



International Technical Support Organization and Authoring Services

2012 Parallel Sysplex and High Availability Update

IBM Redbooks

Frank Kyne
kyne@us.ibm.com

Introduction

Thank you for coming!

Who am I?

- Started in IT as an operator in VM/VSE customer in Ireland
- Joined IBM Ireland as trainee MVS system programmer
- Currently Project Leader in ITSO Poughkeepsie, responsible for RedBooks, workshops, and consultancy on sysplex, high availability, and performance

Questions? **PLEASE** ask as I go along.

- And remind me when I start speaking too fast

Evaluation forms - **PLEASE** complete the evaluation form so we can know what you like and improve what you don't

Agenda

- 09:00-10:00 zEC12 sysplex considerations
- 10:00-10:30 New IXCNOTE facility
- 10:30-10:50 Coffee
- 10:50-11:10 Hands-on implementation of BCPii and SSDPP
- 11:10-12:10 Understanding and optimizing Hiperdispatch
- 12:10-13:15 Lunch
- 13:15-13:30 JES2 Dynamic Proclib support
- 13:30-14:00 Sysplex enhancements in MQ 7.1
- 14:00-15:00 IBM Poughkeepsie's test systems and facilities
- 15:00-15:20 Coffee
- 15:20-16:00 MVS Workload Manager and dynamic workload routing
- 16:00-16:30 Extended distance end-to-end design
- 16:30-17:00 Miscellaneous

Acknowledgements

Want to thank the following for their help in creating this material:

- Paul Dennis
- Gary Fisher
- Jim Gualtieri
- Jeff Kubala
- Alain Maneville
- Iain Neville
- Ray Newsom
- Bernie Pierce
- Rich Prewitt
- Dan Rosa
- Pete Siddall
- Horst Sinram
- Bart Steegmans
- Ralf Streit
- Dave Surman





International Technical Support Organization and Authoring Services

EC12 Sysplex Considerations

IBM Redbooks

zEC12 CF link support

Support for ICB4 links was removed on z196.

- Infiniband HCA3 adapters in IFB3 mode offer superior performance to ICB4 links.

The only ways to get ISC3 links on an zEC12 are:

- Carry forward on an upgrade from an earlier processor
- Possibly via an RPQ on H66, H89, HA1 - discuss with your IBM representative
- Stmt of direction that zC12 is last System z generation to support ISC3

Note that InfiniBand provides a superior alternative to ISC3:

- Up to 150 metres, HCA3 12X provides vastly better performance
- Beyond 150 metres, HCA3 1X provides greater flexibility and better performance with fewer links

zEC12 CF link support

Link Type	Maximum ports H20	Maximum ports H43, H66, H89, HA1
1X IFB HCA3-O LR	32	64 (32 on z196 GA1, 48 on z196 GA2)
12X IFB,12X IFB3 HCA3-O	16	32
1X IFB * HCA2-O LR	16	32
12X IFB * HCA2-O	16	32
ISC-3	48	48
ICP	32	32

- * HCA2 are carry-forward only

These are maximum port numbers. Maximum number of HCA adapters is 16. Max number of external coupling links cannot exceed 112 (96 on z196). Max number of Coupling CHPIDs is 128.

See hardware day material for more detailed info.

zEC12 Coupling links

Different link types are aimed at different applications and have different performance characteristics.

- HCA3-O IFB3 mode (up to 150 metres) is designed for data-centre-distance and very high performance/throughput
- HCA3-O 1X is designed for long distance, so supports more subchannels per CHPID:
 - Up to 175 km, depending on DWDM - needs RPQ 8P2340 for distances > 100km
 - Up to 10 km unrepeated - needs RPQ 8P2340 to extend to 20km

	ISC3	PSIFB 1X	PSIFB 12X	PSIFB 12X IFB3	ICB4	ICP
z10						
Lock	20-30	14-18	11-15	N/A	8-12	3-8
Cache/List	25-40	18-25	15-20	N/A	10-16	6-10
z196						
Lock	20-30	14-17	10-14	5-8	N/A	2-8
Cache/List	25-40	16-25	14-18	7-9	N/A	4-9
zEC12						
Lock	20-30	12-16	10-14	5-8	N/A	2-6
Cache/List	25-40	14-24	13-17	7-9	N/A	4-8

Coupling z/OS CPU cost

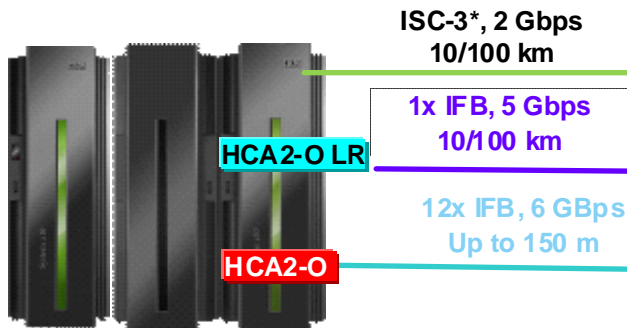
CFHost	z10 BC	z10 EC	z114	z196	zEC12
z10 BC ISC3	16	18	17	21	24
z10 BC IFB 1X	13	14	14	17	19
z10 BC IFB 12X	12	13	13	15	17
z10 BC ICB4	10	11	N/A	N/A	N/A
z10 EC ISC3	16	17	17	21	24
z10 EC IFB 1X	13	14	14	17	19
z10 EC IFB 12X	11	12	12	14	16
z10 EC ICB4	10	10	N/A	N/A	N/A
z114 ISC3	16	18	17	21	24
z114 IFB 1X	13	14	14	17	19
z114 IFB 12X	12	13	12	15	17
z114 IFB3 12X	N/A	N/A	10	12	13
z196 ISC3	16	17	17	21	24
z196 IFB 1X	13	14	13	16	18
z196 IFB 12X	11	12	11	14	15
z196 IFB3 12X	N/A	N/A	9	11	12
zEC12 ISC3	16	17	17	21	24
zEC12 IFB 1X	13	13	13	16	18
zEC12 IFB 12X	11	11	11	13	15
zEC12 IFB3 12X	9	9	9	10	11

Based on 9 CF Req/Sec/MIPS

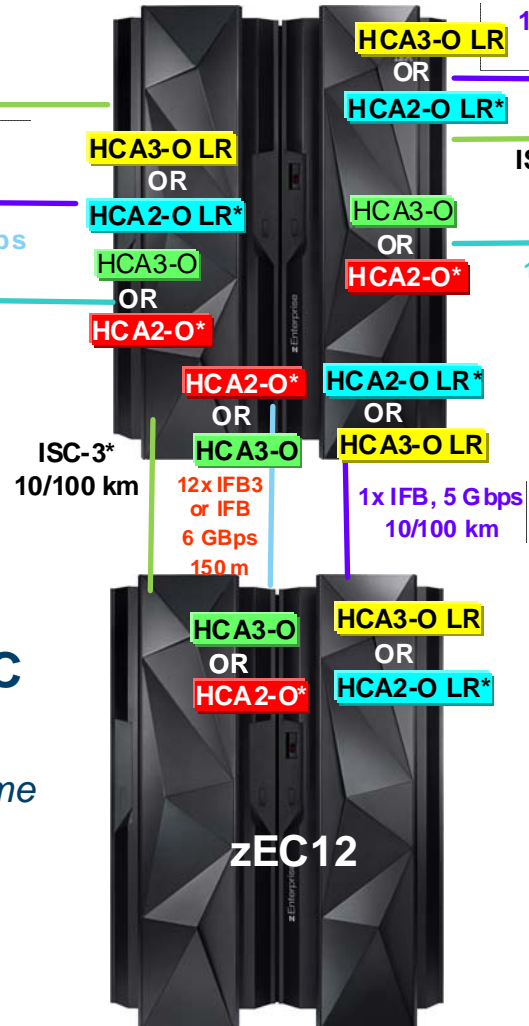
In practice, XES heuristic algorithm caps overhead at about 18%

zEC12 Coupling links

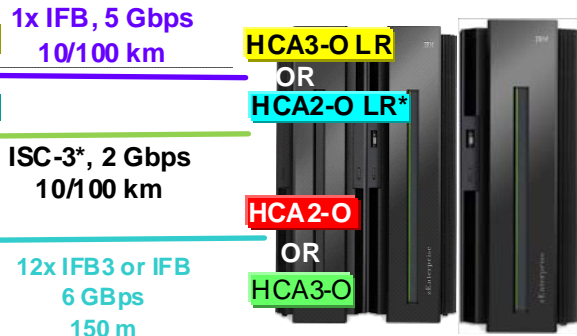
z10 EC and z10 BC
12x IFB, 1x IFB & ISC-3



zEC12



z196 and z114
12x IFB, 12x IFB3, 1x IFB, & ISC-3



z9 EC and z9 BC
z890, z990

Not supported in same Parallel Sysplex or STP CTN with zEC12

*HCA2-O, HCA2-O LR carry forward only on zEC12

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.

zEC12 Coupling LPAR support

Prior to zEC12, a System z LPAR supported a maximum of 16 ICF engines.

However, a System z CPC also supported a maximum of only 16 ICF engines.

- This could be an issue if you want to define multiple large CF LPARs on one CPC.

On zEC12, a CF LPAR still supports a maximum of 16 ICF engines.

But the maximum number of ICF engines on the CPC is now 101, meaning that you can now have multiple large or backup CFs on the same CPC.

zEC12 Coupling links

Statements of direction:

- zEC12 is the last generation of System z that will support ISC3 links.
 - MESSAGE - if you are not already on InfiniBand links, create a migration plan NOW. HCA3 12X in IFB3 mode are much better than ICB4, and HCA3 1X are much better than ISC3, so there is no reason to delay...
- zEC12 is the last System z generation that will support a mixed mode STP CTN. z10 is the last generation that supports direct Sysplex Timer connection, so next generation after zEC12 will not co-exist in a CTN with a CPC that supports Sysplex Timer connection.
 - MESSAGE - if you are not yet in an STP-only CTN, prepare plans to complete your migration from mixed mode CTN.
 - STP redbooks (SG24-7280, SG24-7281, SG24-7380) currently being updated to reflect the latest STP enhancements.

zEC12 Coupling links

Prior to installing ANY new processor type (or upgrading to a new processor type) or installing first InfiniBand adapters, ensure that you bring the microcode level on all processors in the configuration up to date.

Prior to installing ANY new processor type (or upgrading to a new processor type) or installing first InfiniBand adapters, ensure that you bring the microcode level on all processors in the configuration up to date.

Prior to installing ANY new processor type (or upgrading to a new processor type) or installing first InfiniBand adapters, ENSURE THAT YOU BRING THE MICROCODE LEVEL ON ALL PROCESSORS IN THE CONFIGURATION UP TO DATE.

ESPECIALLY, get CFCC levels on z10 and z196/z114 up to date before connecting to a zEC12

zEC12 Coupling links

Sources of information:

zEnterprise EC12 System Overview, SA22-1088

IBM zEnterprise EC12 Technical Introduction, SG24-8050

IBM zEnterprise EC12 Technical Guide, SG24-8049

IBM zEnterprise EC12 Configuration Setup, SG24-8034

**Implementing and Managing InfiniBand Coupling Links on System z,
SG24-7539**

Server Time Protocol Planning Guide, SG24-7280

Server Time Protocol Recovery Guide, SG24-7380

Server Time Protocol Implementation Guide, SG24-7281



International Technical Support Organization and Authoring Services

CFCC Level 18

IBM Redbooks

zEC12 Coupling Facility

zEC12 delivers new CF Level - CFLevel 18

- Only available on zEC12
- Delivers enhancements in performance, usability, and RAS
- As usual, there are structure size considerations

zEC12 Coupling Facility

Performance Improvements

Enhanced Delete_Name performance

- In DB2 V9, when a pageset/partition becomes non-GBP-dependent, the Delete_Name process deleted both data and directory entries. DB2 V10 was changed so that only data entries were deleted. This was intended to avoid cross invalidation processing at that time (cleanup of the directory entries would be done later when other pages are registered). However, the CFCC processing was not as efficient when only data entries are deleted, resulting in this process taking MORE rather than LESS time in some cases.
- CF Level 18, together with changes in XES and DB2, addresses this to deliver the originally-intended improvements.

Delivered by APAR OA38419 (XES), PM67544 (DB2 V9 and V10), and CF Level 18 (and will be rolled back to CF Level 17)

zEC12 Coupling Facility

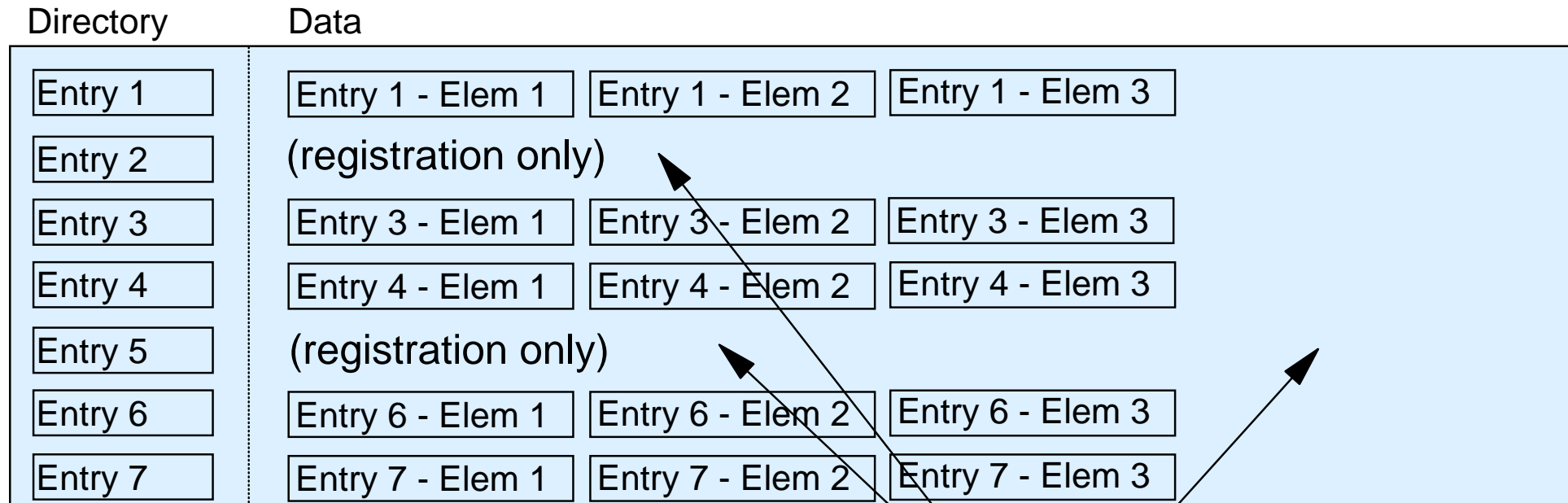
Performance Improvements

Elapsed time improvements when dynamically altering the size or entry/element ratios of a cache structure



zEC12 Coupling Facility

CF Cache structure - why you might want to alter the ratio between entries and elements

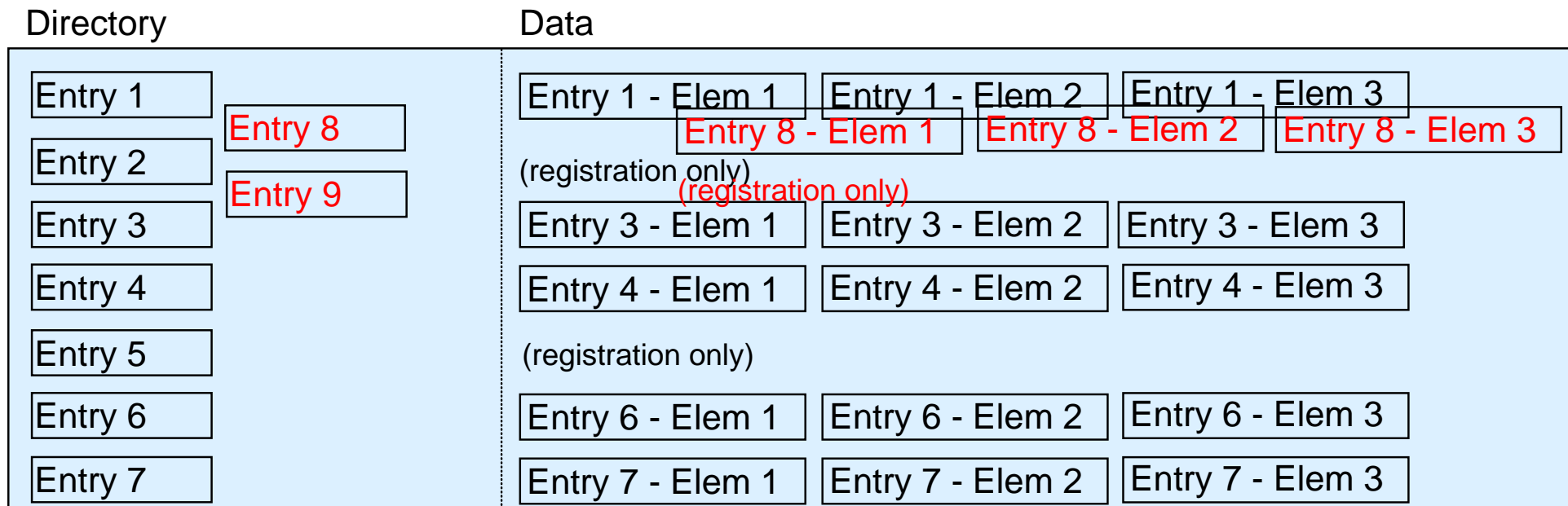


All space for
entries is used
up

But still empty
space for
elements

zEC12 Coupling Facility

Auto Alter can: make structures larger, make structures smaller, change ratio between entries and elements



Adjust percent of storage reserved for entries so structure can now manage more data

zEC12 Coupling Facility

CFCC 18 changes the way storage within a cache structure is managed, meaning that the elapsed time to perform either of the following types of structure changes should be reduced:

- An alter to reduce the size of the structure
- An alter to change the ratio between entries and elements
- This change does not impact structure Alters that increase the overall structure size.

Larger structures should benefit to a greater extent than smaller structures

All changes are self-contained within CFCC, no need for supporting APARs in z/OS or subsystems

It is expected that anyone that has disabled Alter should be able to re-enable it once they move to CF Level 18.

zEC12 Coupling Facility

Performance Improvements

"Cache write-around" enhancement

When:

- DB2 detects heavy writers to a GBP structure (DB2 utilities, batch jobs for example)
- *And* the GBP structure or storage class is above the castout threshold
- *And* the page being written is not already in the GBP
- *And* the writing system is the only one with an interest in that page

DB2 can now protect the GBP working set by writing updated pages directly to disk rather than flooding the GBP structure with pages that probably will not be re-referenced.

zEC12 Coupling Facility

Performance Improvements - "Cache write-around" enhancement

Performance measurements in IBM showed DB2 batch improvements of up to 50% in environments that were previously experiencing GBP-full conditions

- Obviously results will vary

Planned to be in DB2 V11 and rolled back to DB2 V10 with DB2 APAR PM70575

Requires XES APAR OA37550

- Function available on z/OS 1.12 and 1.13
- REQUIRES toleration support on older z/OS Levels (same APAR number)

Support also rolled back to CF Level 17, Service Level 10.15

zEC12 Coupling Facility

Performance Improvements

Throughput enhancements for parallel cache castout processing

In order to avoid delays in writing to a GBP, it is vital that the GBP doesn't fill up.

CF Level 18 adds an enhancement, to be exploited by DB2 V11, that will provide greater overlap between reading pages that are to be castout from the GBP, with writes to the disk data sets, thereby reducing the risk of the GBP filling during peak activity.

- All the changes are in CFCC and in DB2 - no changes required in XES.

zEC12 Coupling Facility

Performance Improvements

The rate at which changed pages can be read from a GBP is far higher than the write rate for a disk. To optimize castout rates, DB2 attempts to drive castouts for multiple data sets concurrently.

As the number of open DB2 data sets continues to increase, this puts even more pressure on castout processing. Extremely busy GBPs have encountered contention due to castout class-level serialization. Increasing the number of castout classes and associated CFCC control blocks will allow DB2 to spread pagesets/partitions over more classes, providing more parallelism.

- All the changes are in CFCC and in DB2 - no changes required in XES.

zEC12 Coupling Facility

Usability enhancements

Coupling link characteristics reporting to z/OS when running on zEC12

- Identifies underlying InfiniBand hardware characteristics for CIB CHPIDs to help with sysplex monitoring and tuning
 - Requires RMF and XES APARs OA37826 and OA38312
 - Also includes improved reporting for peer CF links from zEC12 CF
- Enables D CF command, RMF Monitor III, and RMF PostProcessor to report additional information:
 - InfiniBand Link type and protocol: 12x IFB, 12x IFB3 and 1x IFB
 - CHPID mapping to physical links - HCA IDs and port numbers
 - Link fiber optic distance (in km)
 - If link is running in degraded status

zEC12 Coupling Facility

New Display command support

Works for any type of InfiniBand adapter when z/OS is running on a zEC12

- Must be z/OS 1.12 or later with supporting APARs
- Other end of the link can be any CPC that supports InfiniBand connection to zEC12

Fields show N/A when command is entered on z/OS running on z196 or older

zEC12 Coupling Facility

Display CF command without the support:

```

D CF,CFNAME=FACIL04
IXL150I 11.26.53 DISPLAY CF 625
COUPLING FACILITY 002817.IBM.02.0000000B3BD5
                PARTITION: 2F CPCID: 00
                CONTROL UNIT ID: FFF9

NAMED FACIL04
COUPLING FACILITY SPACE UTILIZATION
  ALLOCATED SPACE          DUMP SPACE UTILIZATION
  STRUCTURES:              0 M          STRUCTURE DUMP TABLES:      0 M
  DUMP SPACE:              20 M          TABLE COUNT:              0
  FREE SPACE:              3635 M        FREE DUMP SPACE:           20 M
  TOTAL SPACE:             3655 M        TOTAL DUMP SPACE:         20 M
                                MAX REQUESTED DUMP SPACE:      0 M
  VOLATILE:                YES          STORAGE INCREMENT SIZE:    1 M
  CFLEVEL:                 17
  CFCC RELEASE 17.00, SERVICE LEVEL 04.18
  BUILT ON 10/26/2011 AT 13:31:00
  COUPLING FACILITY HAS 0 SHARED AND 1 DEDICATED PROCESSORS
  DYNAMIC CF DISPATCHING: OFF

```

CF REQUEST TIME ORDERING: REQUIRED AND ENABLED

COUPLING FACILITY SPACE CONFIGURATION

	IN USE	FREE	TOTAL
CONTROL SPACE:	20 M	3635 M	3655 M
NON-CONTROL SPACE:	0 M	0 M	0 M

SENDER PATH	PHYSICAL	LOGICAL	CHANNEL TYPE
A4 / 0199	ONLINE	ONLINE	CFP
A6 / 0190	ONLINE	ONLINE	CFP
A8 / 0111	ONLINE	ONLINE	CFP
AA / 0119	ONLINE	ONLINE	CFP

zEC12 Coupling Facility

Display CF command with the support:

```

D CF,CFNAME=FACIL05
IXL150I 14.31.03 DISPLAY CF 976
COUPLING FACILITY 002817.IBM.02.0000000B3BD5
PARTITION: 2E CPCID: 00
CONTROL UNIT ID: FFD4

NAMED FACIL05
COUPLING FACILITY SPACE UTILIZATION
ALLOCATED SPACE          DUMP SPACE UTILIZATION
STRUCTURES:              128 M          STRUCTURE DUMP TABLES:          0 M
DUMP SPACE:              20 M          TABLE COUNT:                    0
FREE SPACE:              7602 M         FREE DUMP SPACE:                  20 M
TOTAL SPACE:             7750 M         TOTAL DUMP SPACE:                 20 M
                                MAX REQUESTED DUMP SPACE:          0 M
VOLATILE:                YES           STORAGE INCREMENT SIZE:          1 M
CFLEVEL:                 17
CFCC RELEASE 17.00, SERVICE LEVEL 10.15
BUILT ON 07/18/2012 AT 13:09:00
COUPLING FACILITY HAS 0 SHARED AND 1 DEDICATED P
DYNAMIC CF DISPATCHING: OFF
COUPLING FACILITY IS NOT STANDALONE

CF REQUEST TIME ORDERING: REQUIRED AND ENABLED
    
```



```

COUPLING FACILITY SPACE CONFIG
IN
CONTROL SPACE:
NON-CONTROL SPACE:
    
```

Type and mode

Adapter ID

Port #

PATH	PHYSICAL	LOGICAL	CHANNEL TYPE	AID	PORT
B1 / 072B	ONLINE	ONLINE	CIB 12X-IFB3	000A	02
B5 / 072C	ONLINE	ONLINE	CIB 12X-IFB3	001A	02
B8 / 0737	ONLINE	ONLINE	CIB 1X-IFB	000D	01
B9 / 0738	ONLINE	ONLINE	CIB 1X-IFB	000D	02
BA / 0739	ONLINE	ONLINE	CIB 1X-IFB	001D	01
BB / 073A	ONLINE	ONLINE	CIB 1X-IFB	001D	02

zEC12 Coupling Facility

And the peer CF links (this command was issued on a z/OS running on a z196):

```
D CF,CFNM=FACIL06
IXL150I 17.20.53 DISPLAY CF 216
COUPLING FACILITY 002827.IBM.02.00000000B8D7
PARTITION: 1D CPCID: 00
LP NAME: A1D CPC NAME: SCZP401
CONTROL UNIT ID: FFE7
```

NAMED FACIL06

PATH	PHYSICAL	LOGICAL	CHANNEL TYPE	AID	PORT
82 / 0720	ONLINE	ONLINE	CIB	N/A	N/A
87 / 0724	ONLINE	ONLINE	CIB	N/A	N/A
90 / 0728	ONLINE	ONLINE	CIB	N/A	N/A
93 / 0729	ONLINE	ONLINE	CIB	N/A	N/A
96 / 072A	ONLINE	ONLINE	CIB	N/A	N/A
99 / 072B	ONLINE	ONLINE	CIB	N/A	N/A

REMOTELY CONNECTED COUPLING FACILITIES

CFNAME	COUPLING FACILITY
-----	-----
FACIL05	002817.IBM.02.00000000B3BD5
	PARTITION: 2E CPCID: 00

CHPIDS ON FACIL06 CONNECTED TO REMOTE FACILITY

RECEIVER:	CHPID	TYPE
	B1	CIB 12X-IFB3
	B5	CIB 12X-IFB3
	B8	CIB 1X-IFB
	B9	CIB 1X-IFB
	BA	CIB 1X-IFB
	BB	CIB 1X-IFB

SENDER:	CHPID	TYPE
	B1	CIB 12X-IFB3
	B5	CIB 12X-IFB3
	B8	CIB 1X-IFB
	B9	CIB 1X-IFB
	BA	CIB 1X-IFB
	BB	CIB 1X-IFB

zEC12 Coupling Facility

Support for new link type information AND distance has been added to RMF Monitor III (option S (Sysplex), then 6 (CFSYS), then select a CF and system)

```

RMF V1R13  CF Systems  - #0$#PLEX  Line 1 of 5
Command ==>  Scroll ==> CSR
Samples: 100  Systems: 2  Date: 09/24/12  Time: 13.30.00  Range: 100  Sec
CF
- RMF Coupling Facility  Id Path
Press Enter to return to the Report
Details for System      : #0$A
Coupling Facility      : FACIL05
Subchannels Generated  : 42
In Use                 : 14
Max                    : 14
Channel Path Details:
ID Type  Operation Mode  Deg Distance PCHID  ID Port  --IOP IDs--
B1 CIB   12x IFB3 HCA3-0  N      <1  072B 000A  02 07
B5 CIB   12x IFB3 HCA3-0  N      <1  072C 001A  02 06
F1=Help  F2=SplitScr  F3=End  F6=RMFhelp  F7=Backward
F8=Forward  F9=SwapScr  F12=Return
    
```

Degraded?

