



International Technical Support Organization

## ITSO 2007 Parallel Sysplex Update

[www.ibm.com/redbooks](http://www.ibm.com/redbooks)

Frank Kyne ([kyne@us.ibm.com](mailto:kyne@us.ibm.com))



© 2007 IBM Corporation. All rights reserved.

[ibm.com/redbooks](http://ibm.com/redbooks)

International Technical Support Organization



**My background.....**

**Your handouts vary slightly from the presentation - sorry!**

**The latest handouts will be available to IBM'ers in PDF format - go to [w3.itso.ibm.com](http://w3.itso.ibm.com), then Additional Materials, then ITSO Materials Repository.**

**PLEASE complete the evaluation forms.**

- Especially, if you are not happy, please say **WHY**
- If you feel additional clarification is needed on any topics, please indicate that as well

**Questions?? Please ask as I go along. Also, if you can't understand my strange accent, please let me know!**



©2007 IBM Corporation. All rights reserved.

## Agenda:

- Start 09:00
- Coffee about 10:30
- Lunch at 12:30 for 1 hour
- Afternoon coffee about 14:40
- Finish about 17:00

## Topics:

- Summary of sysplex-related changes in z/OS 1.7, 1.8, and 1.9
- CF Level 15-related enhancements
- XCF/XES/SFM/RMF new features and functions
- System Logger enhancements and tips
- SMF support for log streams
- Sysplex-related hardware considerations
- Miscellaneous topics of interest
- Appendix:
  - Enhancements to RRS and VTAM Generic Resources
  - Sysplex-related IBM services offerings

## Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

IBM has two registered trademarks for the branding of ITSO publications. These registered marks are for the text word "IBM Redbooks" and the Redbooks logo. In a nutshell, the term Redbooks must always be used in the plural form (for both text and logo) since IBM only owns the registered mark for the plural form. Usage must follow the guidelines below:

### Using the term Redbooks in written text

Redbooks are only to be referred to in the plural form, NEVER in the singular.

For the initial reference (first occurrence), you must use "IBM Redbooks®" and include "IBM" as well as the ®. For instances thereafter you may use "Redbooks" without "IBM" preceding the word or ® following it.

### Correct usage for written text :

In this IBM Redbooks® publication we will explore.....(® symbol required for 1st usage)  
This Redbooks publication will show you.....(2nd usage or later - no ® or "IBM" needed)

### Using the logo:

Redbooks (logo)



### OTHER ITSO PUBLICATIONS - Marks not yet registered

Trademark registration is a lengthy process and until we are officially registered, we cannot use the ® symbol. For those terms/logos in process, we will be using the ™ symbol. In contrast to the ® symbol (placed in the lower right hand corner), the ™ symbol is placed in the upper right hand corner. Please see examples below:

Redpaper ™  
Redpapers ™  
Redwiki ™  
Redwikis ™



The following terms are trademarks of other companies:

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.



©2007 IBM Corporation. All rights reserved.

## Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

eServer™	DS8000™	RACF®
pSeries®	FICON®	REXX™
z/OS®	GDPS®	RMF™
z/VM®	HyperSwap™	System z™
z/VSE™	IBM®	SystemPac®
z9™	IMS™	Tivoli®
BatchPipes®	MVS™	VTAM®
CICS®	Parallel Sysplex®	WebSphere®
DB2®	PAL™	
DFSMSHsm™	PR/SM™	
DFSMSrmm™	Redbooks®	

The following terms are trademarks of other companies:

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

CA is a trademark of Computer Associates

Other company, product, and service names may be trademarks or service marks of others.



©2007 IBM Corporation. All rights reserved.

## Themes

### Mainframe simplification

Remove constraints for large users - scalability

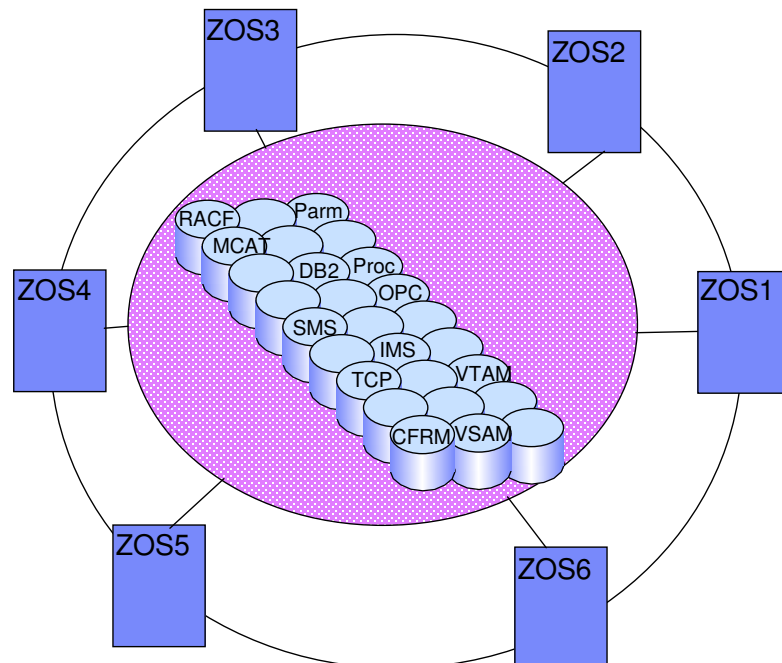
Reduced Total Cost of Ownership

Extend availability leadership:

- Planned outages
- Unplanned outages (includes performance)
- Autonomics:
  - Self tuning
  - Self healing
  - Reporting of exceptions from Best Practices
- Protect sysplex as an entity

