#SHAREorg

Z/OS Parallel Sysplex Update

Mark A Brooks
IBM

March 11, 2014
Session Number 15105

SHARE in Anaheim

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.

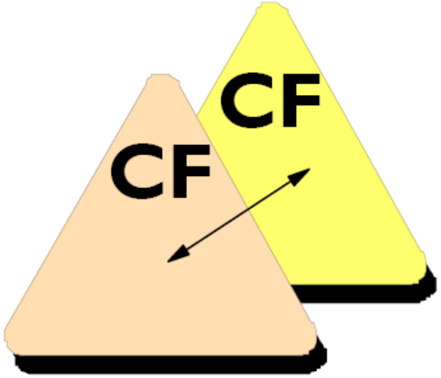
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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.

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



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

CFLEVEL 19

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

- Requirements for CFLEVEL 19
  - z/OS V2.1
  - 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

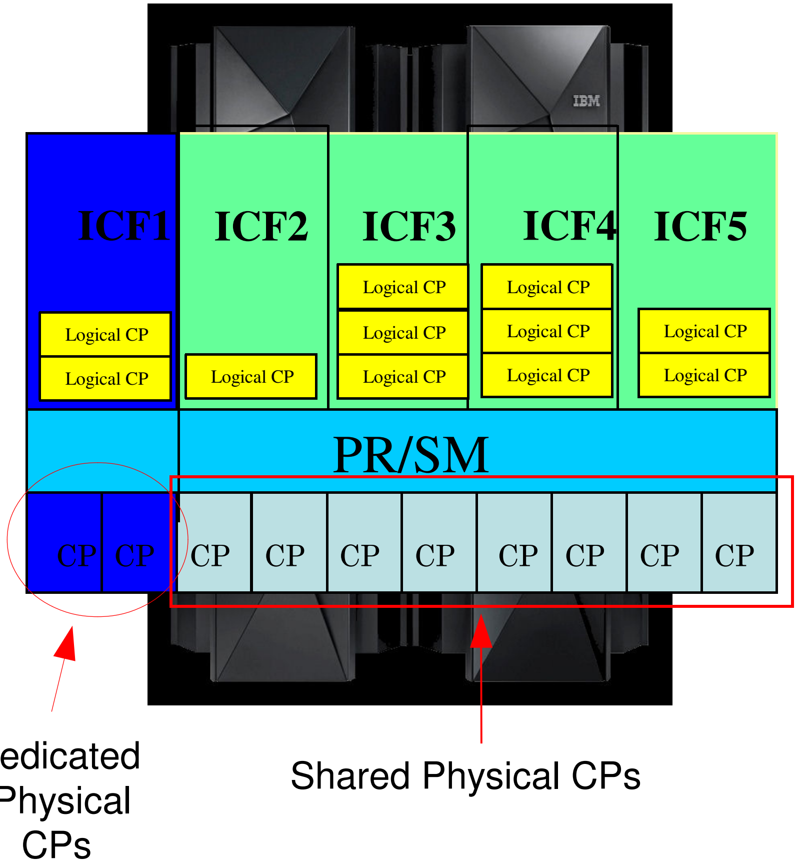
- CFLEVEL 19 provides support for
  - Thin Interrupts to reduce parallel sysplex costs
  - Flash Memory to enhance resiliency (planned 1H2014)*

** Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.*

Review: Dedicated vs Shared CPs for Coupling Facility

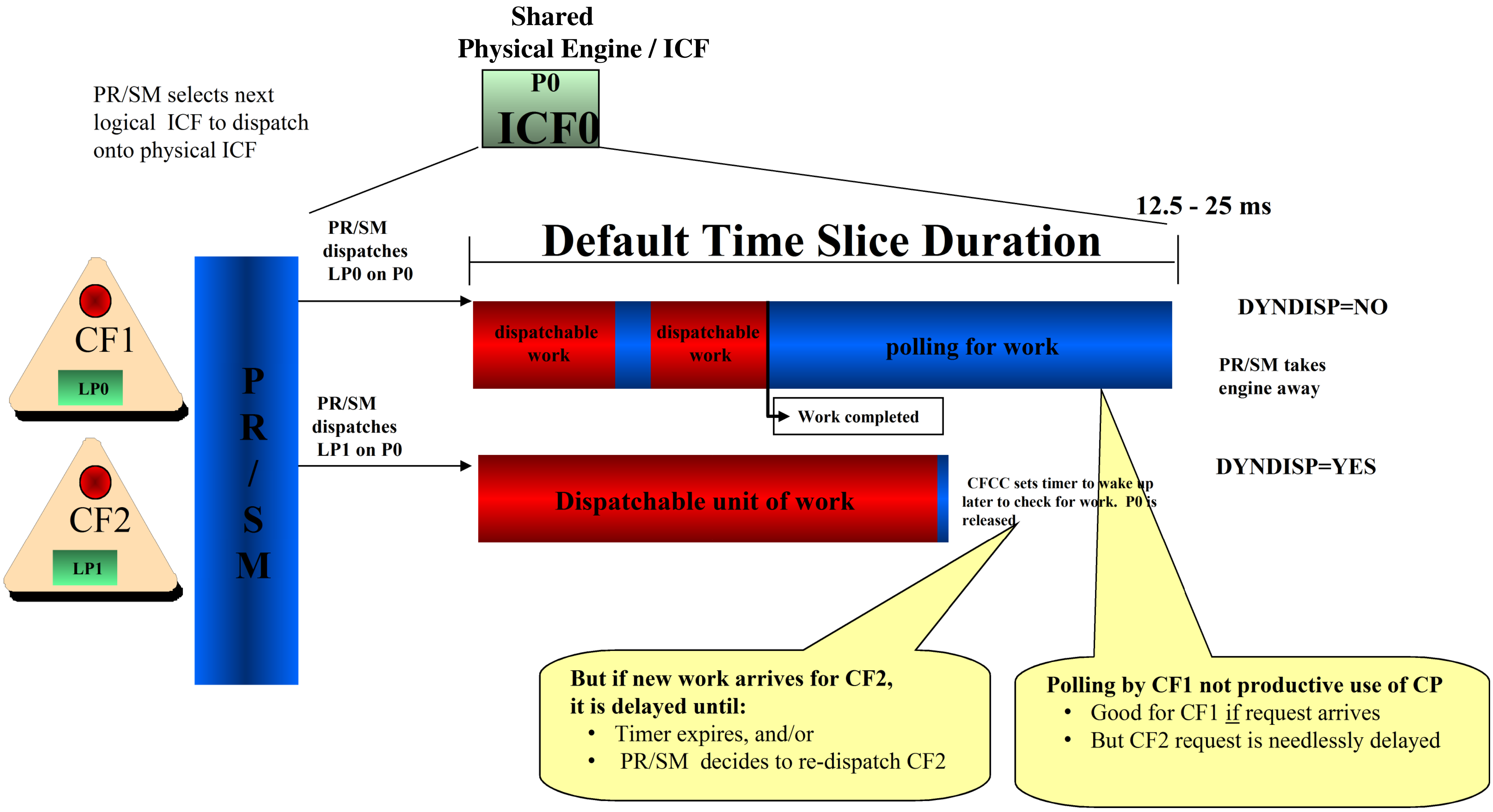
- CFCC uses a polling model
  - Provides consistently good services when dedicated engine assigned to CF
  - But “hurts” when sharing engines since needlessly consumes cycles looking for work when there is none
- With shared engines, CF service times are unpredictable and highly variable
  - PR/SM determines 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

Coupling Facility



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

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

CF1
LP0

CF2
LP1

P R / S M

PR/SM dispatches LP0 on P0

dispatchable work

dispatchable work

polling for work

Work completed

DYNDISP=NO

PR/SM takes engine away

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:

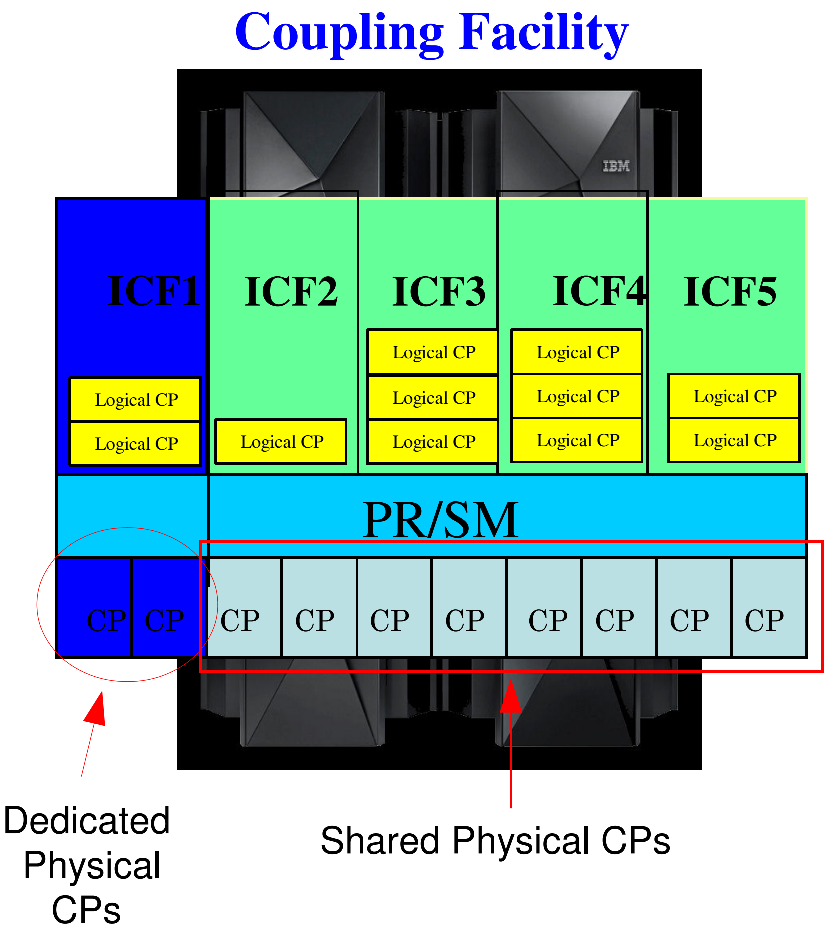
- Timer expires, and/or
- PR/SM decides to re-dispatch CF2

Polling by CF1 not productive use of CP

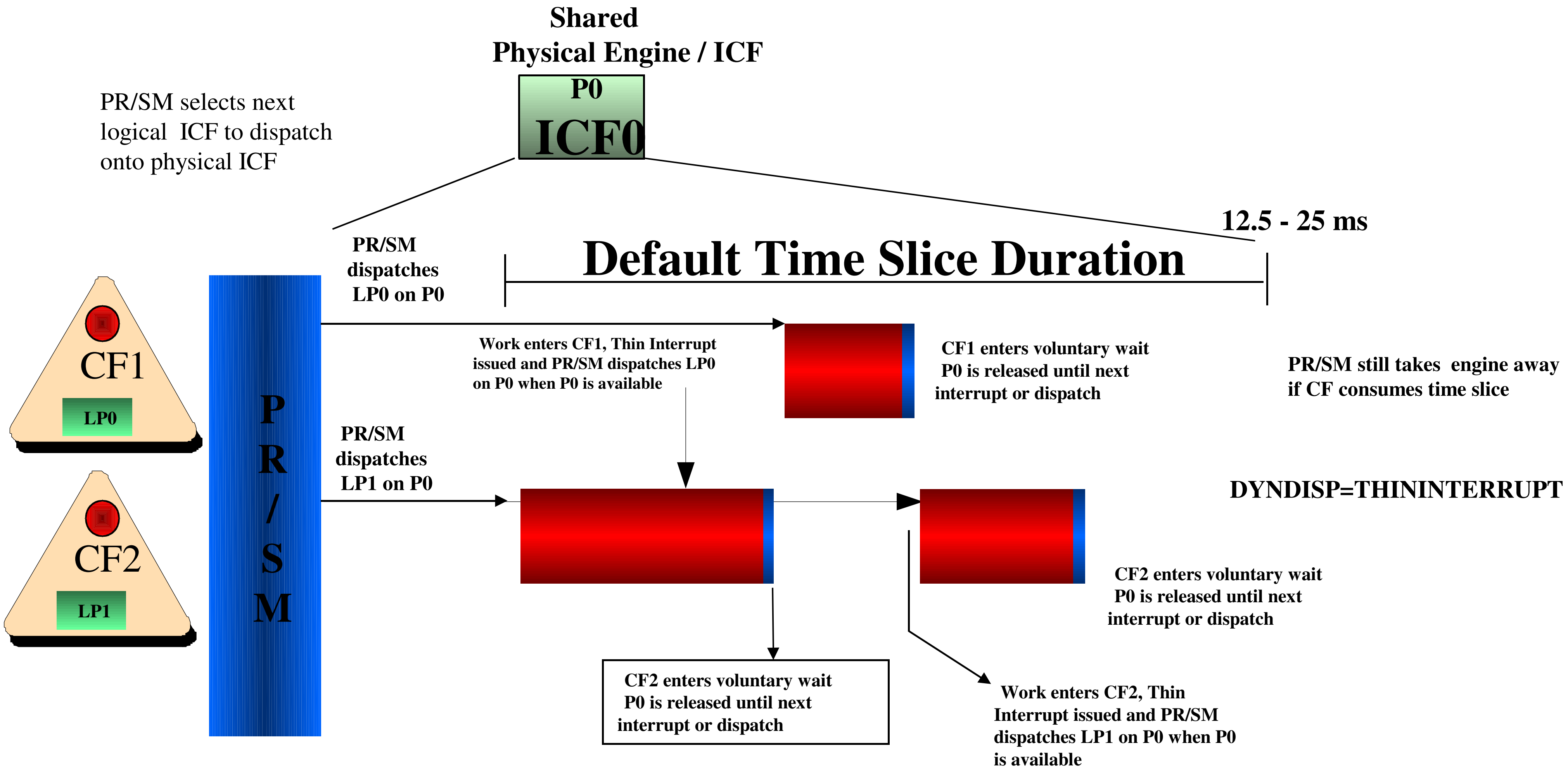
- Good for CF1 if request arrives
- But CF2 request is needlessly delayed

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 ...



New Dynamic CF Dispatch Option – Thin Interrupts



Shared Physical Engine / ICF

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

12.5 - 25 ms

Default Time Slice Duration

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

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)

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

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

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

CFLEVEL 19 - “CF Flash”

Planned for 1H2014*

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 19 with appropriate MCL
  - z/OS V2R1 or z/OS V1R13 (both with appropriate PTFs)
  - A new level of MQ is not required

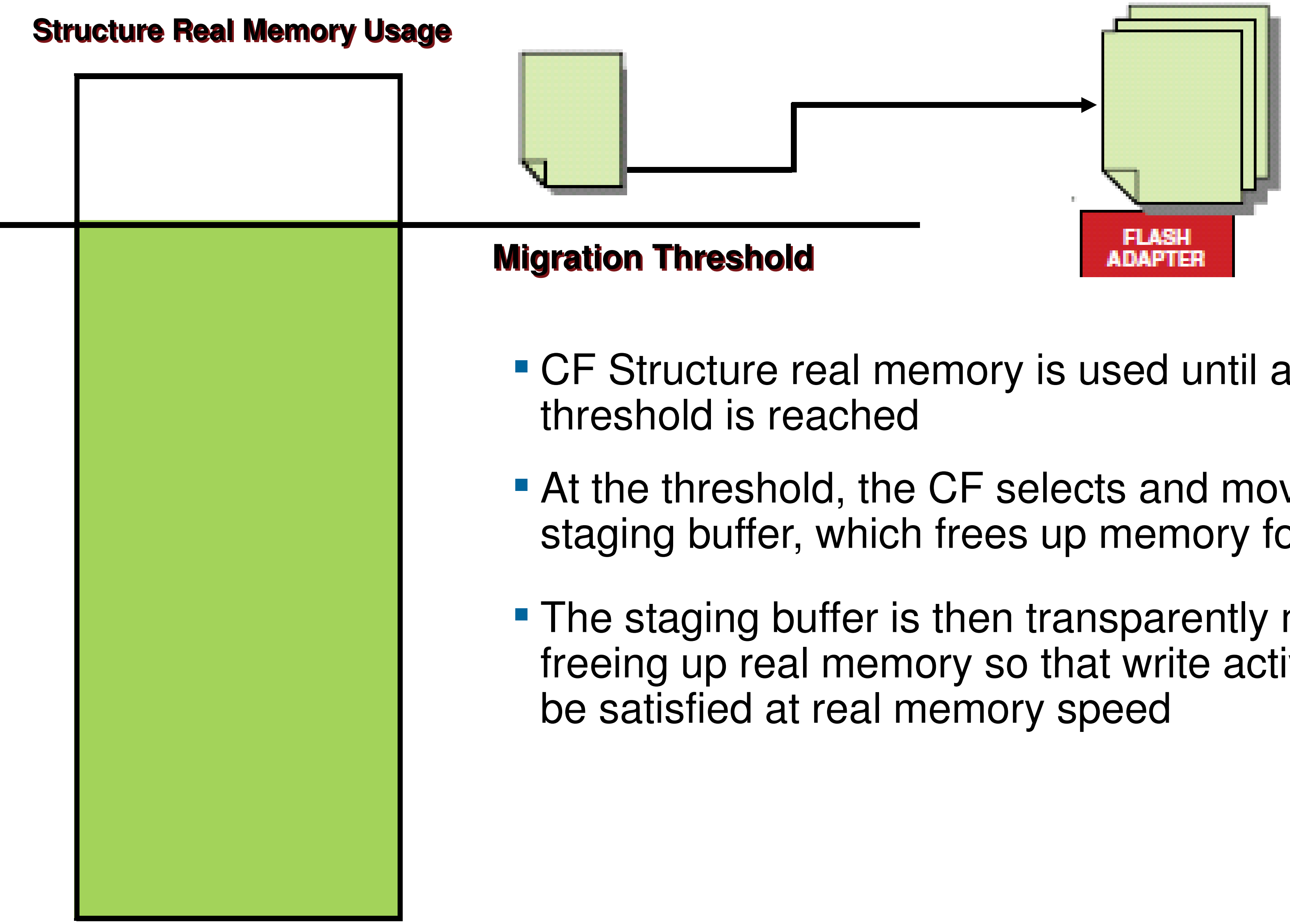
** Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.*

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

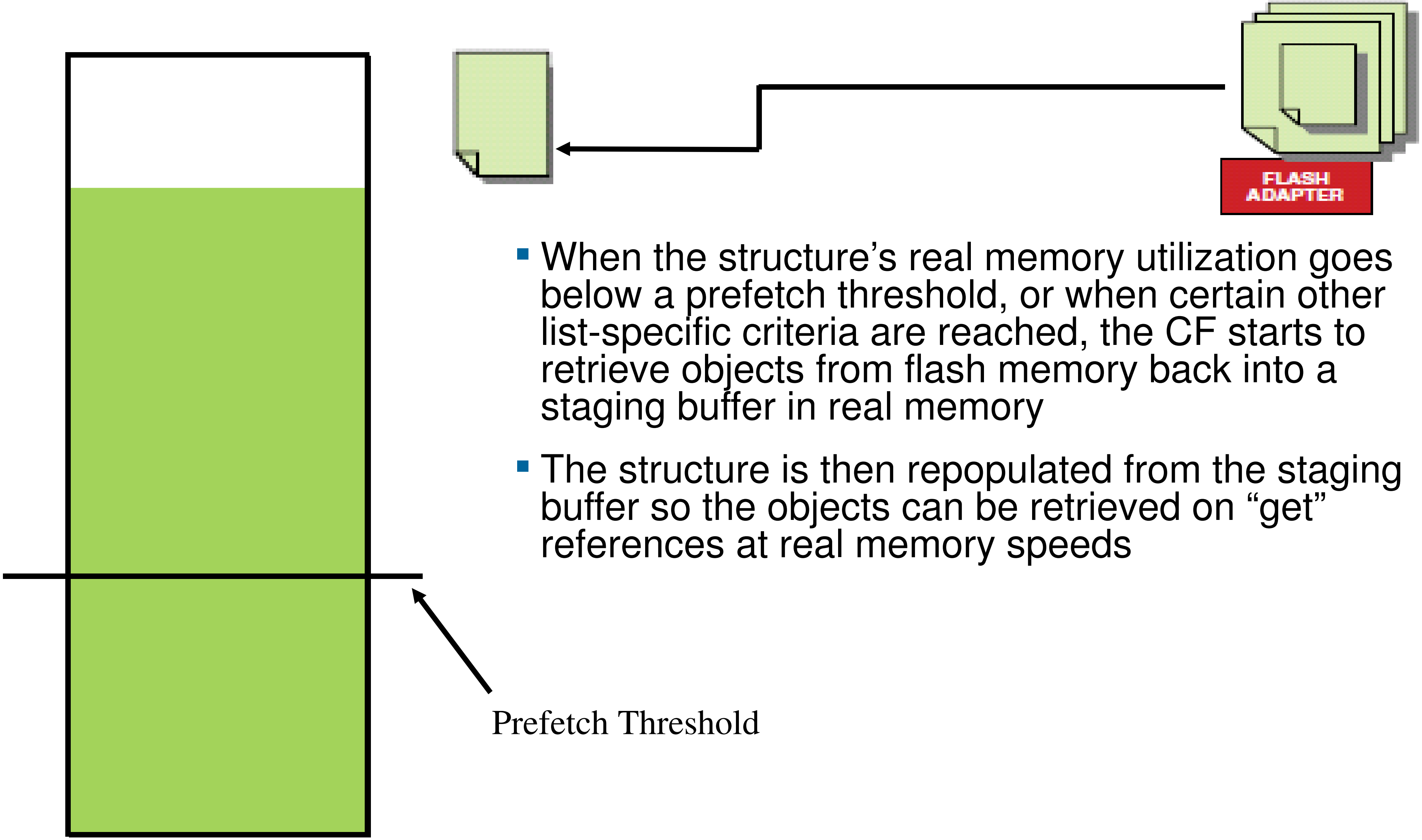
So how will this work ...