Avg distance to CF (in km)

CHPID

Bandwidth

Mode

HCA type

AID Port

F1=HELP
F7=UP

F2=SPLIT
F8=DOWN

F3=END
F9=SWAP

F4=RETURN
F10=BREF

F5=RFIND
F11=FREF

F6=TOGGLE
F12=RETRIEVE

zEC12 Coupling Facility

Press PF1 for an explanation of the new fields in the report....

```

Command ==>          RMF V1R13  CF Systems          - #@$#PLEX          Line 1 of 5
                               Scroll ==> CSR
Samples: 100         Systems: 2      Date: 09/24/12   Time: 13.30.00   Range: 100      Sec
CF
RMF Coupling Facility - Subchannels and Paths
RMF Monitor III CF Systems - Subchannel and Channel Path Details
COMMAND ==> _
More: - +
Deg          Character Y in this column indicates that the channel path
              is operating at reduced capacity (degraded) or not
              operating at all.
Distance     Estimated distance in kilometers. The value is calculated
              as follows:
              -----
              Average round-trip path time in microseconds
              -----
              10 microseconds / kilometer
              A value of zero means that the time was not measured.
PCHID        Physical channel identifier.
HCA ID       The hexadecimal host channel adapter identifier.
F1=Help      F2=SplitScr   F3=End       F6=RMFHelp   F7=PrevPage
F8=NextPage  F9=SwapScr    F12=Return
  
```

Note that "degraded" means a 12X link that has a hardware problem and that is currently running as 8X, 4X, or 1X

Is also shown in the status field on the D CF command if a CHPID is degraded.

zEC12 Coupling Facility

RMF PP CF-to-CF Activity report also enhanced

COUPLING FACILITY NAME = FACIL06

					CF TO CF	
PEER	# REQ	-- CF LINKS --			REQU	
CF	TOTAL	TYPE	USE		#	-SEE
	AVG/SEC				REQ	
FACIL03	0	CIB	2	SYNC	0	
	0.0					
FACIL05	0	CIB	6	SYNC	0	0.0 0.0
	0.0					

Note: Peer mode report does not include AID and Port info

CHANNEL PATH DETAILS						
PEER CF	ID	TYPE	OPERATION MODE		DEGRADED	DISTANCE
FACIL03	90	CIB	12X	IFB HCA3-O	N	<1
	94	CIB	12X	IFB HCA3-O	N	<1
FACIL05	B1	CIB	12X	IFB3 HCA3-O	N	<1
	B5	CIB	12X	IFB3 HCA3-O	N	<1
	B8	CIB	1X	IFB HCA3-O LR	N	<1
	B9	CIB	1X	IFB HCA3-O LR	N	<1
	BA	CIB	1X	IFB HCA3-O LR	N	<1
	BB	CIB	1X	IFB HCA3-O LR	N	<1



zEC12 Coupling Facility

```

D XCF,C
IXC357I 23.29.08 DISPLAY XCF 022
SYSTEM #@$A DATA
  INTERVAL  OPNOTIFY  MAXMSG  CLEANUP  RETRY  CLASSLEN
    165      168      2000    15        10      956

  SSUM ACTION  SSUM INTERVAL  SSUM LIMIT  WEIGHT  MEMSTALLTIME
    ISOLATE    0             900        1       300

  CFSTRHANGTIME
    900

  DEFAULT USER INTERVAL: 165
  DERIVED SPIN INTERVAL: 165
  PARMLIB USER OPNOTIFY: + 3

  MAX SUPPORTED CFLEVEL: 17
  MAX SUPPORTED SYSTEM-MANAGED PROCESS LEVEL: 17

  SIMPLEX SYNC/ASYNCH THRESHOLD: 26
  DUPLEX SYNC/ASYNCH THRESHOLD: 26
  SIMPLEX LOCK SYNC/ASYNCH THRESHOLD: 26
  DUPLEX LOCK SYNC/ASYNCH THRESHOLD: 26

```

Note that this number does NOT change to say CF Level 18, even though the CF Level 18 support is installed

```

D CF,CFNM=FACIL06
IXL150I 23.45.28 DISPLAY CF 035
SY COUPLING FACILITY 002827.IBM.02.00000000B8D7
  PARTITION: 1D CPCID: 00
  LP NAME: A1D CPC NAME: SCZP401
  CONTROL UNIT ID: FFEB

  NAMED FACIL06
SY COUPLING FACILITY SPACE UTILIZATION
  ALLOCATED SPACE          DUMP SPACE UTILIZATION
  STRUCTURES:              275 M          STRUCTURE DUMP TABLES: 0 M
  DUMP SPACE:              20 M          TABLE COUNT:          0
  FREE SPACE:              3363 M        FREE DUMP SPACE:       20 M
  TOTAL SPACE:             3658 M        TOTAL DUMP SPACE:     20 M
                                MAX REQUESTED DUMP SPACE: 0 M
                                STORAGE INCREMENT SIZE: 1 M

  VOLATILE: YES
  CFLEVEL: 18
  CFCC RELEASE 18.00, SERVICE LEVEL 00.27
  BUILT ON 08/17/2012 AT 10:54:00
  COUPLING FACILITY HAS 0 SHARED AND 2 DEDICATED PROCESSORS
  DYNAMIC CF DISPATCHING: OFF
  COUPLING FACILITY IS NOT STANDALONE

```

zEC12 Coupling Facility

There are also significant changes to the RMF Type 74.4 SMF record to contain all this new information:

- New fields
- New data section

Refer to RMF APAR OA37826 for all the details

zEC12 Coupling Facility

Resiliency Improvements

CF Level 18 includes enhanced capabilities to non-disruptively capture and collect extended diagnostic structure data from Coupling Facility structures that have encountered an error.

- Now collects data about broken structure that previously would have resulted in a CF outage.
 - No customer data is collected
 - More triggers to capture a non-disruptive dump
 - Soft-failure cases beyond break-duplexing
 - Rolled back to CFLEVEL 17 (Service Level 10.15) on z196 and z114.
- Requires support in XES APAR OA37550.

zEC12 Coupling Facility

Resiliency Improvements

Verification of local cache controls for a Coupling Facility cache structure connector

- Performed during registration of connection interest in a data item against lost cross-invalidation signals.
- Rolled back to CFLEVEL 17 (Service Level 10.15) on z196 and z114.
- Requires support in XES APAR OA37550.

zEC12 Coupling Facility

Resiliency Improvements

Enhanced CFCC tracing support

- Significantly enhanced trace points, especially in potentially 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.

zEC12 Coupling Facility

Resiliency Improvements - Enhanced CFCC tracing support

Quantity gathered

- Trace buffer size increased

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
- Would only be activated under direction of IBM Service

Tracing enhancements rolled back to CF Level 17, Service Level 10.15

zEC12 Coupling Facility

Structure and CF Storage Sizing with CFCC level 18

- Cache structure sizes may increase when moving from CF Level 17 (or below) to CF Level 18
 - List and Lock structures may actually DEcrease in size when moving from CF Level 17 to CF Level 18.
 - CFCC requires about 440MB in CF Level 18 versus 442MB for Level 17
- Download and run SIZER batch job AFTER CF Level 18 CF is accessible to sysplex and BEFORE you repopulate it.
 - Download from <http://www-947.ibm.com/systems/support/z/cfsizer/altsize.html>
- Use of the CF Sizer Tool is recommended:
<http://www.ibm.com/systems/z/cfsizer/>
- **IMPORTANT** - Remember to update INITSIZEs in CFRM policy to reflect new structure sizes in CF Level 18 CF.

For more information

Refer to the cover letters for APARs OA37826, OA37550, OA38312, and OA38419

Monitor the zEC12 PSP Bucket

(http://www.ibm.com/support/docview.wss?uid=isg1_2827DEVICE_2827-ZOS)
for the latest news on EC12-related APARs for z/OS and related products



International Technical Support Organization and Authoring Services

IXCNOTE Facility

IBM Redbooks

IXCNOTE Facility

Challenge: How to provide flexibility and performance of placing data in a CF, without the complexity of using XES interfaces

Solution: XCF-provided IXCNOTE facility

IXCNOTE Facility

Lets programs read and write notes (list entries) in a sysplex-wide XCF note pad (which resides in a CF list structure).

- Supports unauthorized callers
- Each note pad can contain a finite number of 1K notes
- One or more note pads can reside in the same list structure

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 because there is no option to inform interested parties when a note pad is updated
- Does not provide the full functionality of list structures

IXCNOTE Facility

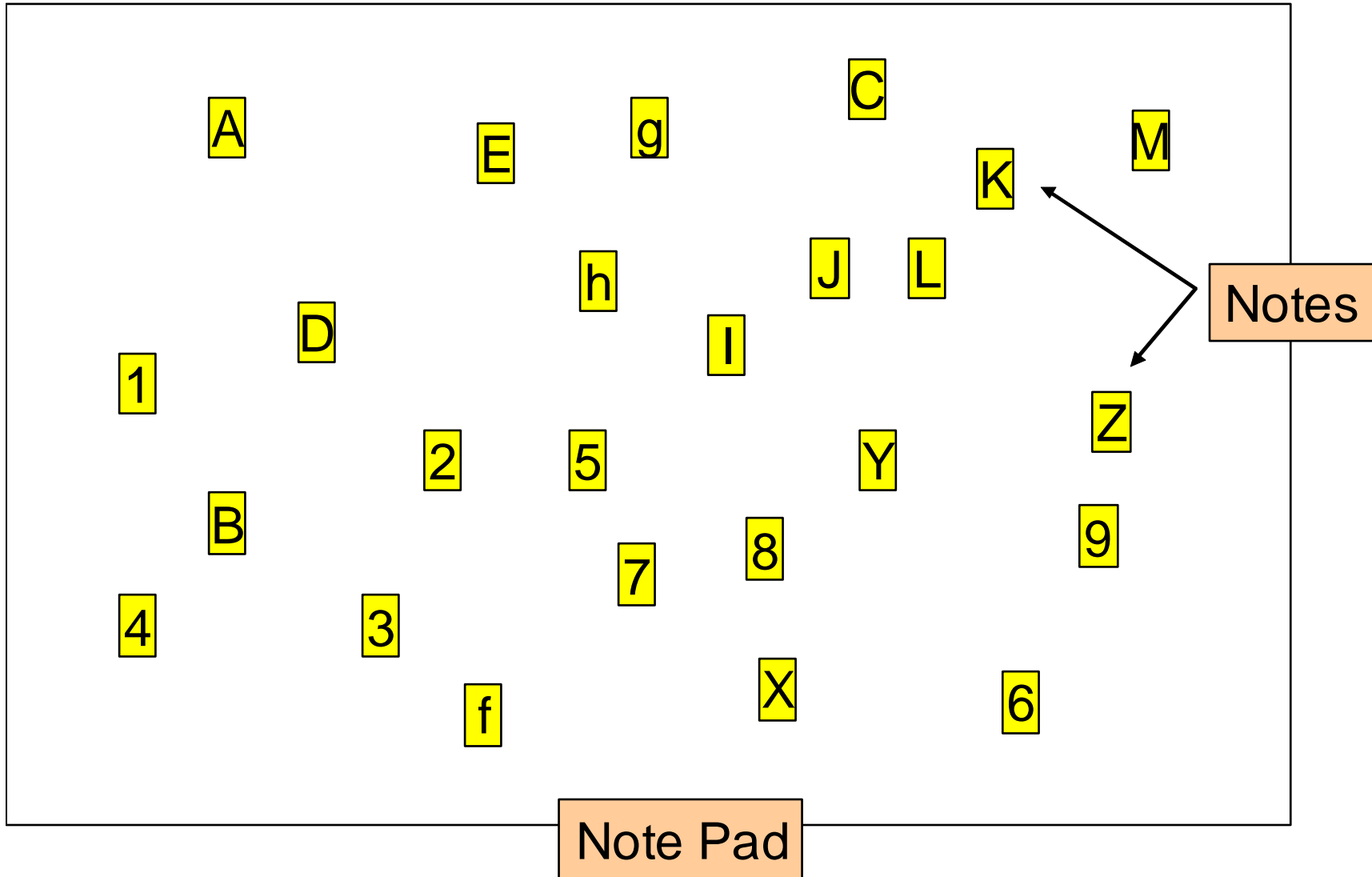
XCF connects to CF structure and deals with various XES exits and protocols - no need for the application program to be aware of any of that complexity.

Simplifies development, reduces complexity, decreases implementation and support costs by masking most of the traditional CF exploitation overhead

- All of the complexity of coding the interface is done once (by XCF) and can be exploited by many

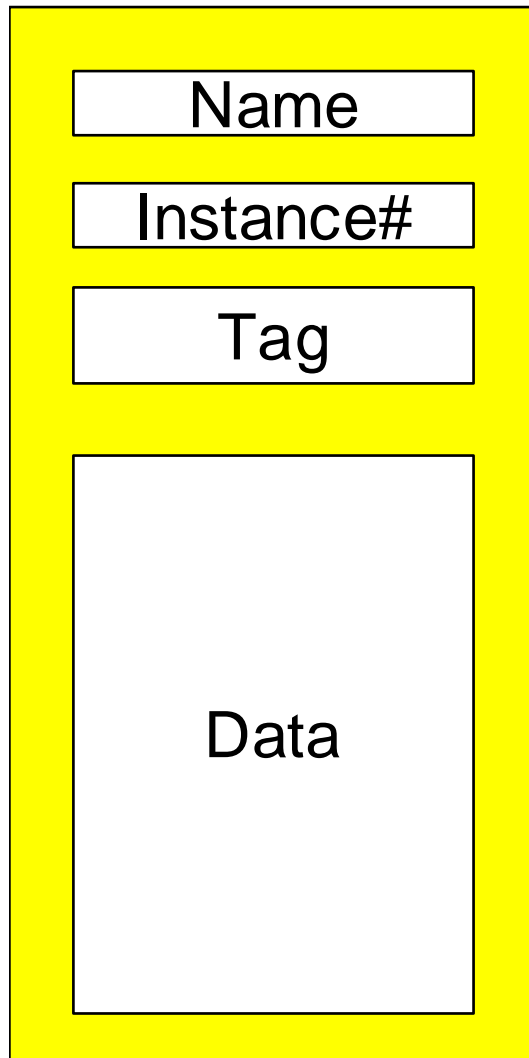
Delivered on z/OS 1.13 by APAR OA38450

IXCNOTE Facility



IXCNOTE Facility

What a note looks like



8 byte user note name

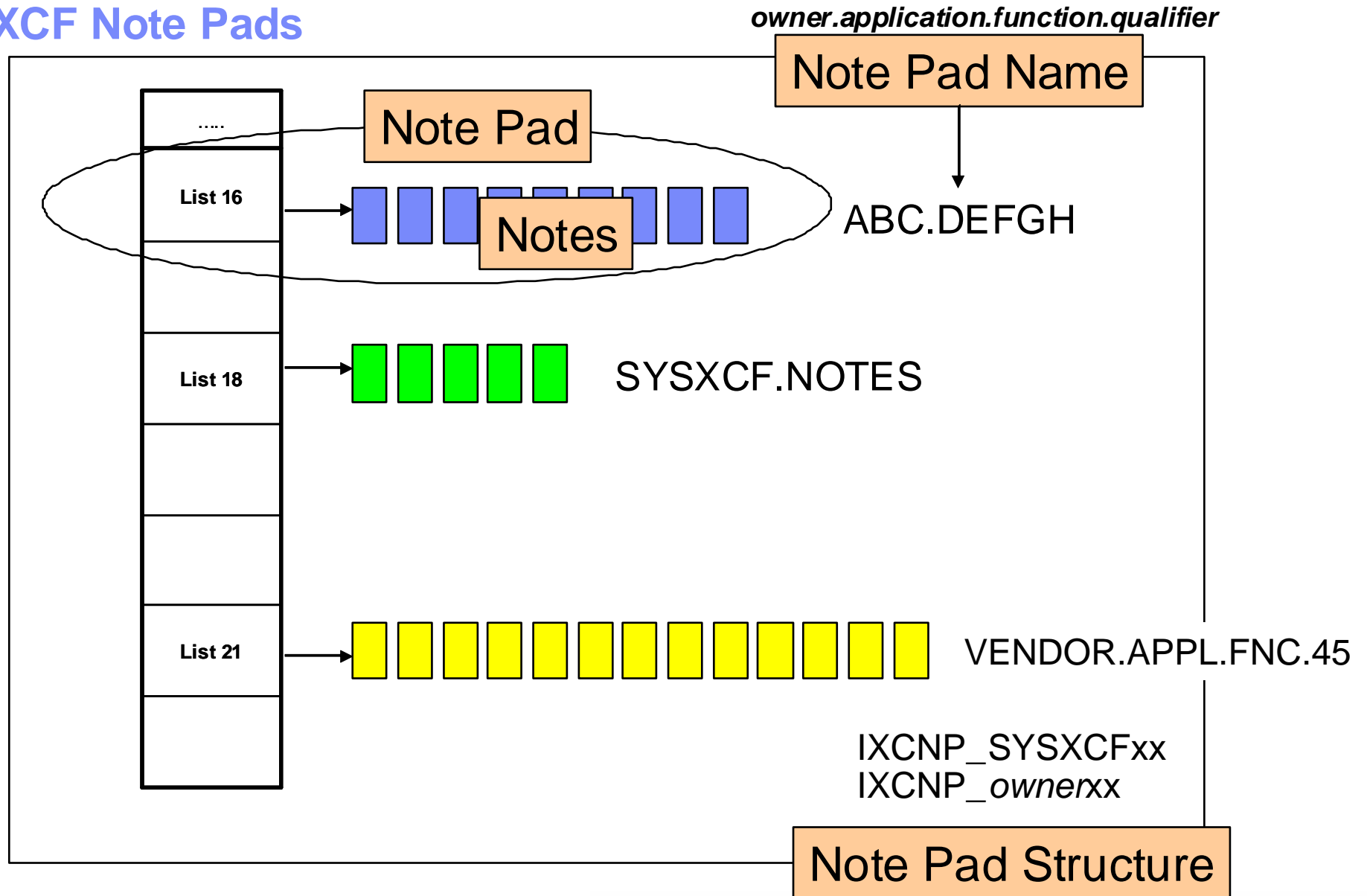
8 byte XCF Seq# for C/S

16 bytes of user metadata

1024 bytes of user data
(or none)

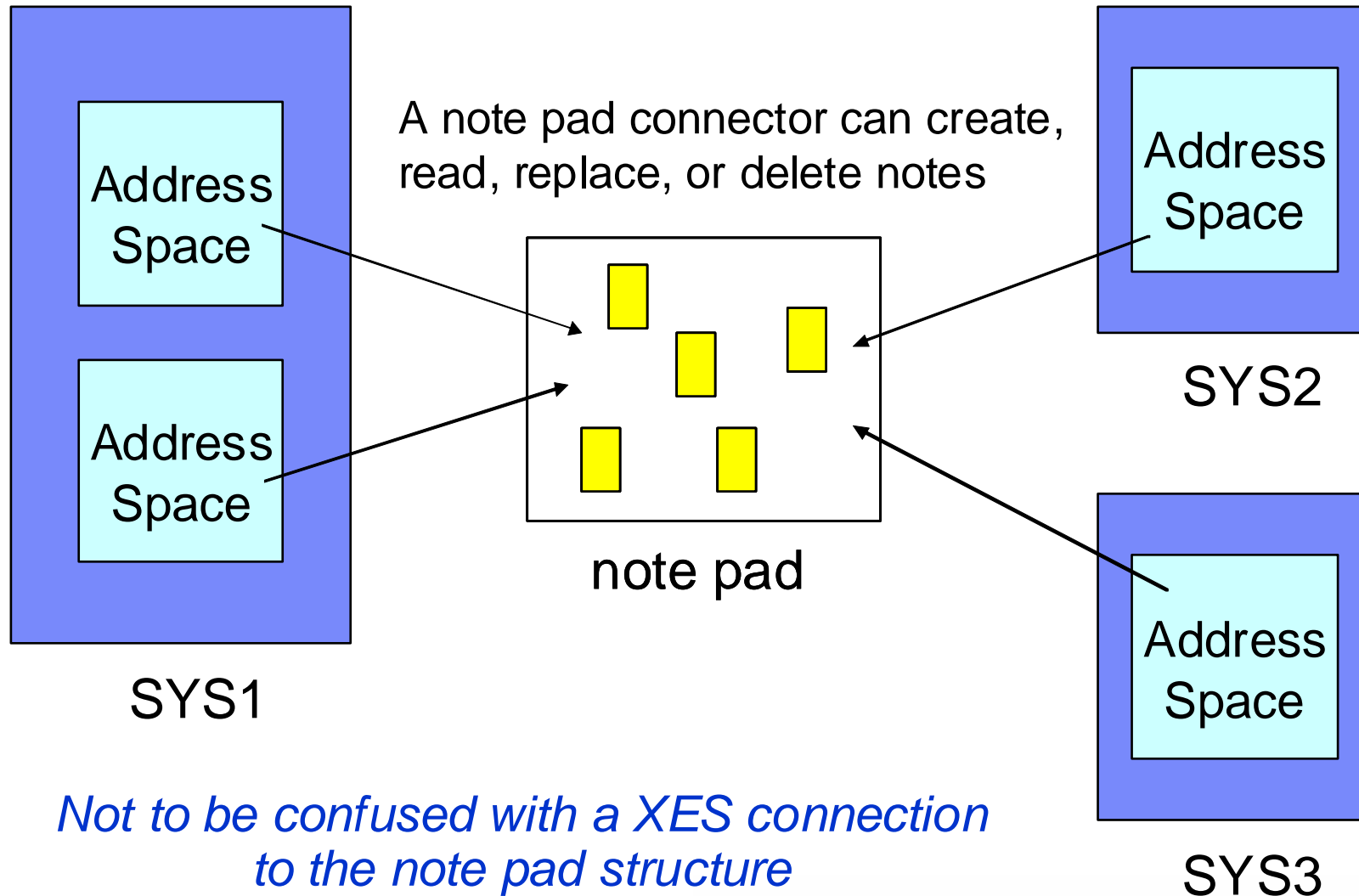
IXCNOTE Facility

XCF Note Pads



IXCNOTE Facility

Connections to a Note Pad



IXCNOTE Facility

XCF itself does not contain the information required to recreate the contents of a notepad structure if the CF is lost.

If the information in the notepad must be persistent the application can request a duplexed structure on the IXCNOTE Create request

IXCNOTE Facility

Notepad name format is owner.application.function.qualifier

Structure name format is:

- IXCNP_SYSXCFxx
- IXCNP_ownerxx

XCF allocates notepads to structures based on:

- The notepad owner name
 - If there is one or more structures matching that owner (owner-specific structure), XCF will select one of those structures if it is suitable
 - If there is no structure matching the owner, XCF will select one of the IXCNP_SYSXCFxx structures (community structures)
- Request for duplexed structure

IXCNOTE Facility

FACILITY Class Resource IXCNOTE.owner.application

CONTROL access

- Create or delete a note pad

UPDATE access

- Create connection with write access
- Write notes (when not recognized as valid user)

READ access

- Query note pad
- Create connection with read access
- Read notes (when not recognized as valid user)

Authorized if no resource profile or no SAF

IXCNOTE Facility

New D XCF,Notepad command

```
D XCF, { NOTEPAD | NP }  
    [ ,{NOTEPADNAME | NPNAME | NPNM}=notepadname | ALL ]  
    [ ,{STRNAME | STRNM}=hoststrname | ALL ]  
    [ ,SCOPE={SUMMARY | SUM} | {DETAIL | DET}
```

- Get list of note pads that have been defined
- Get detailed information about a note pad

- Can filter by note pad name/pattern
- Can filter by CF structure name

Use D XCF,STR,STRNAME=IXCNP_* to list note pad structures

IXCNOTE Facility

```
VMTOOL1
File Edit View Communication Actions Window Help
| SY1 *HZSSTMON: Frames currently in use by Health Checker: 13.652M
- SY1 d xcf,np
*SY1 *HZSSTMON: Frames currently in use by Health Checker: 13.699M
IXC442I 12.42.06 DISPLAY XCF          FRAME LAST  F      E      SYS=SY1
NOTEPAD NAME                          HOST STRUCTURE
SAP.APPL1.CHECKOUT.XCJN$B01           IXCNP_SAP01
SAP.APPL2.CHECKOUT.XCJN$B01           IXCNP_SAP01
SAP.APPL3.CHECKOUT.XCJN$B01           IXCNP_SAP01

IEE612I CN=SY1          DEVNUM=03E0 SYS=SY1          CMDSYS=SY1
```

IXCNOTE Facility

```

H - XA1_24X80.WS
File Edit View Communication Actions Window Help
Display Filter View Print Options Search Help
-----
SDSF SYSLOG      21.103 SY1  SY1  04/22/2012 3W          899      COLUMNS 52- 131
COMMAND INPUT ==>
SCROLL ==> CSR
0290  D XCF,NP,SCOPE=DET
0000  IXC443I  19.23.19  DISPLAY XCF 261
0000      INFO FOR NOTE PAD FCT.APPL1.CHECKOUT.XCJNM001
0000      DESCRIPTION: NOTEPAD1 DESCRIPTION XCJNM001
0000      HOST STRUCTURE: IXCNP_FCT01
0000      STATUS: CREATED
0000      SYSTEMS CONNECTED: SY1          SY2
0000      CREATED: 04/22/2012 19:22:48.361282
0000      LIST NUMBER: 784
0000      MAX TAG: E200D500C3000000      | S N C      |
0000      0000000000000000082          |          b      |
0000      CURRENT NUMBER OF NOTES: 0
0000
0000      NOTE PAD DEFINITION
0000      REQUIRED NUMBER OF NOTES: 20
0000      TAGGING: USER
0000      TRACK TAG: LIFETIME
0000      MULTIWRITE: YES
0000      INFO:          D5D6E3C5D7C1C4F1      | NOTEPAD1      |
0000      40C9D5C6D640E7C3      | INFO XC      |
0000      D1D5D4F0F0F14040      | JNM001      |
0000      4040404040404040
0000      4040404040404040
0000      4040404040404040
0000      4040404040404040
0000      4040404040404040
0000      4040404040404040
0000
0000      INFO FOR NOTE PAD FCT.APPL2.CHECKOUT.XCJNM001
0000      DESCRIPTION: NOTEPAD2 DESCRIPTION XCJNM001
0000      HOST STRUCTURE: IXCNP_FCT01
0000      STATUS: CREATED
0000      SYSTEMS CONNECTED: SY1          SY2
0000      CREATED: 04/22/2012 19:22:49.701680
0000      LIST NUMBER: 785
0000      MAX TAG: E200D500C3000000      | S N C      |
0000      00000000000000000A0          |          |
0000      CURRENT NUMBER OF NOTES: 5
F1=HELP      F2=SPLIT      F3=END      F4=RETURN      F5=IFIND      F6=BOOK
F7=UP      F8=DOWN      F9=STOP      F10=LEFT      F11=RIGHT      F12=RETRIEVE

```


IXCNOTE Facility

Summary:

- Greatly simplified mechanism to exploit the performance and flexibility of CF list structures.
- Available to unauthorized programs.
- Comprehensive operator interface to understand and control what is going on.
- Ability to provide any required level of security control over access to note pads.
- Delivered by APAR OA38450 on z/OS 1.13. Does not require CF Level 18. Is not rolled back to earlier releases.
- For more information, see Sysplex Services Guide and <http://publibz.boulder.ibm.com/zoslib/pdf/OA38450.pdf>



International Technical Support Organization and Authoring Services

Setting up BCPii and SSDPP Live Demo

IBM Redbooks

BCPii and SSDPP

In this session we will show you how to:

- Set up z/OS BCPii
 - Note that this is NOT the same as the BCPii function provided with Tivoli System Automation
 - z/OS BCPii is a pre-req for SSDPP
- Implement System Status Detection Partitioning Protocol (SSDPP)
 - Including a demo of the difference in how long it takes to partition a failed system time without and with SSDPP

Why BCPii and SSDPP?