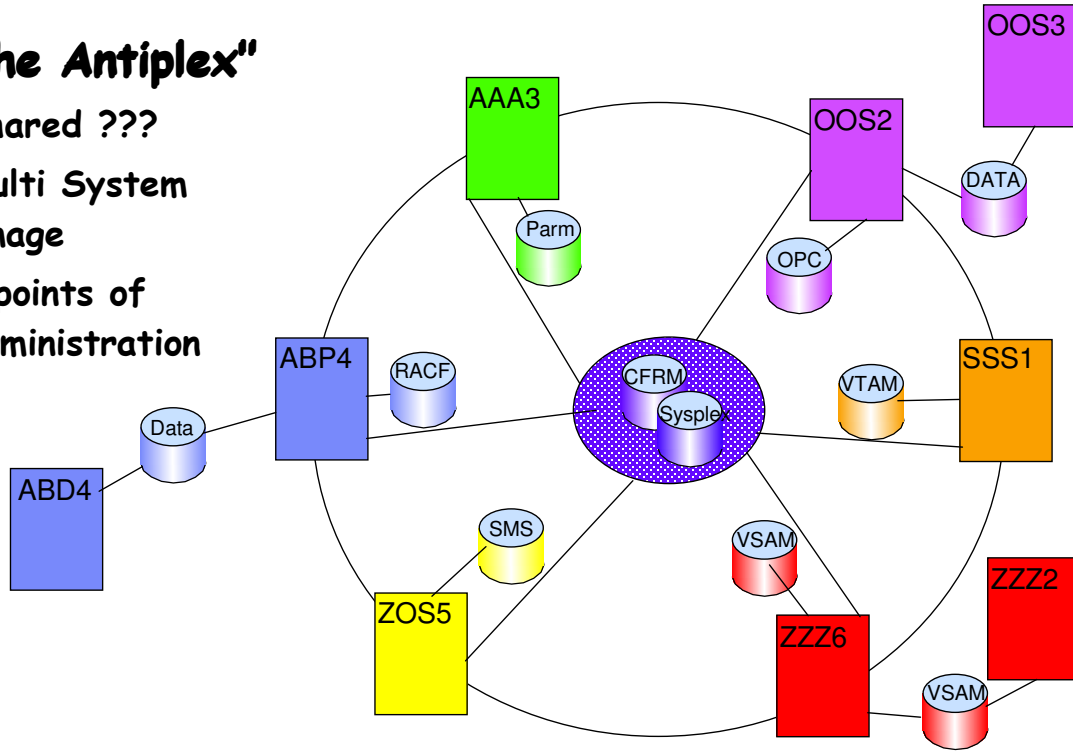
## "The Sysplex"

- Shared everything
- Single System Image
- Extensive use of system symbols
- Single points of administration

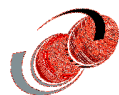


# "The Antiplex"

- Shared ???
- Multi System Image
- n points of administration



## Summary of recent sysplex-related enhancements



# Redbooks Workshop

IBM ITSO - International Technical Support Organization

## z/OS R7

- **Maximum number of locks per lock structure connector increased by about 16 times**
- **Number of connected DASDONLY log streams increased**
- **Increase from 1 up to 256, the number of Logger tasks to manage all DASDONLY log streams**
- **Add XRC+ support for GDPS/XRC**
- **Add ability to NOT start System Logger address space**
- **DEFAULT XCF MAXMSG value increased from 750 to 2000KB**
- **Add IXCDELET utility to delete members of an XCF group**
- **Introduce support for STP**

## z/OS R8

- **CFRM Performance Enhancements Stage 2**
- **SFM for XCF Sympathy Sickness (MEMSTALLTIME)**
- **Enhanced parallelism for CDS access**
- **Enhancements to SETXCF REALLOCATE command**
- **System Logger separation of Prod and Test log streams and certain associated Logger tasks**
- **System Logger Log Stream rename**
- **Throttle unauthorized Logger users**
- **"CF Hint" support in GDPS - SITE keyword in CFRM policy**
- **Enhancements to Distributed Byte Range Lock Manager to improve availability following a system failure**

## z/OS R8

- Ability to dynamically move the GRS Contention Notification System role between systems (SETGRS CNS= command)
- Improvements in balancing of WLM Managed Initiators (RA)
- Parallel Sysplex and ARM support for LDAP
- New routing option for Sysplex Distributor - "Local First"
- Enhancements to sysplex autonomics in TCP
- WLM enhancements for use with Sysplex Distributor (Health)
- VTAM and TCP subplex support
- Multiple APPC log streams per sysplex
- RRS SHUTDOWN command for clean shutdown of RRS
- New GRS, RRS, and Logger Health Checks

## Displaying "Health" value for a server

```
Netstat VDPT/-O DETAIL - display OPTLOCAL, health indicators, and active connections
Long Format:
MVS TCP/IP NETSTAT CS V1R8          TCPIP Name: TCPCS
Dynamic VIPA Destination Port Table:
Dest:          201.2.10.11..8000
DestXCF:       193.9.200.1
TotalConn: 0000000050 Rdy: 001 WLM: 00 TSR: 100
Flg: ServerWLM
TCSR: 100 CER: 100 SEF: 100
Abnorm: 1000 Health: 100
ActConn:       00000042
QoSPlcAct:    *DEFAULT*
W/Q: 0
QoSPlcAct:    Gold-Service
W/Q: 0
```

## z/OS R9

- Support for SM Duplexing enhancements in CF Level 15 (\*)
- Support for more granular CF reporting (\*)
- New SETXCF command to place CF in "maintenance mode" (\*)
- RMF Spreadsheet Reporter and RMF Data Portal extended to include support for XCF Postprocessor reports (\*)
- Externalization of sysplex information to CIM server by XCF
- SFM support for status-updt-missing-but-not-dead mbrs (\*)
- Support for system symbols in data set names in ISPF panels
- Share ISPF variables in a sysplex, eliminating need for multiple ISPF profile data sets
  - See ["Customizing for profile sharing" in ISPF Planning and Cust...](#)
  - **AND, you can disassemble a load module using ISRDDN!**

## z/OS R9

- Message Flood Automation function included with z/OS
- Concurrent offload data set recall for Logger (\*)
- SMF support for writing to log streams (\*)
- Support for multiple RMMplexes in one sysplex
- System symbol support in DFSMSrmm parmlib member
- Improved support for D XCF,C,TYPE=BPXMCDs command
- Reduced initialization time for USS when multiple systems in the plex IPL at the same time
- Batch equivalent of RRS ISPF interface (\*)
- Ability to force "UNSET" in RRS of a Resource Manager that is stuck in limbo (\*)



## z/OS R9

- **SLIP** now has ability to request a dump on another member of the Parallel Sysplex
- **WLM routing services** now include **zIIP** and **zAAP** capacity so that info can be used by **TCP**, **DB2**, **WAS** when deciding where to send work in a sysplex
- **GRS Storage Constraint Relief** - move all ENQ information to 64-bit storage
- **ARM support for CIM server**
  - See [CIM Users Guide](#) for required RACF commands
- **More new Health Checks**, plus support for writing your own checks in System REXX. For more info, see:
  - [http://www.ibm.com/servers/eserver/zseries/zos/hchecker/check\\_table.html](http://www.ibm.com/servers/eserver/zseries/zos/hchecker/check_table.html)