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

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

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”)

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

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

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 ...

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

CF Flash – Migration and Coexistence Considerations

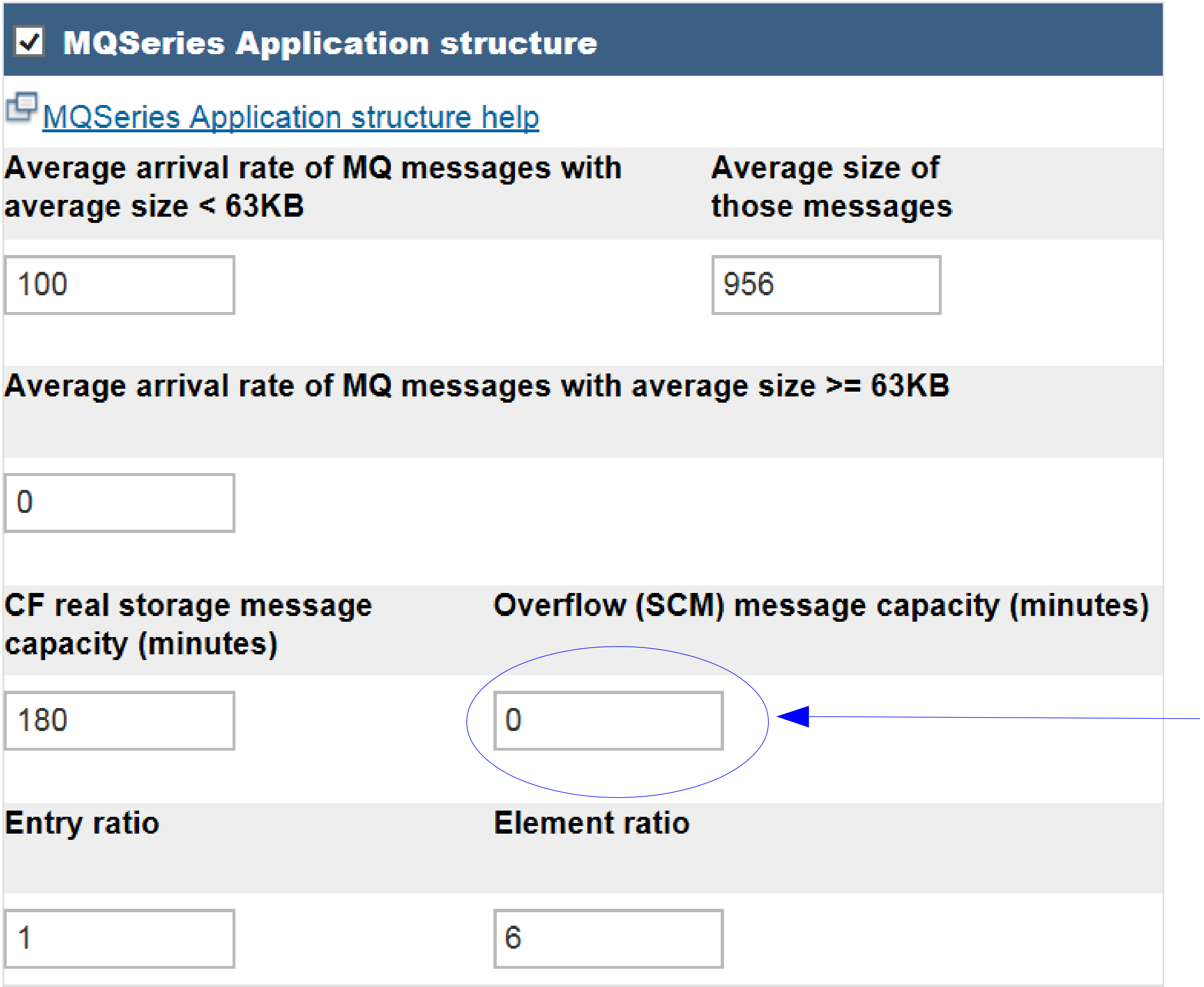
- All systems in sysplex must have appropriate z/OS support before CF flash will be exploited
- Down-level systems cannot connect to, rebuild, alter, display, or dump a structure that is capable of using flash

- 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

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

New CFSIZER input panel for MQ



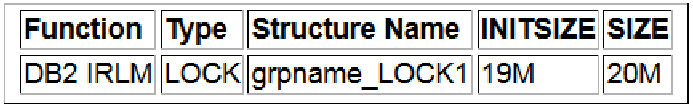
☑ MQSeries Application structure

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

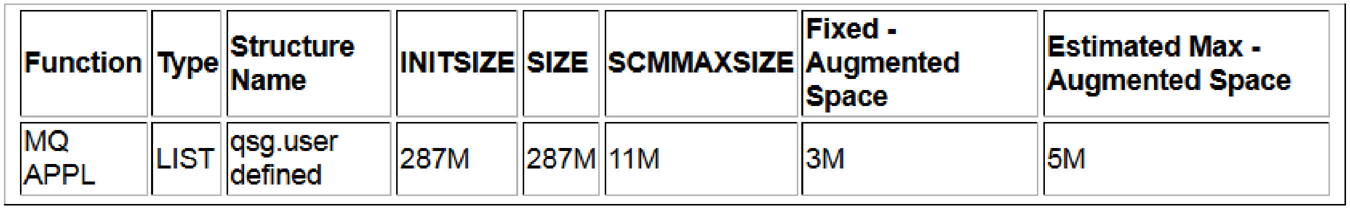
Nonzero value triggers SCM calculations

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).

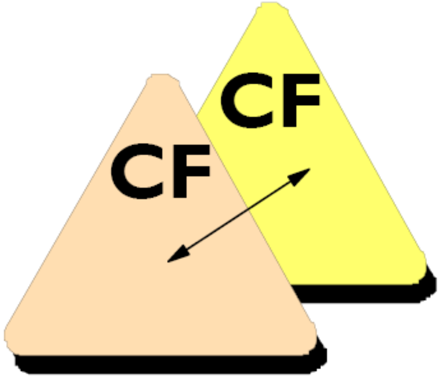


Function	Type	Structure Name	INITSIZE	SIZE	SCMMAXSIZE	Fixed - Augmented Space	Estimated Max - Augmented Space
MQ APPL	LIST	qsg.user defined	287M	287M	11M	3M	5M

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

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



CFLEVEL 18

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

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

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)

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..*

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.

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 (planned* for future DB2 release)
- Also note
  - Rolled back to CFCC Release 17
  - Rolled back to CFCC Release 16

** Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.*

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)

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

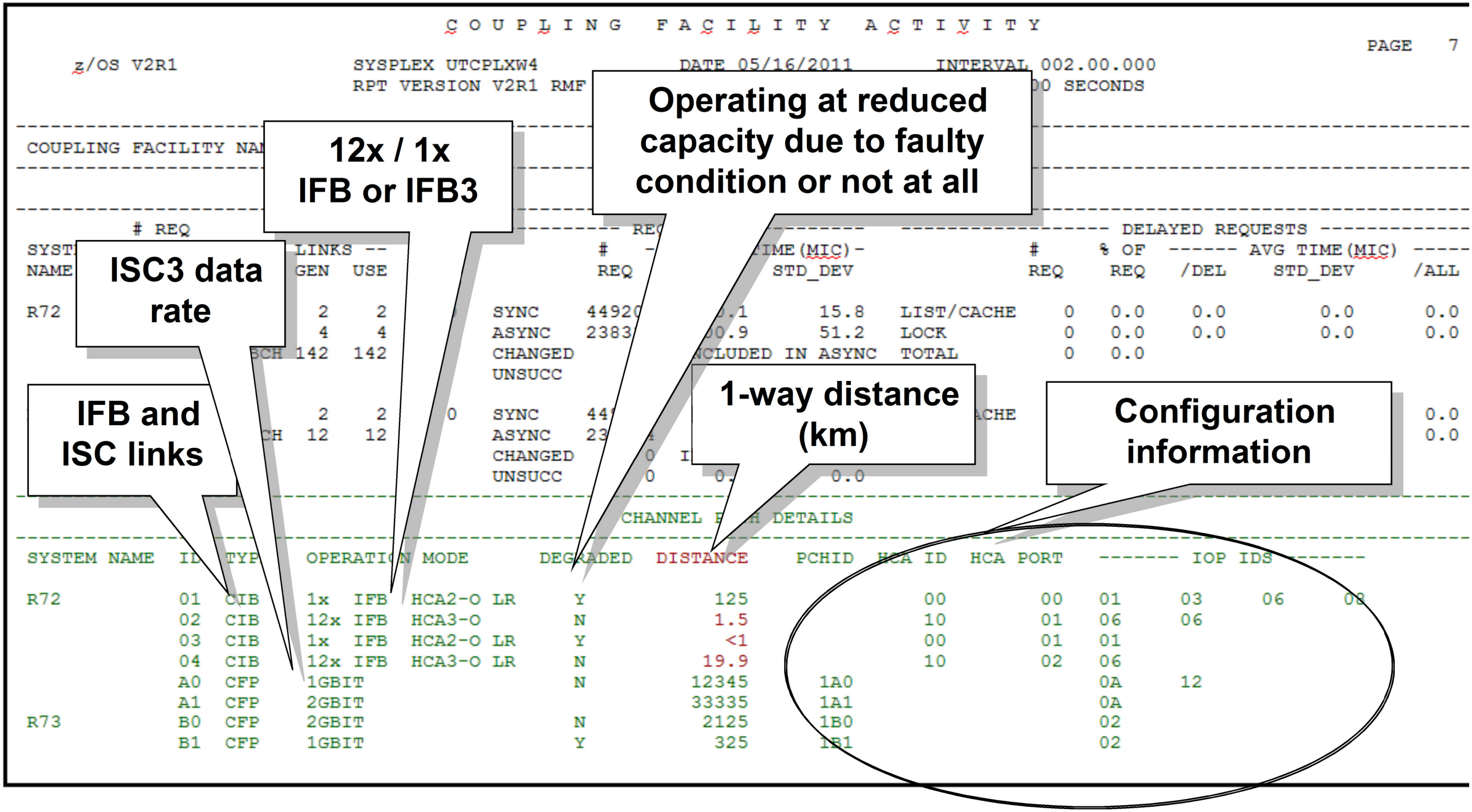
CFCC Level 18 – RAS Enhancements

- 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

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

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 NAM
------------------------------------------------------------------------------------------------------

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

R72                    2     2       SYNC    44920       .1       15.8   LIST/CACHE    0    0.0    0.0         0.0         0.0
                       4     4       ASYNC   2383       0.9       51.2   LOCK          0    0.0    0.0         0.0         0.0
                 SCH 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