If a member of a sysplex dies, it is probably holding resources that will be required by other members of the sysplex.

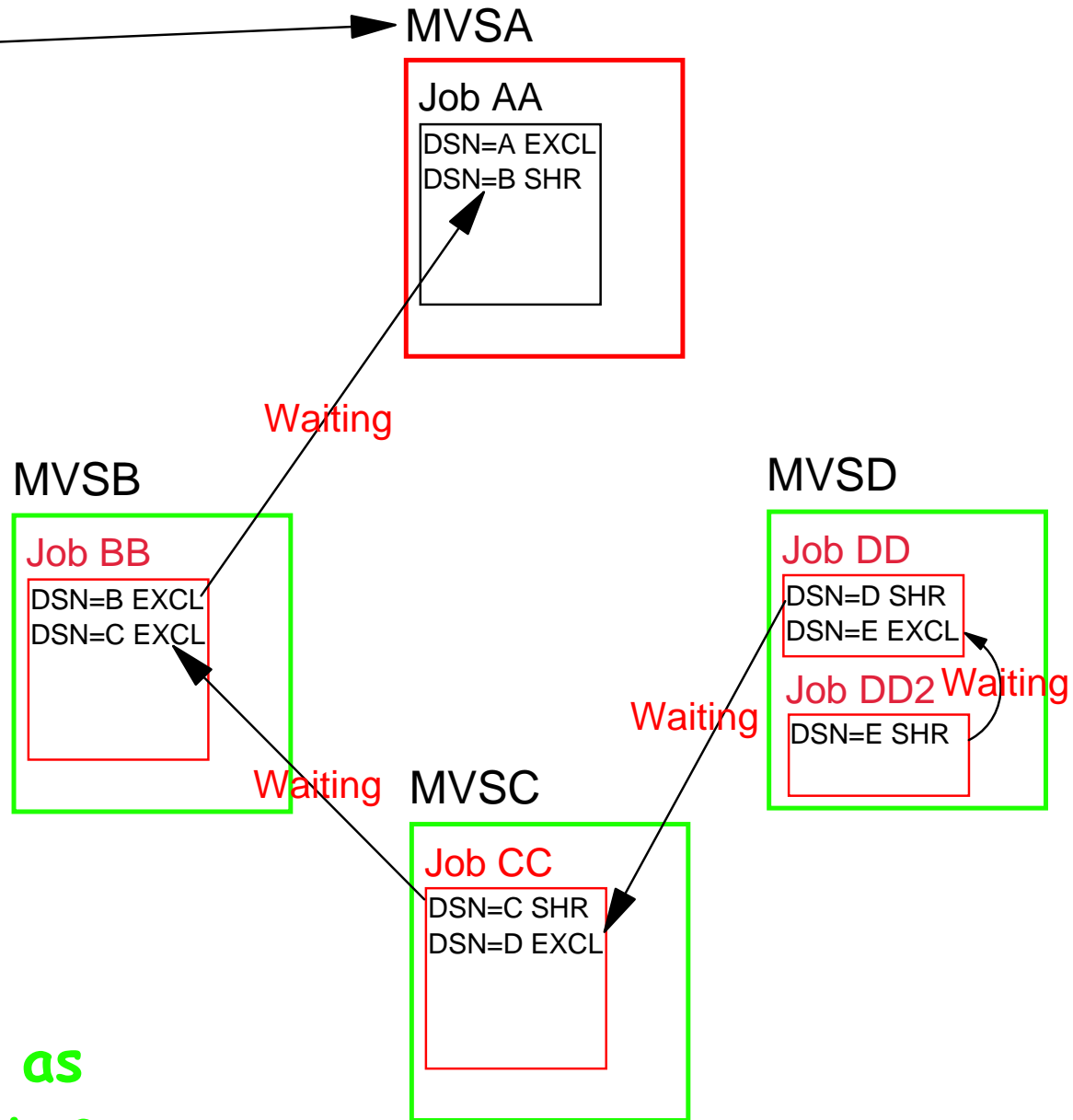
The longer this situation is allowed to go on for, the more units of work will be impacted.

If a system stops:

- It is probably holding resources that will be needed by another member of the sysplex.
- It will not release those resources until it recovers or is removed from plex

The longer a stalled system remains in the sysplex (holding resources), the larger is the impact on other systems.

So, OBVIOUSLY, the answer is to partition MVSA out of the plex as quickly as possible, right?



SSDPP benefits

Prior to z/OS 1.11, the only mechanism that z/OS had to determine the status of another member of the sysplex was to check that system's heartbeat in the sysplex CDS.

- If a system is going through recovery, it might not be able to update its heartbeat in the CDS. This means that you need to give a system some "reasonable" amount of time to recover before the system partitions the sick system out of the sysplex.
 - An IPL might take 30 minutes. Would you rather give a little more time for recovery to work, or kill it now and face an IPL? Your answer is probably "it depends on whether the system is dead or is in the middle of recovery".
 - Prior to z/OS 1.11, z/OS had no way to know whether another system was dead or trying to recover.
- SSDPP (and BCPii) changed that.

SSDPP benefits

First, let's see how long it takes to partition a failed system out of the sysplex **WITHOUT SSDPP....**

For our demo, we will use our little 2-way sysplex. The systems are called #@\$2 (LPAR A21 on z196) and #@\$A (LPAR A19 on zEC12). Both are running z/OS 1.13.

SSDPP benefits

First, let's see how the system is currently set up:

Failure detection interval (FDI)
(increased from 85 to 165) seconds
by z/OS 1.11

SFM action when FDI is exceeded

Does sysplex CDS support SSDPP?

These fields would be populated
if BCPii was working

```

D XCF,C
IXC357I 13.30.33 DISPLAY XCF 214
SYSTEM #@$2 DATA
INTERVAL  OPNOTIFY  MAXMSG  CLEANUP  RETRY  CLASSLEN
      165      168      2000      15      10      956

SSUM ACTION  SSUM INTERVAL  SSUM LIMIT  WEIGHT  MEMSTALLTIME
      ISOLATE      0      900      90      300

CFSTRHANGTIME
      900

...

SYSTEM STATUS DETECTION PARTITIONING PROTOCOL ELIGIBILITY:
SYSTEM CANNOT TARGET OTHER SYSTEMS.
REASON: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE PROTOCOL
SYSTEM IS NOT ELIGIBLE TO BE TARGETED BY OTHER SYSTEMS.
REASON: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE PROTOCOL

SYSTEM NODE DESCRIPTOR: 002817.IBM.02.0000000B3BD5
PARTITION: 21 CPCID: 00

SYSTEM IDENTIFIER: 3BD52817 21000008

NETWORK ADDRESS: N/A
PARTITION IMAGE NAME: N/A
IPL TOKEN: N/A
    
```



SSDPP benefits

BCPii is currently only running on #@\$A system (on zEC12)

Session B - gdps mop whitescreen.ws - [43 x 80]

File Edit View Communication Actions Window Help

Host: 9.12.6.50 Port: 23 LU Name: Disconnect

Display Filter View Print Options Search Help

SDSF	DA	#@\$2	(ALL)	PAG	0	CPU/L/Z	3/	2/	0	LINE	37-72	(112)
NP	JOBNAME	SIO	CPU%	ASID	ASIDX	EXCP-Cnt	CPU-Time	SR	Status	SysName	S	
	HIS	0.00	0.00	42	002A	6	0.00			#@\$A		
	HIS	0.00	0.00	82	0052	7	0.00			#@\$2		
	HWIBCPII	0.00	0.00	26	001A	2089	0.01			#@\$A		
	HZSPROC	0.00	0.00	307	0133	904	0.30			#@\$A		
	HZSPROC	0.00	0.00	44	0089	950	0.44			#@\$2		
	IEFSC			2		42	0.00			#@\$A		
	IEFSC			2		49	0.00			#@\$2		
	IOSAS			6		534	0.07			#@\$A		
	IOSAS			6		405	0.12			#@\$2		
	IXGLO			7		880	0.17			#@\$2		
	IXGLO			7		542	0.14			#@\$2		
	JESXC			3		858	0.03			#@\$A		
	JESXC			3		962	0.09			#@\$2		
	JES2					22536	1.56			#@\$A		
	JES2AUX	0.00	0.00	39	0027	16	0.00			#@\$A		
	JES2MON	0.00	0.00	38	0026	1	0.18			#@\$A		
	JES2S001	0.00	0.00	52	0034	106	0.00			#@\$A		
	JES3	0.00	0.50	34	0022	1442	0.25			#@\$2		
	JES3AUX	0.00	0.00	38	0026	13	0.00			#@\$2		
	JES3DLOG	0.00	0.00	37	0025	19	0.04			#@\$2		
	LLA	0.00	0.00	28	001C	9069	0.16			#@\$2		
	LLA	0.00	0.00	28	001C	6842	0.20			#@\$2		
	NET	0.00	0.00	32	0020	1655	0.22			#@\$A		
	NET	0.00	0.00	32	0020	1750	0.32			#@\$2		
	OMVS	0.00	0.00	16	0010	838	1.79			#@\$A		
	OMVS	0.00	0.00	16	0010	907	2.11			#@\$2		
	PAGENT	0.00	0.00	31	001F	4077	0.02		DW	#@\$A		
	PCAUTH	0.00	0.00	2	0002	23	0.00			#@\$A		
	PCAUTH	0.00	0.00	2	0002	24	0.00			#@\$2		
	PFA	0.00	0.00	308	0134	752	0.04		LW	#@\$A		
	PFA	0.00	0.00	39	0027	736	0.05		LW	#@\$2		
	RACF	0.00	0.00	45	002D	429	0.01			#@\$A		
	RACF	0.00	0.00	44	002C	452	0.01			#@\$2		
	RASP	0.00	0.00	3	0003	2	0.03			#@\$A		
	RASP	0.00	0.00	3	0003	2	0.01			#@\$2		
	RESOLVER	0.00	0.00	17	0011	289	0.00			#@\$A		

COMMAND INPUT ==>

F1=HELP F2=SPLIT F3=END F4=RETURN F5=IFIND F6=BOOK F7=UP F8=DOWN F9=SWAP nex F10=LEFT F11=RIGHT F12=RETRIEVE

SCROLL ==> CSR

MA B

Connected to remote server/host 9.12.6.50 using lu/pool SC38TCD7 and port 23

SSDPP benefits

Now let's wait-state the system and see how long we have to wait until we see the IXC105I Partitioning complete message... (should be a little under 3 minutes...)

Now we will set up BCPii and SSDPP and then repeat this exercise

BCPii setup

Address space (HWIBCPII) that provides authorized programs running on z/OS with the ability to query, change, and perform HMC-like functions against the System z processors on the HMC network.

Program communication from z/OS directly to HMC - no need for TCP access from z/OS to HMC, so may help address security concerns about exposing HMC network beyond the machine room.

Delivered with z/OS 1.11, and rolled back to z/OS 1.10 with APAR OA25426.

BCPii setup

Starting with z/OS 1.11, system automatically tries to start BCPII address space at IPL time.

- You don't need to add anything to `COMMNDxx`, or automation.

Successful start requires that certain setup has been carried out:

- Setup on the HMC:
 - Enable Cross Partition Authority for every LPAR that you want to be able to issue or be the target of BCPii commands.
 - Enable SNMP and define the Community Name.
 - Both of these can be changed non-disruptively if you wish
- Setup in z/OS
- SAF Security authorizations (in z/OS)

Need to give LPARs authority to issue commands to other LPARs...

Select CPC you want to set up BCPii on

Hardware Management Console

Operating System Messages kynef1 | Help | Logoff

Systems Management > Systems > SCZP301

Images | Topology

Select	Name	Status	Activation Profile	Last Used Profile	OS Name	OS Type	OS Level
<input type="radio"/>	A1D	Operating	A1D	A1D			
<input type="radio"/>	A1E	Not activated	A1E	A1E			
<input type="radio"/>	A1F	Not activated	A1F	A1F			
<input type="radio"/>	A21	Operating	TRAINER13	TRAINER13	#@\$2	z/OS	V1R13
<input type="radio"/>	A22	Operating	TRAINER13	TRAINER13	#@\$3	z/OS	V1R13
<input type="radio"/>	A23	Not Operating	ITSOZVM1	ITSOZVM1			
<input type="radio"/>	A24	Operating	ITSOZVM2	ITSOZVM2	ITSOZVM2	z/VM	6.2.0 - 1101
<input type="radio"/>	A25	Operating	A25	LBSIPL	SC90	z/OS	V1R12

Max Page Size: 500 Total: 54 Filtered: 54 Selected: 0

Tasks: SCZP301

- CPC Details
- Toggle Lock
- Daily
- Recovery
- Single Object Operations
- Service
- Change
- Remote Customization
- Operational Customization
- Definition
- Configuration
- Energy Management
- Monitor

Status: Exceptions and Messages

https://sczhmc7.itso.ibm.com/hmc/bon...estamp=138f2ddc699#tableTop_16a7f1f0

Select Single Object Operations

You are logged on to the SE

Support Element

System Management > SCZP301

Select	Name / ID	Status	Type	Description
<input type="checkbox"/>	Processors	OK		All Processors of the Server
<input type="checkbox"/>	Channels	Exceptions		All Physical Channel Identifiers of the Server
<input type="checkbox"/>	Cryptos	OK		All Crypto Channels of the Server
<input type="checkbox"/>	Partitions			All Partitions of the Server

Tasks: SCZP301

- CPC Details
- Toggle Lock
- Daily
- CPC Recovery
- Service
- Change Management
- CPC Remote Customization
- CPC Operational Customization
 - Automatic Activation
 - Change LPAR Controls
 - Change LPAR Group Controls
 - Change LPAR I/O Priority Queuing
 - Change LPAR Security
 - Customize/Delete Activation Profiles
 - Customize Scheduled Operations
- CPC Configuration
- Channel Operations
- Energy Management
- Monitor

Status: Exceptions and Messages

Select the CPC

Expand "CPC Operational Customization"

Select "Change LPAR Security"

Change Logical Partition Security - SCZP301

Input/output configuration data set (IOCDs): a2 IODF00

Logical Partition	Active	Performance Data Control	I/O Config Control	Cross Partition Authority	Partition Isolation	Basic Counter	Problem State Counter	Crypto Activity Counter	Extended Counter	G C
A16	No	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
A17	No	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
A18	No	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
A19	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
A2A	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
A2B	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A2E	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A2F	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A21	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
A22	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
A23	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A24	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A25	Yes	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A28	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A3E	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A3F	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A31	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A34	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A35	Yes	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
A1A	Yes	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Buttons: Save and Change, Change Running System, Save to Profiles, Reset, Cancel, Help

Remember that this must be done for every LPAR that will exploit BCPii

Enable "Cross Partition Authority"

Select Save and Change

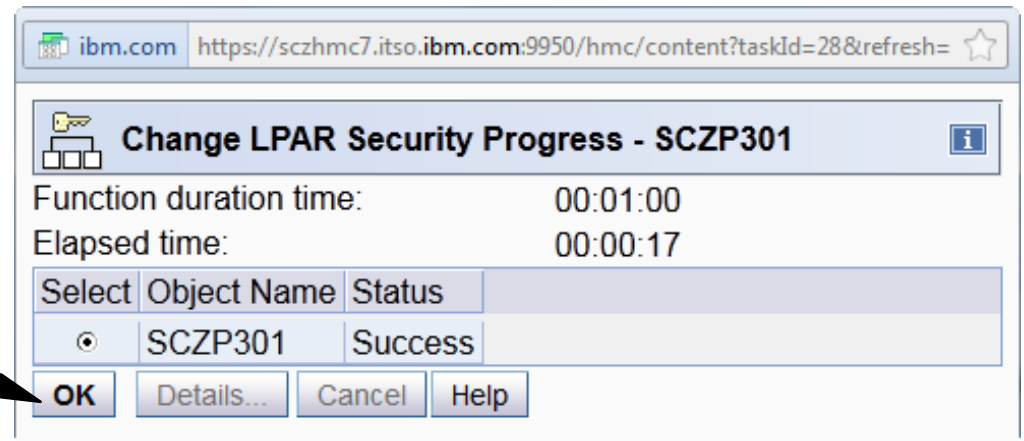
This should update activation profiles and implement change on active LPAR



This may take a little while

Press OK when finished

Recommend verifying that Activation Profiles were actually updated



BCPii setup

Next step is to add the SNMP definitions:

- These must be added in Single Object Operations for every CPC to be managed
- SE userid must have ACSADMIN authority to be able to do this....

Support Element

ibm.com https://sczhmc7.itso.ibm.com:9950/hmc/connects/mainuiFrameset.jsp

KYNEF | Help | Logoff

Welcome

System Management

SE Management

Service Management

Tasks Index

Status: Exceptions and Messages

Transferring data from sczhmc7.itso.ibm.com...

Create, customize, or verify the password rules assigned to the system users

User Profiles

Manage your system users that log onto the Hardware Management Console

User Patterns

Create, edit and remove user pattern definitions

Object Locking Settings

Change the automatic locking of managed objects.

Domain Security

Change console's domain name or password.

Configuration

Console Default User Settings

Customize the default appearance of the workplace

Customize API Settings

Customize the Application Programming Interface for the console

Customize Network Settings

View current network information and change settings

Migrate Channel Configuration Files

Migrate Channel Configuration Files

Define, customize and remove managed resource roles and task roles

User Templates

Create, edit and remove user template definitions

Manage Enterprise Directory Server Definitions

Create, edit and remove enterprise directory server definitions

Manage SSH Keys

Manage SSH Keys used for Secure FTP access

User Settings

Customize the appearance of the workplace

Customize Console Services

Customize the enablement of various console services

Customize Support Element Date/Time

Set time of day clocks of support elements for selected CPCs

Select SE Management

Then select "Customize API Settings"

Select "Enable
SNMP APIs"

Customize API Settings

SNMP

Enable Allow capacity change API requests

SNMP agent parameters:

Community Names

Select	Name	Address	Network Mask / Prefix	Access Type
--------	------	---------	-----------------------	-------------

Add... Change... Delete

SNMPv3 Users

Select	User Name	Access Type
--------	-----------	-------------

Add... Change... Delete

Event Notification Information

Specify any additional locations where SNMP trap messages will be sent.

Select	TCP/IP Address
--------	----------------

Add... Change... Delete

OK Cancel Help

Then click on Add in
Community Names
section

Fill in exactly as shown here.
Remember to select Read/Write
Then press OK

Community Name Information

Name: BCPII

Address: 127.0.0.1

Network mask / Prefix: 255.255.255.255

Access Type

Read only

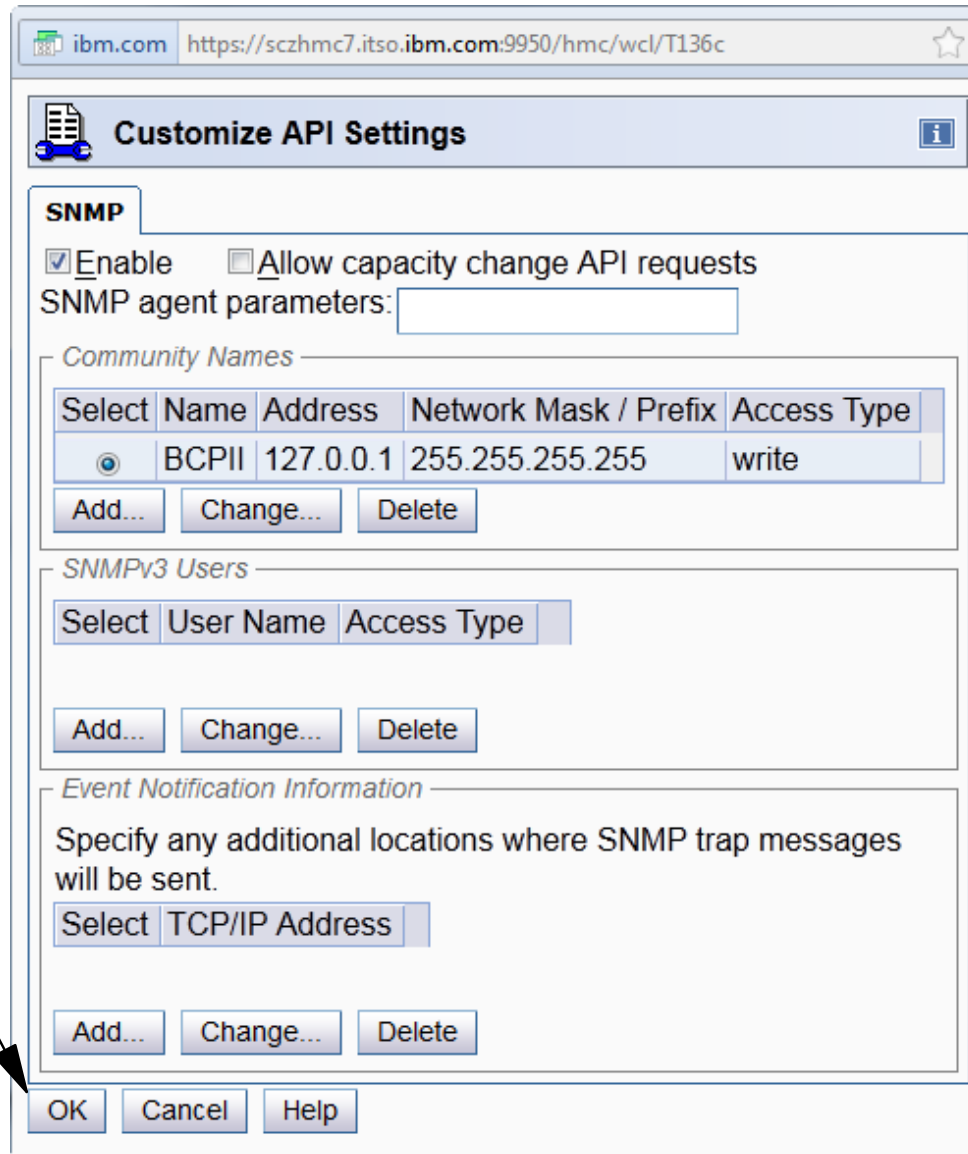
Read/write

OK Cancel Help

Name must be 1-16 chars, alphanumeric, no lower case.
Value you specify here must match name used in SAF CPC profile for this CPC

The Name value can be the same on every CPC, or different on every CPC. It is NOT necessary for each CPC to have a different Name value if you don't wish to.

Finally, click OK to apply and save the changes



The screenshot shows a web browser window with the URL `https://sczhmc7.itso.ibm.com:9950/hmc/wcl/T136c`. The page title is "Customize API Settings". The "SNMP" tab is active, showing the following configuration options:

- Enable Allow capacity change API requests
- SNMP agent parameters:
- Community Names table:

Select	Name	Address	Network Mask / Prefix	Access Type
<input checked="" type="radio"/>	BCPII	127.0.0.1	255.255.255.255	write
- SNMPv3 Users table:

Select	User Name	Access Type
--------	-----------	-------------
- Event Notification Information section with a "Specify any additional locations where SNMP trap messages will be sent." instruction and a "TCP/IP Address" table.

Buttons for "Add...", "Change...", and "Delete" are present for each table. At the bottom of the form are "OK", "Cancel", and "Help" buttons. A green callout box with an arrow points to the "OK" button.

The hardware setup for BCPii is now complete.....

BCPii setup

hlq.SCEERUN and hlq.SCEERUN2 must be in LNKLST.

Program authority:

- Program that will be calling BCPii services must reside in an APF-authorized library.

Issuing BCPii commands:

- The profile HWI.APPLNAME.HWISERV in the FACILITY resource class controls which applications can use BCPii services.
 - Anyone wishing to use BCPii must at least have READ access to this profile.
 - For XCF, simply have to ensure that the XCFAS started task is defined in RACF with the TRUSTED attribute - this is nearly always the case, but check to be sure.
- The FACILITY class must be RACLISTed.

BCPii setup

A BCPii application needs to have authority to the particular resource (CPC, Image, Capacity Record, Activation Profile) that it is trying to access (This is **IN ADDITION** to having access to the **HWISERV FACILITY** profile).