## z/OS R9

- **Comms Server enhancements:**
  - **YET ANOTHER(!)** option for distributing requests from Sysplex Distributor! **WEIGHTEDACTIVE** lets you control the percent of sessions to be allocated to each server
  - The new **SYSPLEX,QUIESCE,PORT=** (and corresponding **RESUME**) command enables you to quiesce individual applications from receiving new sysplex distributed workload. The application is identified in the quiesce command by its port and optionally, its job name, **ASID**, or both.
  - A new option, **DELAYSTART**, delays procedures configured in the **AUTOLOG** profile statement from automatically starting until **TCPIP** has joined (or rejoined) the sysplex group and processed its dynamic **VIPA** configuration.

## z/OS R9

### ▪ Comms Server enhancements:

- Unpronounceable enhancements related to Source IP addresses and VIPAs in a sysplex
- Automatically remove dynamic XCF definitions to a system when the last stack on that system shuts down
- Exploits information about zIIP and zAAP capacity from WLM when deciding where to route sessions:
  - For BaseWLM, you specify balance of zIIP and zAAP for a server
  - For ServerWLM, this information is provided automatically from WLM

### ▪ For VTAM:

- Significantly enhanced mechanism for controlling generic resource resolution rules (\*)

### ▪ See *Comms Server New Function Summary* for more info

## z/OS R9

### DB2 V9:

- Enhancement to remove directory information from secondary GBPs (\*)
- 14 other DB2 changes specifically to improve performance and availability and reduce data sharing overhead:
  - For more information, see the section entitled "Data Sharing" in "DB2 9 for z/OS Technical Overview", SG24-7330



International Technical Support Organization

[ibm.com/redbooks](http://ibm.com/redbooks)

## CF Level 15 and related z/OS enhancements



© 2007 IBM Corporation. All rights reserved.

[ibm.com/redbooks](http://ibm.com/redbooks)

International Technical Support Organization



## Summary

### ▪ New features in CF Level 15:

- More detail about CF config provided in D CF output and RMF reports
- Structure-level CF CPU reporting in RMF PP reports and Monitor III
- Increase maximum number of concurrent CF tasks from 48 to 112
  - Results in increases in structure sizes
- Improved performance for System-Managed Duplexing
  - Reduce number of CF-to-CF interactions

## Enhanced response to D CF command.....

```

D CF,CFNM=FACIL05
IXL150I 11.07.43 DISPLAY CF 984
COUPLING FACILITY 002094.IBM.02.00000002991E
PARTITION: 1E CPCID: 00
CONTROL UNIT ID: FFF3

NAMED FACIL05
COUPLING FACILITY SPACE UTILIZATION
ALLOCATED SPACE          DUMP SPACE UTILIZATION
STRUCTURES:             284160 K          STRUCTURE DUMP TABLES:          0 K
DUMP SPACE:             2048 K          TABLE COUNT:                   0
FREE SPACE:             1719808 K        FREE DUMP SPACE:                 2048 K
TOTAL SPACE:            2006016 K        TOTAL DUMP SPACE:                2048 K
                                MAX REQUESTED DUMP SPACE:          0 K
                                STORAGE INCREMENT SIZE:          512 K

VOLATILE:                YES
CFLEVEL:                 15
CFCC RELEASE 15.00, SERVICE LEVEL 00.19
BUILT ON 04/10/2007 AT 17:09:00
COUPLING FACILITY HAS 0 SHARED AND 1 DEDICATED PROCESSORS
DYNAMIC CF DISPATCHING: OFF

CF REQUEST TIME ORDERING: NOT-REQUIRED AND NOT-ENABLED
...
    
```

Previously this was only available on the CF HMC console

## Enhanced CF Information in RMF PP report.....

PROCESSOR SUMMARY							
COUPLING FACILITY	2094	MODEL S18	CFLEVEL 15	DYNDISP OFF			
AVERAGE CF UTILIZATION (% BUSY)	0.1	LOGICAL PROCESSORS:	DEFINED 1	EFFECTIVE 1	0		
			SHARED 0	AVG WEIGHT	0.0		

Doesn't directly report on the number of dedicated processors, but DEDICATED = DEFINED - SHARED

Had to guess at these previously, based on "Defined" and "Effective" numbers

## RMF PP now reports CF CPU utilization information at the structure level....

VERY important for accurate CF capacity planning as different structure types use differing amounts of CF CPU per request AND helps problem determination if CF CPU unexpectedly high

```

COUPLING FACILITY ACTIVITY
z/OS V1R8          SYSPLEX #@$#PLEX      START 07/31/2007-09.00.00   INTERVAL 000.30.00
                   RPT VERSION V1R8 RMF      END   07/31/2007-09.30.00   CYCLE 01.000 SECONDS
-----
COUPLING FACILITY NAME = FACIL05
TOTAL SAMPLES (AVG) = 60 (MAX) = 60 (MIN) = 59
-----
COUPLING FACILITY USAGE SUMMARY
-----
STRUCTURE SUMMARY
-----

```

TYPE	STRUCTURE NAME	STATUS	CHG	ALLOC SIZE	% OF CF STOR	# REQ	% OF ALL REQ	% OF CF UTIL	AVG REQ/SEC	LST/DIR TOT/CUR	DATA ELEMENTS TOT/CUR	LOCK ENTRIES TOT/CUR	DIR REC/DIR REC XI'S
LIST	IXC_BIG_1	ACTIVE		13M	0.7	12491	52.8	16.8	6.94	1225	1207	N/A	N/A
....	IXC_DEFAULT_2	ACTIVE		13M	0.7	1690	7.1	2.8	0.94	1225	1207	N/A	N/A

This column *usually* sums to 100. So, if overall CF utilization is 30%, and one structure shows 16.8% in this column, it used roughly 5.04% of CF CPU during this interval

## RMF Mon III also provides near-realtime CF CPU utilization information at the structure level....

```

RMF V1R8 CF Activity - #@$#PLEX Line 1 of 27
Command ==> Scroll ==> CSR
Samples: 120 Systems: 2 Date: 07/31/07 Time: 11.17.00 Range: 120 Sec
CF: FACIL06