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

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]
```

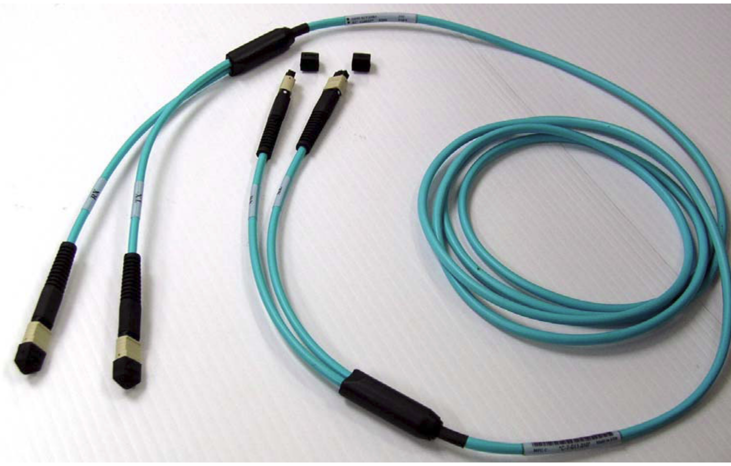
Whenever you migrate to a new CFLEVEL

- 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/

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



Coupling Link Choices - Overview

- **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
- **IC (Internal Coupling Channel / Internal CPC)**
  - Microcode - no external connection
  - Only between partitions on same processor

Statement of Direction

- 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.

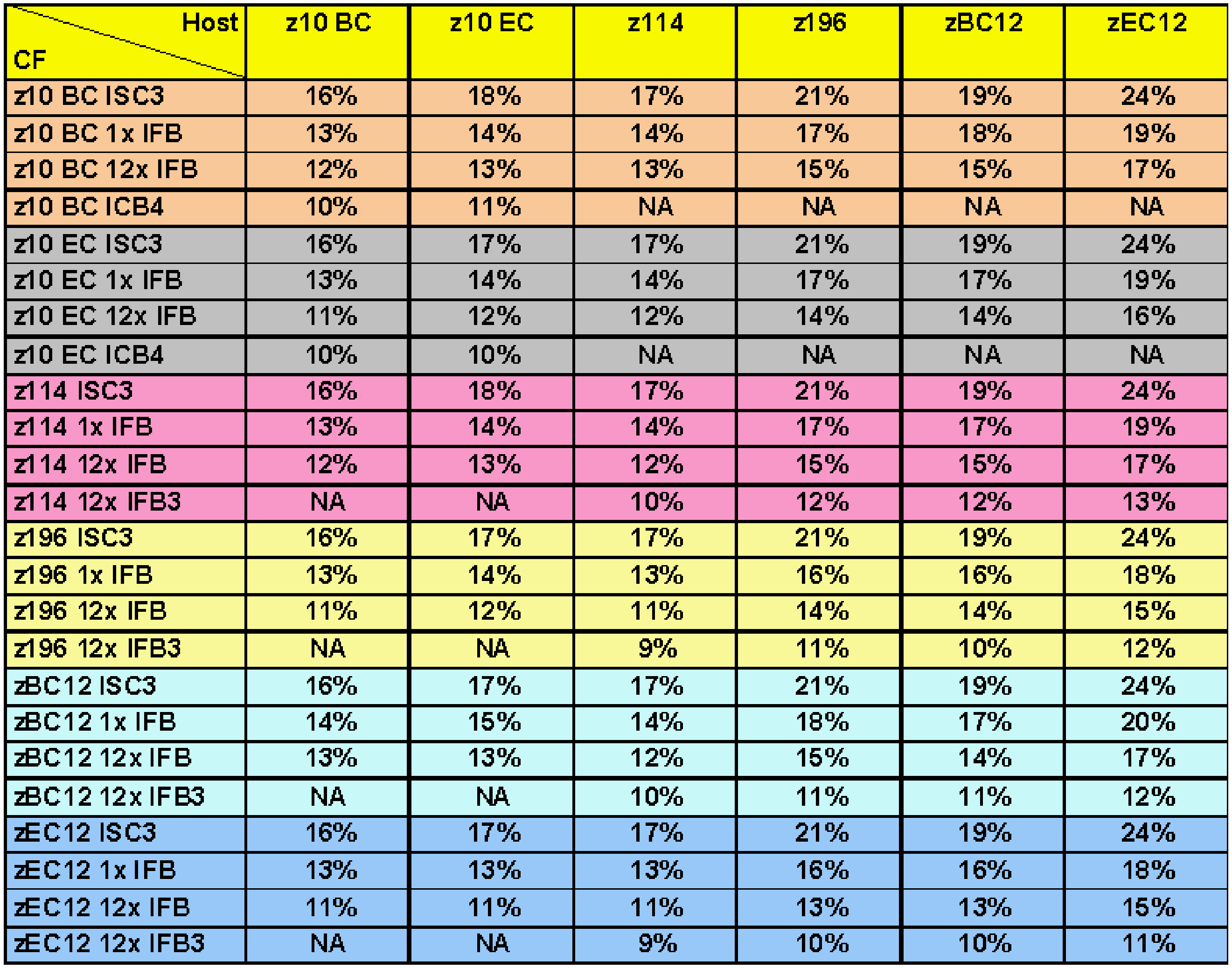
- Removal of support for the HCA2-O fanouts for 12x IFB and 1x IFB InfiniBand coupling links

  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).

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



Host CF	z10 BC	z10 EC	z114	z196	zBC12	zEC12
z10 BC ISC3	16%	18%	17%	21%	19%	24%
z10 BC 1x IFB	13%	14%	14%	17%	18%	19%
z10 BC 12x IFB	12%	13%	13%	15%	15%	17%
z10 BC ICB4	10%	11%	NA	NA	NA	NA
z10 EC ISC3	16%	17%	17%	21%	19%	24%
z10 EC 1x IFB	13%	14%	14%	17%	17%	19%
z10 EC 12x IFB	11%	12%	12%	14%	14%	16%
z10 EC ICB4	10%	10%	NA	NA	NA	NA
z114 ISC3	16%	18%	17%	21%	19%	24%
z114 1x IFB	13%	14%	14%	17%	17%	19%
z114 12x IFB	12%	13%	12%	15%	15%	17%
z114 12x IFB3	NA	NA	10%	12%	12%	13%
z196 ISC3	16%	17%	17%	21%	19%	24%
z196 1x IFB	13%	14%	13%	16%	16%	18%
z196 12x IFB	11%	12%	11%	14%	14%	15%
z196 12x IFB3	NA	NA	9%	11%	10%	12%
zBC12 ISC3	16%	17%	17%	21%	19%	24%
zBC12 1x IFB	14%	15%	14%	18%	17%	20%
zBC12 12x IFB	13%	13%	12%	15%	14%	17%
zBC12 12x IFB3	NA	NA	10%	11%	11%	12%
zEC12 ISC3	16%	17%	17%	21%	19%	24%
zEC12 1x IFB	13%	13%	13%	16%	16%	18%
zEC12 12x IFB	11%	11%	11%	13%	13%	15%
zEC12 12x IFB3	NA	NA	9%	10%	10%	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

System z – Maximum Coupling Links and CHPIDs

Server	1x IFB (HCA3-O LR)	12x IFB 12x IFB3 (HCA3-O)	1x IFB (HCA2-O LR*)	12x IFB (HCA2-O*)	IC	ICB-4	ISC-3*	Max External Links	Max Coupling CHPIDs
zEC12	64 H20 – 32* H43 – 64*	32 H20 – 16* H43 – 32*	32(4) H20 – 16* H43 – 32*	32(4) H20 – 16* H43 – 32*	32	N/A	48(4)	112(1) H20 – 72*(2) H43 – 104*(1)	128
zBC12	H13 – 32* H06 – 16*	H13 – 16* H06 – 8*	H13 – 16* H06 – 8*	H13 – 16* H06 – 8*	32	N/A	32 (4)	H13-72 (2) H06-56 (3)	128
z196	48 M15 – 32*	32 M15 – 16* M32 – 32*	32 M15 – 16* M32 – 32*	32 M15 – 16* M32 – 32*	32	N/A	48	104 (1) M15-72*(2) M32-100*(1)	128
z114	M10 – 32* M05 – 16*	M10 – 16* M05 – 8*	M10 – 12* M05 – 8*	M10 – 16* M05 – 8*	32	N/A	48	M10-72* (2) M05-56* (3)	128
z10 EC	N/A	N/A	32 E12 – 16*	32 E12 – 16**	32	16 (32/RPQ)	48	64	64
z10 BC	N/A	N/A	12	12	32	12	48	64	64

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

System z – Maximum Coupling Links and CHPIDs (notes)

(1) A zEC12 H66, H89 or HA1 supports a maximum 112 extended distance links (64 1x IFB and 48 ISC-3) with no 12x IFB links

A zEC12 H43 supports a maximum 104 extended distance links (56 1x IFB and 48 ISC-3) with no 12x IFB links or other I/O.

A zEC12 H20 or z196 M15 supports a maximum 72 extended distance links (24 1x IFB and 48 ISC-3) with no 12x IFB links or other I/O.

A z196 M49, M66 or M80 supports a maximum of 96 extended distance links (48 1x IFB and 48 ISC-3) with 8 12x IFB links

A z196 M32 supports a maximum of 96 extended distance links (48 1x IFB and 48 ISC-3) with 4 12x IFB links and no other I/O

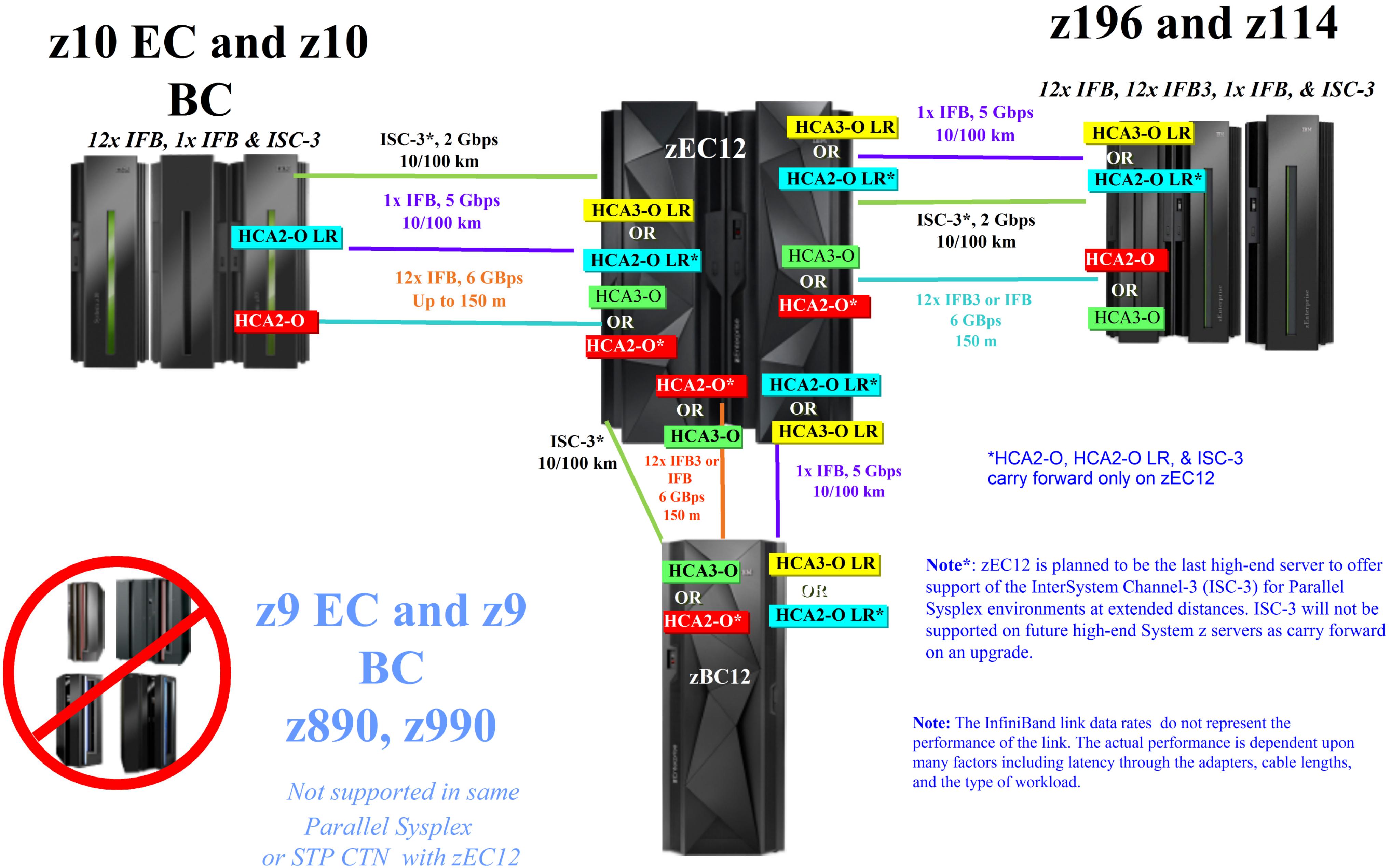
(2) zEC12 H20**, zBC12 H13,** z196 M15 or z114 M10 support a maximum of 72 extended distance links (24 1x IFB and 48 ISC-3) with no 12x IFB links or I/O.

(3) **zBC12 H06** or z114 M05 supports a maximum of 56 extended distance links (8 1x IFB and 48 ISC-3) with no 12x IFB links or I/O.

(4) zEC12 H20 and H43 and **zBC12 H06 and H13** support ISC-3, HCA2-O and HCA2-O LR as carry forward only, not on new build

zEC12 H89 and HA1 (only) support ISC-3 as carry forward and on new-build by RPQ when 16 PSIFB fanout features are also configured

zEC12/zBC12 Parallel Sysplex Coupling Connectivity