Profile names are:

- CPC: HWI.TARGET.netid.nau
- Image: HWI.TARGET.netid.nau.imagename
- Capacity Record: HWI.CAPREC.netid.nau.caprec
- Activation Profile: HWI.TARGET.netid.nau
- netid.nau is the 3-17 character SNA name for CPC (defined when you first define the SE to the HMC)

Level of access that is required depends on what you are trying to do - See Callable Services manual for details

BCPii setup

When defining the CPC profiles, APPLDATA must match the community name you specified on the SE:

- RDEFINE FACILITY HWI.TARGET.USIBMSC.SCZP301 UACC(NONE)
APPLDATA('BCPII')

```

BROWSE - RACF COMMAND OUTPUT----- LINE 00000000 COL 001 080
***** Top of Data *****
CLASS          NAME
-----
FACILITY      HWI.TARGET.USIBMSC.SCZP301

LEVEL  OWNER      UNIVERSAL ACCESS  YOUR ACCESS  WARNING
-----
00     KYNEF              NONE              NONE          NO

INSTALLATION DATA
-----
NONE

APPLICATION DATA
-----
BCPII

AUDITING
-----
FAILURES(READ)

NOTIFY
-----
NO USER TO BE NOTIFIED
***** Bottom of Data *****

```

You will need one of these for EACH CPC that will be managed using BCPii

COMMAND ==> F1=HELP F2=SPLIT F3=END F4=RETURN F5=RFIND F6=RCHANGE F7=UP F8=DOWN F9=SWAP nex F10=LEFT F11=RIGHT F12=RETRIEVE

BCPii setup

System automatically tries to start BCPII address space at every IPL:

- Address space name is HWIBCPII.
- Address space shows up in SDSF DA, but not in D A,L output.

Address space can be stopped using P HWIBCPII command:

- Once the address space is stopped, no BCPII calls will be processed.
- ENF signal is broadcast to let any interested parties know that the interface is stopping.
- If P command doesn't work, you can use a CANCEL HWIBCPII

**Address space can be started again using S HWISTART
(HWISTART is delivered in SYS1.PROCLIB)**

BCPii setup

There is currently no console command to check the status of BCPii.

If the SE/HW definitions are not in place at IPL time, address space will start and then stop.

So, if address space is active, that is at least a positive sign.

- Check for message HWI001I BCPII IS ACTIVE among IPL messages AND check for any messages immediately after that message.
- Doesn't guarantee that every CPC has been set up to support BCPII
- Currently the only way to check is from a program that uses the BCPII API

BCPii setup

Having completed the setup work on our CPC and in RACF, we now start BCPii address space:

```

Display Filter View Print Options Search Help
-----
SDSF OPERLOG DATE 08/04/2012 0 WTORS 1 FILTER COLUMNS 52- 131
-----
000210 -JOBNAME STEPNAME PROCSTEP RC EXCP CPU SRB VECT VAFF
CLOCK SERV PG PAGE SWAP VIO SWAPS
000210 -HWISTART STARTING HWISTART 00 0 1 .00 .00 .00 .00
.0 39 0 0 0 0 0 0
000210 -HWISTART ENDED. NAME- TOTAL CPU TIME= .00
TOTAL ELAPSED TIME= .0
000010 $HASP395 HWISTART ENDED
000200 IEA989I SLIP TRAP ID=X33E MATCHED. JOBNAME=*UNAVAIL, ASID=012D.
000201 IEF196I 1 //IEESYSAS JOB MSGLEVEL=1
000201 IEF196I 2 //HWIBCP11 EXEC IEESYSAS,PROG=HWIAMIN2
000201 IEF196I STMT NO. MESSAGE
000201 IEF196I 2 IEFC001I PROCEDURE IEESYSAS WAS EXPANDED USING
SYSTEM
000201 IEF196I LIBRARY SYS1.PROCLIB
000201 IEF196I 3 XXIEESYSAS PROC PROG=IEFBR14
000201 IEF196I 4 XXIEFPROC EXEC PGM=&PROG
000201 IEF196I XX* THE IEESYSAS PROCEDURE IS SPECIFIED IN THE
000201 IEF196I XX* PARAMETER LIST TO IEEMB881 BY MVS COMPONENTS
000201 IEF196I XX* STARTING FULL FUNCTION SYSTEM ADDRESS SPACES.
000201 IEF196I IEFC653I SUBSTITUTION JCL - PGM=HWIAMIN2
000200 IEE252I MEMBER CTIHWI00 FOUND IN SYS1.IBM.PARMLIB
000201 IEF196I IEF285I SYS1.PARMLIB KEPT
000201 IEF196I IEF285I VOL SER NOS= #@#$M1.
000201 IEF196I IEF285I SYS1.IBM.PARMLIB KEPT
000201 IEF196I IEF285I VOL SER NOS= Z1DRE1.
000010 HWI016I THE BCP11 COMMUNICATION RECOVERY ENVIRONMENT IS 962
000010 NOW ESTABLISHED.
000210 HWI007I BCP11 IS ATTEMPTING COMMUNICATION WITH THE LOCAL CENTRAL 963
000210 PROCESSOR COMPLEX (CPC).
000010 HWI001I BCP11 IS ACTIVE.
000000 IXC104I SYSTEM STATUS DETECTION PARTITIONING PROTOCOL ELIGIBILITY: 965
000000 SYSTEM CANNOT TARGET OTHER SYSTEMS.
000000 REASON: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE PROTOCOL
000000 SYSTEM IS NOT ELIGIBLE TO BE TARGETED BY OTHER SYSTEMS.
000000 REASON: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE PROTOCOL
***** BOTTOM OF DATA *****
COMMAND INPUT ==> SCROLL ==> CSR
F1=HELP F2=SPLIT F3=END F4=RETURN F5=IFIND F6=BOOK
F7=UP F8=DOWN F9=SWAP nex F10=LEFT F11=RIGHT F12=RETRIEVE

```

BCPii setup

Software:

- z/OS 1.11 (included in the base)
- z/OS 1.10 with APAR OA25426

Hardware:

- The program issuing the BCPii calls must be running on any CPC supported by z/OS 1.11 (z900 or later)
- It is always wise to keep CPCs (even old ones) at current microcode levels

The HWICMD function can only be used against z9 or later with the following microcode levels:

- z9: G40965.133
- z10: F85906.116

BCPii setup

z/OS 1.11 MVS Programming: Callable Services for High-Level Languages:

- Primary BCPii documentation including installation instructions and BCPii API documentation.

z/OS 1.11 MVS System Commands:

- START HWISTART and STOP HWIBCPII commands.

z/OS 1.11 MVS Diagnosis: Tools and Service Aids:

- BCPii's CTRACE documentation.

z/OS MVS Programming: Authorized Assembler Services Reference, Volume 2 (EDT-IXG):

- BCPii's ENF68 documentation.

Various SHARE presentations - see www.share.org

SSDPP setup

Now that BCPii is up and running (**NON-DISRUPTIVELY!!**), next step is to set up SSDPP.

System Status Detection Partitioning Protocol (SSDPP) is an enhancement to failed-system handling designed to partition a failed system from the sysplex in a more timely way and with improved data integrity.

SSDPP achieves this by exploiting the z/OS BCPii support to communicate with the SE to obtain the current status of an LPAR.

SSDPP setup

When a z/OS 1.11 or later system is IPLed using a correctly formatted Sysplex CDS, it writes new information about itself into the CDS. It gets this information from BCPii:

- The network name of the CPC it is running on (netid.nau).
- The name of the LPAR it resides in.
- An IPL Token.

Both the hardware and the software know the IPL Token:

- The IPL token is valid for the life of the IPL, as long as the system is still functioning.
- If the LPAR is RESET, the IPL Token in the hardware will change.
- If the LPAR waitstates (non-restartable), the IPL Token in the hardware will change.
- If the LPAR is IPLed, the IPL token will change.

All of this information is available to the other members of the sysplex via the Sysplex CDS and the BCPii.


```
D XCF,C
IXC357I 13.30.33 DISPLAY XCF 214
SYSTEM #@$2 DATA
```

...

```
SYSTEM STATUS DETECTION PARTITIONING PROTOCOL ELIGIBILITY:
SYSTEM CANNOT TARGET OTHER SYSTEMS.
REASON: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE PROTOCOL
SYSTEM IS NOT ELIGIBLE TO BE TARGETED BY OTHER SYSTEMS.
REASON: SYSPLEX COUPLE DATA SET NOT FORMATTED FOR THE PROTOCOL
```

```
SYSTEM NODE DESCRIPTOR: 002817.IBM.02.0000000B3BD5
PARTITION: 21 CPCID: 00
```

```
SYSTEM IDENTIFIER: 3BD52817 21000008
```

```
NETWORK ADDRESS: N/A
```

```
PARTITION IMAGE NAME: N/A
```

```
IPL TOKEN: N/A
```

Obtained via BCPii (if
SSD is active)

SSDPP setup

What do I need to do to enable SSDPP?

- The systems that will drive the System Status Detection Partitioning Protocol processing, or be the target of such processing, **MUST** be running on z10 EC GA2 or z10 BC GA1 or later.
- BCPii must be configured and functioning.
- XCFAS must be defined as TRUSTED to RACF or must have access to the required BCPii SAF profiles.
- Only z/OS 1.11 or later systems can exploit SSDPP, but previous levels can tolerate the new Sysplex CDS format that is required for SSDPP.

SSDPP setup

Let's check the format of our current sysplex CDS....

```

D XCF,C,TYPE=SYSPLEX
IXC358I 15.24.12 DISPLAY XCF 977
SYSPLEX COUPLE DATA SETS
PRIMARY   DSN: SYS1.XCF.CDS03
          VOLSER: #@$#X1      DEVN: D20F
          FORMAT TOD          MAXSYSTEM MAXGROUP(PEAK) MAXMEMBER(PEAK)
          04/12/2012 14:31:32      4      500    (42)      303    (8)
          ADDITIONAL INFORMATION:
          ALL TYPES OF COUPLE DATA SETS ARE SUPPORTED
          GRS STAR MODE IS SUPPORTED
ALTERNATE DSN: SYS1.XCF.CDS04
          VOLSER: #@$#X2      DEVN: D30F
          FORMAT TOD          MAXSYSTEM MAXGROUP          MAXMEMBER
          04/12/2012 14:31:36      4      500          303
          ADDITIONAL INFORMATION:
          ALL TYPES OF COUPLE DATA SETS ARE SUPPORTED
          GRS STAR MODE IS SUPPORTED

```

No mention of SSDPP support, so we need to move to correctly formatted Sysplex couple data sets.

SSDPP setup

Format 3 new Sysplex CDSs (primary, alternate, and spare) using the SSTATDET keyword:

```
//DEFCOUP JOB (0,0),'DEF XCF CDSS',NOTIFY=&SYSUID,  
// CLASS=A,MSGCLASS=X,REGION=0M  
//STEP1 EXEC PGM=IXCL1DSU  
//STEPLIB DD DSN=SYS1.MIGLIB,DISP=SHR  
//SYSPRINT DD SYSOUT=*  
//SYSIN DD *  
        DEFINEDS SYSPLEX(##$#PLEX)  
                DSN(SYS1.XCF.CDS05) VOLSER(##$#X1)  
                MAXSYSTEM(4)  
                CATALOG  
        DATA TYPE(SYSPLEX)  
                ITEM NAME(GRS) NUMBER(1)  
                ITEM NAME(GROUP) NUMBER(500)  
                ITEM NAME(MEMBER) NUMBER(303)  
                ITEM NAME(SSTATDET) NUMBER(1)  
...  
/*
```

SSDPP setup

Enabling SSD (cont)...

- Issue the SETXCF COUPLE,ACOUPL=dsn and SETXCF COUPLE,PSWITCH commands to roll the new CDSs into production.
- Note that after you activate a new CDS formatted for SSD, it may take a few seconds before you see:

```
IXC103I SYSTEM IDENTIFICATION INFORMATION 033
CONNECTION STATUS:    CONNECTED
SYSTEM NAME:         #@$2
SYSTEM NUMBER:       0100000E
IMAGE NAME:          A21
NODE DESCRIPTOR:     002817.IBM.02.0000000B3BD5
PARTITION NUMBER:    21
CPC ID:              00
NETWORK ADDRESS:     USIBMSC.SCZP301
IPL TOKEN:           C9F849E0 890FC7A5
IXC104I SYSTEM STATUS DETECTION PARTITIONING PROTOCOL ELIGIBILITY: 034
SYSTEM CAN TARGET OTHER SYSTEMS.
SYSTEM IS ELIGIBLE TO BE TARGETED BY OTHER SYSTEMS.
IXC111I LOGICAL PARTITION REMOTE CONNECTION INFORMATION 035
CONNECTION STATUS:    CONNECTED
SYSTEM NAME:         #@$3
SYSTEM NUMBER:       0200000F
IMAGE NAME:          A22
NETWORK ADDRESS:     USIBMSC.SCZP301
IPL TOKEN:           C9F84E37 44695DEB
DIAG INFO:           N/A
```

SSDPP setup

Check Sysplex CDS format now:

```

D XCF,C,TYPE=SYSPLEX
IXC358I 15.43.54 DISPLAY XCF 046
SYSPLEX COUPLE DATA SETS
PRIMARY   DSN: SYS1.XCF.CDS05
          VOLSER: #@$#X1      DEVN: D20F
          FORMAT TOD          MAXSYSTEM MAXGROUP(PEAK) MAXMEMBER(PEAK)
          08/04/2012 15:33:31      4      500      (42)      303      (8)
          ADDITIONAL INFORMATION:
          ALL TYPES OF COUPLE DATA SETS ARE SUPPORTED
          GRS STAR MODE IS SUPPORTED
          SYSTEM STATUS DETECTION PROTOCOL IS SUPPORTED
ALTERNATE DSN: SYS1.XCF.CDS06
          VOLSER: #@$#X2      DEVN: D30F
          FORMAT TOD          MAXSYSTEM MAXGROUP          MAXMEMBER
          08/04/2012 15:33:33      4      500          303
          ADDITIONAL INFORMATION:
          ALL TYPES OF COUPLE DATA SETS ARE SUPPORTED
          GRS STAR MODE IS SUPPORTED
          SYSTEM STATUS DETECTION PROTOCOL IS SUPPORTED

```

Remember to update COUPLExx to reflect new CDS names

SSDPP setup

Time to wait-state #@\$2 again and see how long recovery takes this time.....

```

2012217 15:49:07.51 JOB19311 00000010 $HASP373 LOADWAIT STARTED - INIT 1 - CLASS A - SYS #@$2
2012217 15:49:07.51 JOB19311 00000010 ZTT JOB#=00000001: LOADWAIT EXECUTION STARTED -- LEVEL ZOS1C.06.001
                                08/30/10 19.23
2012217 15:49:07.57                00000201 IEF196I IEF237I D057 ALLOCATED TO SYS00076
2012217 15:49:07.57                00000201 IEF196I IEF285I MSPCT.ZOS1CZTT.LOADLIB KEPT
2012217 15:49:07.57                00000201 IEF196I IEF285I VOL SER NOS= #@$#W1.
2012217 15:49:11.75 INTERNAL 00000010 IST1494I PATH SWITCH STARTED FOR RTP CNR00003 TO USIBMSC.#@$2M 284
                                284 00000010 IST1818I PATH SWITCH REASON: SHORT REQUEST RETRY LIMIT EXHAUSTED
                                284 00000010 IST314I END
2012217 15:49:16.52                00000000 IXC106I SYSTEM #@$2 285
                                285 00000000 RESET OR NEW IMAGE LOADED
2012217 15:49:16.52                00000000 IXC101I SYSPLEX PARTITIONING IN PROGRESS FOR #@$2 REQUESTED BY 286
                                286 00000000 XCFAS. REASON: SYSTEM RESET OR NEW IMAGE LOADED
2012217 15:49:16.53                00000200 IXC113I BCPII CONNECTION TO SYSTEM #@$2 RELEASED 287
                                287 00000200 DISCONNECT REASON: SYSTEM REMOVED FROM SYSPLEX
                                287 00000200 IMAGE NAME: A21
                                287 00000200 NETWORK ADDRESS: USIBMSC.SCZP301
                                287 00000200 SYSTEM NUMBER: 0100000E
                                287 00000200 IPL TOKEN: C9F849E0 890FC7A5

```

So it took about 30 minutes to implement and it saved about 2.5 minutes on every unplanned outage

SSDPP setup

Anything else?

You can turn the use of SSDPP on or off dynamically at the system level using the SETXCF FUNCTIONS command and/or in COUPLExx member if you wish:

- Default is ENABLED - this is the recommended setting
- If you DISABLE SSDPP on a system, that system cannot be the target of any BCPii-related actions and will not use BCPii to initiate actions against any other systems.

SSDPP

Summary:

- Prereqs:
 - z10 GA2 or later
 - z/OS 1.11
 - Correctly formatted Sysplex CDS
 - Implement BCPii
- System Status Detection Partitioning Protocol is a significant step forward. This is the most fundamental change to handling of system failures since the introduction of SFM.
- Easy to implement.
- You can start to enable it as soon as your first z10 z/OS system moves to z/OS 1.11 - no need to wait for the whole sysplex to be upgraded.



International Technical Support Organization and Authoring Services

Understanding and exploiting HiperDispatch

IBM Redbooks

Topics

Modern System z processor cache architecture

- To explain WHY we invented HiperDispatch

HiperDispatch description

- What HiperDispatch does

Controlling HiperDispatch

- Where, when, and how to use HiperDispatch

Monitoring HiperDispatch

- RMF support for HiperDispatch

Background

How do you make a computer chip faster?

- The speed of electricity is more or less fixed, so we can't make that go faster.
- If you can't make it go faster, then make its journey shorter.....

The fastest-growing inhibitor to ever-faster processor speeds is how long it takes light to travel from one side of a chip to the other.

Compared to accessing the on-chip cache, having to travel ALL the way out to memory to get an instruction or some data is an eternity

To try to reduce the time the processor is waiting to get its next instruction or a piece of data, IBM provides various levels of cache

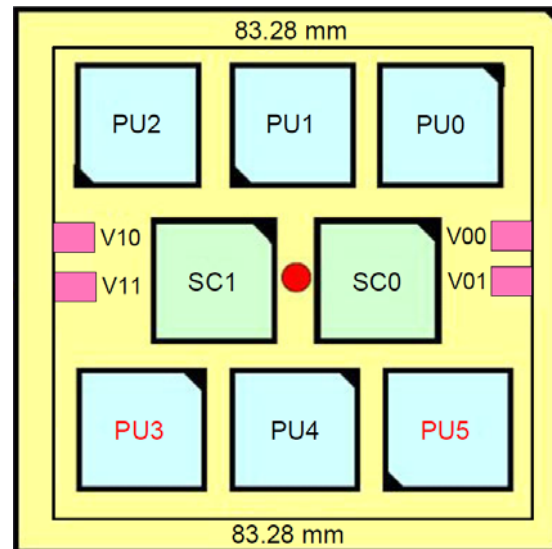
- The fastest (most expensive) cache is closest to the processing core on the chip.
- As you get further from the chip, the cache size gets larger, and it is shared by more processors.

z196 architecture

A z196 can have from 1 to 4 books.

Each book contains one Multi Chip Module (MCM)

Each MCM contains 6 processor unit chips and 2 controller chips

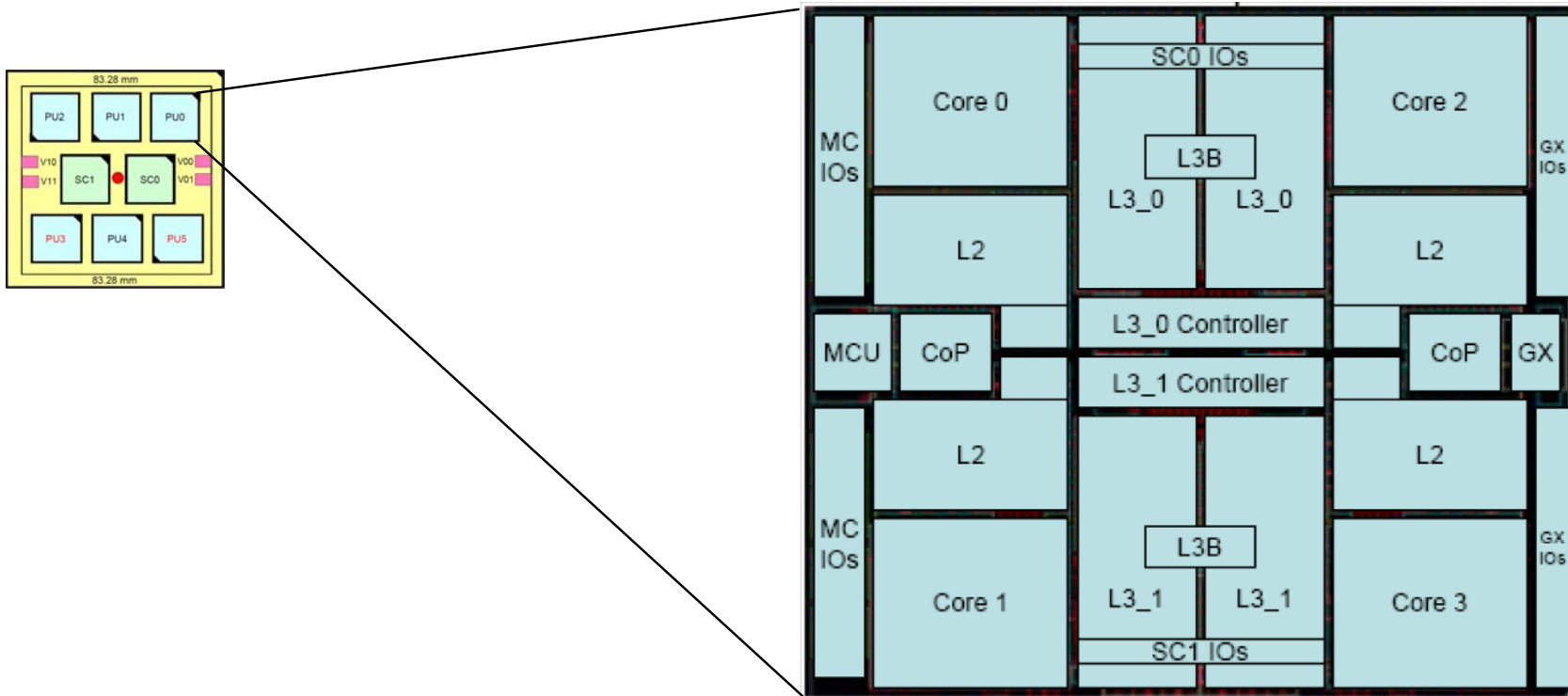


Each processor unit chip contains 4 cores (up to 6 on zEC12)

- Each core appears as a physical processor to PR/SM and z/OS

z196 architecture

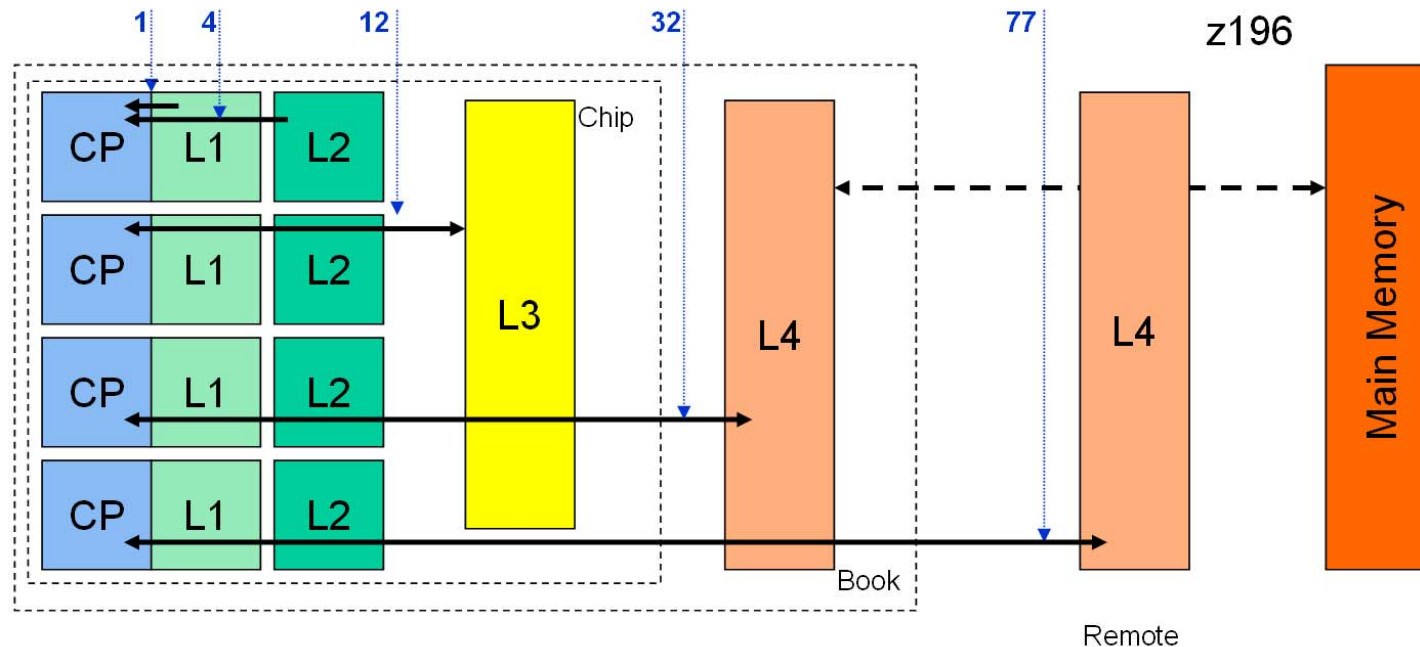
Each z196 PU chip looks something like this:



Level 1 cache resides on core. Each Level 2 cache is dedicated to its associated core. Level 3 cache is shared by all cores on that chip. Level 4 caches are on SC chips and shared between all PUs in that book.

Background

Relationship between processor caches and the PU in a z196



You can see the big difference in access time depending on where the data and instructions need to be retrieved from

Background

Prior to HiperDispatch, PR/SM would *try* to re-dispatch a logical processor on the same physical processor (and therefore beside the same caches) that it was dispatched on last time.

If HiperDispatch is not enabled, z/OS has a single queue of work waiting to use a general purpose CP, and attempts to balance the load across all its online logical processors. **NO ATTEMPT IS MADE TO DISPATCH A PIECE OF WORK ON THE SAME LOGICAL PROCESSOR THAT IT LAST RAN ON.** So even if the logical processor **DOES** get re-dispatched on the same physical processor, it is probably not running the same work as the last time it was dispatched.

The result of this is that as the number of chips and books grows, the impact of having to wait to get data and instructions from "a long" way away increases.

Background

Because Moore's Law is coming to an end, all computer manufacturers are being forced to grow horizontally (MORE chips) rather than vertically (FASTER chips)



Even PHONES are coming with multi-core chips now!!



Background

A little more background.....