```

Structure Name	Type	ST	System	CF Util %	Sync Rate	Avg Serv	Async Rate	Avg Serv	Chng %	Del %
DB8QU_SCA	LIST	A	*ALL	11.2	0.0	0	2.0	1008	0.0	0.0
	LIST		#@\$2	0.0	0.0	0	0.0	0	0.0	0.0
	LIST		#@\$3	0.0	0.0	0	2.0	1008	0.0	0.0
IRRXCF00_B001	CACHE	A	*ALL	0.7	0.0	0	0.0	0	0.0	0.0
	CACHE		#@\$2	0.0	0.0	0	0.0	0	0.0	0.0
	CACHE		#@\$3	0.0	0.0	0	0.0	0	0.0	0.0
ISGLOCK	LOCK	A	*ALL	20.7	1.3	217	9.8	1151	0.0	0.0
	LOCK		#@\$2	0.9	0.9	216	6.1	1196	0.0	0.0
	LOCK		#@\$3	0.4	0.4	217	3.7	1077	0.0	0.0
ISTGENERIC	LIST	A	*ALL	1.6	0.0	0	0.8	994	0.0	0.0
	LIST		#@\$2	0.0	0.0	0	0.4	1029	0.0	0.0
	LIST		#@\$3	0.0	0.0	0	0.4	959	0.0	0.0

## Plus subchannel utilization information....

```

Command ==>>          RMF V1R8   CF Systems      - #@$#PLEX          Line 1 of 4
                                                                Scroll ==>> CSR

Samples: 120      Systems: 2      Date: 07/31/07   Time: 11.17.00   Range: 120   Sec

CF Name  System  Subchannel  -- Paths --  -- Sync ---  ----- Async -----
          Delay  Busy  Avail  Delay  Rate  Avg  Rate  Avg  Chng  Del
          %      %      %      %      %    %    %    %    %    %
FACIL05  #@$2    0.0  0.0  3      0.0  0.0  0      5.9  553  0.0  0.0
          #@$3    0.0  0.0  2      0.0  <0.1  245  <0.1  859  0.0  0.0
FACIL06  #@$2    0.0  0.1  2      0.0  0.9  217  10.2  1062  0.0  0.0
          #@$3    0.0  0.0  4      0.0  0.4  217  9.8  916  0.0  0.0
  
```

"Subchannel Busy" column is what we call Subchannel Utilization.

- Unfortunately this same information is NOT provided in the RMF PP reports

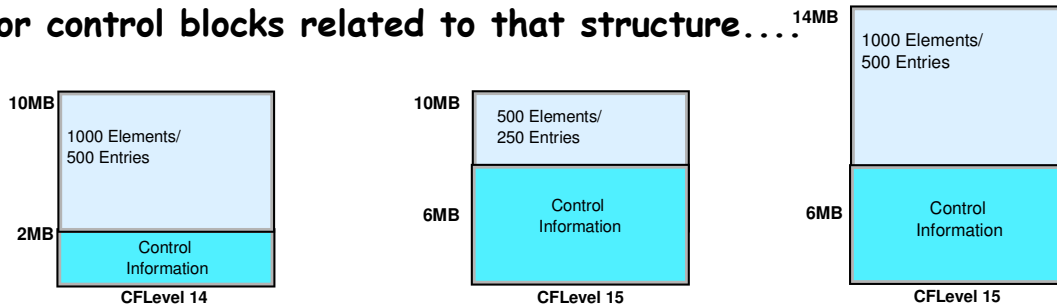
"Paths" in this report is the number of *defined* CF links from this z/OS to this CF - includes offline paths.

## Reporting granularity enhancements

- **Supporting APARs OA17070 (RMF), OA17055 (XCF) - PTFs available now:**
  - PTFs go back to z/OS 1.6
  - Requires CF at CF Level 15 to get the new information
- **Not necessary to have all CFs at the new level**
  - Additional information is available for CFs running Level 15 or higher, older levels will still report same information as before
- **Not necessary to have all systems at the new level**
  - Information comes from the CF, so any systems with the required PTFs will be able to see this, other systems will not.

## Increased maximum number of CF tasks

- Remember that some of the space in each structure is used for control blocks related to that structure....



- And CF Level 15 increases the maximum number of concurrent tasks per CF (means more control information)
- Structure size increase for CF Level 15 is a fixed amount per structure, *not* related to current structure size
- Vital that you adjust structure sizes to allow for this, *especially* for very small structures...

## Increased maximum number of CF tasks

- Following structures (with 64KB max data entry size) will increase by 4MB:
  - XCF Signalling
  - WLM IRD, Enclaves
  - Logger (CICS, SA, IMS, OPERLOG, LOGREC, RRS, HealthChecker, WAS, APPC, IMS CQS, other logstreams)
  - IMS EMH, CQS shared message queue
  - MQ shared queues application and administrative
  - VTAM MNPS (multi-node persistent sessions)
  - TCP/IP sysplex wide security associations
  - BatchPipes

## Increased maximum number of CF tasks

- **Following structures (with 32KB max data entry size) will increase by 2MB:**
  - DB2 GBPs (w/32K page size)
  - VSAM RLS cache
  - IMS cache (various types)
  - CICS temp storage, shared data tables

## Increased maximum number of CF tasks

- **Following structures (with 16KB max data entry size) will increase by 1MB:**
  - DB2 GBPs (w/4K, 8K, 16K pagesize), SCA
  - VTAM GR (ISTGENERIC)
  - RACF cache
  - Enhanced Catalog Sharing (ECS)
  - HSM common recall queue
  - CICS named counter server
  - IMS VSO
  - JES2 checkpoint
  - TCP/IP sysplex ports