z10 EC and z10 BC

12x IFB, 1x IFB & ISC-3

HCA2-O LR

HCA2-O

ISC-3*, 2 Gbps
10/100 km

1x IFB, 5 Gbps
10/100 km

12x IFB, 6 GBps
Up to 150 m

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

z196 and z114

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

1x IFB, 5 Gbps
10/100 km

ISC-3*, 2 Gbps
10/100 km

12x IFB3 or IFB
6 GBps
150 m

HCA3-O LR
OR
HCA2-O LR*

HCA2-O
OR
HCA3-O

ISC-3*
10/100 km

12x IFB3 or IFB
6 GBps
150 m

1x IFB, 5 Gbps
10/100 km

zBC12

HCA3-O
OR
HCA2-O*

HCA3-O LR
OR
HCA2-O LR*

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

z9 EC and z9 BC
z890, z990

Not supported in same Parallel Sysplex or STP CTN with 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.

For More Information

- “IBM System z Connectivity Handbook” (SG24-5444)
- “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

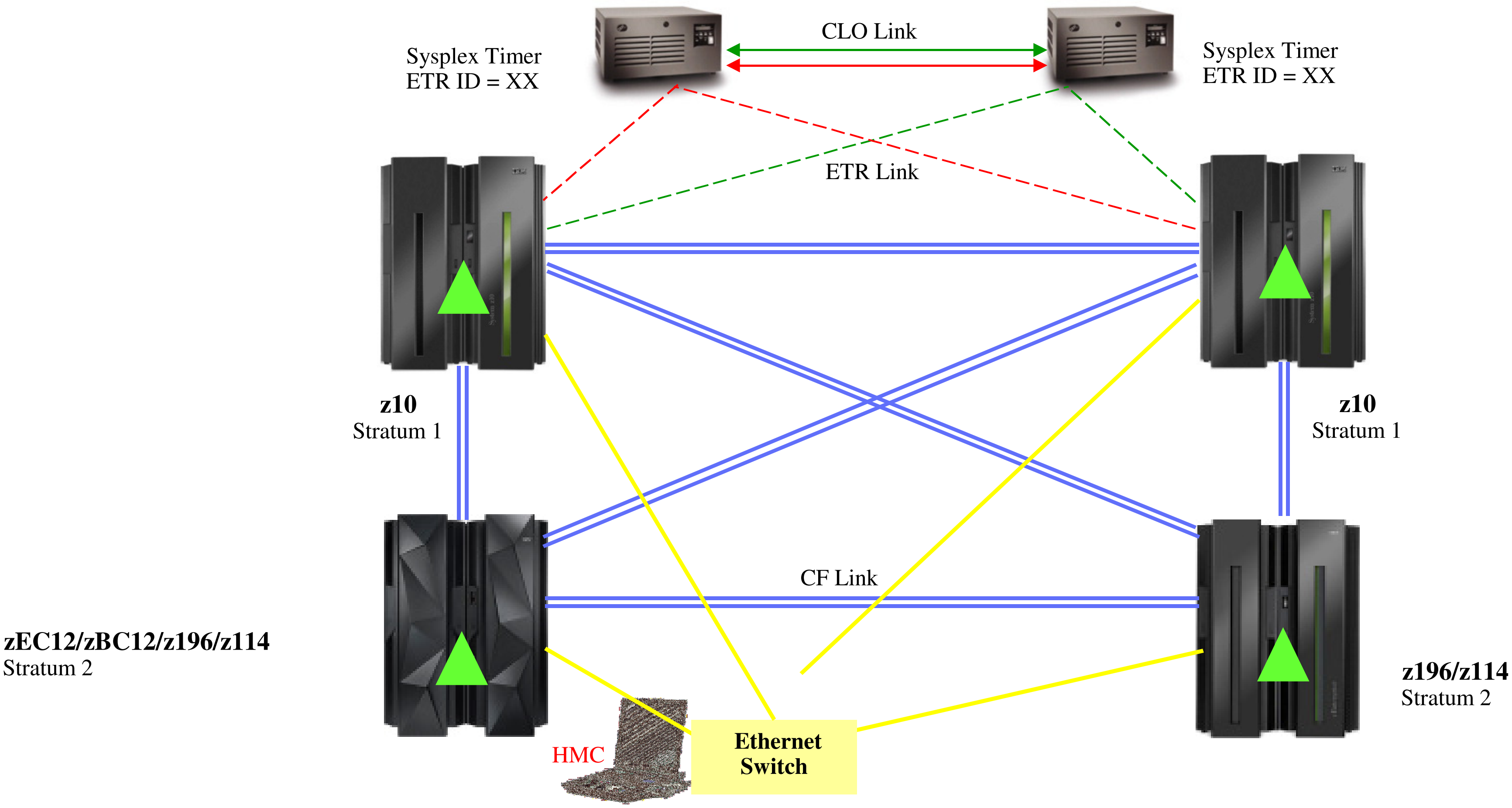
Agenda

- 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

Glossary for System z Server Time Protocol (STP)

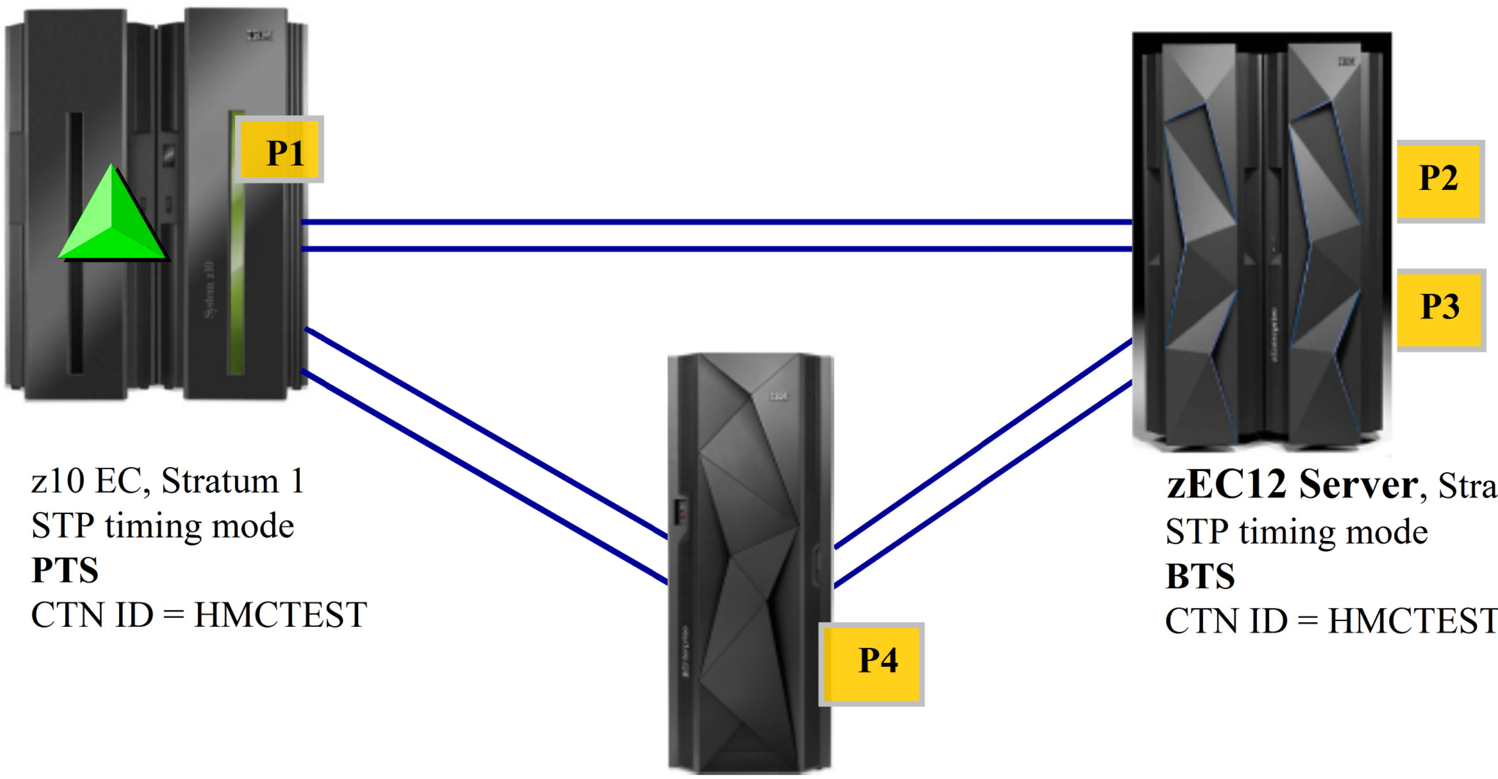
Acronym	Full name	Comments
Arbiter	Arbiter	Server assigned by the customer to provide additional means for the Backup Time Server to determine whether it should take over as the Current Time Server.
BTS	Backup Time Server	Server assigned by the customer to take over as the Current Time Server (stratum 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.

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



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.

STP-only CTN Example with System zEC12/zBC12 Servers



z10 EC, Stratum 1
STP timing mode
PTS
CTN ID = HMCTEST

zEC12 Server, Stratum 2
STP timing mode
BTS
CTN ID = HMCTEST

zBC12, Stratum 2
STP timing mode
Arbiter
CTN ID = HMCTEST

Server Time Protocol Enhancements

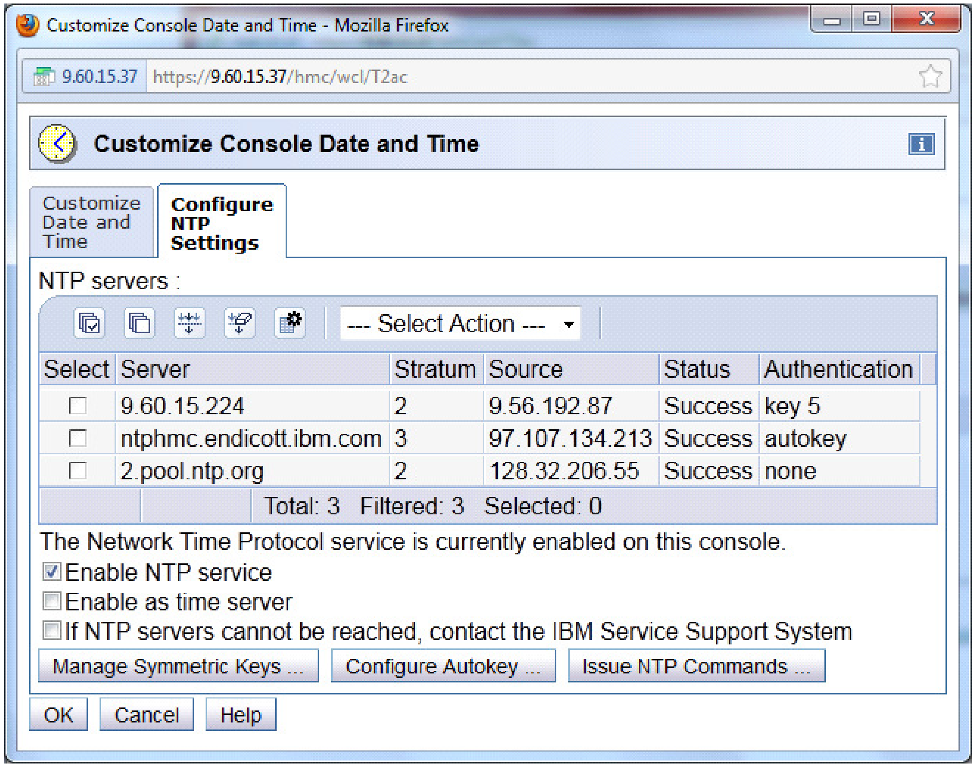
- Improved SE Time Accuracy – zEC12 GA2 and zBC12
  - Optionally, the SE can be configured to connect to an external time source periodically to maintain highly accurate time that can be used, if required, to initialize CTN time during POR.
- Broadband Security Improvements for STP
  - Authenticates 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

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)

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.

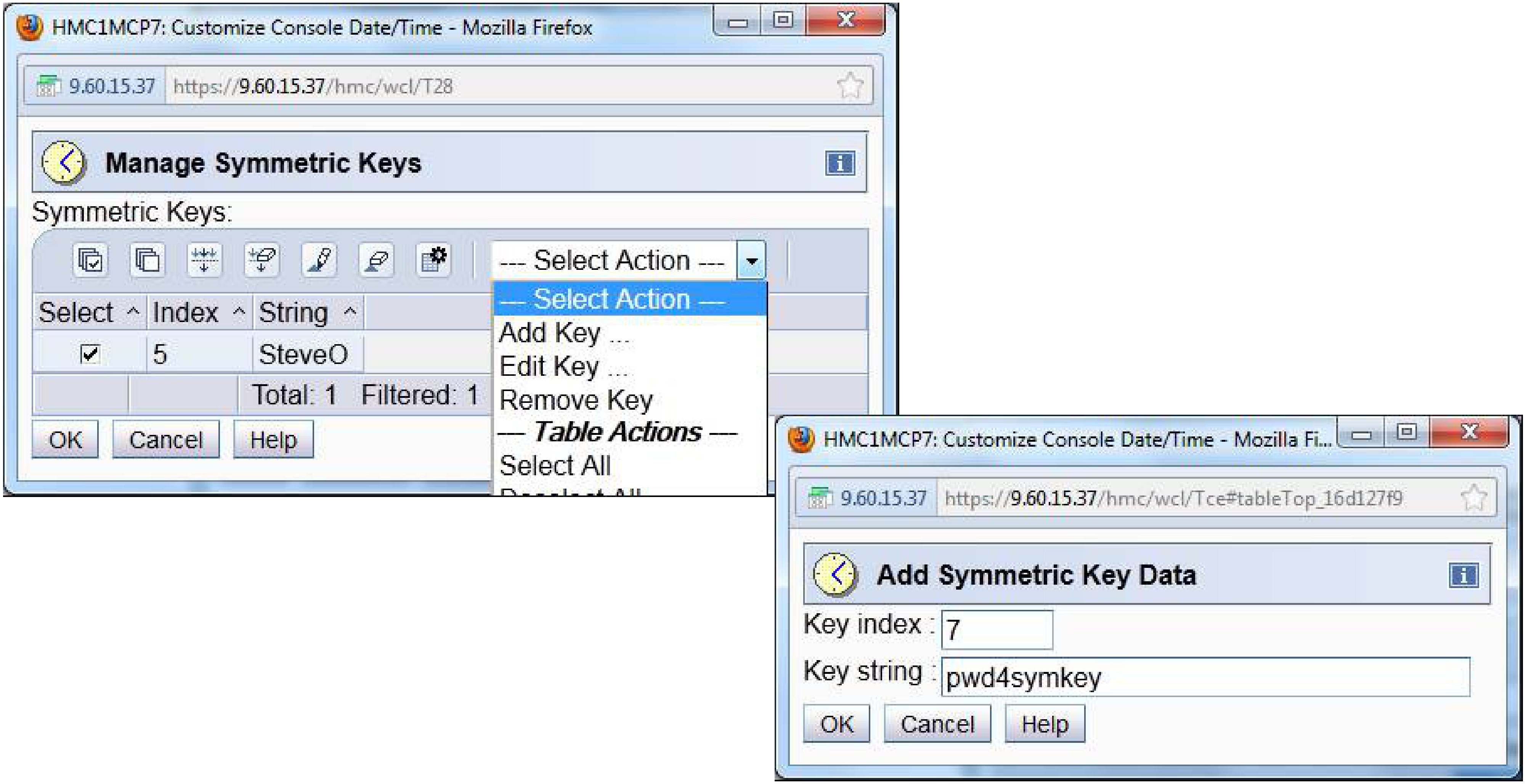
☑ Enable NTP service

□ Enable as time server

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

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

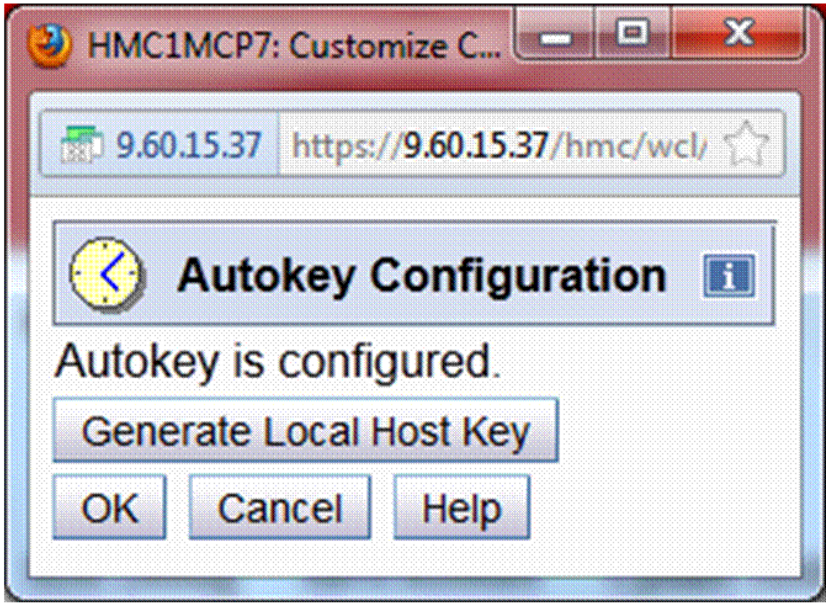
Symmetric Keys:

Select	Index	String
☑	5	SteveO
		Total: 1 Filtered: 1

Key index : 7

Key string : pwd4symkey

NTP Authentication - AutoKey



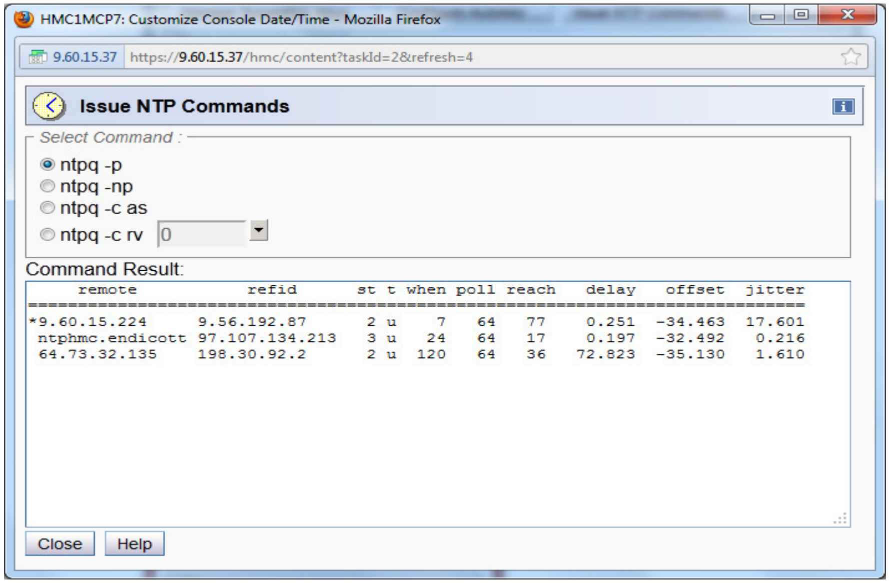
Autokey Configuration

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 Authentication – Issue NTP Commands Panel

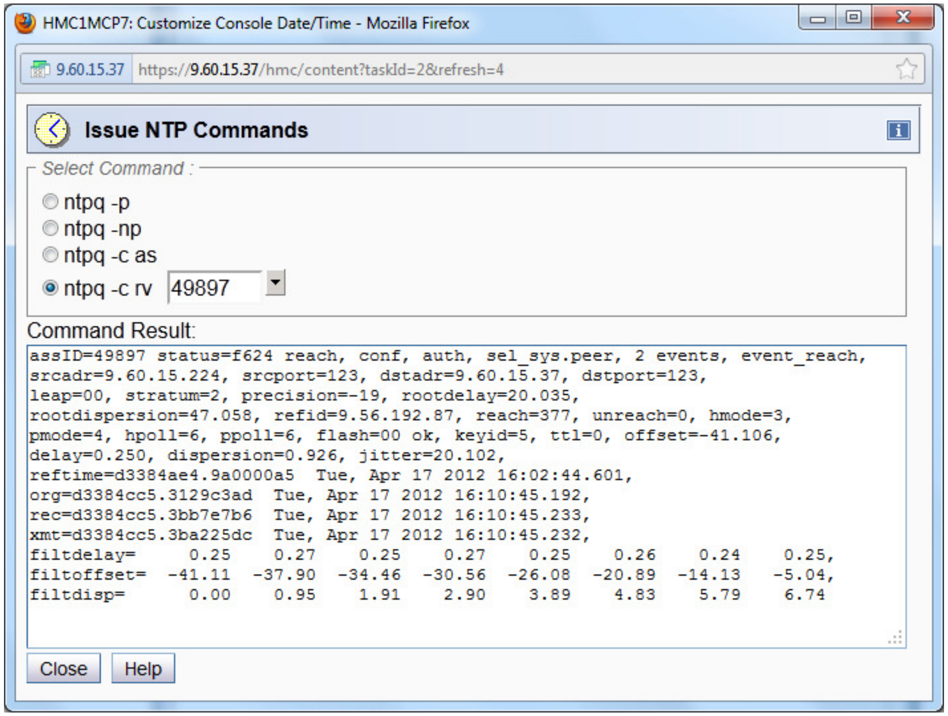
- One customer complaint in regard to the HMC NTP server panels, as they stand today, has been the status of the connection to target NTP servers.
- With the addition of NTP authentication, this display will aid in the determination of failures during configuration.


HMC1MCP7: Customize Console Date/Time - Mozilla Firefox
9.60.15.37 https://9.60.15.37/hmc/content?taskId=2&refresh=4
Issue NTP Commands
Select Command :
Close
Help

● 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=2&refresh=4
Issue NTP Commands
Select Command :
Close
Help

○ 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

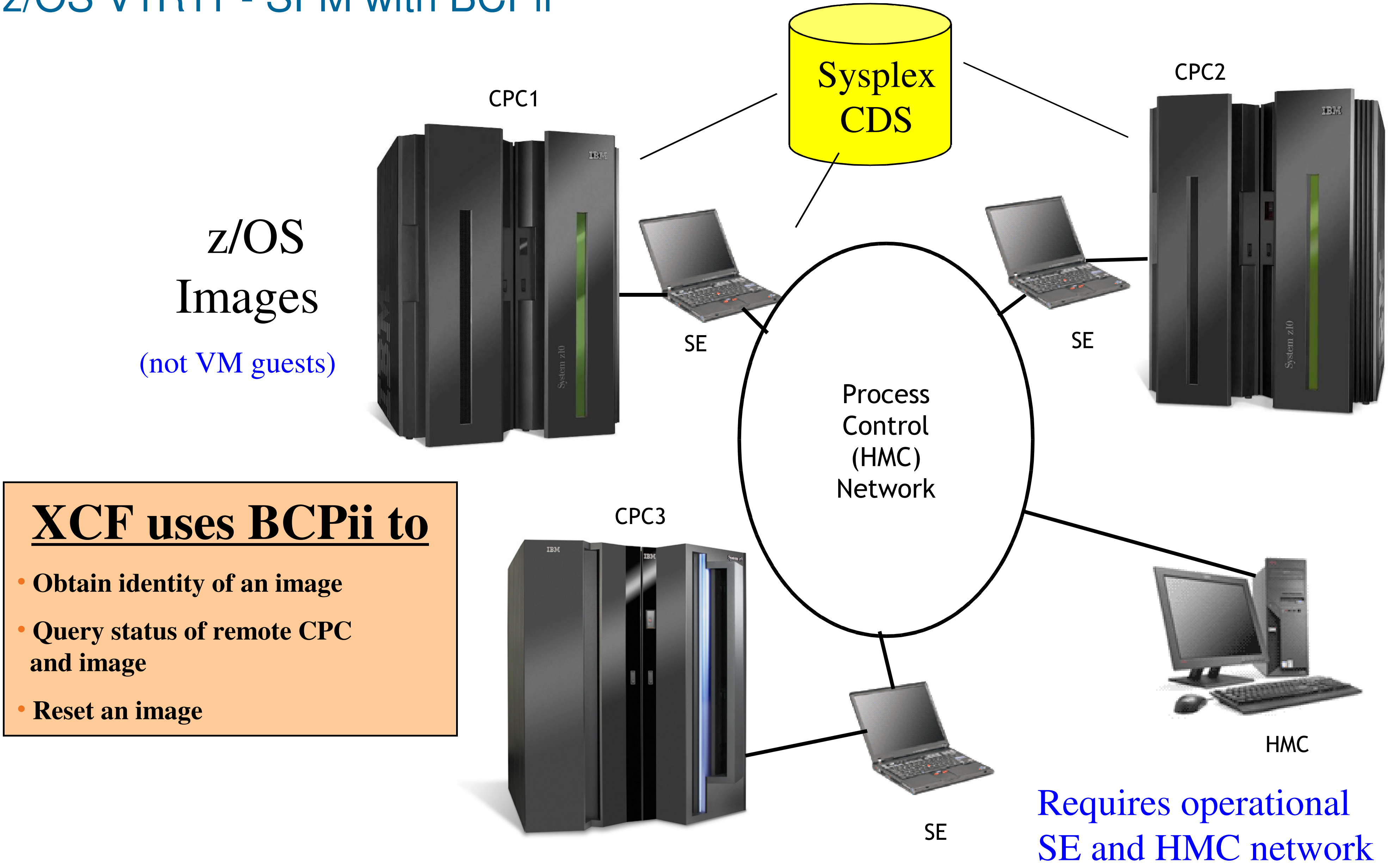
STP References for Additional Information

- **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

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

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

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”

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
  - **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 !

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.

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”

APAR OA41203 ...