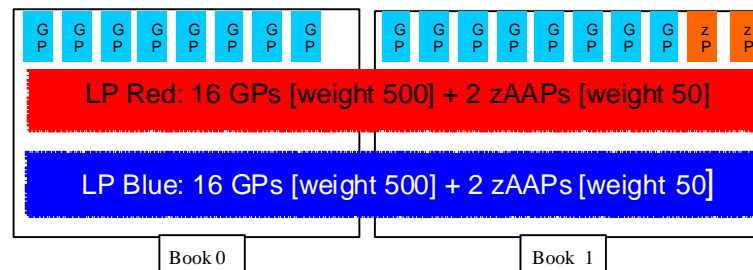
What determines how much CPU service an LPAR will receive?

- PR/SM attempts to guarantee that an LPAR will get at least the share of the total capacity based on its weight (this is called its "fair share")
 - Critical to remember that weight is NOT a relative dispatching priority. An LPAR that is using less than its fair share will have a higher queue priority than an LPAR with a larger weight that is using more than its fair share.
- The LPAR must have enough CPs defined. If I want the LPAR to get 50% of the capacity on a 6-way CPC, I have to define the LPAR with at least 3 logical CPs.
- If I have enough work to consume more than my fair share, AND other LPARs are not using all of their fair share, then PR/SM will distribute the unused capacity to those LPARs that want it, again based on their relative weights.
 - The weight is not a cap, unless you enable Capping for that LPAR on the HMC

Background

In order to be able to get more CPU than your fair share, it is normal to define an LPAR with more logical CPs than are required to get your fair share.

- If multiple LPARs are defined in this way, PR/SM has more logical CPs to manage than it has physical CPs to dispatch them on.



Background

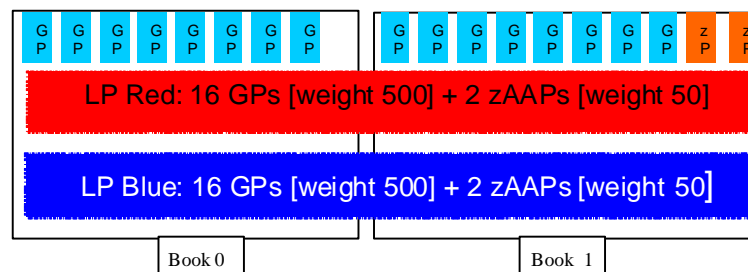
Traditionally, PR/SM distributes a partition's fair share evenly across all the online logical processors in the LPAR

- Actual usage is tracked at the logical CP level, and the relative priority of that logical CP when looking for a physical CP reflects whether the logical CP is ahead of or behind its fair share.
- So, the more logical CPs there are in the LPAR, the lower will be the share of each logical CP, and therefore the lower will be its priority when waiting to get dispatched again

Background

This means:

- Each logical CP will only get a portion of a physical CP - this makes it look like z/OS is running on a slower processor than it actually is.
- Extra work for PR/SM because it has more logical CPs to manage (higher "LPAR Overhead").
- An ever-decreasing chance of a unit of work getting redispached on the same physical CP that it was running on previously, thereby losing the benefit of the contents of that cache.



HiperDispatch

To address this, and position System z for future generations of multi-book CPCs, IBM introduced a new capability called HiperDispatch together with the z10 generation of CPCs.

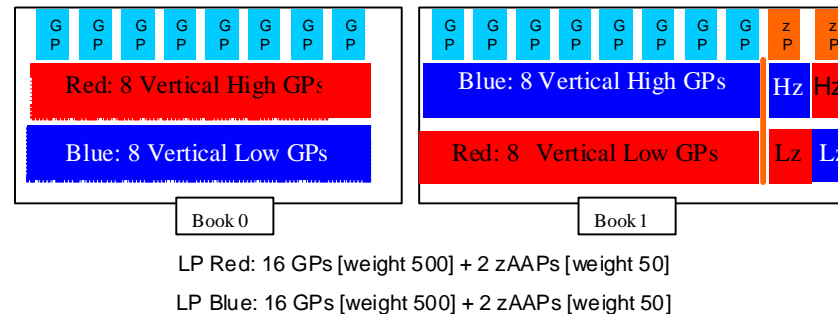
The objective of HiperDispatch is to try to maximize the number of "MIPS" delivered by a given configuration by optimizing the value of processor cache. It achieves this by increasing the likelihood of work being re-dispatched closer to the cache it was using previously.

It consists of support in PR/SM and in z/OS (and supporting products).

HiperDispatch

So what does HiperDispatch do?

- It tries to more closely align the number of logical CPs that z/OS will actively use, with the amount of capacity that is guaranteed by its weight.



- It attempts to pseudo-dedicate logical CPs to physical CPs (cores) where possible, thereby greatly increasing the chance of a logical CP being able to benefit from the cache contents
- It groups logical CPs into "affinity nodes" (based on topology information provided by PR/SM) and attempts to re-dispatch work on the same affinity node that it was running on previously

HiperDispatch

HiperDispatch has three classifications of logical processors:

- Ones where the logical processor is guaranteed nearly 100% of a physical processor.
 - So every time that logical processor is dispatched, it will be dispatched on the same physical processor (using the same cache that it used when it was last dispatched). These are called "Vertical High (VH)" logical CPs.
- Ones where the logical processor is guaranteed between 1 and 100% of a pool of physical processors.
 - These are called "Vertical Medium (VM)" logical CPs.
- Ones that are currently unneeded, based on the current capacity requirements of the LPAR.
 - These are online, however they currently are in a long wait so they will not get dispatched on a physical processor.
 - These are called "Vertical Low (VL)". Also called "Parked" processors.

HiperDispatch

How does z/OS and PR/SM decide how many of each type of engine the LPAR will have?

Remember that the objective is to try to dedicate logical CPs to physical CPs AND try to optimize the buffer value for vertical medium logical CPs and to avoid "short CPs"...

PR/SM and z/OS work cooperatively to create an optimal configuration

HiperDispatch

PR/SM uses the LPAR's weight and number of logical CPs to determine how many logical CPs worth of capacity the LPAR is guaranteed and categorize logical CPs based on the following rules:

- Attempts to have as many VHS as possible.
- However, every VM should have at least 50%.
- Every LPAR should have at least 1 VM.
- He will assign VHS and at most two VMs until the LPAR reaches the number of logical CPs that can deliver its fair share
 - Remaining (VL) engines will be Parked

HiperDispatch

An example.....

Logical partition share is 640% (6.4 CPUs) and LPAR is defined with 8 logical processors

- 5 LPs are VH with 100% share
- 2 LPs are VM with 140% share between them - mediums should have at least 50% share
- 1 LP is VL - it was parked 85.78% and busy 13.84%

C P U A C T I V I T Y											
z/OS V1R8					SYSTEM ID UNKN		DATE 11/26/2007				
					RPT VERSION V1R8 RMF		TIME 22.33.43				
CPU	2097	MODEL	732	H/W MODEL	E40	SEQUENCE	CODE	000000000000DC6CE	HIPERDISPATCH=	YES	
---	CPU---	----- TIME % -----				LOG PROC		--I/O INTERRUPTS--			
NUM	TYPE	ONLINE	LPAR	BUSY	MVS	BUSY	PARKED	SHARE	%	RATE	% VIA TPI
0	CP	100.00	96.33	97.34	0.00	100.0	5.80	48.75			
1	CP	100.00	95.96	97.07	0.00	100.0	4.59	55.30			
2	CP	100.00	95.79	96.84	0.00	100.0	5.10	55.18			
3	CP	100.00	95.46	96.68	0.00	100.0	2.40	53.75			
4	CP	100.00	95.08	96.41	0.00	100.0	8435	10.05			
5	CP	100.00	73.92	96.86	0.00	70.0	20.74	4.95			
6	CP	100.00	74.33	97.13	0.00	70.0	14.15	19.39			
7	CP	100.00	13.84	98.89	85.78	0.0	0.00	0.00			
TOTAL/AVERAGE			80.09	96.94	640.0		8488	10.14			

HiperDispatch

How about an LPAR with a weight < 1 CP and defined with 1 logical CP?

- It will have one VM Logical CP

How about an LPAR with a weight = 1 CP and defined with 1 logical CP?

- It will have one VM Logical CP

What about an LPAR with a weight = 1 CP and defined with 3 logical CPs?

- It will have 2 VM Logical CPs and 1 parked CP

What about an LPAR with a weight = 1.2 CPs and defined with 3 logical CPs?

- It will have 2 VM Logical CPs and 1 parked CP

HiperDispatch

What is a "short CP"?

VM and unparked VL logical CPs are queued for access to physical CPs based on whether they are ahead of, or behind, their fair share (for that logical CP).

- A logical CP that is ahead of its fair share will have a very low priority
- A logical CP with a low weight it likely to be ahead of its fair share, and therefore have a low priority and have to queue for a long time.
- What happens to a task that was running on that logical CP when its time slice expired? It will not be able to complete until the logical CP is dispatched again. Imagine the impact if the waiting task was running some critical task or holding some critical resource.....

This is why HiperDispatch tries to avoid having logical CPs with low weights.....

HiperDispatch

What is the "Warning-Track Interruption Facility"?

Let's say that you have some piece of work running on a VL or VM LCP and that LCP gets to the end of its time slice....

- Physical engine gets ripped away from that LCP.
- As far as z/OS is concerned, that work is still running.
- That work can't complete until the LCP gets dispatched again.
- Depending on the CPC utilization and the relative priority of that LCP (which is probably low if it used all its timeslice), it could take a long time before it gets dispatched again.
- Any other work in the system that is waiting for that work to complete now has to wait.

HiperDispatch

To avoid this situation, zEC12 provides a new type of interrupt

- Warning Track Interrupt
- Indicates that the physical CP is about to be taken away from the LCP

PR/SM sends a Warning Track interrupt to a VL or VM LCP in one of the following situations:

- The LCP is getting close to the end of its timeslice.
- The LCPs is running on a core that belongs to a VH LCP, and that VH LCP now wants to use the processor.

Assuming that the LCP is enabled for interrupts, z/OS then requeues that work and releases the physical engine back to PR/SM.

Rather than having to wait for the LCP to get dispatched again, the interrupted work can run on another LCP and complete

HiperDispatch

Warning-Track Interruption Facility is only available on zEC12.

z/OS support delivered in APAR OA37186

RMF APAR OA37803 adds information about the use of Warning-Track Interruption Facility to its Type 70 SMF records (in the CPU data section):

- SMF70WTS Number of times PR/SM issued a warning-track interruption to a logical processorz/OS was able to return the logical processor within the grace period.
- SMF70WTU Number of times z/OS was NOT able to return the logical processor within the grace period.
- SMF70WTI Amount of time in milliseconds that a logical processor was not dispatched on a physical CP.

These fields are NOT reported by RMF PP or Mon III, but they CAN be viewed using RMF exception or overview reports

HiperDispatch

How do the various components support HiperDispatch?

– PR/SM

- Supplies topology information/updates to z/OS
- Ties vertical high logical processors to physicals (gives 100% share)
- Distributes remaining share to VM and unparked VL LCPs
- Distributes unused share to unparked VL LCPs

z/OS

- Ties tasks to affinity nodes
- Dispatches work to affinity nodes
- "Parks" vertical low processors that are no longer needed
- Hardware cache optimization occurs when a given unit of work is consistently dispatched on the same physical CPU or at least a core in the same PU chip (same L3/L4)

HiperDispatch

Who will benefit the most?

- LPARs with many more logical CPs than are needed to deliver the LPAR's fair share.
- Configurations with high logical to physical CP ratios
- Configurations with multiple books
- LPARs that use both general purpose CPs and zIIPs and or zAAPs

Are there drawbacks?

- In order to optimize use of the cache, HiperDispatch will try to minimize the number of active logical CPs. With fewer servers, queue time (waiting for CP) might increase.
- Because there are likely to be fewer in-use LCPs, appropriate assignment of WLM importances becomes more important

HiperDispatch

Does PR/SM dedicate a physical CP to a vertical high logical CP?

- If the logical CP associated with the physical has no work to do, PR/SM can give that physical to another LCP. However, as soon as the LCP is ready to run, PR/SM will pre-empt whatever is running on the physical CP and give control back to the VH LCP.
- This behavior only applies to VH LCPs. Nothing will be pre-empted if a VM LCP comes ready.

If I am not using all my fair share, will HiperDispatch turn a VH into a VM?

- No, unless the LPAR weight or the number of online LCPs is changed, the number of VH LCPs will not change. However other LCPs will probably be dispatched on the physical CPs that are pseudo-dedicated to the VH LCPs.

HiperDispatch

How does MVS dispatching change when HiperDispatch is enabled?

- Work is now re-queued to the affinity node that it was using previously. Rather than trying to balance work across all online logical CPs, the Dispatcher will now try to balance work across the available affinity nodes, taking account of the number of available logical CPs in each node.

Does PR/SM do anything for special purpose engines (zIIP and zAAP)?

- If the LPAR has enough of a share of a special purpose engine to have a VH logical zAAP or zIIP, HiperDispatch will (if possible) use a core in the same book as the VH CPs to deliver the VH physical zAAP or zIIP

HiperDispatch

How does HiperDispatch decide to park or unpark an engine?

Reason for Un-Parking:

- CP
 - MVS Busy > 95% and enough free capacity available on CEC
- zXXP
 - MVS Busy > 80%

Reason for Parking:

- CP
 - MVS Busy < 80%
- zXXP
 - MVS Busy < 66%



HiperDispatch

How does this relate to Vary CPU management in IRD?

- HiperDispatch puts parked engines into a long wait. The overhead of taking it out of that wait is nearly zero.
- IRD Varies CPs on and offline - this takes a lot more work (AND more time) than putting the CPs into a wait.
- IRD only works with general purpose CPs (doesn't do anything with special purpose engines).
- HiperDispatch revisits its decision about how many logical CPs it needs every 2 seconds. Vary CPU management does this every 10 seconds.
- If you enable both VARYCPU and HIPERDISPATCH in IEAOPTxx, the system will automatically turn VARYCPU off.

Setting up HiperDispatch

Starting with z/OS 1.13, HiperDispatch will automatically be enabled on z196 and later CPCs.

For prior releases, HiperDispatch is enabled using the HIPERDISPATCH keyword in the IEAOPTxx member.

IF you experience response time increases for a high priority server address space which you feel is intolerable, *and* your application is one with many short requests for processor, you might try assigning that address space to SYSSTC

- SYSSTC units of work can be dispatched anywhere - they are not tied to an affinity node.

Avoid discretionary service classes unless the work really is discretionary

The “control global performance data” security setting must be enabled for proper operation of HiperDispatch (this is the default)

Monitoring HiperDispatch

How can you see what HiperDispatch is doing?

The RMF CPU report has been updated to show the weight of each logical CP (LOG PROC SHARE %) and the percent of time that the logical CP was parked (PARKED)

The calculation of MVS Busy Time has been revised to take Parked time into account:

$$\begin{array}{l}
 - \quad \text{Online Time} - (\text{Wait Time} + \text{Parked Time}) \\
 - \text{MVS BUSY TIME \%} = \frac{\text{-----}}{\text{Online Time} - \text{Parked Time}} * 100 \\
 -
 \end{array}$$

SMF Type 113 records (created by the HIS address space) contain information about how effectively cache is being used. Type 99 (WLM) records also contain valuable information.

Further information

z/OS: Planning Considerations for HiperDispatch Mode White Paper
available on Techdocs

*Setting Up and Using the IBM System z CPU Measurement Facility
with z/OS*, REDP-4727

z/OS Intelligent Resource Director, SG24-5952

IBM zEnterprise 196 microprocessor and cache subsystem, article in
February 2012 issue of IBM Journal of Research and Development

Alain Maneville's HD paper and tool, available on WLM home page:

<http://www.ibm.com/systems/z/os/zos/features/wlm/tools/WLMsetupdesigntools.html>

Various SHARE and other user group presentations

FEEDING TIME!!





International Technical Support Organization and Authoring Services

JES2 Dynamic Proclib

IBM Redbooks

JES2 Dynamic Proclib

Who amongst you can honestly say that you never had a JCL error in your JES2 proc?

Did you ever have someone delete a JES2 proclib and only find out that it is gone the next time you tried to IPL?

How much fun is it to do a MAS-wide restart of JES2 so you can add a proclib to JES2?

The answer to your problems is here (and has been here for the last 10 years!) thanks to those nice JES2 Development people - Dynamic Proclib support

JES2 Dynamic Proclib

What can you do with Dynamic Proclib?

- Change proclib concatenations without touching JES2 JCL.
- Bypass errors in proclib definitions.
- Display PROCxx definitions.
- Dynamically add PROCxx definitions.
- Dynamically MODIFY existing PROCxx definitions.
- Add a new PROCxx definition, test it, and then rename it.
- Delete PROCxx definitions.

Let's see some examples...

JES2 Dynamic Proclib

Here is what we started with:

```
//JES2      PROC M=J2USECF
//IEFPROC EXEC PGM=HASJES20 ,TIME=1440 ,DPRTY=(15 ,14)
//HASPLIST  DD DDNAME=IEFRDER
//HASPPARM  DD DSN=SYS1.PARMLIB(&M) ,DISP=SHR
//PROC00    DD DSN=SYS1.DIST.PROCLIB ,DISP=SHR
//          DD DSN=SYS1.PROCLIB ,DISP=SHR
//          DD DSN=SYS1.IBM.PROCLIB ,DISP=SHR
```

JES2 Dynamic Proclib

To add a new data set to PROC00, we need to update JCL:

```
//JES2      PROC M=J2USECF
//IEFPROC  EXEC PGM=HASJES20,TIME=1440,DPRTY=(15,14)
//HASPLIST  DD DDNAME=IEFRDER
//HASPPARM  DD DSN=SYS1.PARMLIB(&M),DISP=SHR
//PROC00    DD DSN=SYS1.DIST.PROCLIB,DISP=SHR
//          DD DSN=SYS1.PROCLIB,DISP=SHR
//          DD DSN=SYS1.IBM.PROCLIB,DISP=SHR
//          DD DSN=SYS1.KYNEF.PROCLIB,DISP=SHR
```

And do a MAS-wide JES2 restart.

JES2 Dynamic Proclib

What happens if we mess up the JCL?

```

S JES2,PARM='NOREQ'
IEF196I          1 //JES2      JOB MSGLEVEL=1
IEF196I          2 //STARTING EXEC JES2,PARM='NOREQ'
IEF196I STMT NO. MESSAGE
IEF196I          2 IEF001I PROCEDURE JES2 WAS EXPANDED USING SYSTEM
IEF196I LIBRARY SYS1.PROCLIB
IEF196I          3 XXJES2      PROC M=J2USECF
IEF196I          4 XXIEFPROC EXEC PGM=HASJES20,TIME=1440,DPRTY=(15,14)
IEF196I          5 XXHASPLIST DD DDNAME=IEFRDER
IEF196I          6 XXHASPPARM DD DSN=SYS1.PARMLIB(&M),DISP=SHR
IEF196I          IEF0653I SUBSTITUTION JCL - DSN=SYS1.PARMLIB(J2USECF
),
IEF196I DISP=SHR
IEF196I          7 XXPROC00    DD DSN=SYS1.DIST.PROCLIB,DISP=SHR
IEF196I          8 XX          DD DSN=SYS1.PROCLIB,DISP=SHR
IEF196I          9 XX          DD DSN=SYS1.IBM.PROCLIB,DISP=SHR
IEF196I         10 XX          DD DSN=SYS1.KYNEFPROCLIB,DISP=SHR
IEF196I         10 IEF642I EXCESSIVE PARAMETER LENGTH IN THE DSNAME
FIELD
IEF677I WARNING MESSAGE(S) FOR JOB JES2      ISSUED
IEF196I          10 IEF686I DDNAME REFERRED TO ON DDNAME KEYWORD IN
PRIOR
IEF196I STEP WAS NOT RESOLVED
IEF452I JES2      - JOB NOT RUN - JCL ERROR
IEE122I START COMMAND JCL ERROR

```

Oops.....

JES2 Dynamic Proclib

So how would we do this using Dynamic Proclib?

- This is what we had in the JES2 Proc:

```
//JES2      PROC M=J2USECF
//IEFPROC  EXEC PGM=HASJES20,TIME=1440,DPRTY=(15,14)
//HASPLIST DD DDNAME=IEFRDER
//HASPPARM DD DSN=SYS1.PARMLIB(&M),DISP=SHR
//PROC00   DD DSN=SYS1.DIST.PROCLIB,DISP=SHR
//         DD DSN=SYS1.PROCLIB,DISP=SHR
//         DD DSN=SYS1.IBM.PROCLIB,DISP=SHR
//         DD DSN=SYS1.KYNEF.PROCLIB,DISP=SHR
```

- This is how we do the same thing in the JES2 Parm member

```
PROCLIB(PROC00) DD(1)=(DSN=SYS1.DIST.PROCLIB),
                DD(2)=(DSN=SYS1.PROCLIB),
                DD(3)=(DSN=SYS1.IBM.PROCLIB),
                DD(4)=(DSN=SYS1.KYNEF.PROCLIB)
```


JES2 Dynamic Proclib

What happens if we mess up the JES2 parm?

```
PROCLIB(PROC00) DD(1)=(DSN=SYS1.DIST.PROCLIB),
                DD(2)=(DSN=SYS1.PROCLIB),
                DD(3)=(DSN=SYS1.IBM.PROCLIB),
                DD(4)=(DSN=SYS1.KYNEFPROCLIB)
```

Automatic replies

```
$HASP466 PARMLIB      STMT      11 DD(4)=(DSN=SYS1.KYNEFPROCLIB)
$HASP003 RC=(03),DD(4) - INVALID PARAMETER STATEMENT
REPLY 13,END
013 $HASP469 REPLY PARAMETER STATEMENT, CANCEL, OR END
IEE600I REPLY TO 013 IS;END
IEF196I IEF285I      SYS1.PARMLIB                      KEPT
IEF196I IEF285I      VOL SER NOS= @$#M1.
$HASP451 ERROR ON JES2 PARAMETER LIBRARY
REPLY 14,Y
014 $HASP441 REPLY 'Y' TO CONTINUE INITIALIZATION OR 'N' TO TERMINATE
IEE600I REPLY TO 014 IS;Y
IEF196I IEF237I D056 ALLOCATED TO SYS00007
$HASP478 INITIAL CHECKPOINT READ IS FROM CKPT1 779
          (STRNAME JES2CKPT_1)
          LAST WRITTEN MONDAY,  6 AUG 2012 AT 21:41:21 (GMT)
$HASP493 JES2 MEMBER-@$2 HOT START IS IN PROGRESS - z11 MODE
```

JES2 Dynamic Proclib

How do we add a new PROCxx concatenation?

```
$ADD PROCLIB(PROC02),DD1=DSN=SYS1.KYNEF.PROCLIB
$HASP319 PROCLIB(PROC02) DD(1)=(DSNAME=SYS1.KYNEF.PROCLIB)
RO #@$2,$D PROCLIB(PROC02)
$D PROCLIB(PROC02)
$HASP319 PROCLIB(PROC02) DD(1)=(DSNAME=SYS1.KYNEF.PROCLIB)
```

JES2 Dynamic Proclib

For more information, refer to:

- JES2 Commands
- JES2 Initialization and Tuning Reference
- z/OS 1.2 Implementation, SG24-6235



International Technical Support Organization and Authoring Services

Sysplex enhancements in WebSphere MQ 7.1

IBM Redbooks

Topics

Intro to MQ and MQ Shared Queues

How MQ reacted to structure-related failures before 7.1

Structure-related enhancements in MQ 7.1

Use of SMDS for large MQ messages

Intro to MQ

MQ provides application developers with a simple way to move messages between programs:

- The programs might be running on the same z/OS system
- Or, one program might be running on z/OS in New York and the other could be on a smartphone in Singapore - in either case, it should be transparent to the program

The application programmer decides whether the message should be persistent (meaning that MQ does logging so it can recover from any failures and guarantee that the message is delivered) or non-persistent:

- Non-persistent reduces overhead because there is no logging
- But the application then has to take responsibility for the situation where the message is lost before it is delivered.

Intro to MQ

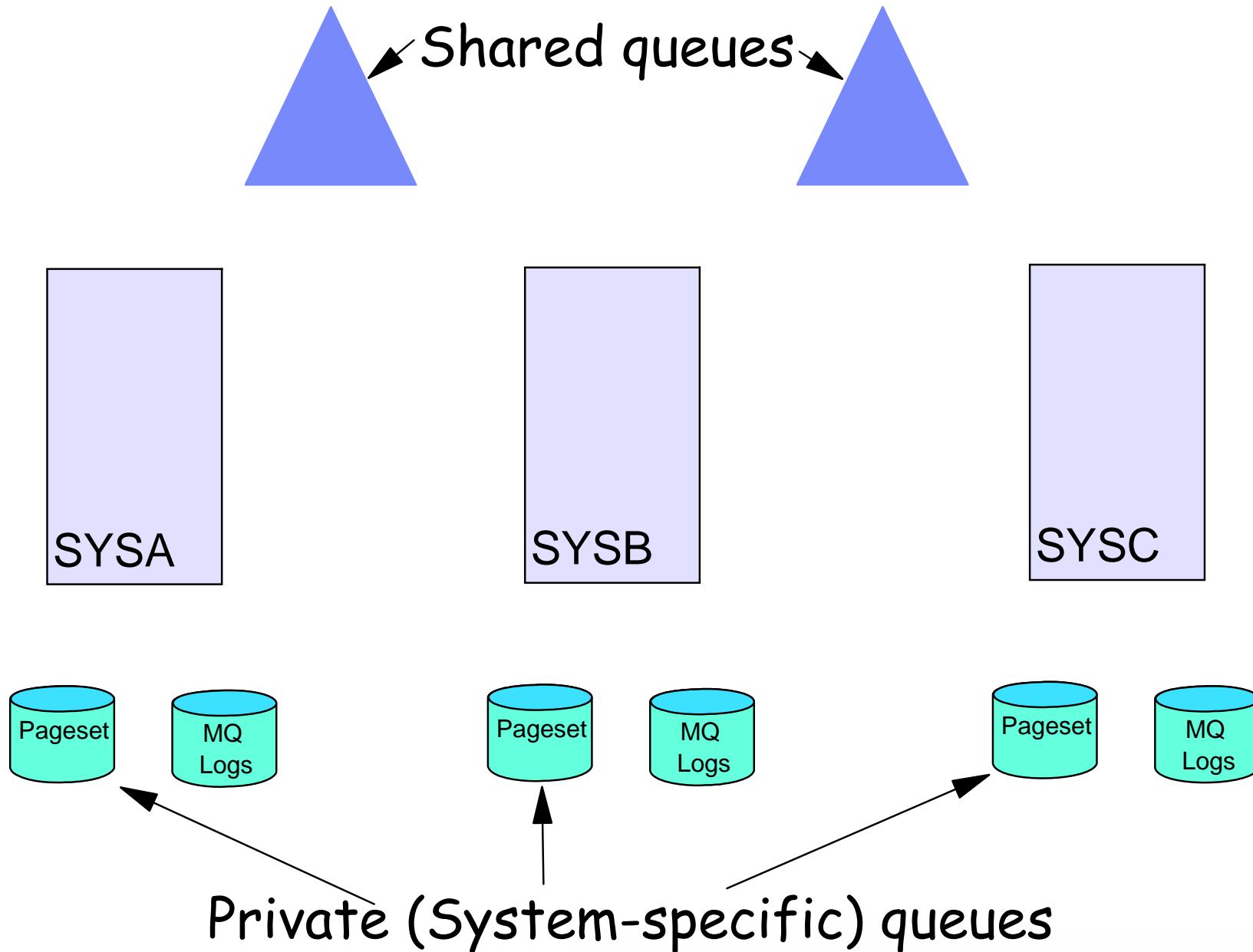
The connection between MQs is most likely to be via SNA or TCP/IP. Depending on your configuration and the applications, messages might be sent directly from client-side MQ queue to z/OS MQ, or they might go through several hops - in every case, message will reside on a queue waiting to be processed or forwarded.