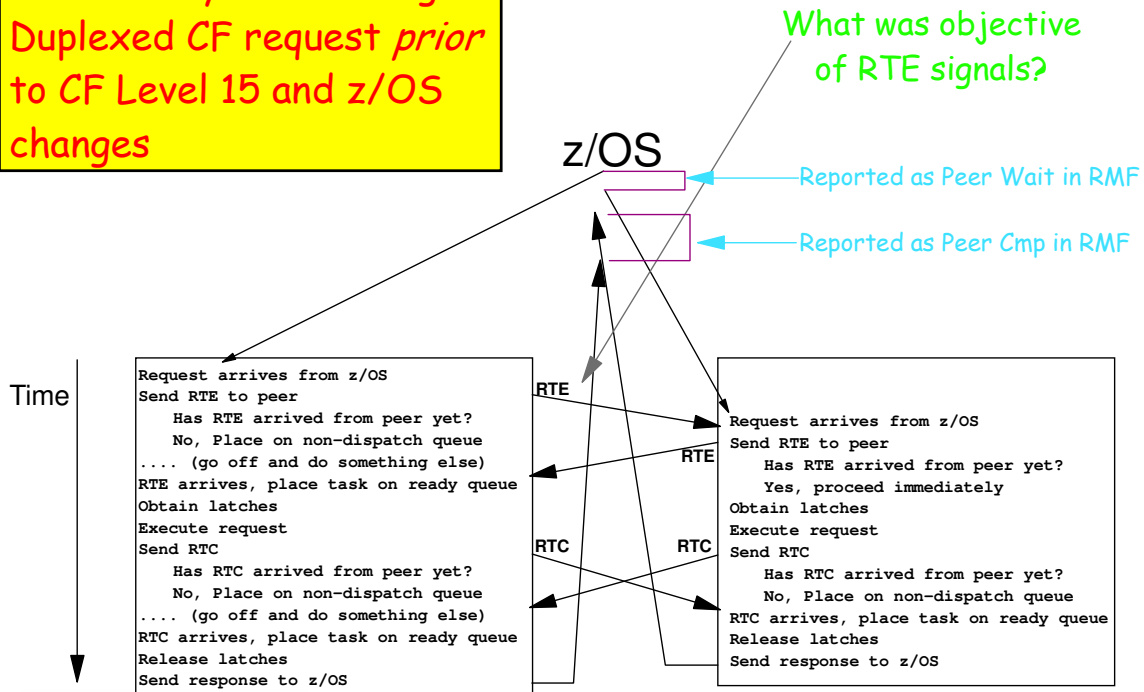
## Increased maximum number of CF tasks

- **Following structures will increase by .5 or .25 MB:**
  - All lock structures (GRS STAR ISGLOCK, IMS IRLM, DB2 IRLM, VSAM RLS IGWLOCK00, others)
- **Also, segment size increases from 256KB in CF Level 14 to 512KB in CF Level 15**
- **We'll talk a little more about handling structure size changes at the end of this section....**
  
- **Additionally - installations may see a small increase in synchronous response time and CF utilization when moving from CF Level 14 to 15**
  - This applies to *all* requests to a CF Level 15 CF

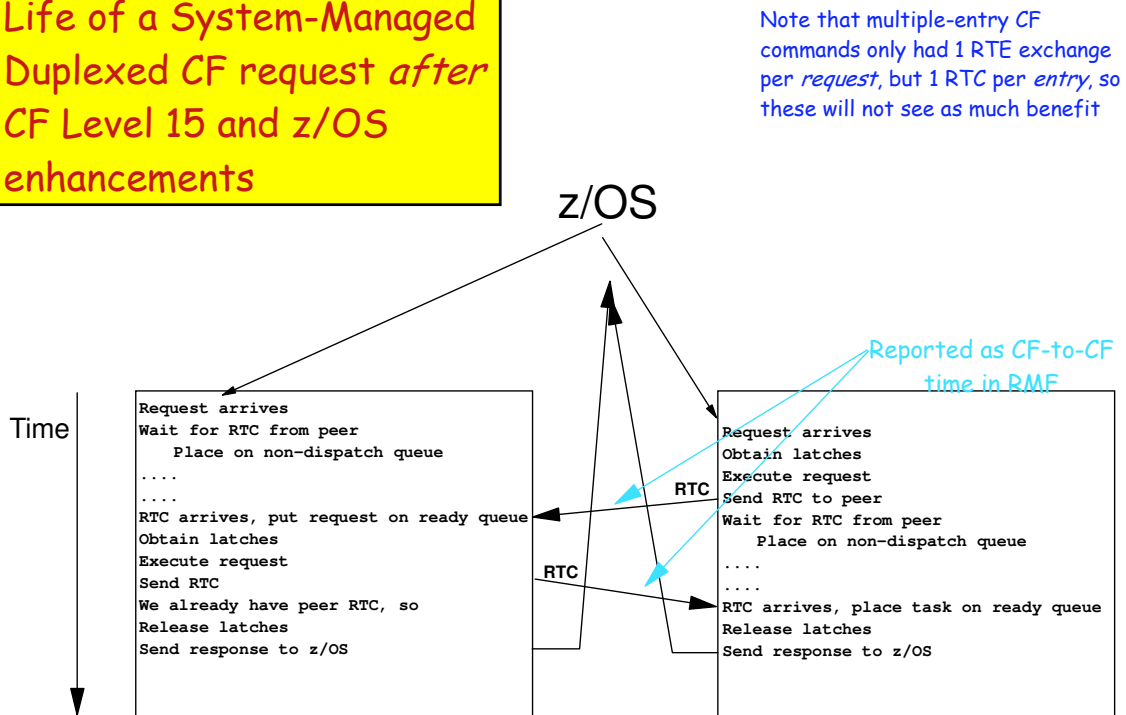
## CF Level 15 SM Duplexing enhancements

- **System-Managed Duplexing was intended as "the other half" of System-Managed Rebuild:**
  - If structures are duplexed, rebuild doesn't have to support recovery from structure or CF failure
- **However the performance impact of the current implementation means that SM Duplexing hasn't been as widely deployed as we originally expected it would be:**
  - The root cause of the performance issue is the CF-to-CF signals that are used to synchronize processing in the two CFs
- **To try to improve performance, XES and CF Level 15 have been changed to reduce the number of CF-to-CF interactions where possible**

### Life of a System-Managed Duplexed CF request prior to CF Level 15 and z/OS changes



### Life of a System-Managed Duplexed CF request after CF Level 15 and z/OS enhancements



## System-Managed Duplex enhancements

- **System-Managed Duplexing enhancements require:**
  - APAR OA21913, OA17055 (XES), and OA17070 (RMF) on top of z/OS 1.6 to 1.9
  - Announced to be available in 1Q2008
  - CF Level 15 - Shipped with z9 GA3 (Driver 67). Will require new (yet-to-be identified) CF Service Level.
  - Same HW requirements as previous implementation of SM Duplexing
    - Still need CF-to-CF links

## System-Managed Duplex enhancements

- **Implementation of new function is transparent once support is installed:**
  - No externals to turn on or off
  - In a plex where some systems have the supporting PTFs and some don't, systems with new function support will operate in the new way, while those without the support will continue to operate as before:
    - Remember that each pair of duplexed requests come from same z/OS image
      - the intent of RTE and RTC is to coordinate between two instances of the same update request, so it is OK if requests from one system work the old way and requests from another system work the new way
  - In a plex with Level 15 and pre-level 15 CFs, new function will not come into effect (both structure instances must reside in CFCC Level 15 CF).