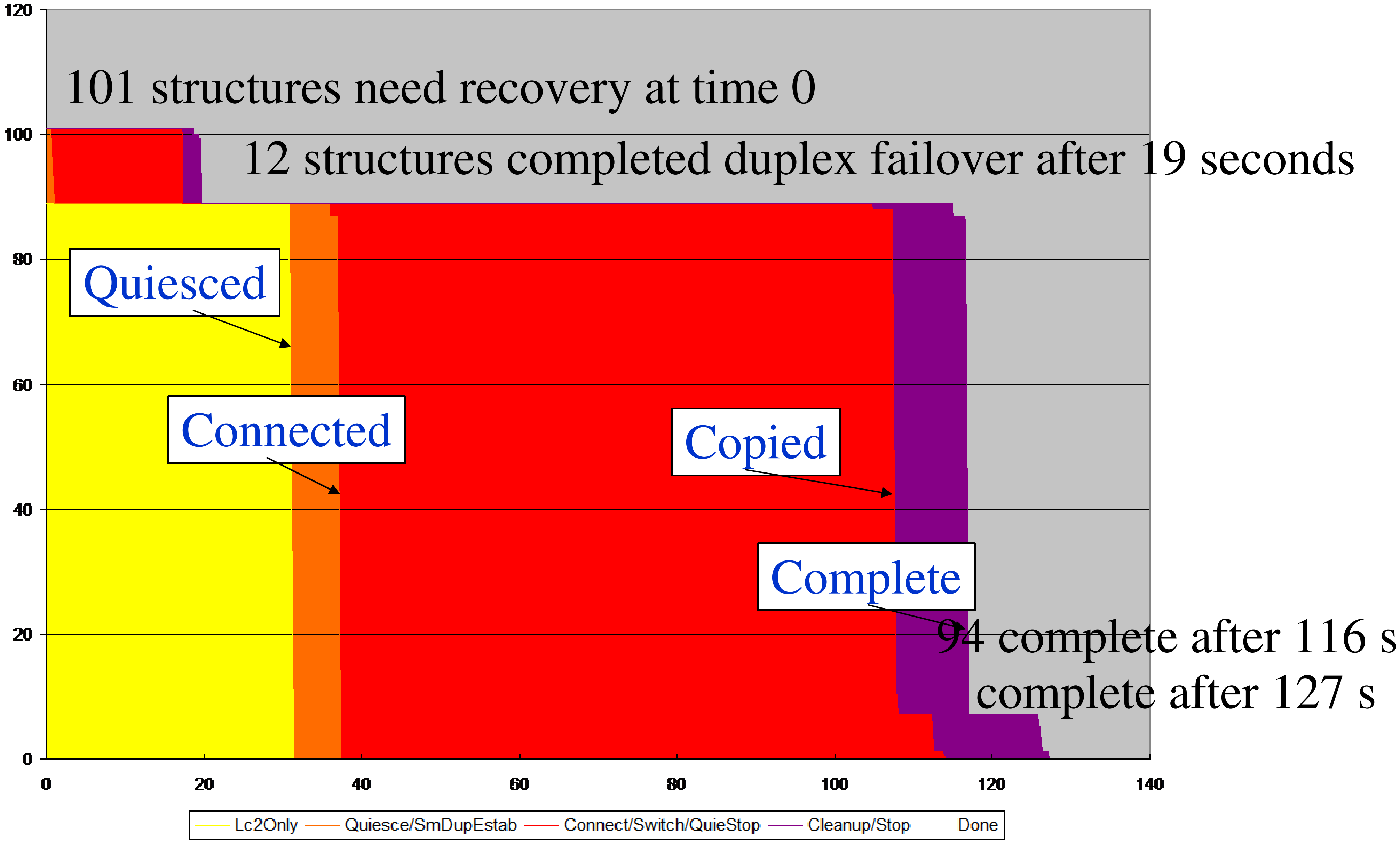
- Also reworked XES selection algorithm for queued CF requests
  - In effect as soon as the APAR is installed
  - Independent of the the CFSTRQMON switch
- Old: FIFO
  - A spike in requests for one structure induces delay for unrelated structures
  - Delay is a function of the number of predecessor requests
- 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

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 …

Parallel Rebuild - Test Results
Policy Based Event Processing

Not formal performance test measurements



101 structures need recovery at time 0

12 structures completed duplex failover after 19 seconds

Quiesced

Connected

Copied

Complete

94 complete after 116 s

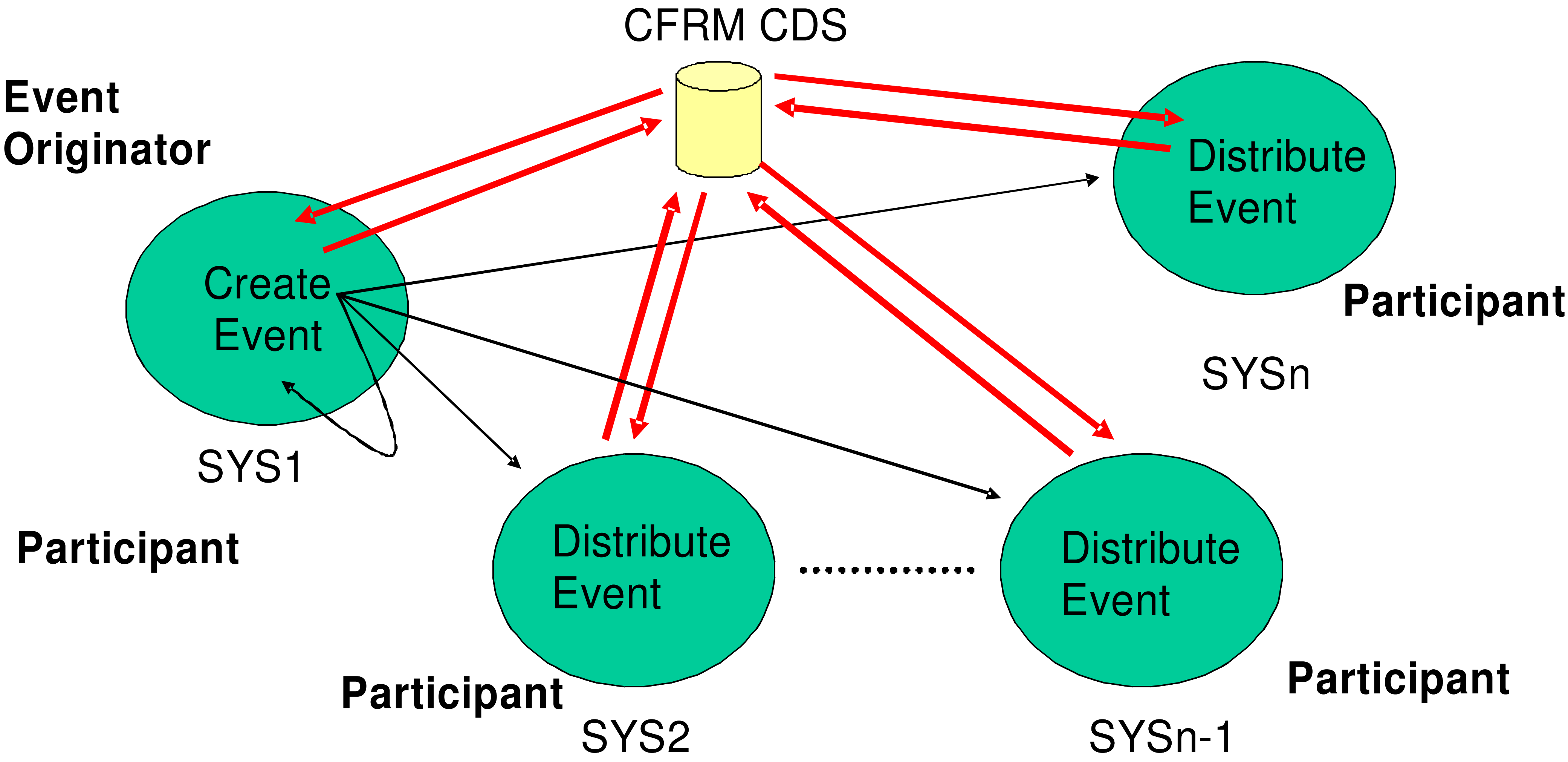
complete after 127 s

Policy Based Event Processing

→ CDS I/O

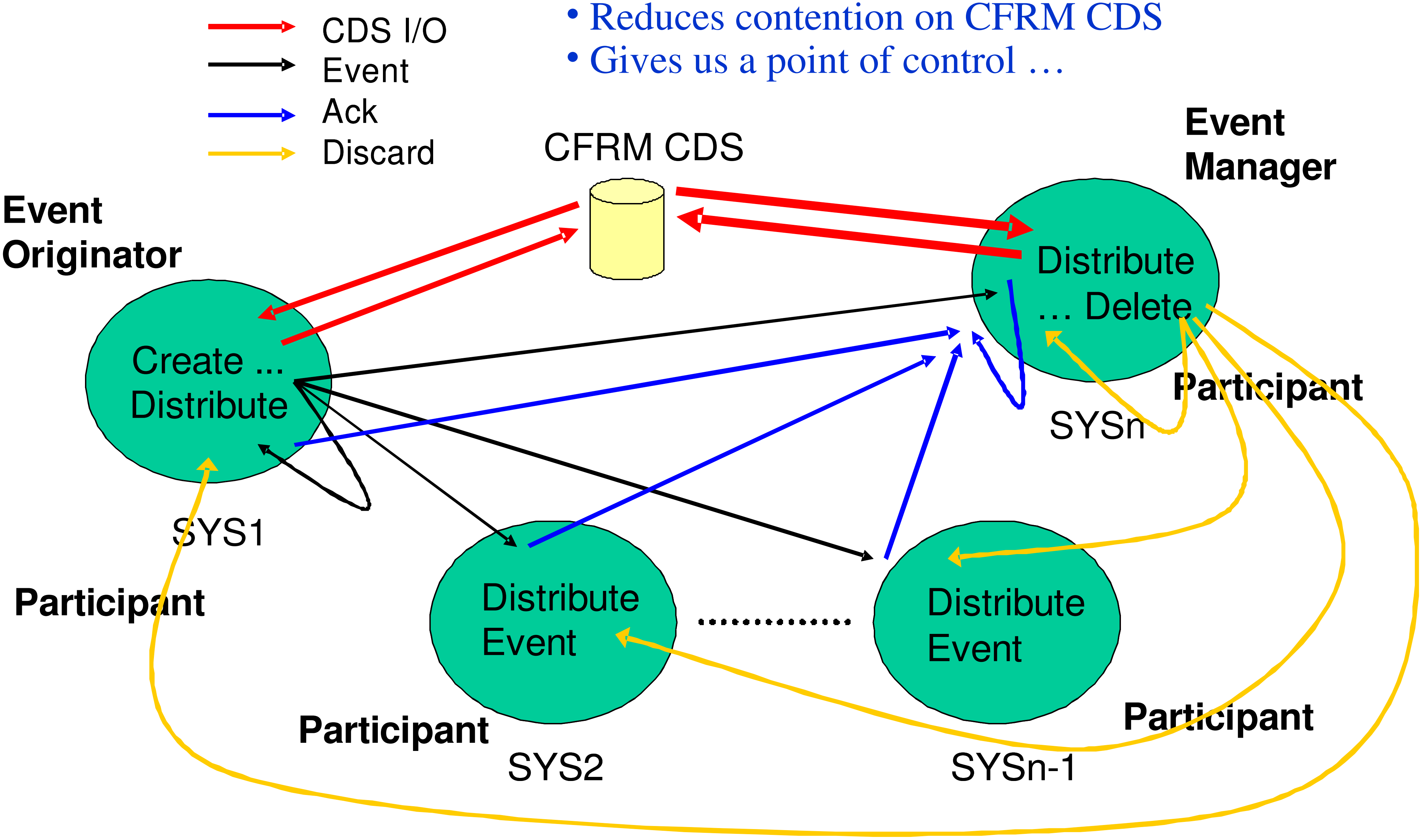
→ GAT

- Lots of contention on CFRM CDS
- Systems work independently



Message Based Event Processing

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

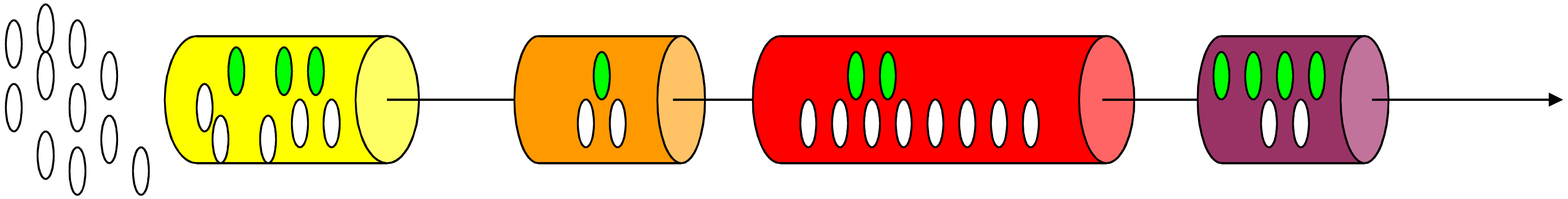


Serial Rebuild Improves Availability by Reducing MTTR

- 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

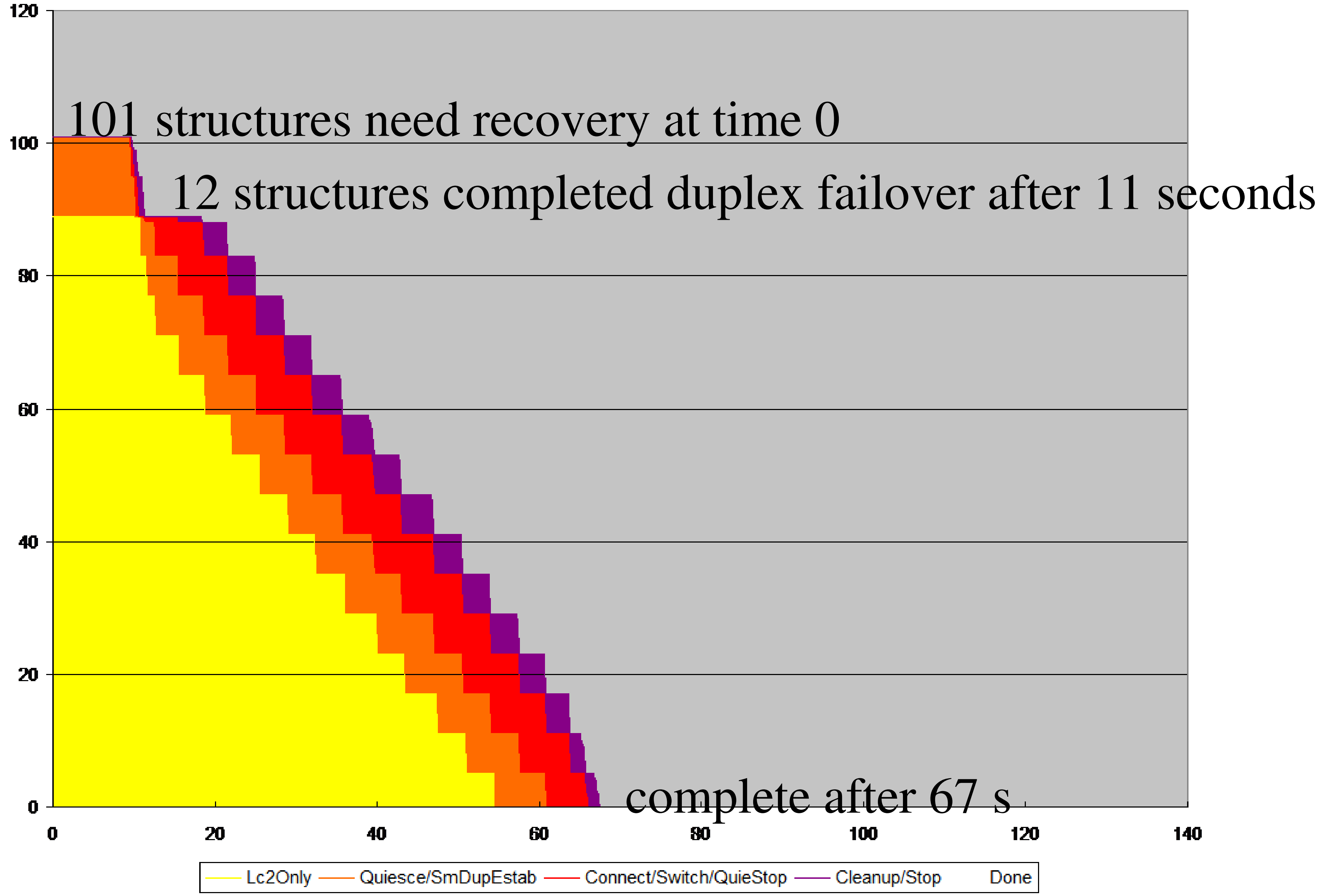
Serial Rebuild – New Behavior

- 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”



Serial Rebuild - Test Results

Not formal performance test measurements



101 structures need recovery at time 0

12 structures completed duplex failover after 11 seconds

complete after 67 s

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

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

Serial Rebuild – Exploitation

- All systems in the sysplex must be z/OS V2R1
- 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”

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

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

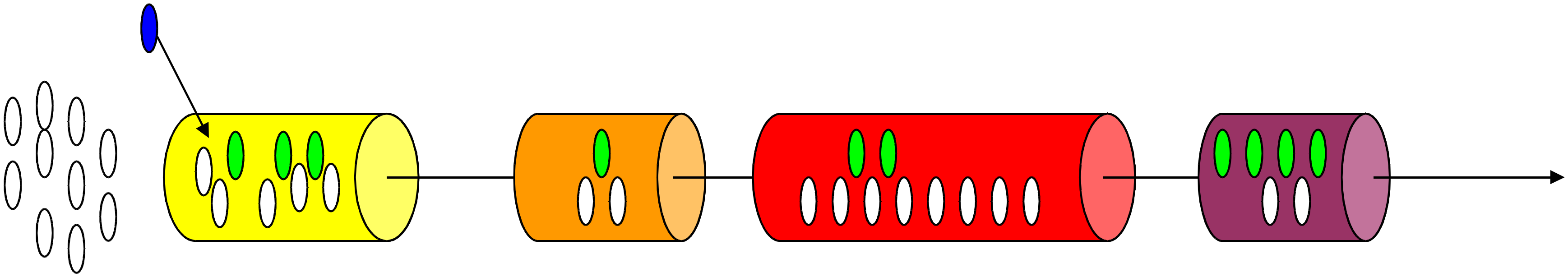
- Work of lesser importance is deferred if there is more important work …

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

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



Recent results from more customer like test environment

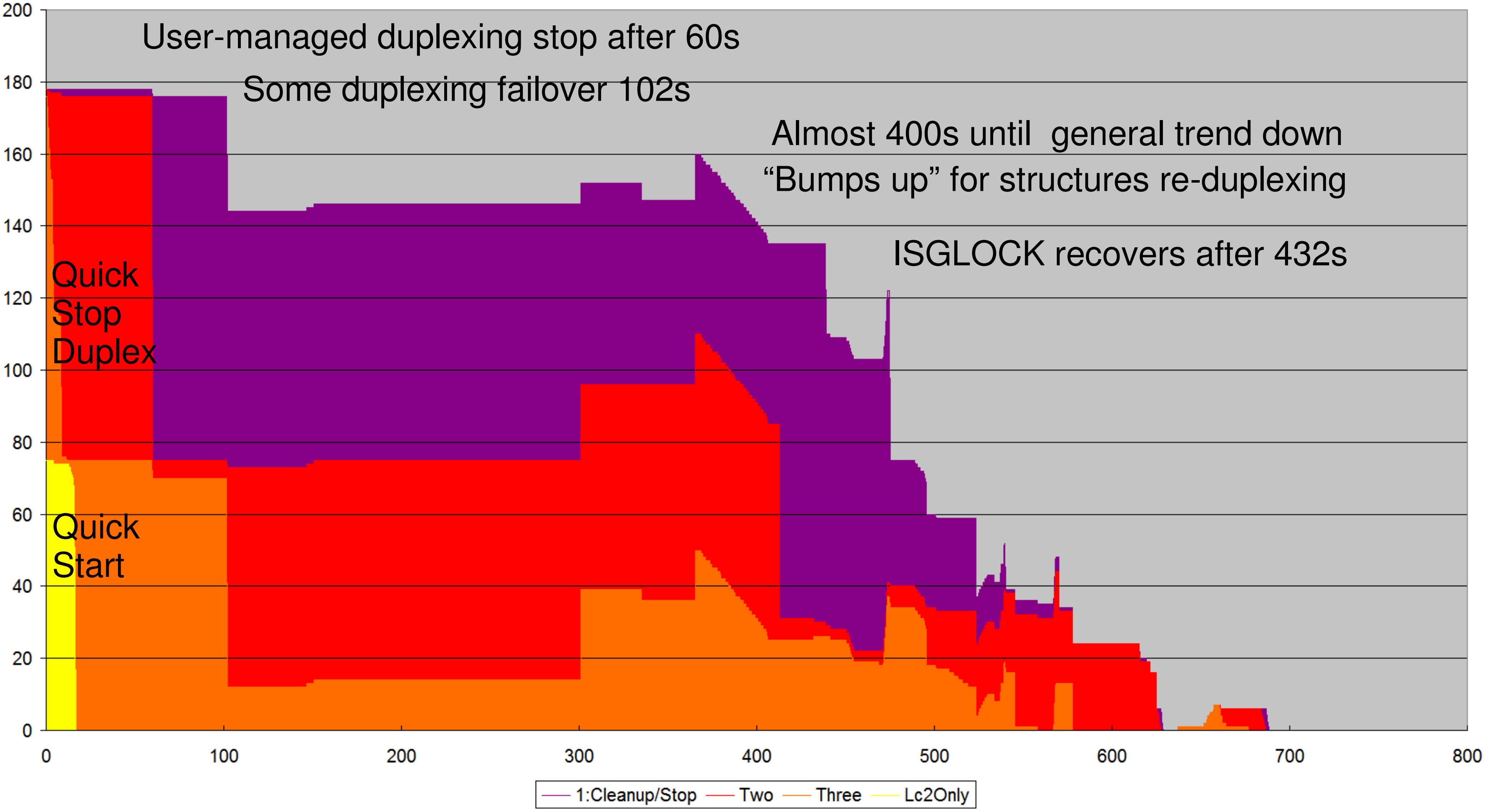
- Roughly 180 structures
- “Bouncing the CF”
  - Restored to service about 2 minutes into the test

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

- But looks promising ...

CFLCRMGMT DISABLED

Not formal performance test measurements



User-managed duplexing stop after 60s

Some duplexing failover 102s

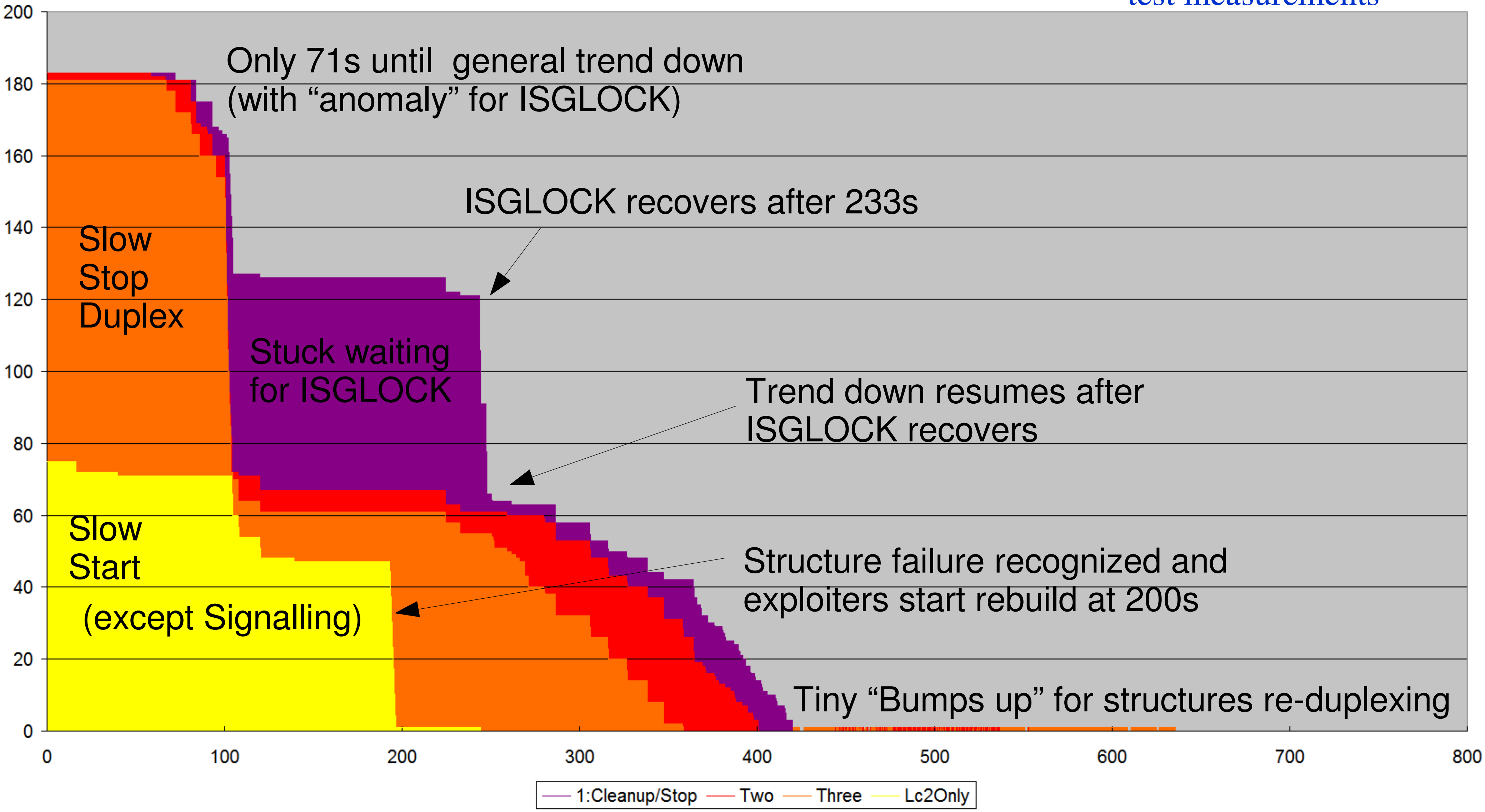
Almost 400s until general trend down

“Bumps up” for structures re-duplexing

ISGLOCK recovers after 432s

Serial Rebuild (CFLCRMGMT ENABLED)

Not formal performance test measurements



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

ISGLOCK recovers after 233s

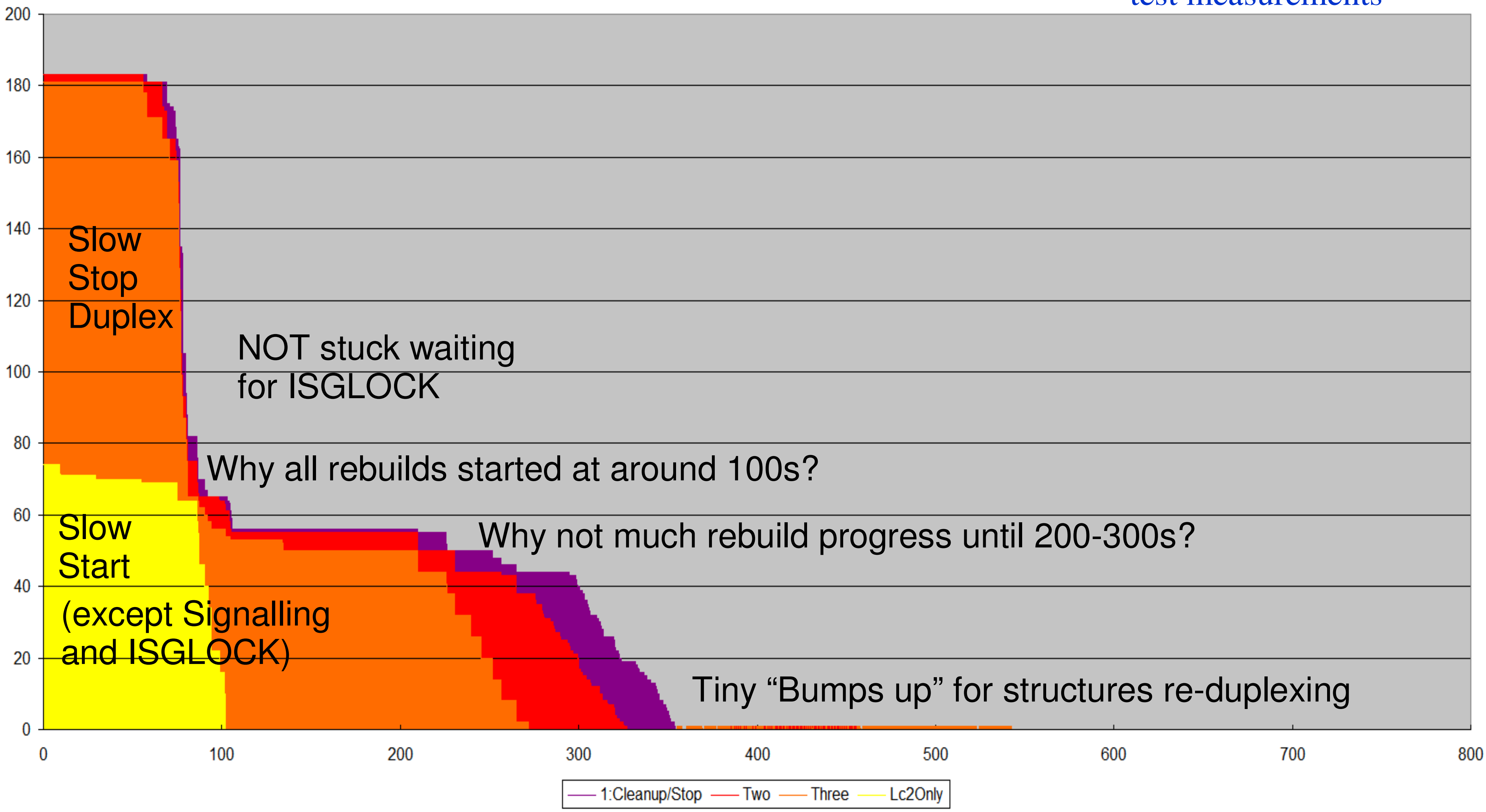
Trend down resumes after ISGLOCK recovers

Structure failure recognized and exploiters start rebuild at 200s

Tiny “Bumps up” for structures re-duplexing

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?

Slow Start

Why not much rebuild progress until 200-300s?

(except Signalling and ISGLOCK)

Tiny “Bumps up” for structures re-duplexing

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

Comparing these particular runs ...

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

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

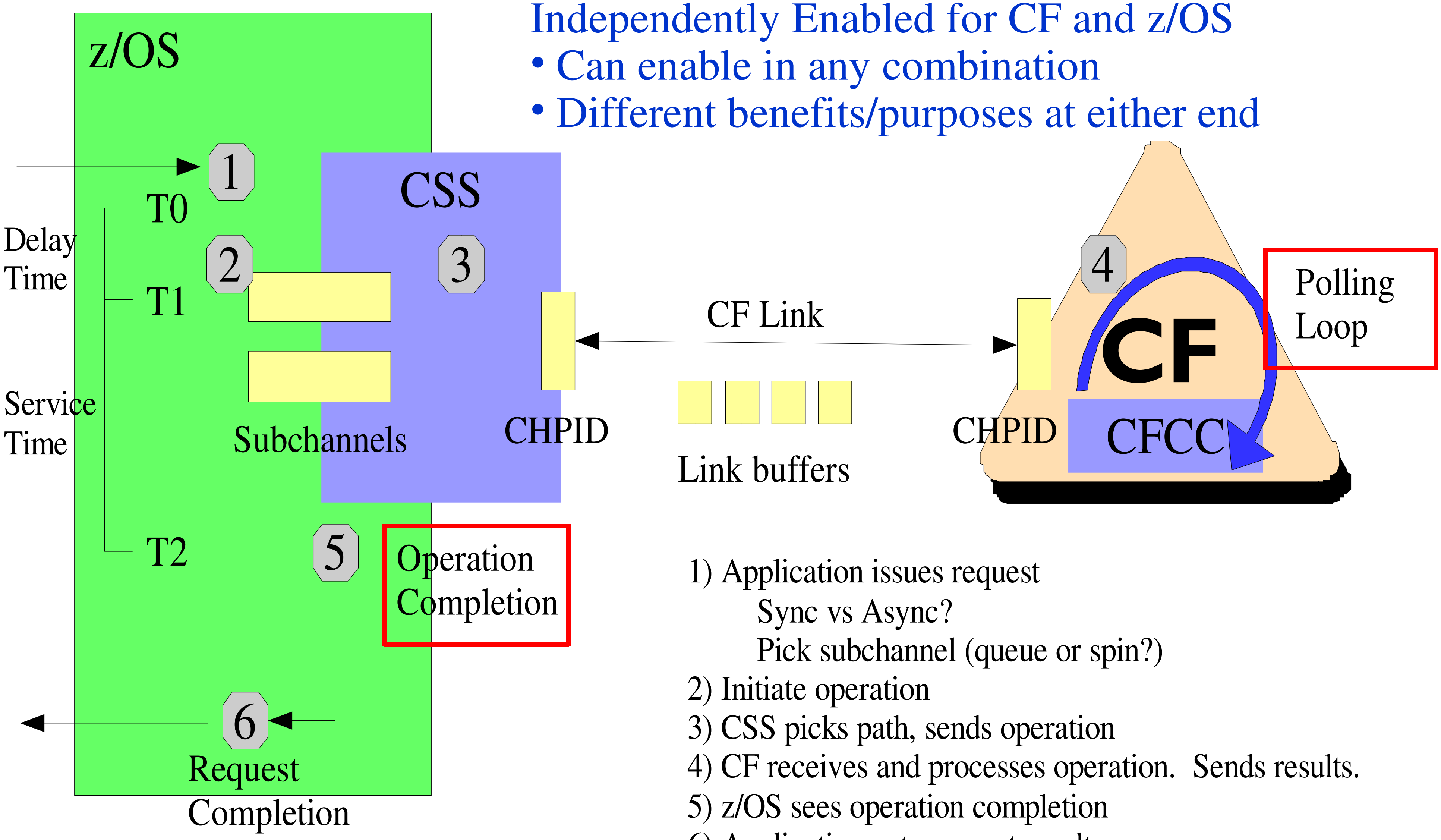
Should help any exploiter with significant async request activity or list notifications

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

z/OS V2R1 – Thin Interrupts

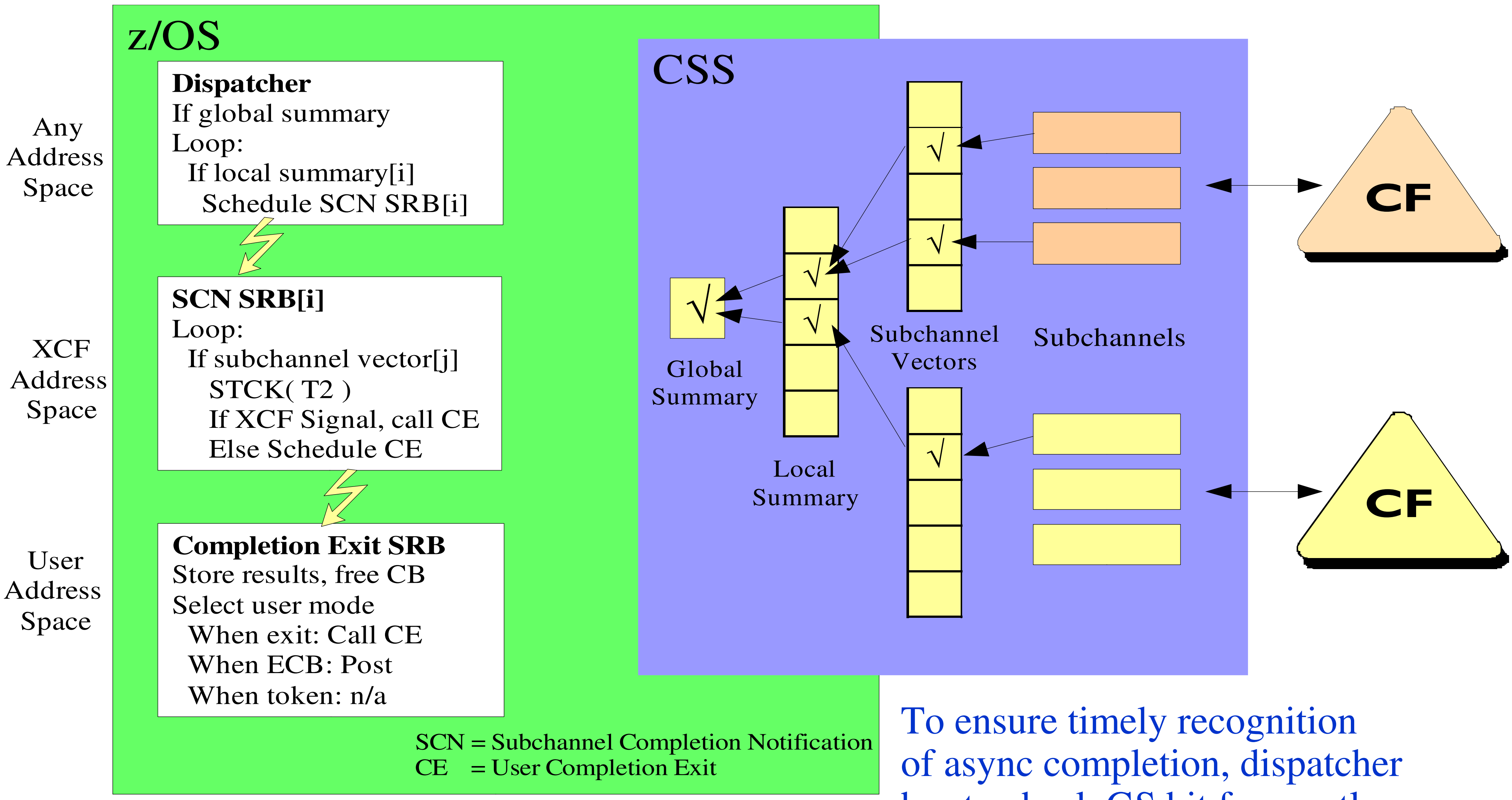
Independently Enabled for CF and z/OS

- 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

Asynchronous Operation Completion - Today



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

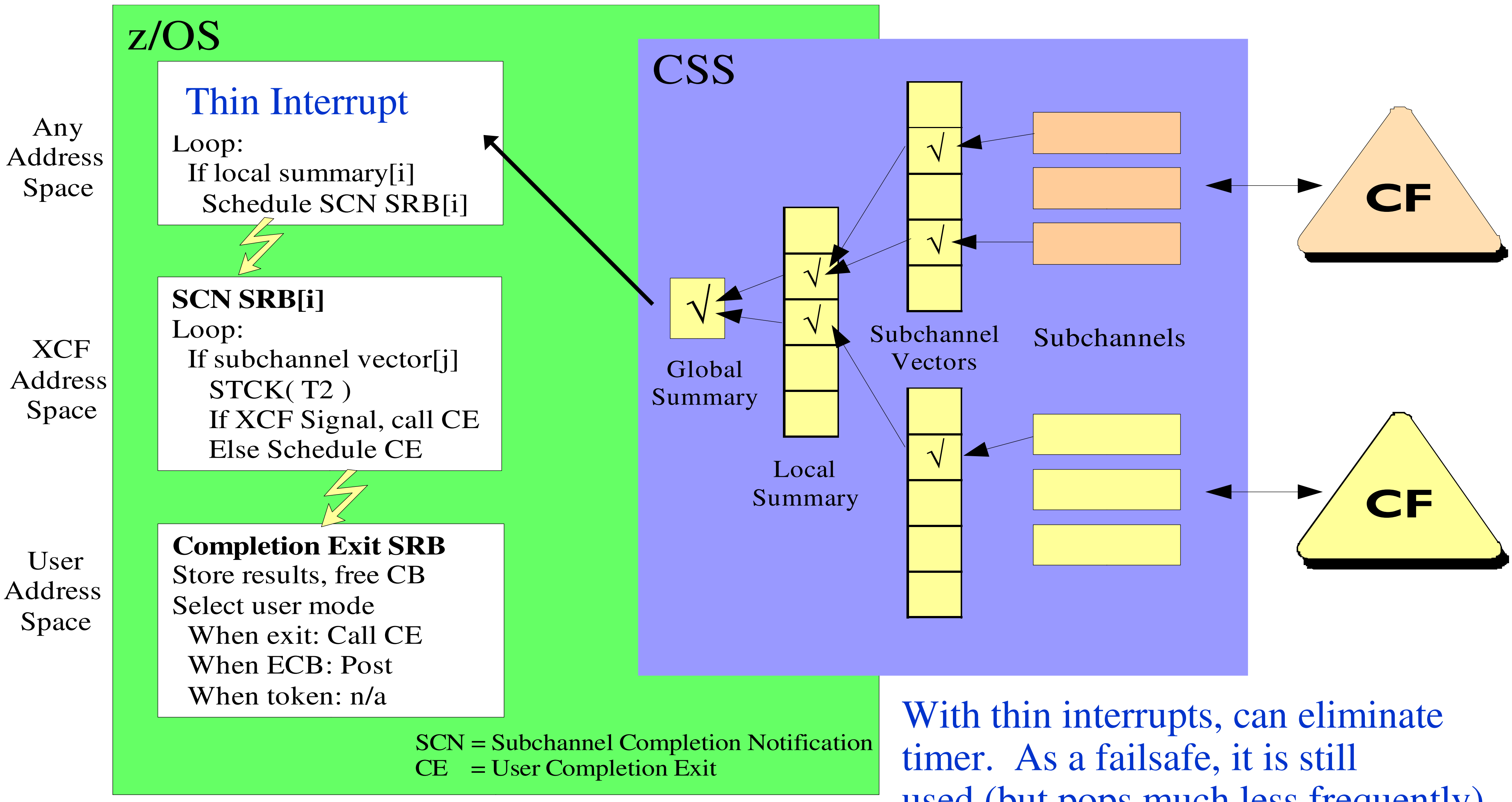
```
SCN SRB[i]
Loop:
  If subchannel vector[j]
    STCK( T2 )
    If XCF Signal, call CE
    Else Schedule CE
```