When the message reaches the target MQ, it is stored in a "queue" until it is retrieved by the target application.

After the application processes the message, it might put the response back on a queue, and from there back to the client.

z/OS supports two types of queues - private and shared...

Intro to MQ queues



Intro to MQ queues

Private queues:

- Can contain both persistent and non-persistent messages
- Are made up of buffer pools and data sets (pagesets)
 - Messages are logged (if persistent) then written to a buffer pool
 - Messages may also be written to pagesets
 - However non-persistent messages are discarded if MQ is restarted
 - Either after an abend OR a planned stop

Intro to MQ queues

Shared queues:

- Queues that reside in a CF structure that is accessible to all members of the MQ Queue Sharing Group:
 - Depending on how the structure is defined to MQ, it can contain both persistent and non-persistent messages.
 - If defined as non-recoverable, it can only contain non-persistent messages
 - Unlike private queues, non-persistent messages are NOT deleted from shared queues if one or even all queue managers are restarted

Intro to MQ queues

What are the advantages of shared queues?

- Workload balancing - queues are accessible from all members of the QSG.
- Improved availability - if one queue manager is unavailable, messages are still available via the other members of the QSG.
- Improved message availability - if a queue manager is restarted, non-persistent messages in private queues are discarded, but non-persistent messages in a shared queue are retained.
- Improved availability - if a queue manager fails, another member of the QSG can perform peer recovery for messages in a shared queue.
- Scalability - easy to add another queue manager to process the same set of message queues.
- Lower cost to route messages between queue managers in the same sysplex

Intro to MQ queues

MQ Shared Queue structure types

There are actually two types of MQ structure if queue sharing is being used:

- Administration structure
 - This structure holds information required for unit of work recovery and for coordinating MQ internal activity across the QSG
 - There is one of these per QSG.
- Application structure
 - Can have up to 63 of these per QSG.
 - Each application structure can contain up to 512 queues.

MQ handles failures differently depending on what type of structure failed (and whether the structure was duplexed or not).

Intro to MQ queues

MQ does not support user-managed rebuild.

Prior to MQ 7.1, the recommended way of protecting a structure from CF failure was to System-Manage Duplex it.

- This is effective, however there is a performance and capacity impact.
- Especially unattractive option if you have to duplex across two sites

Intro to MQ queues

What are the recovery considerations when using shared queues?
What types of failure can you encounter?

- MQ Queue Manager address space
- The connection from ONE system to the MQ admin structure fails (partial connectivity failure)
- The CF containing the MQ admin structure fails (total connectivity failure)
- The connection from ONE system to the MQ application structure fails (partial connectivity failure)
- The CF containing the MQ application structure fails (total connectivity failure)

MQ resilience

Queue manager failure

	Private queue	Shared queue
Persistent messages	Unavailable while queue manager is down. Available when queue manager restarts.	Continue to be available while queue manager is down.
Non-persistent messages	Unavailable while queue manager is down. Messages are purged when queue manager restarts.	Continue to be available while queue manager is down. Even if ALL queue managers are stopped, messages will still be there when they are restarted.

Very important to remember that a queue manager could be using BOTH private and shared queues

MQ resilience

Admin queue structure partial connectivity failure

- Prior to MQ 7.1, if a queue manager loses access to the admin structure, that queue manager would abend.
 - Refer back to the impact of a queue manager abend
 - And remember that the queue manager might have private queues, so not only will the queue manager stop working, all non-persistent messages in the private queues for that queue manager will be lost.

MQ resilience



Admin queue structure total connectivity failure

- Prior to V7.0.1, if the Admin Structure was lost for some reason (DR situation, loss of power to the CF etc), then each queue manager had to recreate its own Admin Structure entries. As the admin structure needs to be complete before application structure recovery to take place, it was necessary in a DR situation to start up all the queue managers in a QSG before application structure recover could take place.
- In V7.0.1, a single queue manager is able to recover the admin structure entries for all the other queue managers in the QSG. If a V7.0.1 (or higher) queue manager notices that the admin structure entries are missing for another queue manager then it will attempt to recover them on behalf of the other queue manager. In a DR situation this means that it is only necessary to start a single queue manager at V7.0.1 (or higher) before being able to recover the application structures.

MQ resilience



Admin queue structure connectivity failure

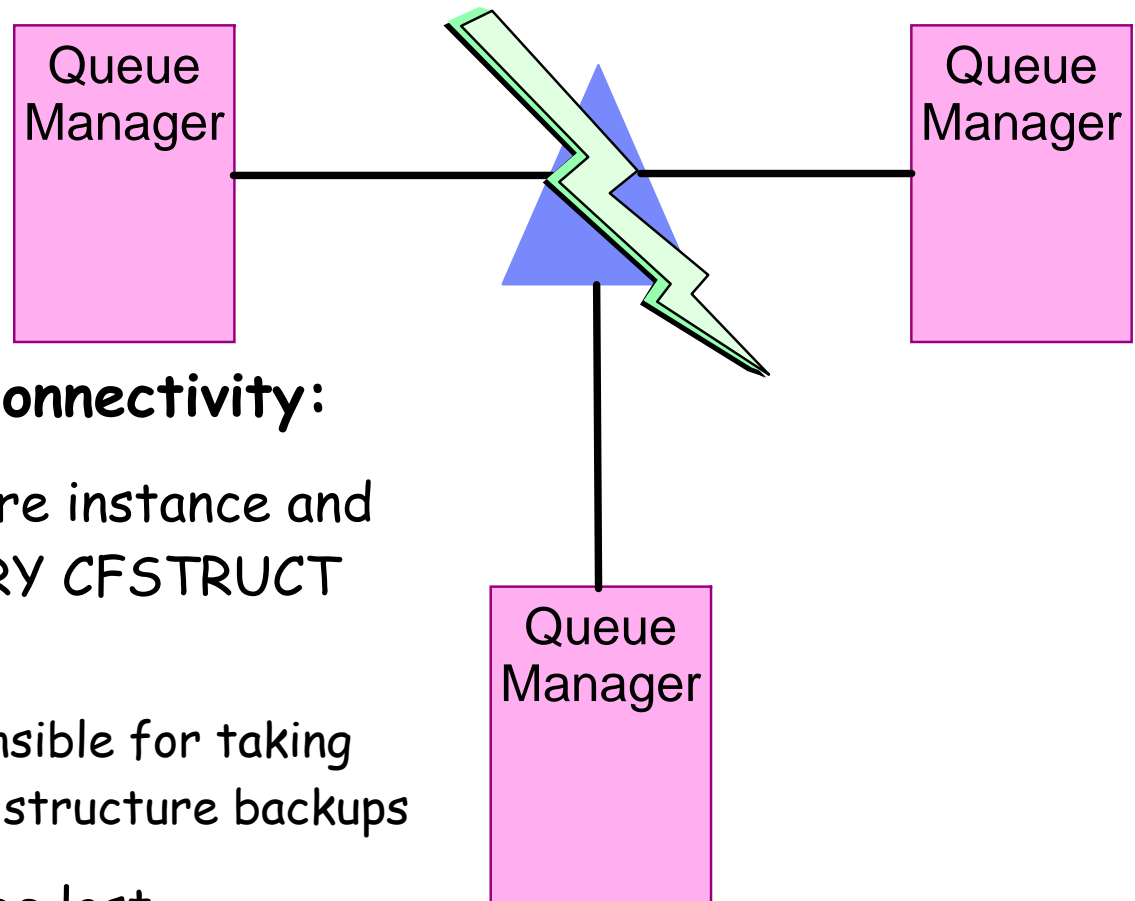
- Queue managers will tolerate loss of connectivity to the admin structure without terminating if: the QMGR CFCONLOS attribute is set to **TOLERATE** and all the queue managers in the QSG are at V7.1
- All queue managers in the QSG will disconnect from the admin structure, then attempt to reconnect and rebuild their own admin structure data.
- If a queue manager cannot reconnect to the admin structure, for example because there is no CF available with better connectivity, some shared queue operations will remain unavailable until the queue manager can successfully reconnect to the admin structure and rebuild its admin structure data.
- The queue manager will automatically reconnect to the admin structure when a suitable CF becomes available to that system.
- Failure to connect to the admin structure during queue manager startup is not tolerated, regardless of the value of CFCONLOS.

MQ resilience

Application queue structure failure

- Prior to MQ 7.1, any queue manager that loses access to an application structure would abend.
- Starting with MQ 7.1:
- All queue managers that lose connectivity to an application structure will disconnect from the structure.
- Queue managers will tolerate loss of connectivity to application structures if:
 - They are defined in MQ as CFLEVEL(5) and
 - The CFCONLOS attribute is set to TOLERATE
- The next action depends on whether it is a partial or total loss of connectivity (according to MQ's definition)...

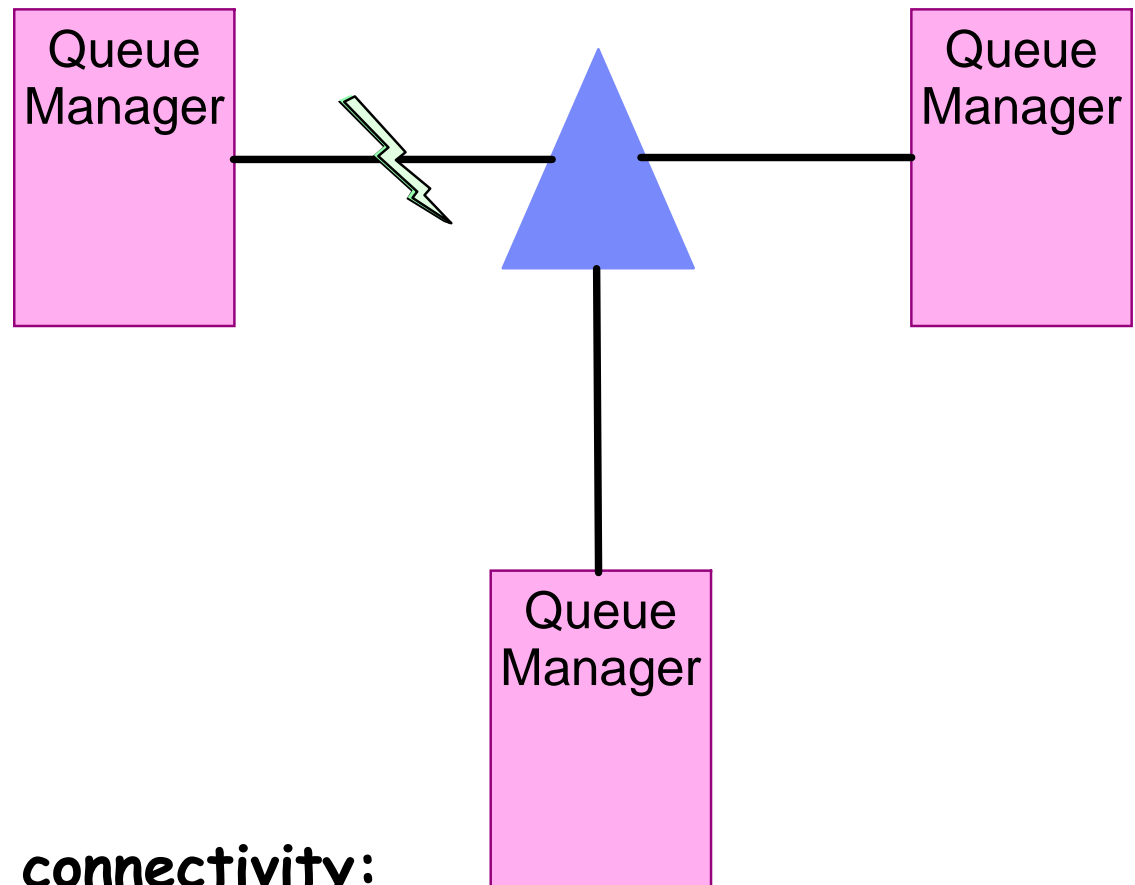
MQ resilience



In the case of total loss of connectivity:

- MQ will allocate a new structure instance and *automatically* issue a RECOVERY CFSTRUCT command for that structure.
 - This means that *YOU* are responsible for taking frequent (every 30-60 minutes) structure backups
- Non-persistent messages will be lost.

MQ resilience



In the case of partial loss of connectivity:

- A System-Managed rebuild will be automatically initiated by the QMGRs that lost connectivity to rebuild the structures into a more available CF. This means that both persistent and non-persistent messages will be retained.

MQ resilience

- QMGR CFCONLOS(TERMINATE|TOLERATE)
 - Specifies whether loss of connectivity to the admin structure should be tolerated
 - Default is TERMINATE
 - Can only be altered to TOLERATE when all QSG members are at 7.1
- CFSTRUCT CFCONLOS(TERMINATE|TOLERATE|ASQMGR)
 - Specifies whether loss of connectivity to application structures should be tolerated
 - Only available at CFLEVEL(5)
 - Default is ASQMGR for new CFLEVEL(5) structures, and TERMINATE for structures altered to CFLEVEL(5)
- CFSTRUCT RECAUTO(YES|NO)
 - Specifies whether application structures should be automatically recovered
 - Only available for CFLEVEL(5) structures
 - Default is YES for new CFLEVEL(5) structure, and NO for structures altered to CFLEVEL(5)

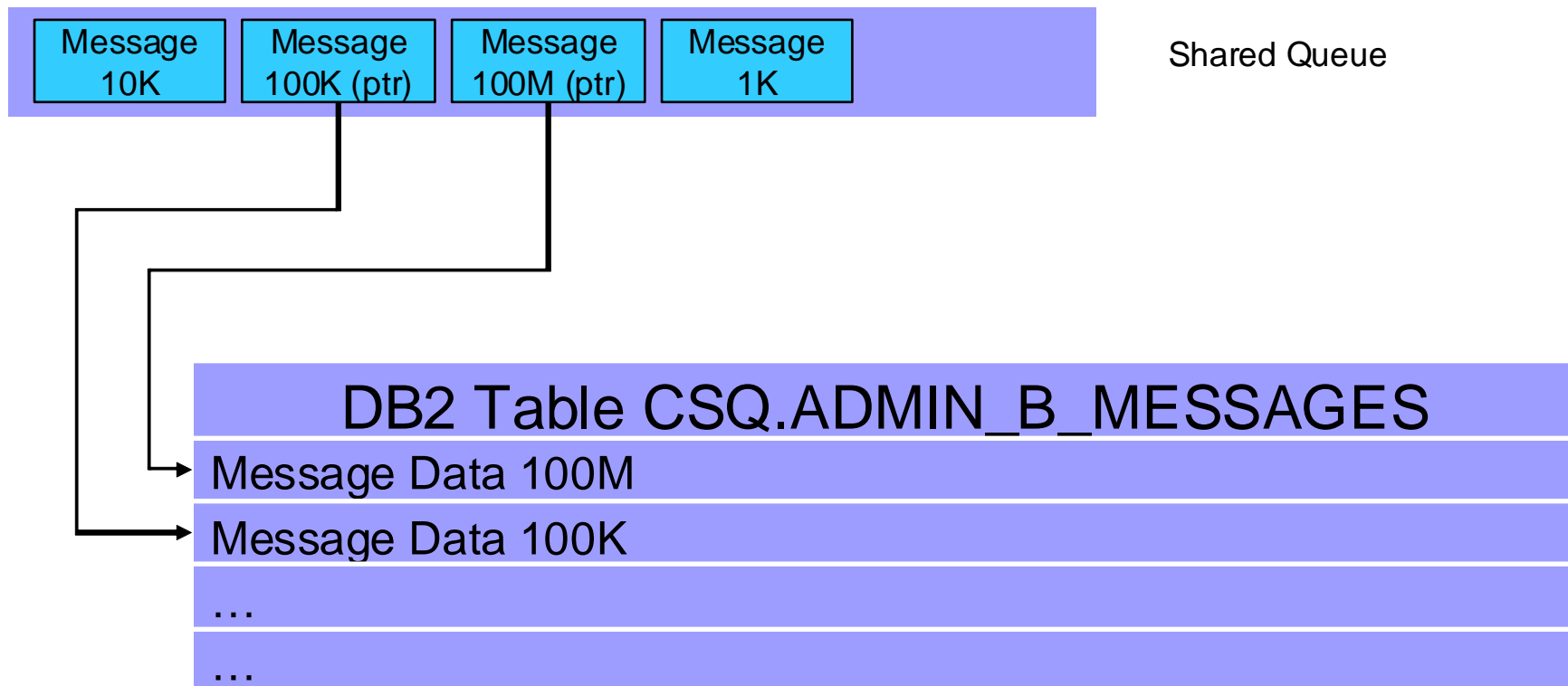
MQ resilience

So how do these changes relate to use of SM Duplexing for MQ structures?

	Exploiting 7.1 capabilities	SM Duplexing
Protect persistent messages in case of CF failure?	Yes	Yes
Protect non-persistent messages in case of CF failure?	No	Yes
Performance	No impact	Higher response times and more CF and z/OS CPU utilization
Recovery time	Varies with number of objects in structure and elapsed time since last BACKUP	Minimal
Queue manager survives loss of structure?	Yes	Yes

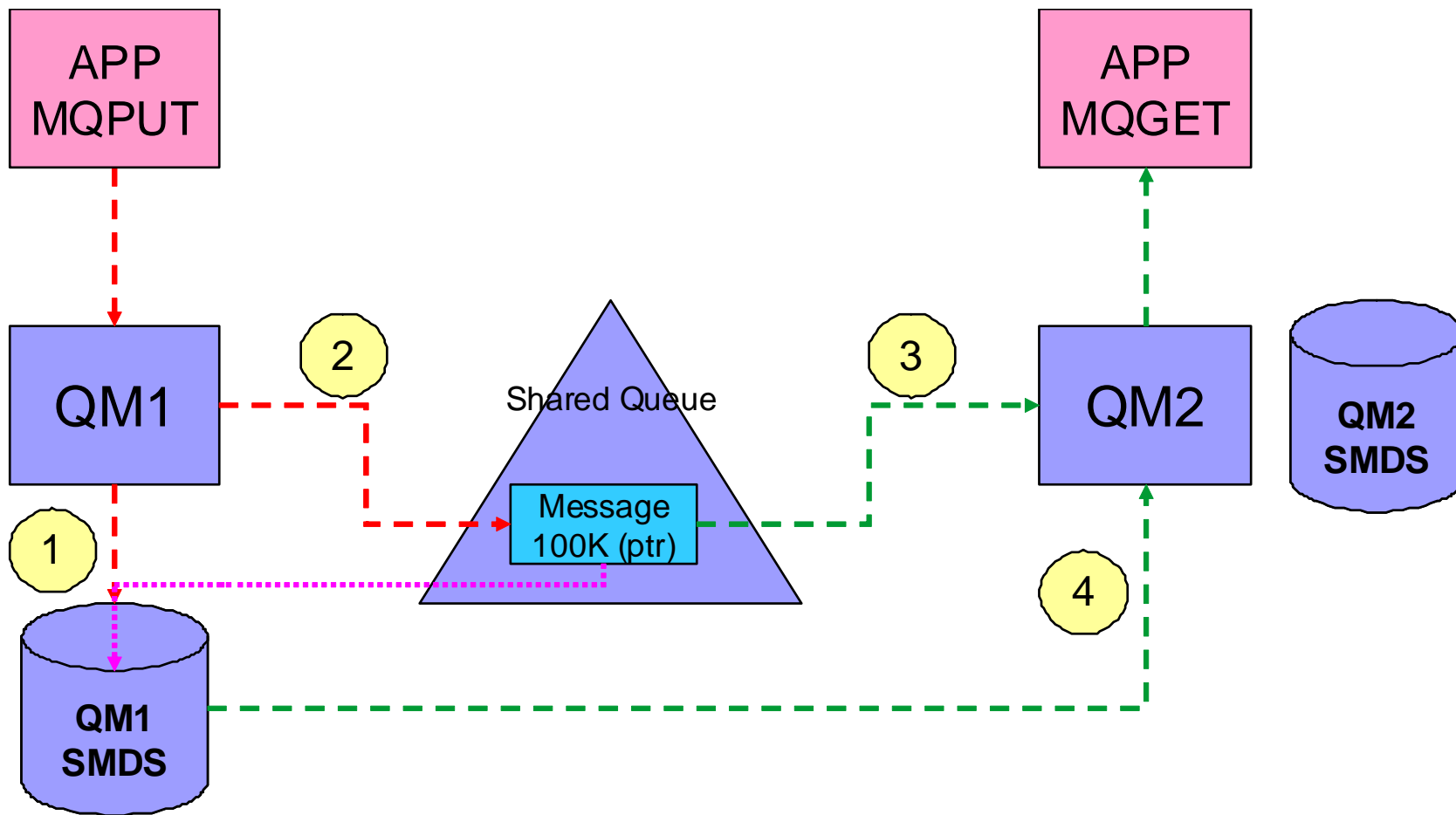
MQ 7.1 SMDS

If a message larger than 63K is placed on a shared queue, only part of the message is actually kept in the CF. Most of the message is kept in a DB2 large object table.



MQ 7.1 SMDS

MQ 7.1 adds the option of placing the messages in a VSAM linear Shared Message Data Set (SMDS) instead of a shared DB2 table



MQ 7.1 SMDS

Offloaded message data for shared messages is stored in data sets.

Each application structure has an associated group of shared message data sets, with one data set per queue manager.

- DSN specified on DSGROUP parameter on CFSTRUCT definition.

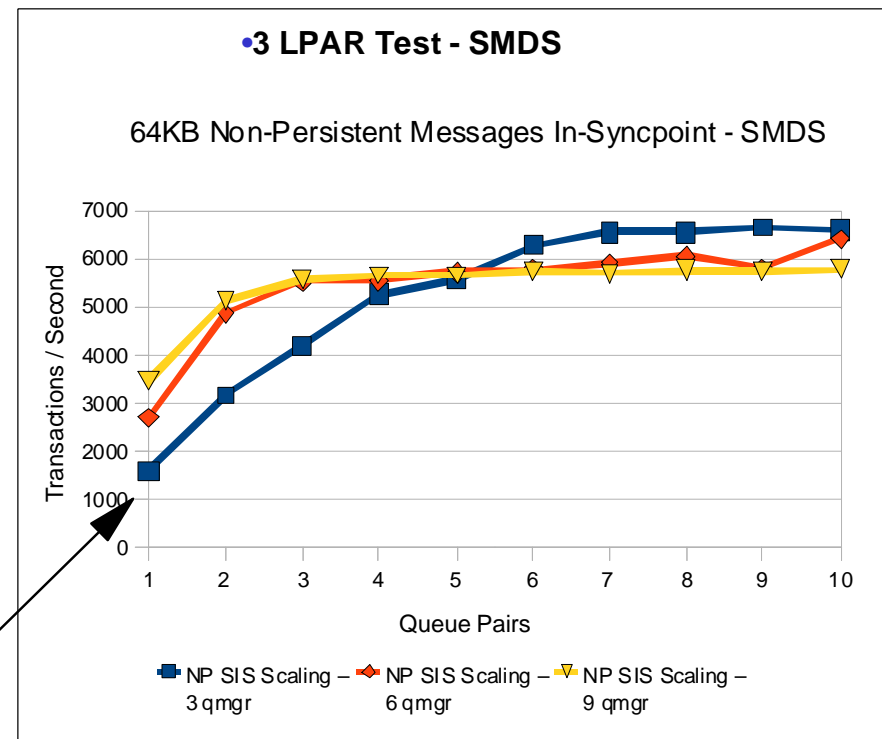
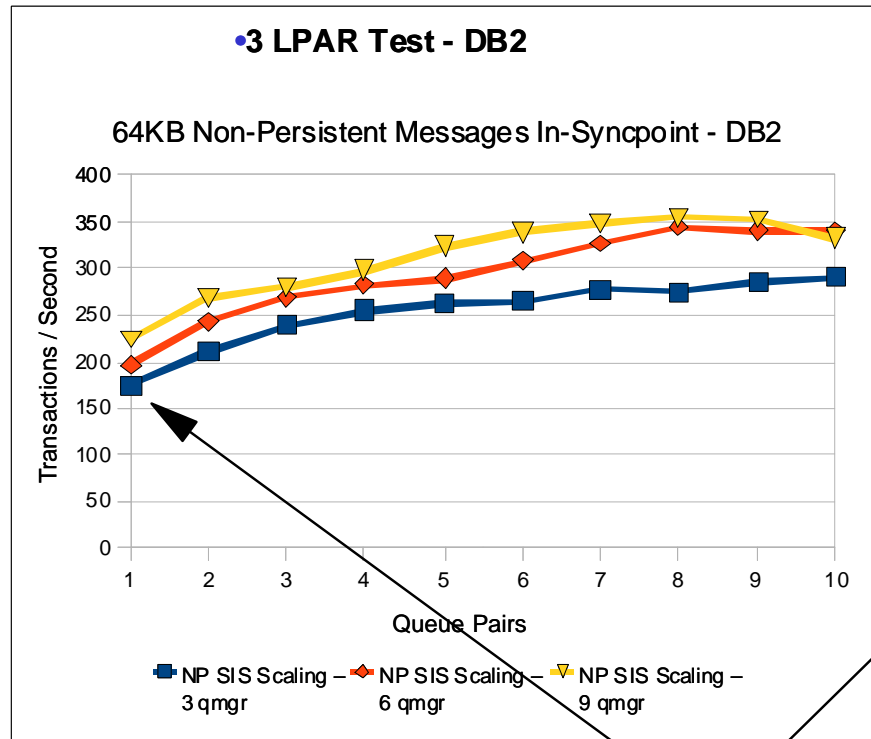
Each queue manager owns a data set for each structure, opened for read/write access, which it uses to write new large messages.

