY - Disable structure alter processing
- SDSF – Sysplex wide data gathering without MQ
- Runtime Diagnostics – Detects more contention
- zFS – Direct access to shared files throughout sysplex

© 2015 IBM Corporation

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

© 2015 IBM Corporation

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
```

© 2015 IBM Corporation

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

© 2015 IBM Corporation

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
```

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

© 2015 IBM Corporation

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

© 2015 IBM Corporation

z/OS V1R13 – New XCF API for Message Passing (XCF Client/Server)

- 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
- Messages delivered to a **task** instead of an SRB
  - **Server** is collection of tasks identified by **server name**
  - Server exit routine (instead of message exit routine)
  - Various server selection criteria (routing options)
- Response processing
  - Occurs under thread of application's choosing
  - Blocking or non-blocking

© 2015 IBM Corporation

z/OS V1R13 XCF Client/Server

- IXCSEND – send request to one or more servers

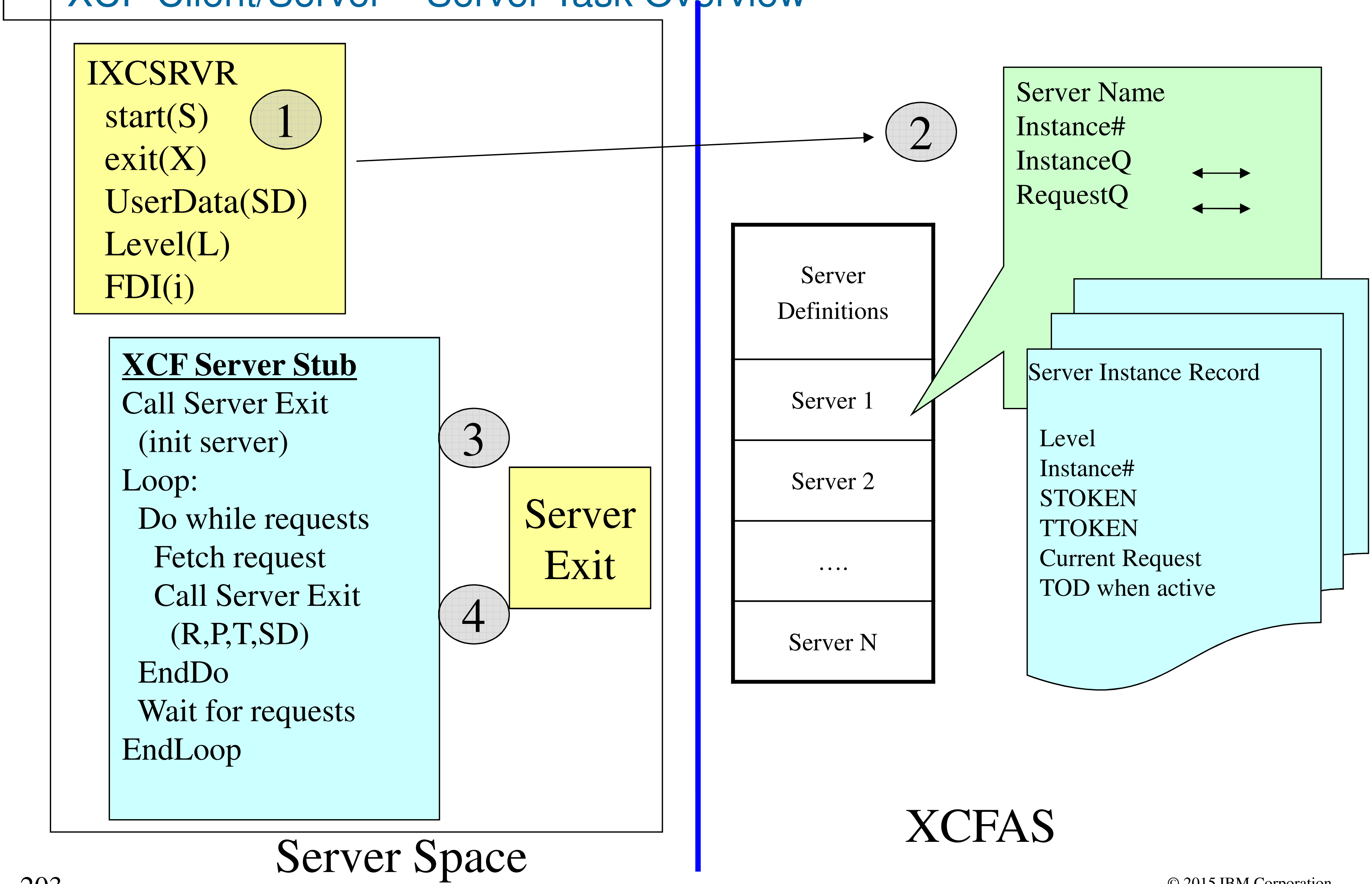
- IXCSRVR – start or stop a server instance
- IXCSEND – send response to client request

- IXCRECV – receive response(s) from server(s)

- IXCYSRVR – data mappings

© 2015 IBM Corporation

XCF Client/Server – Server Task Overview



```
IXCSRVR
  start(S)
  exit(X)
  UserData(SD)
  Level(L)
  FDI(i)
```

XCF Server Stub

```
Call Server Exit
  (init server)
Loop:
  Do while requests
    Fetch request
    Call Server Exit
      (R,P,T,SD)
  EndDo
  Wait for requests
EndLoop
```

Server Exit

Server Space

Server Definitions

Server 1

Server 2

....