```
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

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).

Thin Interrupts for z/OS - Configuration

- Software Requirements (for z/OS exploitation)
  - z/OS V2R1, 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 or BC12 for z/OS coupling thin interrupts

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

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

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

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

Setting Conversion Threshold

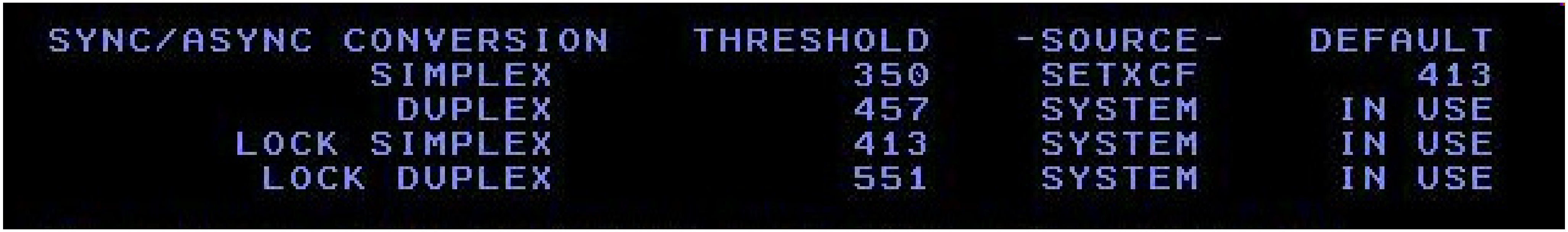
- COUPLExx parmlib member
  - On new SYNCASYNC statement specify: keyword(value)
- SETXCF MODIFY,SYNCASYNC,keyword=value
- “keyword” is one of the following thresholds
  - SIMPLEX - for simplex list and cache requests
  - DUPLEX - for duplexed list and cache requests
  - LOCKSIMPLEX - for simplex lock 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

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.