## System-Managed Duplex enhancements

- **Performance:**
  - Measurement runs ongoing - no published results yet
- **Expected that:**
  - CF utilization may drop (depends on what percentage of requests are duplexed)
  - Response times for SM Duplexed requests should decrease
  - z/OS utilization will be unchanged
- **Please remember that turning on System Managed Duplexing results in a significant increase in subchannel utilization**
  - Structure response times increase
  - **AND, you are now talking to TWO CFs, so two CF links in use**

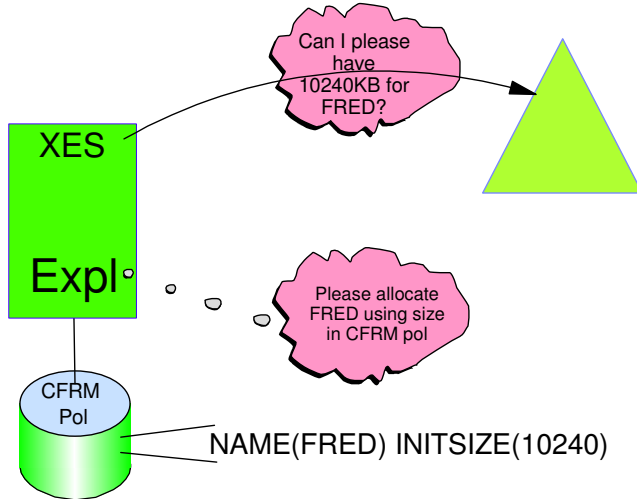
## General CFCC Level upgrade thoughts

- **Don't recommend upgrading all your CFs at the same time**
- **If you have production and test/Dev't CFs in the same CPC, you might want to exploit the ability to run different CF LPARs at different levels, while you test the new function**
- **Be aware that there are currently some small inaccuracies in the CFCC function to calculate required structure sizes based on structure attributes:**
  - May lead to strange results from SIZER batch utility
  - Don't get hung up over differences of .5MB - numbers are accurate enough to be useful
- **Need to think about your process for changing structure sizes, in terms of moving structures and CFRM policy changes**

## Structure size considerations

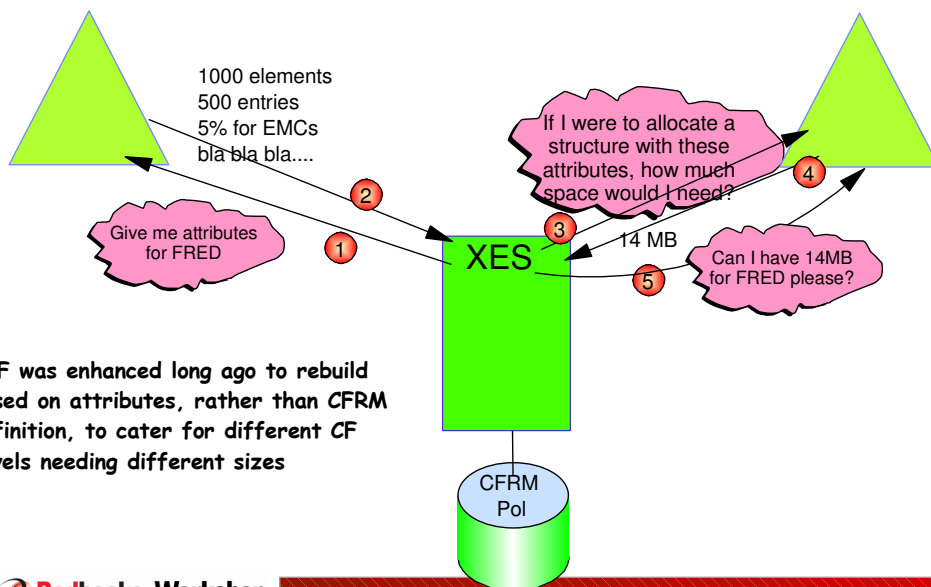
▪ We have 2.2 ways a structure can be allocated:

- Option 1 - structure is not currently allocated



## Structure size considerations

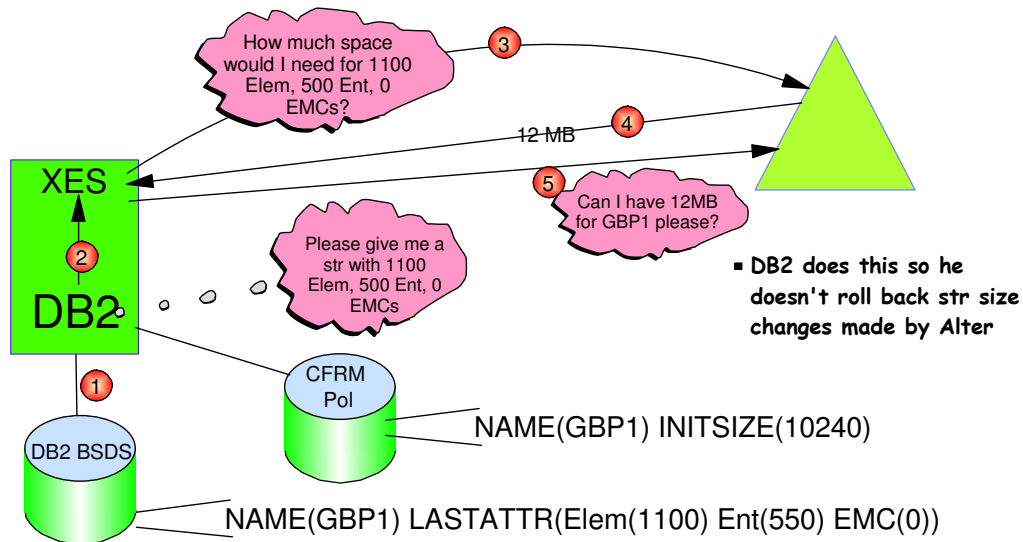
- Option 2 - structure is already allocated, but we want to move it to the other CF