Each queue manager opens the data sets for the other queue managers for read-only access, so it can read their message data.

When a message with offloaded data needs to be deleted, it is passed back to the queue manager which originally wrote it, so that the queue manager can free the data set space when it deletes the message.

MQ 7.1 SMDS

What is the point in this enhancement? Performance....



Look at the scale on the y-axis

- Tests show comparable CPU savings making SMDS a more usable feature for managing your CF storage
- SMDS per CF structure provides better scaling than DB2 BLOB storage

MQ 7.1 SMDS

Messages too large for CF entry (> 63K bytes) are always offloaded.

Other messages may be selectively offloaded using offload rules:

- Each structure has three offload rules, specified on the CFSTRUCT definition.
- Each rule specifies message size in Kbytes and structure usage threshold, using two parameters:
 - OFFLDnSZ(size) and OFFLDnTH(percentage), where n = 1, 2, 3.
- Data for new messages exceeding the specified size is offloaded (as for a large message) when structure usage exceeds the specified threshold.
- Default rules are provided which should be acceptable in most cases.
- Rules can be set to dummy values if not required.

Without offloading data, it is possible to store 1.25M messages of 63KB on a 100GB structure. When offloading all messages, possible to store approx 140M messages on the same structure, irrespective of message size

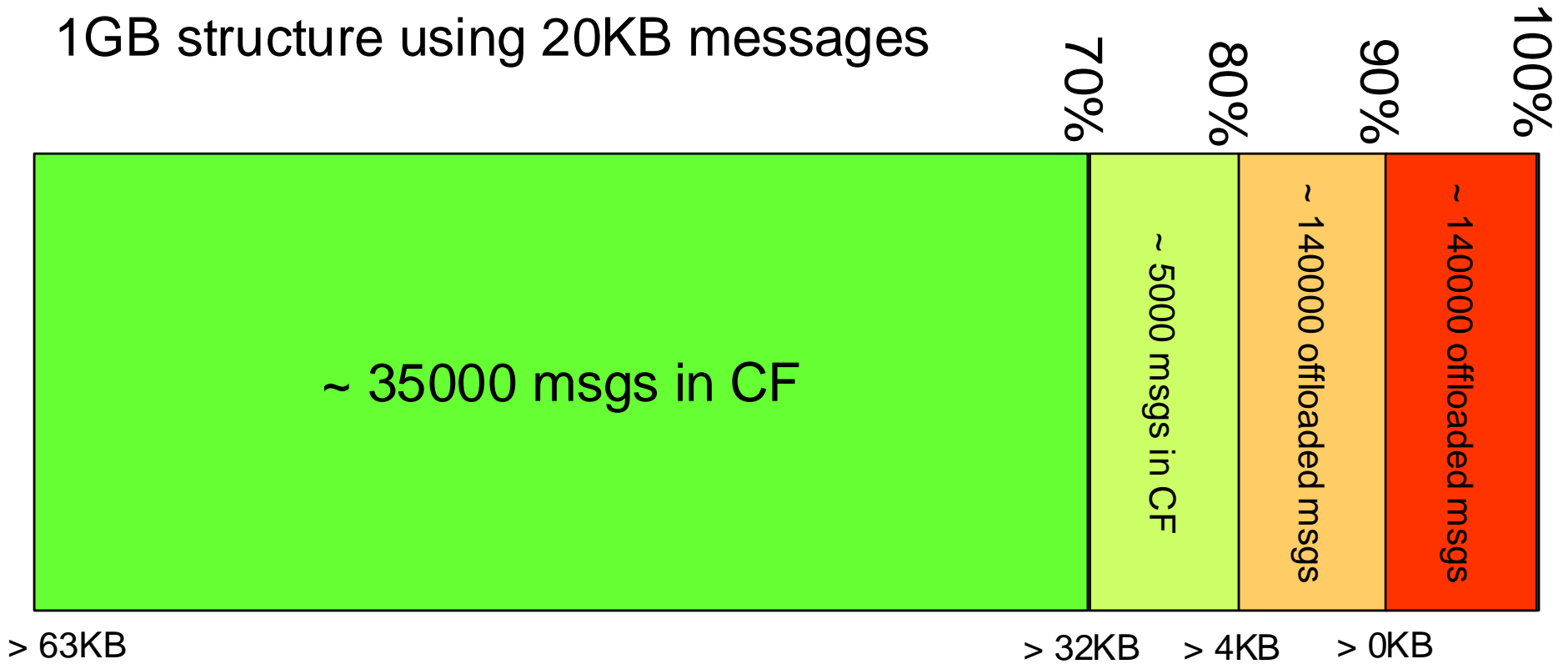
MQ 7.1 SMDS

The three offload rules have no fixed order but are typically intended to be used as follows:

- Rule 1 is used to save space for fairly large messages by offloading them, with little performance impact, even when plenty of space left.
 - SMDS defaults: OFFLD1SZ(32K), OFFLD1TH(70)
- Rule 2 is used as an intermediate step between rules 1 and 3, to start saving more space as the structure usage increases, in exchange for a minor performance impact.
 - SMDS defaults: OFFLD2SZ(4K), OFFLD2TH(80)
- Rule 3 is used to maximize the remaining space when the structure is nearly full, by offloading everything possible.
 - SMDS defaults: OFFLD3SZ(0K), OFFLD3TH(90)

MQ 7.1 SMDS

1GB structure using 20KB messages



~ 320000 msgs using offloading vs ~ 50000 without offloading

MQ 7.1 SMDS

SMDS is defined as a VSAM linear data set using DEFINE CLUSTER:

- Requires LINEAR option.
- Control interval size must be 4096, which is the default for linear.
- Requires SHAREOPTIONS(2 3), allowing one queue manager to write and other queue managers to read at the same time.
- If maximum size may need to exceed 4GB, requires SMS data class which has VSAM extended addressability attribute.
- If automatic expansion is desired, requires an appropriate secondary space allocation (although a default of 20% will be used if an expansion attempt fails because of no secondary allocation).

Can optionally be pre-formatted using CSQJUFMT.

- Otherwise formatted automatically when first opened.

MQ 7.1 SMDS

The **DSGROUP** parameter on the **CFSTRUCT** definition specifies the group of data sets associated with the application structure.

- It is specified as a generic data set name with a single asterisk as the point where the owning queue manager name is to be inserted.
- It is required when the option **OFFLOAD(SMDS)** is specified.

MQ 7.1 SMDS summary

Attractive alternative to placing large messages in DB2 BLOBs

Have additional benefit that they can also be used as an overflow area if MQ structure starts filling up

- Performance is not as good retrieving message from CF, but is a more attractive alternative to running out of space in the structure in the CF.

All queue managers in the QSG must be running MQ 7.1 or later.

- Associated CF structure must be defined in MQ as CFLEVEL(5)



International Technical Support Organization and Authoring Services

IBM Poughkeepsie Testing

IBM Redbooks

Topics

IBM z/OS Test Processes and Phases

- z/OS Unit Test
- z/OS Function/Component Test
- z/OS COMBAT
- z/OS System Test
- z/OS Performance Test
- z/OS Platform Evaluation Test
- z/OS Beta/ESP Test
- z/OS Service Test
- z/OS Consolidated Service Test

Engineering the Test - IBM Best Practices & Recommendations

IBM Poughkeepsie testing

Quality is engineered from design and development time

Multiple large scale tests phases with differing focus, scope and objectives

Continuous enhancements and test methodology improvements

Investments in cross platform efforts, and general solution tests

Increased teaming among testers across the corporation

Culture of test engineering with cross organization quality discussion groups and regular strategy meetings



IBM Poughkeepsie testing

z/OS Development and Function Test with over 500 employees

- Focused on development, unit and function test

STG System Assurance (SVT included) with over 1,100 employees

- Focused on testing and test related activities across the STG Division of IBM

STG System z software test floor:

- Variety of support families/generations
- Over 3.5 PB of data supporting test efforts
- Approximately housed on ~25,000 ft² / 2,323 m²

Overall STG System z Total test floor:

- 5 PB of data on ~67,000 ft² / 6220 m²

IBM Poughkeepsie testing

IBM Poughkeepsie site

Aerial Photography by *Rosspilot*

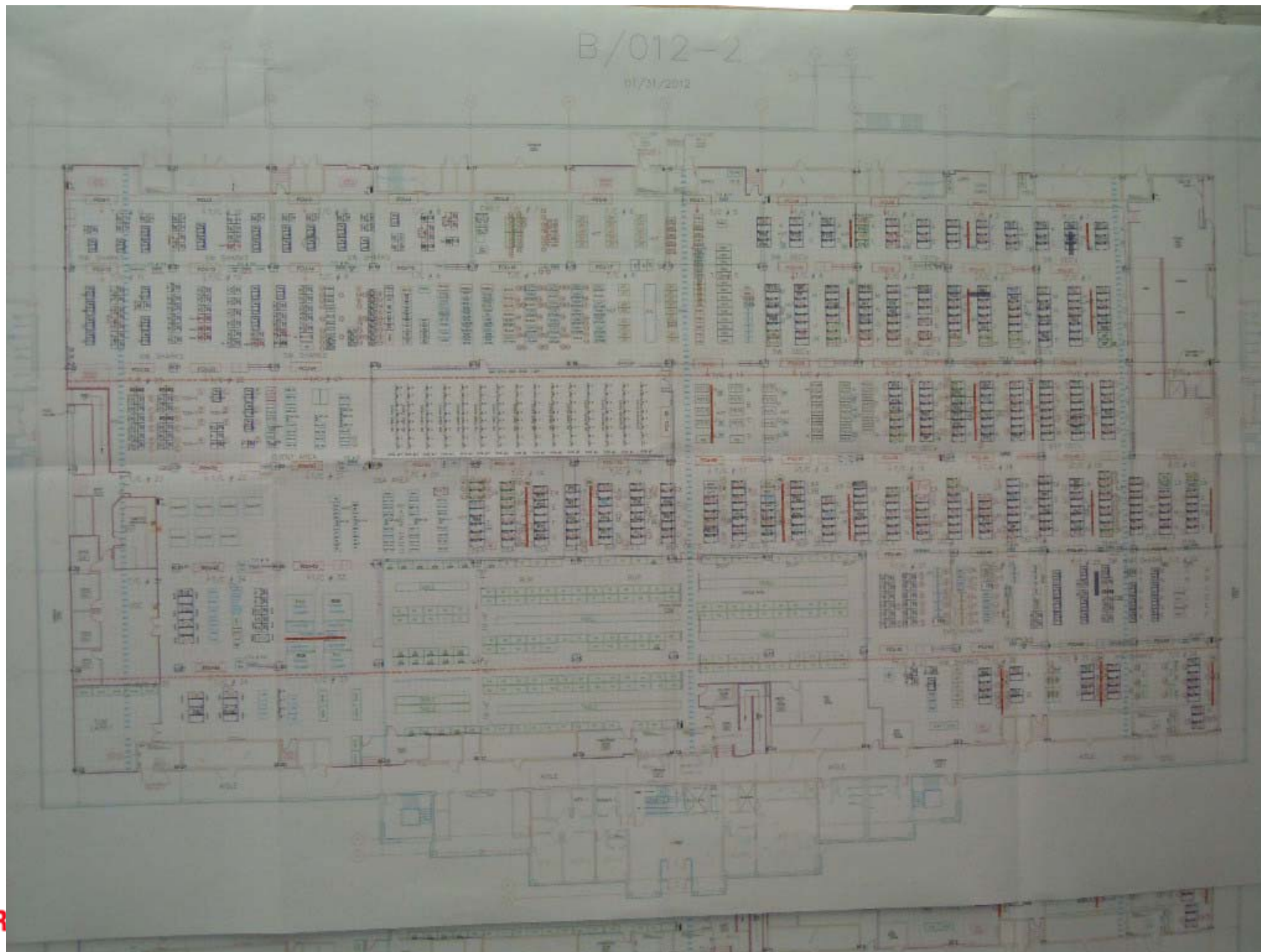
April 30, 2008



IBM Headquarters
Poughkeepsie, NY

IBM Poughkeepsie System z test floor

System z test floor layout



IBM Poughkeepsie System z test floor

Some statistics about the System z test floor layout:

- Number of CPUs: About 200 (z900 up to EC12)
- DASD Storage: 5 PBs (200 boxes)
- Switches: 122
- Number of ESCON/FICON connections: 56,000
- Time to replace a CPU: 8 hours
- Number of connections in cabling room: 100,000

IBM Poughkeepsie System z test floor

Looking down along the length of the System z test floor



IBM Poughkeepsie System z test floor

"Shake and bake" cells



IBM Poughkeepsie System z test floor

The back of one of the switches



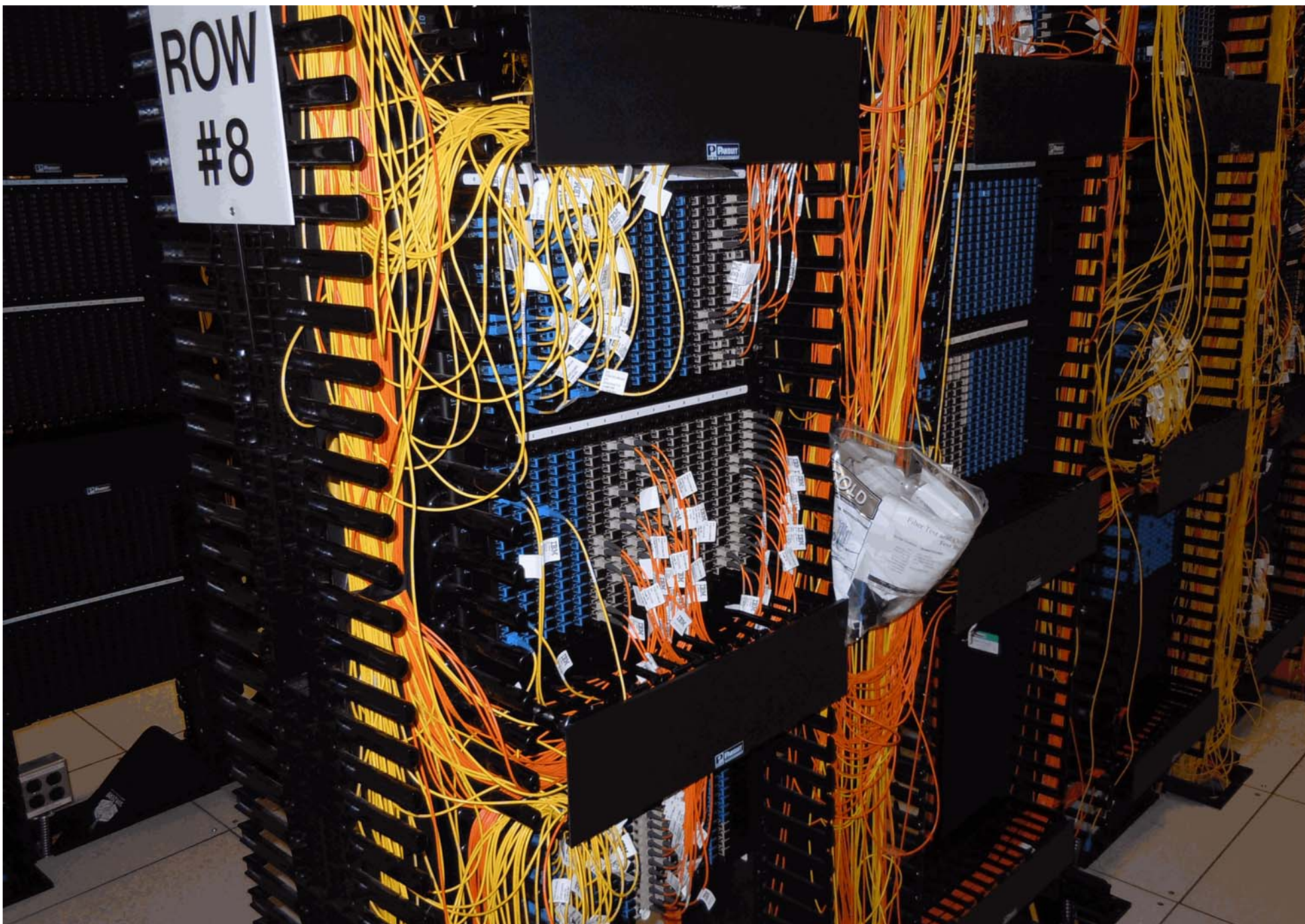
IBM Poughkeepsie System z test floor



IBM Poughkeepsie System z test floor



IBM Poughkeepsie System z test floor



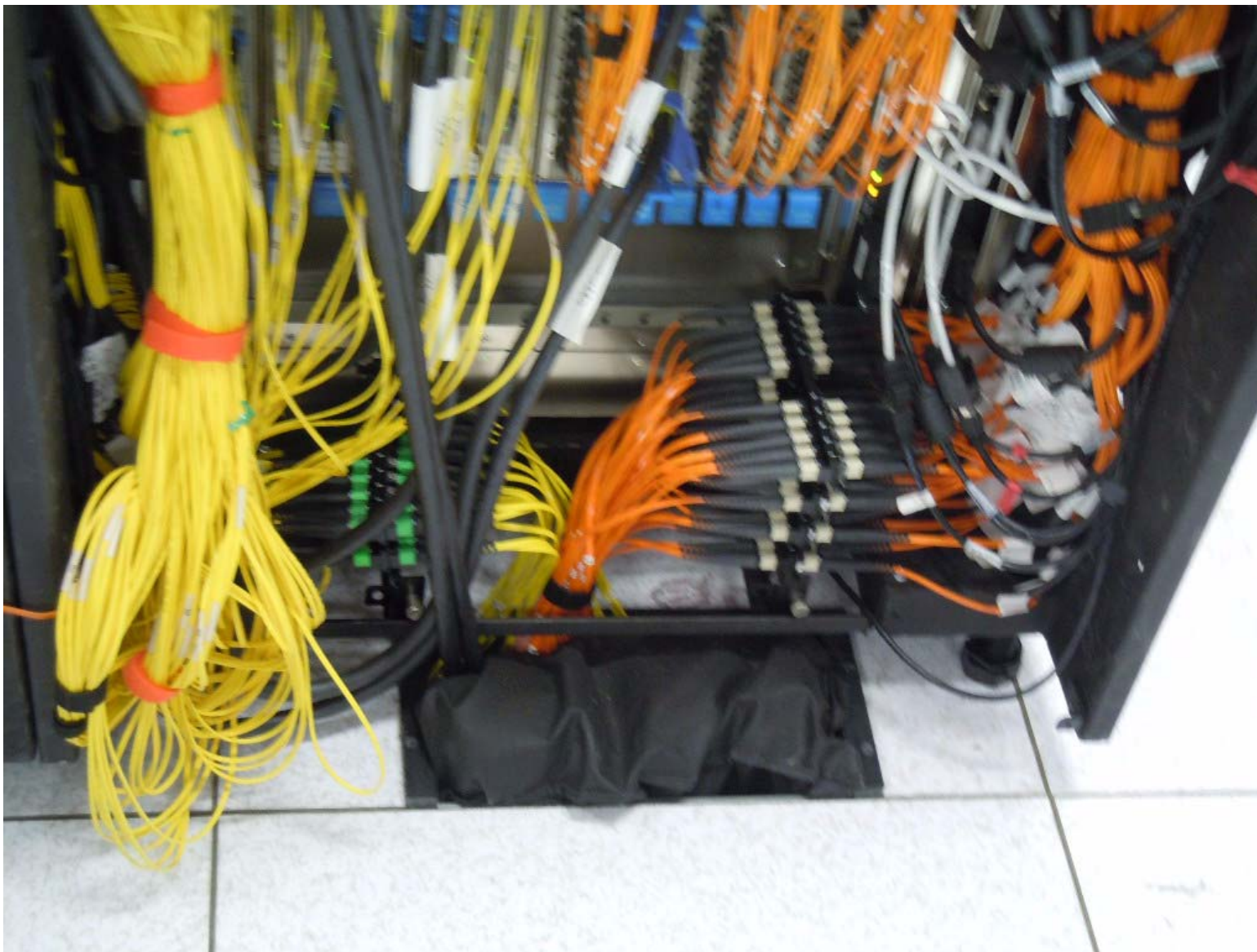
IBM Poughkeepsie System z test floor

"Fiber suitcase" for extending distance on a link



IBM Poughkeepsie System z test floor

Quick connect feature on CPU



IBM Poughkeepsie System z test floor

CPU_s



IBM Poughkeepsie System z test floor

And more CPUs



IBM Poughkeepsie System z test floor

And yet more CPUs



IBM Poughkeepsie System z test floor

IBM Netezza



z/OS Unit test

Initial verification that all new and changed code within a module or macro is error free

- Execute
 - Every line
 - Every branch (both ways)
- Verify error recovery procedures

Performed by the developer

Test environment is typically developer workstation

z/OS Function/Component test

Test modules that comprise a component or function

- External interfaces (e.g., panels, commands, messages)
- Inter-component interfaces
- Intra-component interfaces
- HW/SW interfaces
- Application program interfaces
- Non-message event recording
- RAS characteristics and error diagnosis
- Shared paths (multitasking) and shared resources (locks, files, etc.)
- Function completeness

Test environment is 2nd level (VM guest)

z/OS Community Build And Test (COMBAT)

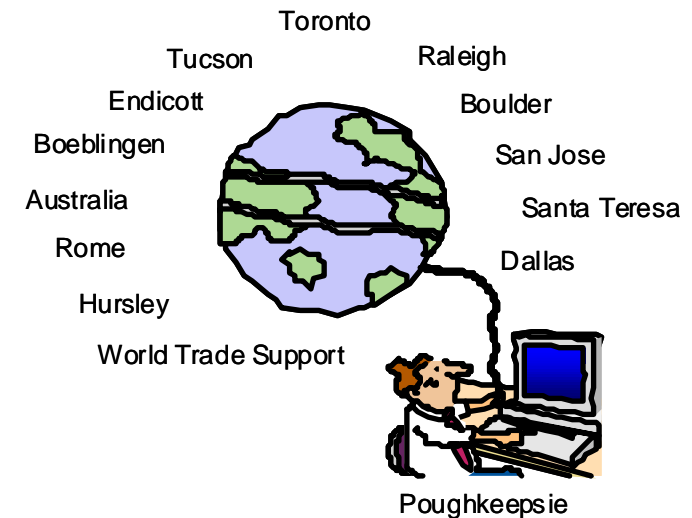
COMBAT is the first point in the z/OS cycle where all z/OS elements are gathered directly from their development organizations, integrated together, and then used as a platform by each of the elements, System Test, Integration Test and Performance Test

COMBAT promotes cross-element testing discussions with the objective to remove cross-element defects and ensure proper integration

All parts in z/OS, test on full z/OS system

Objective is to:

- Manage and test functional dependencies
- Test the elements together
- Test as early as possible
- Manage multi-lab issues



z/OS System Verification Test

Test Goals:

- All software components combined and tested as single unit
- Focus on high levels of users, transactions, and heavy load and stress
- Verify software can withstand memory shortages, CPU and I/O saturation, and that error recovery under those conditions is successful
- Explicit release migration, coexistence, and fall-back testing is verified
- Long running test engagements of 1-3 weeks

Defect Goals:

- Timing and serialization problems
- High stress widens processing windows for data integrity bugs
- Data aging and fragmentation problems over extended runs

z/OS System Verification Test

Environment:

- All testing is performed on native hardware
- z/OS release being developed (N) is primary focus
- Regular mixed testing with N-1, N-2
- Between 7-10 sysplexes and monoplexes are active concurrently
- 4-way sysplex is most common, several 8-way and 16-way sysplexes also available
- Multiple generation of hardware families are active
 - z9, z10 EC, z10 BC, z196, EC12,??
- Subsystems include DB2 v9/v10, MQ v6/V7, CICS/TS v4.1.0, and WAS v7/v8

z/OS System Verification Test

Workloads:

- Legacy and Current Batch Processing
- Online Transaction Processing
- System and Component level Thrashers
- "Organic" customer representative workloads with cycles of improvement and incremental enhancements

z/OS Performance test

Release-to-Release and Processor Performance

Strict Methodology

- Workload characteristics strictly managed
 - Must stand up in a court of law
- Only 1 change at a time
- Repeatability is a must
- Baselines reset when needed

Key Measurements

- Number of transactions per CPU busy
- Number of transactions per elapsed time

z/OS Platform Evaluation Test (PET)

Validate the platform

Implement new parallel solutions

Continuous environment enhancements

Act as z/OS's first customer as the final testing phase before GA (z/OS Integration Test)

- Customer representative workloads added and improved upon regularly

Run 24 x 7 operations

Two Parallel Sysplexes

- 9 system production sysplex
- 4 system test sysplex
- A mix of z9 EC, z10 EC, z10 BC and z196

z/OS Platform Evaluation Test (PET)

Team organized by traditional I/S roles

- Base OS - BCP, JES, SMS, Operations, VTAM, TCP/IP, NFS
- Middleware IMS, DB2, CICS, MQ, DBA, USS, WAS, HFS, zFS
- Security - zPET security portfolio, z/OS Security products (RACF, SSL, PKE, ICSF, etc.)
- Testware - workload development
- Linux - native, z/VM

Document Experiences for customers

- <http://www.ibm.com/systems/services/platformtest/servers/systemz.html>

z/OS ESP and Beta test

Joint Project Development (JPD)

- Focuses on joint development between IBM and a customer
- Normally begins during the 'Plan' phase and continues throughout the development and testing phases.