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

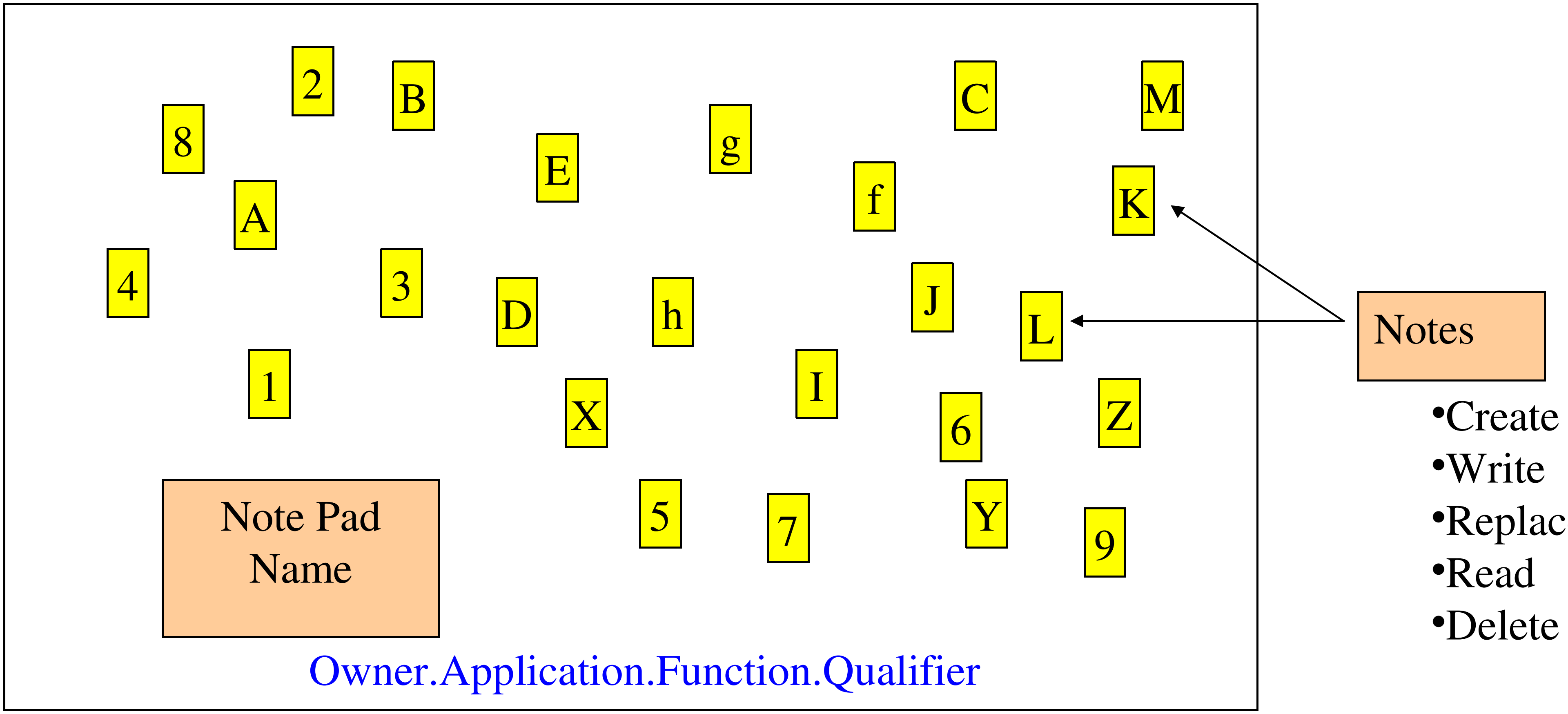
How was it set?

Current system default if not already in use

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

Abstract View of an XCF Note Pad



Note Pad Name

Owner.Application.Function.Qualifier

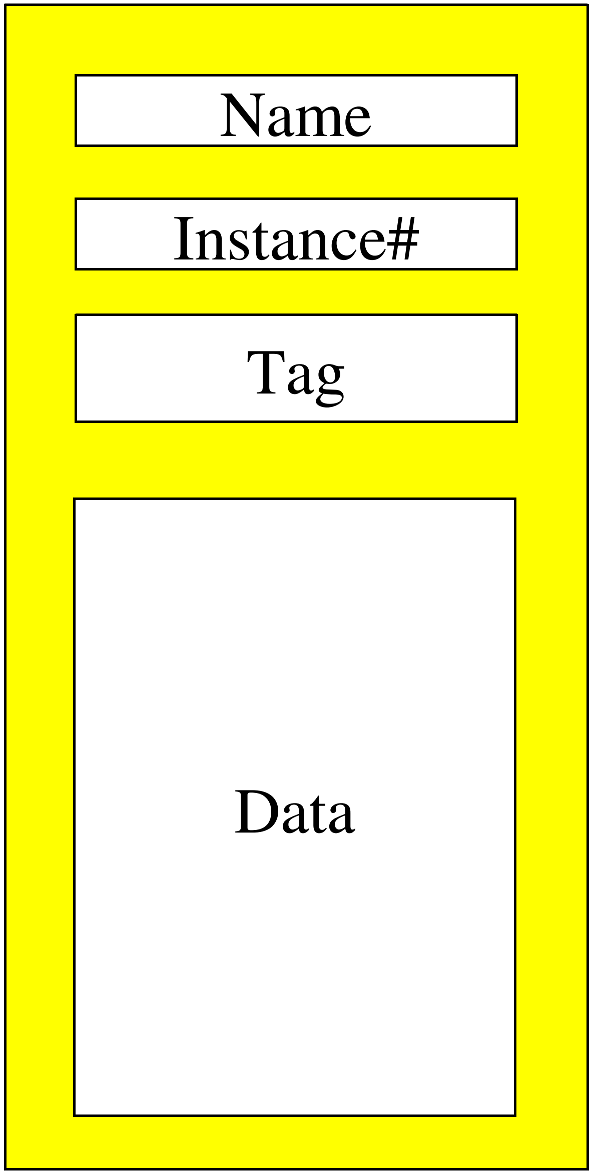
Notes

- Create
- Write
- Replace
- Read
- Delete

Note Pad

- Create
- Query
- Delete

Note in a Note Pad



Name

Instance#

Tag

Data

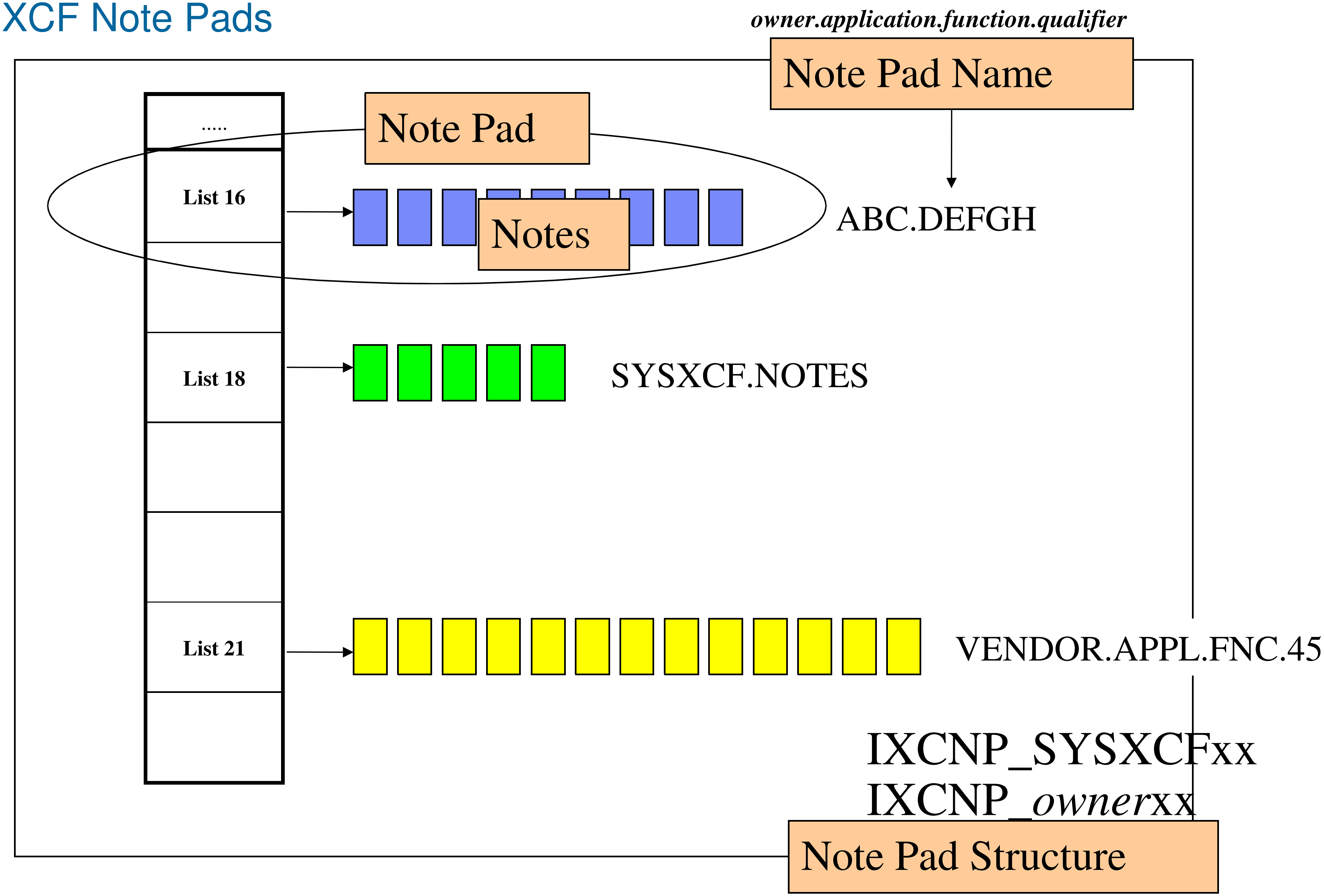
8 byte user note name

8 byte XCF instance number

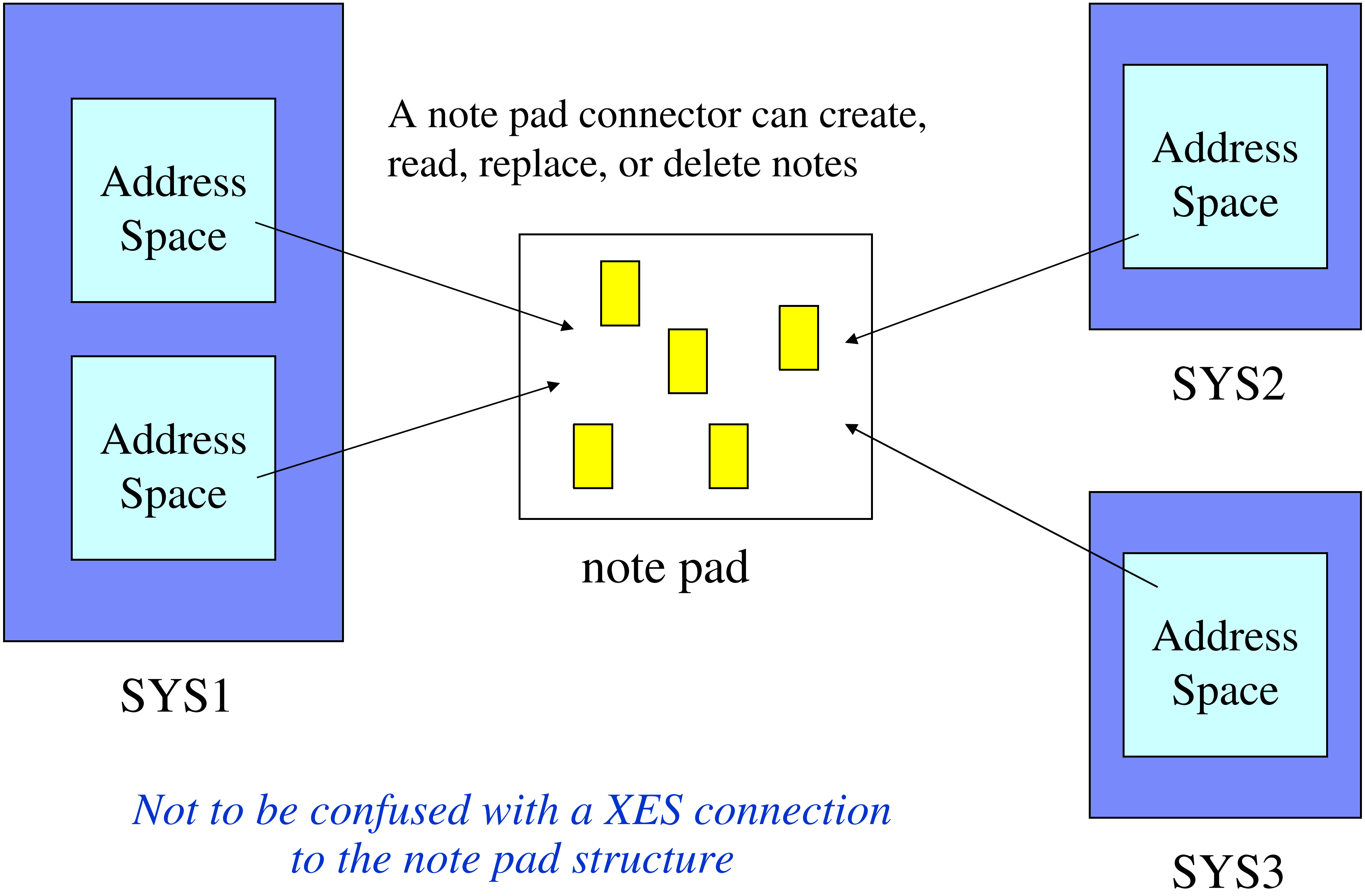
16 bytes of user metadata

1024 bytes of user data
(or none)