▪ XCF was enhanced long ago to rebuild based on attributes, rather than CFRM definition, to cater for different CF Levels needing different sizes

## Structure size considerations

### - Option 2.2 - allocate DB2 GBP that is not currently allocated



## Value summary

### Customer value:

- Duplexing enhancements may allow use of SM Duplexing where it was not an option previously, especially for multi-site sysplexes. This in turn might deliver costs savings by enabling an ICF-only configuration
- Metrics enhancements allow better control, better problem determination, more accurate capacity planning, and easier-to-interpret reports

### Ease of implementation:

- 9 out of 10
  - Planning for structure size changes is the only complexity



International Technical Support Organization

[ibm.com/redbooks](http://ibm.com/redbooks)

XCF/XES/SFM/RMF New Functions



© 2007 IBM Corporation. All rights reserved.

[ibm.com/redbooks](http://ibm.com/redbooks)

International Technical Support Organization



## Coupling Facility Planned Outage Management



©2007 IBM Corporation. All rights reserved.

## Managing Planned CF outages

- **A unique attribute of Coupling Facilities is the ability to take a running CF, move its contents to an alternate CF, upgrade and recycle the CF, and move all the contents back - all without causing a service interruption to any applications using its services**
  - This capability is vital in order to support the 24x7 application availability that more and more businesses require

## Managing Planned CF outages

- **However, a weak link in this capability is the complexity that may be associated with emptying and repopulating a CF:**
  - How to handle duplexed structures - can't "move" them with rebuild?
  - Have to move all structures out of the CF in a manner that causes as little disruption as possible
    - SETXCF START,RB,CFNM=cf1,LOC=OTHER moves every structure in parallel
    - Don't forget that XCF structures may require special treatment
  - And then, having successfully achieved all this, you need to move everything back to where it belongs:
    - Which command to use (LOC=NORMAL, POPCF, or REALLOCATE)?
    - Need to make sure all structures that should be duplexed are back in duplex state
    - Need to make sure that duplexed structures are in the "right" CF - that is, that all primary structures are in the first CF in their PREFLIST



## Managing Planned CF outages

### ▪ How is emptying the CF achieved?

- Maintain multiple CFRM policies with different combinations of CFs and PREFLISTs and start different policies to match your (transient) target configuration, OR
- Move/handle the structures manually, OR
- Just kill the CF and let recovery handle things (hopefully not this option!!), OR
- Use the new support in z/OS 1.9.....

## Managing Planned CF outages

### ▪ To significantly simplify things, z/OS 1.9 adds a new command (SETXCF START,MAINTMODE) to make a CF unavailable for new structure allocations:

- Has similar effect to removing the named CF from the PREFLISTs in the CFRM policy, but without all the manual work to do that

## Managing Planned CF outages

- First, let's see what CFs we have and what is in them...

```

Session B - gdps mop whitescreen.ws - [43 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ==>
000210 D XCF,CF,CFNM=ALL SCROLL ==> CSR
000210 IXC3591 09/11/07 17.33 DISPLAY XCF 345
000210 CFNAME: CF1
000210 COUPLING FACILITY : 002094.IBM.02.00000002991E
000210 PARTITION: 0F CPCID: 00
000210 SITE : N/A
000210 POLICY DUMP SPACE SIZE: 2048 K
000210 ACTUAL DUMP SPACE SIZE: 2048 K
000210 STORAGE INCREMENT SIZE: 512 K
000210
000210 CONNECTED SYSTEMS:
000210 SC63 SC64 SC65 SC70
000210
000210 STRUCTURES:
000210 DB8FU_LOCK1 DB8FU_SCA IGWLOCK00
000210 ISGLOCK ISTMNPS IXC_DEFAULT_1
000210 IXC_DEFAULT_3 RRS_DELAYEDUR_1 RRS_MAINUR_1
000210 RRS_START_1 RRS_RNDATA_1 SYSGGAS_ECS
000210 SYSTEM_LOGREC(OLD) SYSTEM_OPERLOG(NEW) SYSZULM_WORKUNIT
000210 CFNAME: CF2
000210 COUPLING FACILITY : 002094.IBM.02.00000002991E
000210 PARTITION: 1D CPCID: 00
000210 SITE : N/A
000210 POLICY DUMP SPACE SIZE: 2048 K
000210 ACTUAL DUMP SPACE SIZE: 2048 K
000210 STORAGE INCREMENT SIZE: 512 K
000210
000210 CONNECTED SYSTEMS:
000210 SC63 SC64 SC65 SC70
000210
000210 STRUCTURES:
000210 EJESGDS_WTSCPLX4 ISTGENERIC IXC_DEFAULT_2
000210 IXC_DEFAULT_4 LOG_IGWLOG_001 SYSARC_PLEX0_RCL
000210 SYSTEM_LOGREC(NEW) SYSTEM_OPERLOG(OLD) SYSZULM_991E2094
000210 CFNAME: CF3
000210 COUPLING FACILITY : 002094.IBM.02.00000002991E
000210 PARTITION: 0E CPCID: 00
000210 SITE : N/A
000210 POLICY DUMP SPACE SIZE: 512 K