Beta

- Early evaluation of product characteristics (such as quality, functionality, performance, usability, etc.) in a customer environment
- Normally starts after development has begun, but prior to the completion of internal testing

z/OS ESP and Beta test

Early Support Program (ESP)

- Used to confirm that a product is ready for general availability (GA)
- Focus is on testing, installation, documentation, distribution, and service support
- Testing should be completed before the product is installed in customer accounts
- Ordered and supplied through normal production/distribution processes and is supported by the normal support structure

Quality Partnership Program (QPP)

- Characterized by long-term contract relationships (e.g. 5 years) with customers who participate in quality verification testing, often over several releases of a Product
- Focus is improving quality by testing new releases in complex customer environments which cannot be easily replicated in the laboratory testing environment
- QPP customers normally participate in successive ESPs

z/OS Service Test

Ensure high quality maintenance

Test Pre-COR Closed PTFs for 3 levels of z/OS in a Parallel Sysplex

- 5-day cycle of workload and product focus
 - Customer representative workloads, regularly improved
- Special Tests (++APARs, ++USERMODs)
- Customer problem recreates / Critical Situation verification

z/OS Consolidated Service Test

Environment for Post-GA and PTF testing

Provide a single, consistent, installable maintenance recommendation across the z/OS stack (including IBM SWG products)

- Ensure service for one product doesn't impact other products
- Single IBM voice for recommendations

Test first, recommend second

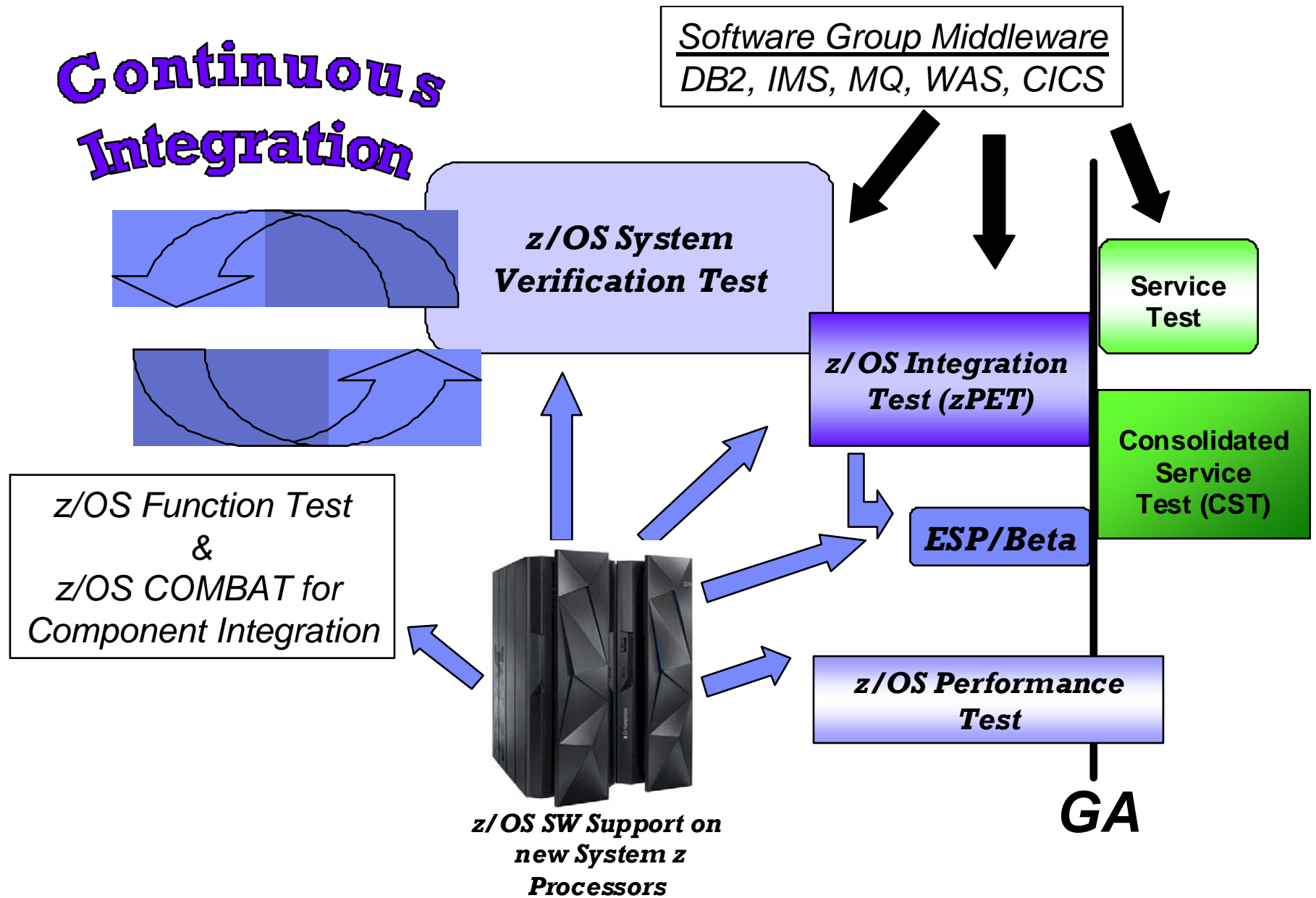
- Multiple releases - N to N-2 of z/OS and N and N-1 of subsystems

Customer representative workloads with continuous development and improvement

Provides complex GDPS environment to protect customers

z/OS CST web site:

- <http://www.ibm.com/systems/z/os/zos/support/servicetest/>



z/OS Test coverage

	<i>Tests PTFs/APARs</i>	<i>Tests New Releases of z/OS</i>	<i>Pre-GA</i>	<i>Post-GA</i>
Unit Test	X	X	X	X
Function / Component Test	X	X	X	X
COMBAT	X	X	X	X
System Test	X (prior release function)	X	X	
Performance Test	X	X	X	X
Platform Evaluation Test	X (limited)	X	X	X
ESP	X (limited)	X	X	X
Service Test	X		X	X
Consolidated Service Test	X			X

Testing - IBM recommendations and best practices

Planning, Planning, Planning!!

Test Environment Management/Utilization

Metrics, Measurements and Quality

Test Advancements

Use Cases of Test and Pre-Production Environments

planning, Planning, PLANNING!

Test Planning and Preparation is the most critical step in the testing process

- A critical dependency of good test planning is a strong requirements and design process w/ associated documentation
 - Requirements come from many places and are tracked via databases
 - Design specifications are reviewed and approved (by test as well) to ensure requirements are properly met and considered
- Detailed understanding of the product and/or application under test.
 - This includes the use cases that the application or product will be expected to process (i.e., how does a transaction work and what data does it manipulate?)
 - In some test phases, this includes the knowledge of how the product or application will perform its processing at the lowest levels, such that if required, you could diagnose problems in the application or product software.
 - This also can include keen understanding for studying performance characteristics and benchmarks

MORE Planning!!

Test Planning and Preparation is the most critical step in the testing process

- Earliest identification of hardware requirements and corresponding software requirements in order to properly be positioned at the test start
- Establish comprehensive and formal test plan for new function and regression testing with appropriate stakeholder representation and approval
- Earliest establishment of comprehensive regression and automation test strategy to ensure dependency stability when there are HW/SW changes
- Develop technical test phase criteria that has the following characteristics:
 - Measurable, Achievable, Meaningful, Discrete, Mutual Agreement
- Plan for strong change control management process because it is critical to ensure that what needs to be tested is actually being tested
- Focus your test planning and execution also on areas of your environment/business that you believe are unique to your business

Test environment management and utilization

Dynamic Test Image Provisioning of LPARs and Resources provides foundation for cross-team utilization

- Main technique used by IBM z/OS Test team for most flexible testing framework
- "One team" approach where test resources are managed and shared across many diverse teams to provide prioritization and necessary capacity

Increased Test Environment complexity as test phases progress

- Starts with partial integration of software elements and progress through large complex integration of hardware and software products and test applications
 - Each phase provides its own business and quality value

Test environment management and utilization

Close resemblance of pre-production test environment to the production environment provides greatest stability

- Changing the size of the system (MIPS, number of CPs, amount of storage, etc) will change the sorts of problems you will find; Need to find the problems that matter in your environment
- Data replication solutions can be used to copy existing, 'real' production environment data into pre-production environment for testing purposes
 - The overall percentage of production data used for testing is important, however, understanding the data and how it is manipulated by the application is more significant
- In many cases, there will be 'size' differences between pre-production (test) and production, therefore you may need to:
 - Scale to the production environment by driving same relative CPU utilization, CF traffic, bottlenecks, workload patterns
- Focus heavy testing on most important/pertinent workload/subsystems - generate more relative work than seen on production systems at peak times of day/year

Metrics, measurement, and quality

Continuous Improvement Analysis

- Root cause escape and outage analysis for internal test phases and external customers; Any problems that escape test to impact your production environments need to be analyzed, and test scenarios created to make sure they don't escape again
- Test phase post-mortem reviews

Defect Management and Analysis

- In-process defect analysis including functional trends; this helps to identify new tests required for this test phase (dynamic test plan)
- Day to day problem management including environmental and non-functional impacts
- Use of Data Analytics to improve our invalid defect rates (i.e., predicting invalid defects at open time) and focus on continuous improvements based on real-time data

Test scenario progress

- Actual test progress versus projected plan

Test advancements

Workload/Operational Profiling

- A proactive approach to test improvement that employs traditional means of understanding characteristics of client production environments, augmented with a process of data mining empirical systems data in order to contrast with IBM Test
- Workload/Operational profiling is an ongoing, iterative and evolving focus on consumable test improvements across IBM
- This technique can also be used as a comparison of internal customer environments. For example, it may be used to see how closely a pre-production environment compares to a production environment with regard to coverage and load/stress of key aspects of the system

Testing advancements

Combinatorial Test Design (CTD)

Measures what combinations of functions or other test attributes will be tested together or have been tested together:

- **Motivations**
 - Too many combinations to easily consider in making a test plan
 - Not enough time to test every possibility
 - Understand the risks of incomplete testing
 - Avoid test omissions
- **Advantages**
 - Systematic planning of tests
 - Reduce the number of required tests
 - Known risk level for reduced test size
 - Omissions obvious during test review

The case for combinatorial test design

Most defects are discovered in tests of the interactions between the values of two variables

Cost vs. risk can be managed by the level of interactions tested (pairs, three-way, etc.)

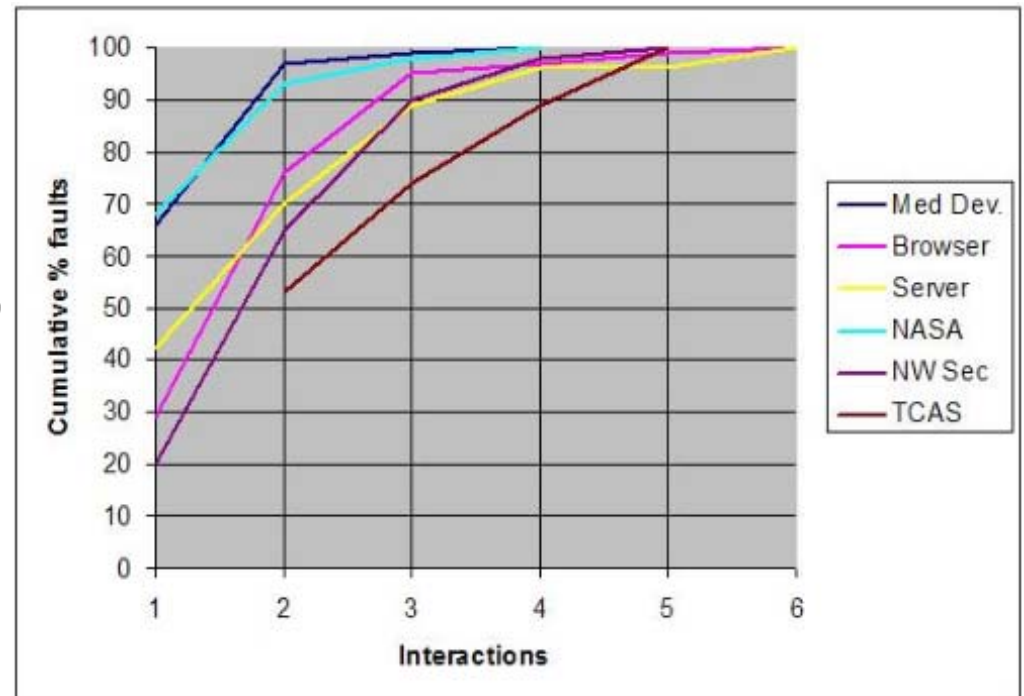


Fig. 1. Number of variables involved in triggering software faults

Source <http://csrc.nist.gov/groups/SNS/acts/ftfi.html>

Sample test environments

Sandbox

- Tests base infrastructure (system, subsystem, ISVs, etc)
- Typically no real workload
- Can also be used as a place to train operators
- Tends to be very small

Application Development and Test

- Focus on verifying application behavior
- May alternate between being used by development and by test

Quality Assurance (QA) System/Environment

- Production-like environment
 - Similar maintenance
 - Similar workload
 - Similar configuration (within reason)

Summary

IBM makes a significant financial investment in the validation of our products

IBM focuses on driving the state of the art testing techniques and practices and continues to evolve them over time

IBM focuses on understanding the client's unique utilization of our products

IBM test strives to improve the overall client experience





International Technical Support Organization and Authoring Services

Miscellaneous

IBM Redbooks

DB2 GBP Structure handling



Important lesson for DB2 GBPs

Scenario - customer empties one CF during batch window to upgrade links. Batch continues to run.. Due to volume of updates, Auto Alter had increased size of one GBP to the SIZE value in CFRM policy. Every 30 seconds, message IXC585E (structure usage threshold exceeded) is issued, stating that XCF was trying to adjust the entry to element ratio.

In the middle of this, upgrade is completed, CF is brought back online, taken out of MAINTMODE, and START,REALLOCATE command is issued.....

Important lesson for DB2 GBPs

Shortly afterwards, message IXL040E (no response to rebuild command) is issued.

```
IXL040E CONNECTOR NAME: DB2_DP12, JOBNAME: DP12DBM1, ASID: 0081 602
HAS NOT RESPONDED TO THE REBUILD CONNECT EVENT, IXLREBLD REQUEST=
COMPLETE EXPECTED.
REBUILD PROCESSING
FOR STRUCTURE DSNBP10_GBP1 CANNOT CONTINUE.
MONITORING FOR RESPONSE STARTED: 09/08/2012 01:44:28.
DIAG: 0000 0000 00000000
```

Then about 20 seconds later, following message is issued:

```
DSNB332I  -DP12 DSNB1PCD THIS MEMBER HAS COMPLETED 612
CASTOUT OWNER WORK FOR GROUP BUFFER POOL GBP1
          PAGES CAST OUT FROM ORIGINAL STRUCTURE =      192407
          PAGES WRITTEN TO NEW STRUCTURE         =           0
DSNB333I  -DP15 DSNB1GBR FINAL SWEEP COMPLETED FOR 845
GROUP BUFFER POOL GBP1
          PAGES WRITTEN TO NEW STRUCTURE         =           63
```

followed by message IXL042I (required response is no longer required).

Q: Do you have any observations based on this history and these messages??

Important lesson for DB2 GBPs

Questions:

- Why did it take so long for DB2 to respond to the rebuild?
- Why did DB2 cast out ALL the pages (to DASD) and ZERO to the new structure?
- Were there any DB2 messages about this GBP before or during this period?
- What was the size and makeup (number of entries and elements) of the primary and secondary GBP structures?

Important lesson for DB2 GBPs

What lessons were learned:

- Implement automation to raise an alert when message IXC585E is issued.
- Do NOT let a structure run at its maximum size. When Auto Alter increases the structure size, increase INITSIZE and SIZE values.
- Implement automation to raise an alert if DSNB319A or DSNB325E is issued.
- Do NOT reduplex a structure when it is very busy.
- Look at all parts of messages - for example, the line in the IXL040E that shows which phase of the rebuild was running, the DSNB332I line that shows that pages were cast out rather than being moved to the secondary GBP, the IXC582I message that shows the size and the number of entries and elements in the secondary GBP.

Sysplex Failure Management Threshold considerations



SFM Threshold considerations

Over the last few z/OS releases there have been a number of new features added to SFM to automatically respond to certain error conditions:

- MEMSTALLTIME z/OS 1.8
- SSUMLIMIT z/OS 1.9
- SFM and AutoIPL z/OS 1.10
- SFM and SSDPP/BCPii z/OS 1.11
- Changed default SSUM action z/OS 1.11
- Critical member support z/OS 1.12
- CFSTRHANGTIME z/OS 1.12

SFM Threshold considerations

Most of these are either turned off, or you specify a time after which SFM should take some action.

IBM provides recommendations for "appropriate" intervals, but what is the right value for you?

You want to strike a balance between avoiding IPLs (don't kill a system that is about to come back to life), and waiting so long that it places your whole sysplex at risk....

SFM Threshold considerations

You don't want to make the number too LOW....

Need to allow time for:

- Whatever process is running to complete.
- Time to do some investigation.
 - Try to figure out what is running, if it is still running, and possible reason why it is taking so long (for example, rebuilding a 10GB cache structure can take a while)..
- Is some recovery process going on? For example, spin loop recovery, or control unit recovery. You don't want to kill something if there is still a chance that it might come back to life soon.

SFM Threshold considerations

You don't want to make the number too HIGH....

The longer you let the sysplex run in a crippled mode, the more tasks and transactions and jobs will be affected.

If you wait too long, it may be difficult to identify the root cause.
All systems appear to have a problem, so who do you shoot?

SFM Threshold considerations

So what do you need to consider?

Look at your experiences with recovery:

- Have you had DASD control units go into recovery? How about hangs while applying maintenance? Speak to vendor about maximum recovery times. What are your experiences?
- What about spin loop recovery? If your systems tend to do a lot of this (look for LOGREC records or message IEE178I), maybe you should consider adjusting the SPINTIME and/or SPINRCVY values. Discuss with IBM before changing.
- Do you have experience of other recovery actions that take a long time, take those times into account. Maybe create some automation to check for these situations automatically.

SFM Threshold considerations

What is more painful for you?

- If all your critical work support data sharing and dynamic transaction routing (meaning that they can run on any system in the sysplex), addressing hung systems/members more aggressively might be appropriate.
 - You don't want sympathy sickness to spread to other systems
 - If the address space or system goes away, users can immediately re-logon to another member of the sysplex and continue working.
- If you have many affinities, meaning that critical work can only run on one system, then you want to do all you can to avoid an IPL.
 - In that case, you probably want to use longer values to give recovery more time to complete successfully or for you to have more time to do problem determination and try to fix the problem.

Identifying impact of dispatching delays



I/O Interrupt Delay

New (only on zEC12) is the ability to see how long you have to wait between when an interrupt is received back from a device, and when z/OS issues the TSCH to retrieve the results.

Provides insight into delays caused by things like:

- LPAR capping and weighting
- Number of virtual CPUs sharing a physical CPU
- Hiperdispatch
- Number of CPUs enabled for I/O interrupts (CPENABLE in IEAOPTxx)

Requires APARs OA37160 (IOS) and OA39993 (RMF) and zEC12

I/O Interrupt Delay

D I R E C T A C C E S S D E V I C E A C T I V I T Y												
z/OS V1R13				SYSTEM ID S5A				DATE 08/23/2012				
				RPT VERSION V1R13 RMF				TIME 13.00.00				
TOTAL SAMPLES =		900	IODF = 6A		CR-DATE: 08/07/2012			CR-TIME: 08.36.34				
STORAGE	DEV	DEVICE	NUMBER	VOLUME	PAV	LCU	DEVICE	AVG	AVG	AVG	AVG	AVG
GROUP	NUM	TYPE	OF CYL	SERIAL			ACTIVITY	RESP	IOSQ	CMR	DB	INT
							RATE	TIME	TIME	DLY	DLY	DLY
VSAMRLS	1101	33909	10017	RL1103	1.0H	0009	0.001	.384	.000	.000	.000	.000
VSAMRLS	1102	33909	10017	RL1104	1.0H	0009	0.078	.580	.000	.037	.000	.044
VSAMRLS	1103	33909	10017	RL1105	1.0H	0009	0.001	.256	.000	.000	.000	.000
VSAMRLS	1104	33909	10017	RL1106	1.0H	0009	0.001	.128	.000	.000	.000	.000
VSAMRLS	1105	33909	10017	RL1107	1.0H	0009	0.001	.384	.000	.000	.000	.000
VSAMRLS	1106	33909	10017	RL1108	1.0H	0009	1.261	.766	.000	.011	.000	.000
VSAMRLS	1107	33909	10017	RL1109	1.0H	0009	0.001	.128	.000	.000	.000	.000
VSAMRLS	1108	33909	10017	RL110A	1.0H	0009	0.428	.617	.000	.016	.000	.000
VSAMRLS	1109	33909	10017	RL110B	1.0H	0009	0.003	.256	.000	.000	.000	.000
VSAMRLS	110A	33909	10017	RL110C	1.0H	0009	3.467	4.28	.000	.033	.000	.001
VSAMRLS	110B	33909	10017	RL110D	1.0H	0009	0.017	.247	.000	.009	.000	.017
VSAMRLS	110C	33909	10017	RL110E	1.0H	0009	0.350	.337	.000	.010	.000	.002
VSAMRLS	110D	33909	10017	RL110F	1.0H	0009	0.001	.128	.000	.000	.000	.000
VSAMRLS	110E	33909	10017	RL1111	1.0H	0009	0.169	.927	.000	.034	.000	.002
VSAMRLS	110F	33909	10017	RL1113	1.0H	0009	0.640	.284	.000	.011	.000	.003

★ Appears in device data sections of SMF type 74 subtype 1 and SMF 79 subtype 9 records

JES3 to JES2 Migration Redbook



JES3 to JES2 migration

Why is ITSO working on a Redbook about JES3 to JES2 migration?

Don't worry, IBM is NOT planning to remove support for JES3!

However, with all the corporate mergers and consolidations, a number of IBM clients now have both JES3 and JES2 and are interested in consolidating onto one job entry subsystem

To help those clients make an informed decision about whether it is cost effective to undertake such a project, we are creating a book that addresses all the things you need to consider



JES3 to JES2 migration

IF you are a JES3 customer, there are a number of things you can do to give yourself more flexibility should you be interested in migrating some time in the future:

- Minimize the use of JES3 JECL - many JECL statements now have a JCL equivalent.
- Identify and, where possible, eliminate JES3 usermods and user exits
- Implement WLM-managed initiators
- Migrate NJE connections from BDT to TCP/IP
- Replace DJC and Deadline scheduling with a batch workload scheduler
- Move away from JES3-managed devices
- Replace JES3 DLOG with OPERLOG

Coupling Facility Upgrade Considerations



CF Upgrade considerations

Recommended procedure for emptying and repopulating CF for normal planned outage:

- Find a quiet time
- SETXCF START,MAINTMODE,CFNM=cf_to_be_upgraded
- SETXCF START,REALLOCATE
- On CF console, SHUTDOWN
- Perform activity
- REACTIVATE CF LPAR
- SETXCF STOP,MAINTMODE,CFNM=cf_to_be_upgraded
- SETXCF START,REALLOCATE

CF Upgrade considerations

Recommended procedure for emptying and repopulating CF for CF upgrade:

- Find a quiet time
- SETXCF START,MAINTMODE,CFNM=cf_to_be_upgraded
- SETXCF START,REALLOCATE
- On CF console, SHUTDOWN
- Update CFRM policy (including change of DUPLEX(ENABLE) to (ALLOWED))
- Perform upgrade
- SETXCF START,MAINTMODE,CFNM=cf_to_be_upgraded
- REACTIVATE CF LPAR
- SETXCF STOP,MAINTMODE,CFNM=cf_to_be_upgraded
- SETXCF START,REALLOCATE

Because upgrading CF resets MAINTMODE

ITSO - the times they are a-changing.....



ITSO Changes

In an effort to:

- Continue to fulfil our role of helping you understand and implement IBM technology
- Address the Generation Y audience

the ITSO is looking at what new things we can do, and what existing things we can do differently/better.

ITSO Changes

Some of the things we have already started doing include:

- Publishing books in ePub format (for use with eReaders, phones, etc).
 - Much nicer than PDFs when reading on a small screen
 - Beware that search is quite slow compared to PDFs
- Creating blog entries describing what we are doing in our residencies
- Putting out some books both as complete books, and also as individual chapters so you can look at just one part of the book if you wish
- Delivering "Solution Guides" - these are summaries of in-progress Redbooks that are aimed at management

QUESTION - do you like info centers?

ITSO Changes

We are also creating short how-to videos to illustrate things we describe in our books:

- See www.youtube.com/ibmredbooks

YouTube



Browse

Movies

Upload



Sign In



IBM Redbooks



Subscribe

167
subscribers

23,899
video views

Featured

Browse videos

Search Channel



About IBM Redbooks

Videos promoting and regarding IBM Redbooks and their publications.

IBM Redbooks web site

by IBMRedbooks

Latest Activity Oct 3, 2012

Date Joined Dec 14, 2009

Country United States

Featured Playlists



Redbooks

QUESTION - Are you allowed to access Youtube in work?

ITSO Changes

Please let us know if you have any other ideas about what we can do to be more valuable to you.

Also, when you use a Redbook, PLEASE take 1 minute to give it a star rating:

- This helps us understand which types of books people like and don't like
- Also helps your peers know which books other people found useful

If you don't like something, PLEASE send us a feedback so we can try to fix it.

If you DO like something, please say that as well, so we can do more of it.....

ITSO Changes

Another request - *please* always pull the latest version of a Redbook from the Redbooks web site:

- There are other web sites that keep copies of Redbooks, however you can't be guaranteed that you are getting the latest version of a book from those sites.
 - We often make corrections or small enhancements to a book and do NOT change the book number, so the only way to be sure that you have the latest version is to get it from www.redbooks.ibm.com.
- One of the criteria we use when deciding whether to update a book is the count of the number of references to the book. We are less likely to update a book that (we think) few people have used. If you take the books from other sites, we don't have that information and therefore don't have a complete picture when deciding which books to update.

And, finally, some unashamed advertising....



This is NOT Poughkeepsie.....



Neither is this.....



ITSO Residencies

Come to Poughkeepsie to take part in a project with other subject matter experts from all over the world to write a Redbook.....

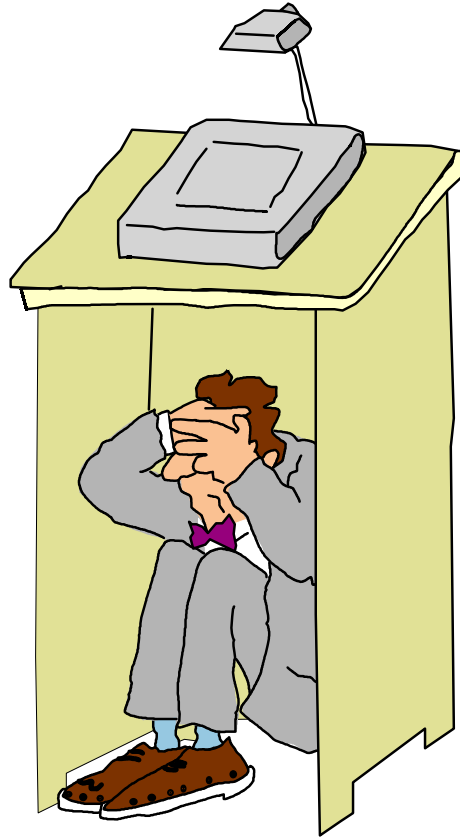
- IBM covers all travel expenses, hotel, meal allowance, car, etc...
- Your mission is to learn as much as you can about the latest and greatest IBM technology and document your experiences
- Gain fame and fortune (well, at least, you will get your name on the front cover of a Redbook)

If you think you might be interested, keep an eye on <http://www.redbooks.ibm.com/residents.nsf/ResIndex/>

or sign up for automatic notification at

<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>

Questions?



Thank you!!

PLEASE remember to complete your session evaluations