XCF Note Pads

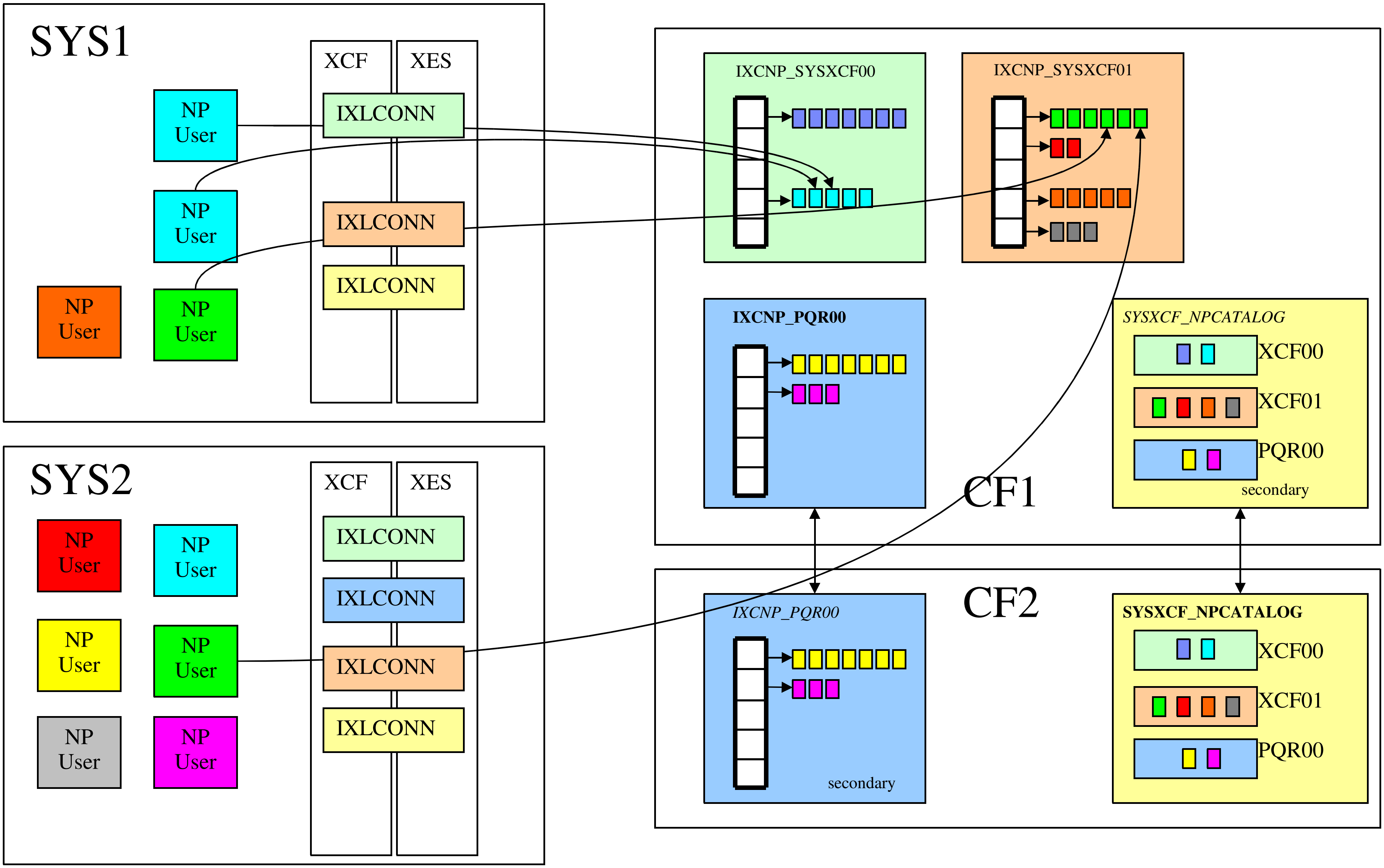


Connections to a Note Pad



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

XCF Note Pads in the Sysplex



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

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

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.
```

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
  - –Roughly 4X improvement over prior releases
  - –Assuming the target member can keep up

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

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 ….

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

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

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

z/OS V1R12 Summary

▪ REALLOCATE	– TEST and REPORT options
▪ Critical Members	– handle sympathy sickness
▪ CFSTRHANGTIME	– handle sympathy sickness
▪ Support for CFLEVEL 17	– Larger, more structures
▪ Health Checks	– Related to XCF, CF
▪ Auto Reply	– Automate during IPL
▪ New XCF APIs	– 64 bit msgs; join parms

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

Highlights

- CFLEVEL 19 for zEC12 GA2, zBC12
- CFLEVEL 18 for zEC12 GA1
  - Serviceability and performance improvements
- STP
  - Enhanced security
- Thin Interrupts:
  - On CF for configuration flexibility
  - On z/OS for better async CF service time
- Serial Rebuild for better MTTR

Complete your session evaluations online at
www.SHARE.org/Anaheim-Eval

z/OS Parallel Sysplex Update

Session 15105

If you have questions or comments,
feel free to contact me:
mabrook@us.ibm.com

z/OS Publications

- *MVS Setting Up a Sysplex*
- *MVS Initialization and Tuning*
- *MVS Systems Commands*
- *MVS Diagnosis: Tools and Service Aids*
- *z/OS V2R1 Migration*
- *z/OS V2R1 Planning for Installation*
- *z/OS MVS Programming: Callable Services for High Level Languages*
  - Documents BCPii Setup and Installation and BCPii APIs

Sysplex-related Redbooks

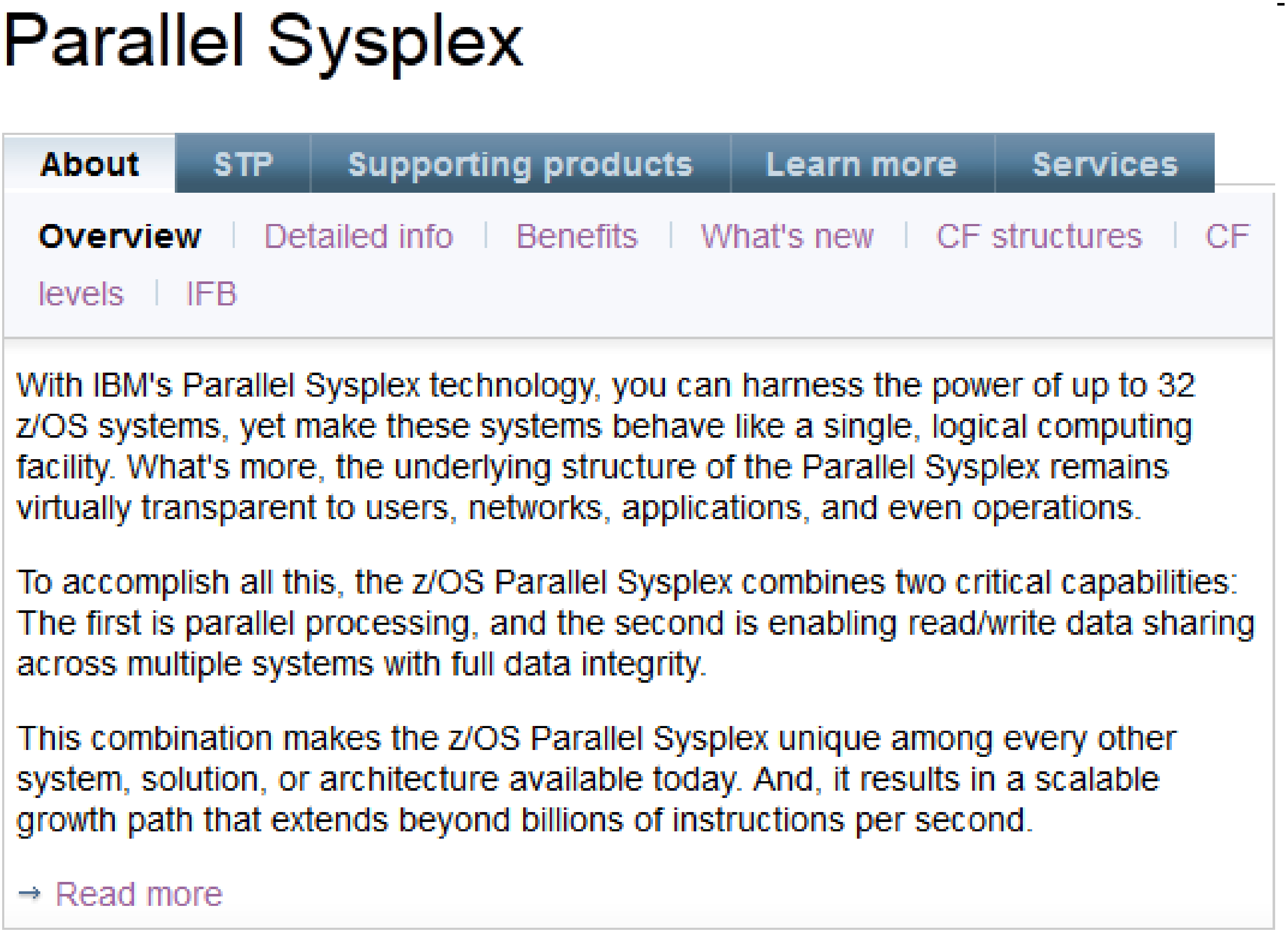
- System z Parallel Sysplex Best Practices, SG24-7817
- Considerations for Multi-Site Sysplex Data Sharing, SG24-7263
- Server Time Protocol Planning Guide, SG24-7280
- Server Time Protocol Implementation Guide, SG24-7281
- Server Time Protocol Recovery Guide, SG24-7380

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

- Available at www.redbooks.ibm.com

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.

→ Read more

Appendix

Sysplex Highlights from z/OS V1R13

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

z/OS V1R13 - DISPLAY XCF,SYSPLEX

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

z/OS V1R13 – D XCF,SYSPLEX,ALL

z/OS 1.12

D XCF,S,ALL

```
IXC335I  12:55:00  DISPLAY XCF       FRAME LAST   F      E   SYS=SY1
SYSPLEX PLEX1
SYSTEM   TYPE SERIAL LPAR STATUS TIME          SYSTEM STATUS
SY1      4381 9F30   N/A  04/22/2011 12:55:00 ACTIVE       TM=SIMETR
SY2      4381 9F30   N/A  04/22/2011 12:54:56 ACTIVE       TM=SIMETR
SY3      4381 9F30   N/A  04/22/2011 12:54:56 ACTIVE       TM=SIMETR

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

z/OS 1.13

D XCF,S,ALL

```
IXC337I  12.29.36  DISPLAY XCF       FRAME LAST   F      E   SYS=SY1
SYSPLEX PLEX1           MODE: MULTISYSTEM-CAPABLE

 SYSTEM SY1             STATUS: ACTIVE
                        TIMING: SIMETR NETID: 0F
                   STATUS TIME: 05/04/2011 12:29:36.000218
                     JOIN TIME: 05/04/2011 10:31:08.072275
                 SYSTEM NUMBER: 01000001
             SYSTEM IDENTIFIER: AC257038 01000001
                   SYSTEM TYPE: 4381  SERIAL: 9F30  LPAR: N/A
               NODE DESCRIPTOR: SIMDEV.IBM.PK.D13ID31
                                PARTITION: 00    CPCID: 00
                       RELEASE: z/OS 01.13.00

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

z/OS V1R13 – CF Structure Placement

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

z/OS V1R13 – CF Structure Placement …

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

Automatic Restart Management (ARM)

- If you have an active ARM policy, then:
  - After system failure, ARM waits up to two minutes for survivors to finish cleanup processing for the failed system
  - If cleanup does not complete within two minutes, ARM proceeds to restart the failed work anyway
- Problem: Restart may fail if cleanup did not complete
- Issue: Two minutes may not be long enough for the applications to finish their cleanup processing

z/OS V1R13 – New ARM Parameter

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

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

- 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

z/OS V1R13 XCF Client/Server

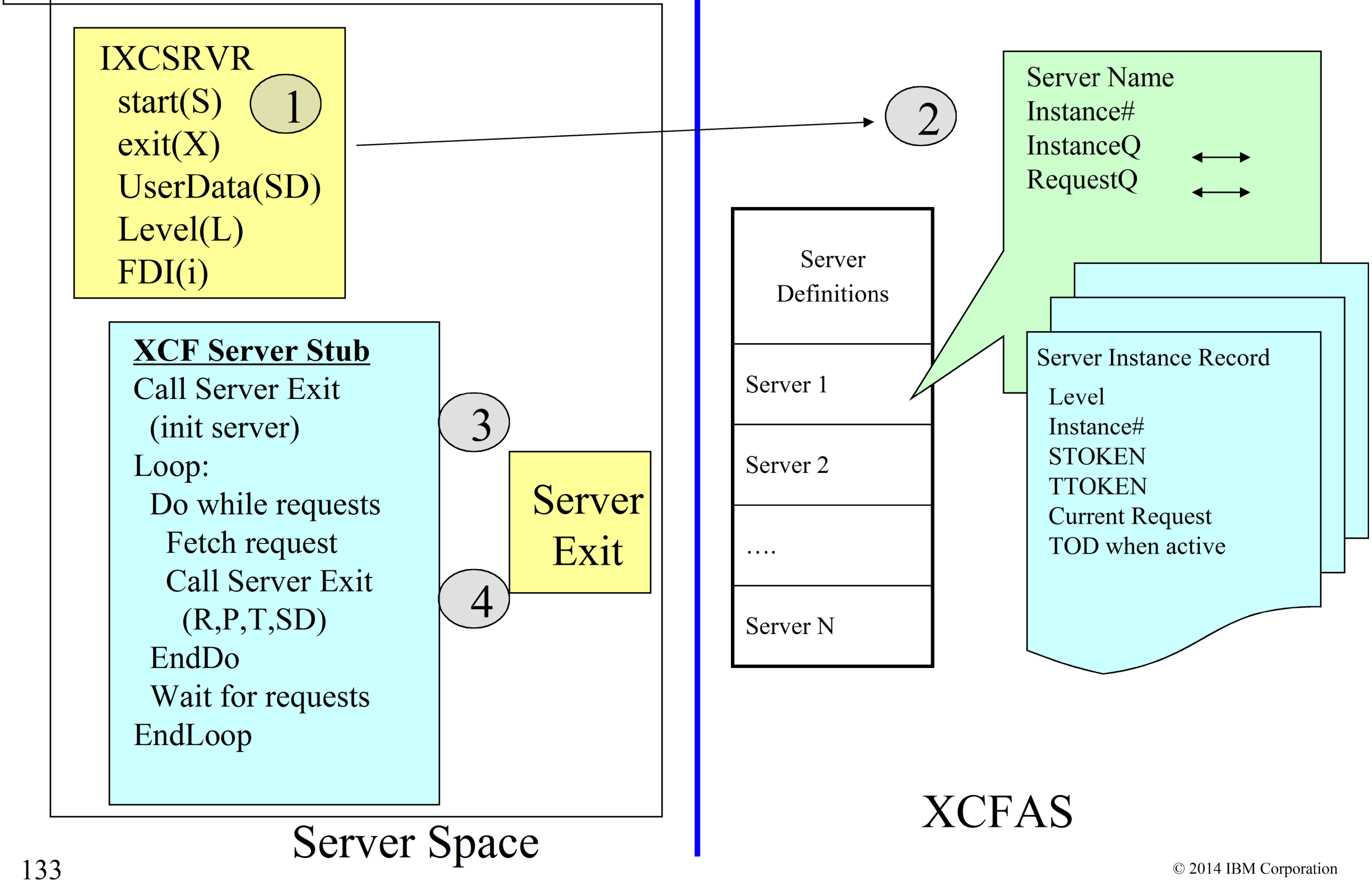
- IXCSEND – send request to one or more servers

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

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

- IXCYSRVR – data mappings

XCF Client/Server – Server Task Overview



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

XCF Server Stub

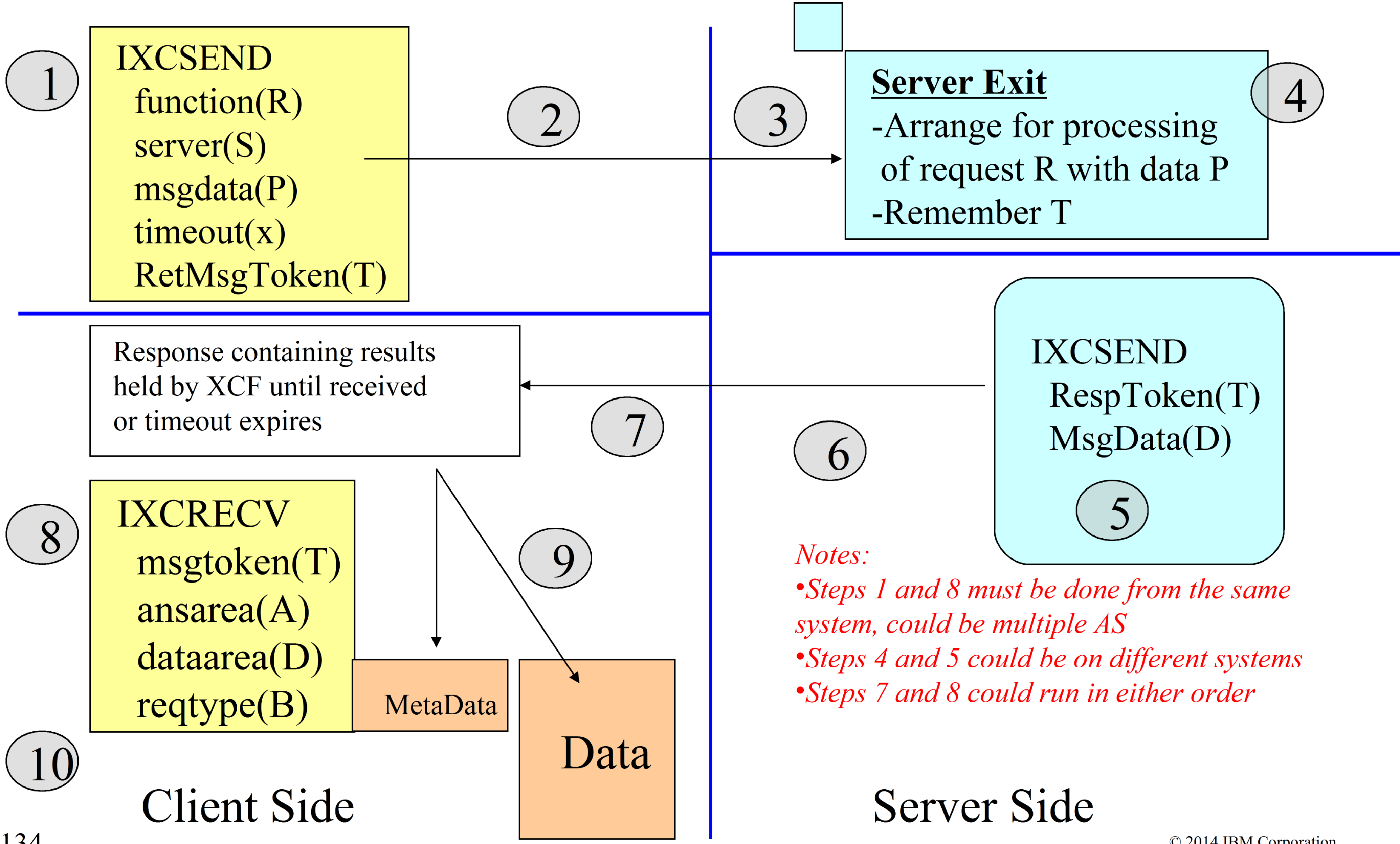
```
Call Server Exit
  (init server)
Loop:
  Do while requests
    Fetch request
    Call Server Exit
      (R,P,T,SD)
  EndDo
  Wait for requests
EndLoop
```

Server Name
Instance#
InstanceQ
RequestQ

Server Definitions
Server 1
Server 2
….
Server N

Server Instance Record
- Level
- Instance#
- STOKEN
- TTOKEN
- Current Request
- TOD when active

XCF Client/Server - Send/Receive Overview



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

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

IXCSEND
RespToken(T)
MsgData(D)

Response containing results held by XCF until received or timeout expires

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

MetaData

Data

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*

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# ] ]
```

CF Structure Alter Processing

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

z/OS V1R13 – Enable/Disable Start Alter Processing

- SETXCF MODIFY,STRNAME=pattern,ALTER=DISABLED
- SETXCF MODIFY,STRNAME=pattern,ALTER=ENABLED
  - STRNAME=strname
  - STRNAME=strprfx*
  - STRNAME=ALL | STRNAME=*
- D XCF,STRUCTURE, ALTER={ENABLED|DISABLED}
- Only systems with support will honor ALTER=DISABLED indicator in the active policy
  - So you may not get the desired behavior until the function is rolled around the sysplex
  - But fall back is trivial since downlevel code ignores it
- APAR OA34579 for z/OS V1R10 and up
  - OA37566 as well

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

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

z/OS V1R13 – Runtime Diagnostics …

Does what you might do manually today:

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

z/OS V1R13 - zFS

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

Appendix

Sysplex Highlights from z/OS V1R12

z/OS V1R12 Summary

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

Background - REALLOCATE

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

Background - REALLOCATE

But…

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

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

z/OS V1R12 - REALLOCATE

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

z/OS V1R12 – REALLOCATE …

Caveats for TEST option

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

z/OS V1R12 – REALLOCATE …

Caveats for REPORT option

- Can be done during or after a real REALLOCATE, but not before a real REALLOCATE is started

- A REPORT is internally initiated by XCF if a REALLOCATE completes with exceptions

z/OS V1R12 - Critical Members

- A system may appear to be healthy with respect to XCF system status monitoring, namely:
  - Updating status in the sysplex CDS
  - Sending signals
- But is the system actually performing useful work?
  - There may be critical functions that are non-operational
  - Which in effect makes the system unusable, and perhaps induces sympathy sickness elsewhere in the sysplex
- Action should be taken to restore the system to normal operation OR it should be removed to avoid sympathy sickness

z/OS V1R12 - Critical Members …

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

z/OS V1R12 - Critical Members …

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

z/OS V1R12 - Critical Members …

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

z/OS V1R12 - Critical Members …

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

XES Connector Hang Detection

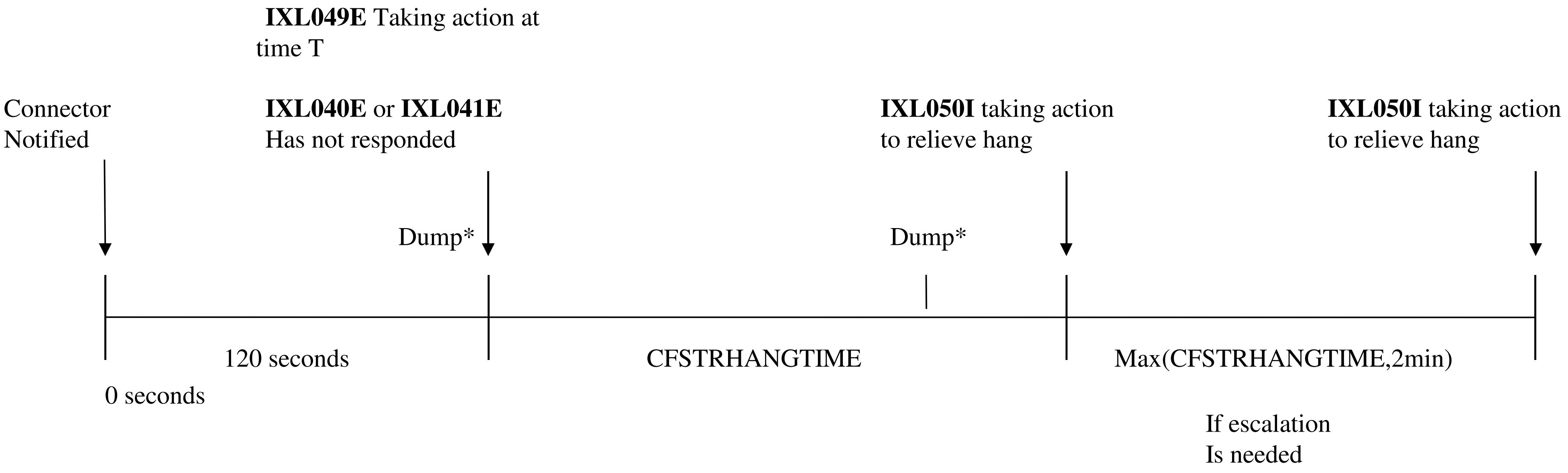
- Connectors to CF structures need to participate in various processes and respond to relevant events
- XES monitors the connectors to ensure that they are responding in a timely fashion
- If not, XES issues messages (IXL040E, IXL041E) to report the unresponsive connector
- Users of the structure may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

- Need a way to resolve this automatically …

z/OS 1VR12 – CFSTRHANGTIME …

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

z/OS V1R12 – CFSTRHANGTIME Processing



IXL049E Taking action at time T

Connector Notified

IXL040E or **IXL041E** Has not responded

IXL050I taking action to relieve hang

IXL050I taking action to relieve hang

Dump*

Dump*

120 seconds

CFSTRHANGTIME

Max(CFSTRHANGTIME,2min)

0 seconds

If escalation Is needed

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

z/OS 1.12 – CFSTRHANGTIME …

New Messages

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

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

z/OS V1R12 – CFSTRHANGTIME …

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

z/OS 1.12 – Support for CFLEVEL 17

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

z/OS 1.12 – Support for CFLEVEL 17 …

- A new version of the CFRM CDS is needed to define more than 1024 structures in a CFRM policy

- May need to roll updated software around the sysplex for any exploiter that wants to request more than 32 connectors to list and lock structures
  - Not aware of any at this point (so really just positioning for future growth)

z/OS 1.12 – Support for CFLEVEL 17 …

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

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

z/OS 1.12 Health Checks

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

z/OS 1.12 Health Checks …

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

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

z/OS 1.12 Auto-Reply

- Fast, accurate, knowledgeable responses can be critical
- Delays in responding to WTOR's can impact the sysplex
- Parmlib member defines a reply value and a time delay for a WTOR. The system issues the reply if the WTOR has been outstanding longer than the delay

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

z/OS 1.12 Auto-Reply ….

- For example:

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

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

z/OS 1.12 - XCF Programming Interfaces

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