Server N

Server Name
Instance#
InstanceQ
RequestQ

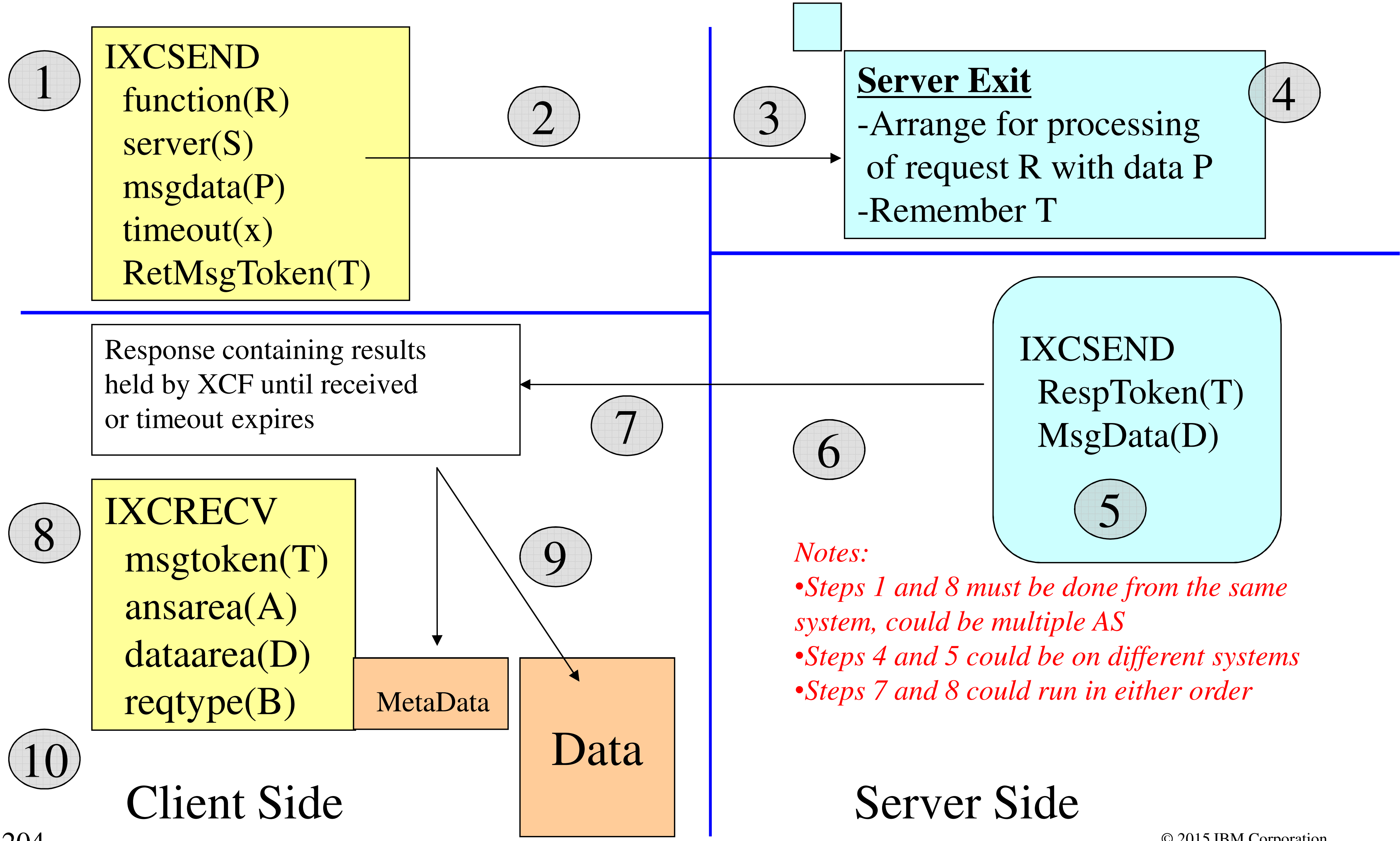
Server Instance Record

Level
Instance#
STOKEN
TTOKEN
Current Request
TOD when active

XCFAS

© 2015 IBM Corporation

XCF Client/Server - Send/Receive Overview



1

IXCSEND
function(R)
server(S)
msgdata(P)
timeout(x)
RetMsgToken(T)

2

3

Server Exit
-Arrange for processing of request R with data P
-Remember T

4

5

IXCSEND
RespToken(T)
MsgData(D)

6

7

Response containing results held by XCF until received or timeout expires

8

IXCRECV
msgtoken(T)
ansarea(A)
dataarea(D)
reqtype(B)

MetaData

9

Data

10

Client Side

Server Side

Notes:

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

© 2015 IBM Corporation

DISPLAY XCF,SERVER

- The DISPLAY XCF command was extended to display information about servers, server instances, and queued work

```
D XCF, { SERVER | SRV }
          [ ,{SYSNAME | SYSNM}={sysname | (sysname [,sysname]. . .) }  ]
          [ ,{SERVERNAME | SRVNAME | SRVNM}={servername}  ]
          [ ,SCOPE={ {SUMMARY | SUM} | {DETAIL | DET} } ]
          [, TYPE=NAME [, STATUS=(STALLED)] |
                  {INSTANCE | INST}
                        [, STATUS=( [{WORKING | WORK}] [, STALLED] ) ]
                        [, {INSTNUM | INST#}=inst# ] ]
```

© 2015 IBM Corporation

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 ….

© 2015 IBM Corporation

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
  - –OA37566 as well

© 2015 IBM Corporation

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

© 2015 IBM Corporation

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

© 2015 IBM Corporation

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

© 2015 IBM Corporation

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

© 2015 IBM Corporation