```

## Managing Planned CF outages

- Now we want to take CF2 down, so we put it in maintmode..

```

Session B - gdps mop whitescreen.ws - [43 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ==>
000210 SETXCF START,MAINTMODE,CFNM=CF2
000210 IXC5691 MAINTENANCE MODE STARTED FOR 350
000210 COUPLING FACILITY 002094.IBM.02.00000002991E
000210 PARTITION: 1D CPCID: 00
000210 NAMED CF2
000210 IXC3691 THE SETXCF START MAINTMODE REQUEST FOR COUPLING FACILITY 351
000210 CF2 WAS SUCCESSFUL
000210
000210 SETXCF START,REALLOCATE
000210 IXC5431 THE REQUESTED START,REALLOCATE WAS ACCEPTED. 353
000210 IXC5741 EVALUATION INFORMATION FOR REALLOCATE PROCESSING 354
000210 OF STRUCTURE IGWLOCK00
000210 SIMPLEX STRUCTURE ALLOCATED IN COUPLING FACILITY: CF1
000210 ACTIVE POLICY INFORMATION USED.
000210 CFNAME STATUS/FAILURE REASON
000210 ----
000210 CF1 PREFERRED CF 1
000210 CF2 ALLOCATION NOT PERMITTED INFO110: 00000028 CC000B00 0000000F
000210 COUPLING FACILITY IS IN MAINTENANCE MODE
000210 INFO110: 00000000 00000000 00004000
000210 WAS NOT ATTEMPTED BECAUSE
000210 STRUCTURE IS ALLOCATED IN PREFERRED CF
000210 IXC5741 EVALUATION INFORMATION FOR REALLOCATE PROCESSING 356
000210 OF STRUCTURE ISGLOCK
000210 SIMPLEX STRUCTURE ALLOCATED IN COUPLING FACILITY: CF1
000210 ACTIVE POLICY INFORMATION USED.
000210 CFNAME STATUS/FAILURE REASON
000210 ----
000210 CF1 PREFERRED CF 1
000210 CF2 ALLOCATION NOT PERMITTED INFO110: 00000028 CC007800 0000000F
000210 COUPLING FACILITY IS IN MAINTENANCE MODE
000210 INFO110: 00000000 00000000 00004000
000210 WAS NOT ATTEMPTED BECAUSE
000210 STRUCTURE IS ALLOCATED IN PREFERRED CF
000210 IXC5741 EVALUATION INFORMATION FOR REALLOCATE PROCESSING 358
000210 OF STRUCTURE ISTGENERIC

```

## Managing Planned CF outages

- CF2 contains simplex structures, duplexed structures, and XCF structures. We can remove ALL these with one command

```

Session B - gdps mop whitescreen.ws - [24 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ==>
000210 SETXCF START,REALLOCATE
000010 IXC543I THE REQUESTED START,REALLOCATE WAS ACCEPTED. 353
000210 IXC574I EVALUATION INFORMATION FOR REALLOCATE PROCESSING 354
000210 OF STRUCTURE IGVLOCK00
000210 SIMPLEX STRUCTURE ALLOCATED IN COUPLING FACILITY: CF1
000210 ACTIVE POLICY INFORMATION USED.
000210 CFNAME STATUS/FAILURE REASON
000210 -----
000210 CF1 PREFERRED CF 1
000210 INFO110: 00000028 CC000B00 0000000F
000210 CF2 ALLOCATION NOT PERMITTED
000210 INFO110: 00000000 00000000 00004000
000210 COUPLING FACILITY IS IN MAINTENANCE MODE
000010 IXC544I REALLOCATE PROCESSING FOR STRUCTURE IGVLOCK00 355
000010 WAS NOT ATTEMPTED BECAUSE
000010 STRUCTURE IS ALLOCATED IN PREFERRED CF
000210 IXC574I EVALUATION INFORMATION FOR REALLOCATE PROCESSING 356
000210 OF STRUCTURE IGVLOCK
F1=HELP F2=SPLIT F3=END F4=RETURN F5=IFIND F6=BOOK
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE
b
Connected to remote server/host wtsnet.itso.ibm.com using lu/pool SC HP DeskJet 890C on LPT1:
  
```

## Managing Planned CF outages

- (Nearly) All structures are moved, DUPLEX(ENABLED) are re-duplexed if possible

```

Session B - gdps mop whitescreen.ws - [24 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ==>
000010 IXC545I REALLOCATE PROCESSING RESULTED IN THE FOLLOWING: 892
000010 7 STRUCTURE(S) REALLOCATED - SIMPLEX
000010 2 STRUCTURE(S) REALLOCATED - DUPLEXED
000010 0 STRUCTURE(S) POLICY CHANGE MADE - SIMPLEX
000010 0 STRUCTURE(S) POLICY CHANGE MADE - DUPLEXED
000010 12 STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF - SIMPLEX
000010 0 STRUCTURE(S) ALREADY ALLOCATED IN PREFERRED CF - DUPLEXED
000010 1 STRUCTURE(S) NOT PROCESSED
000010 36 STRUCTURE(S) NOT ALLOCATED
000010 20 STRUCTURE(S) NOT DEFINED
000010 -----
000010 100 TOTAL
000010 0 ERROR(S) ENCOUNTERED DURING PROCESSING
000010 IXC543I THE REQUESTED START,REALLOCATE WAS COMPLETED. 893
000210 IXC466I INBOUND SIGNAL CONNECTIVITY ESTABLISHED WITH SYSTEM SC70 241
000210 VIA STRUCTURE IXC_DEFAULT_4 LIST 10
000210 IXC466I OUTBOUND SIGNAL CONNECTIVITY ESTABLISHED WITH SYSTEM SC63 755
F1=HELP F2=SPLIT F3=END F4=RETURN F5=IFIND F6=BOOK
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE
b
Connected to remote server/host wtsnet.itso.ibm.com using lu/pool SC HP DeskJet 890C on LPT1:
  
```

## Managing Planned CF outages

### ▪ CF2 is now (nearly) empty (with just two commands)

```

Session B - gdps mop whitescreen.ws - [43 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ===== SCROLL ===== CSR
000010 D XCF CF,CFNM=ALL
000010 IXC362I 19.26.53 DISPLAY XCF 961
000010 CFNAME: CF1
000010 COUPLING FACILITY : 002094.IBM.02.00000002991E
000010 PARTITION: 0F CPCID: 00
000010 SITE : N/A
000010 POLICY DUMP SPACE SIZE: 2048 K
000010 ACTUAL DUMP SPACE SIZE: 2048 K
000010 STORAGE INCREMENT SIZE: 512 K
000010 CONNECTED SYSTEMS:
000010 SC63 SC64 SC65 SC70
000010 STRUCTURES:
000010 DB8FU_LOCK1 DB8FU_SCA IGWLOCK00
000010 ISGL0CK ISTGENERIC ISTMNPS
000010 IXC_DEFAULT_1 IXC_DEFAULT_2 IXC_DEFAULT_3
000010 IXC_DEFAULT_4 LOG_IGWLOG_001 RRS_DELAYEDUR_1
000010 RRS_MAINUR_1 RRS_RESTART_1 RRS_RMDATA_1
000010 SYSARC_PLEX0_RCL SYSARCAS_ECS SYSTEM_LOGREC(OLD)
000010 SYSTEM_OPERLOG(OLD) SYSZULM_WORKUNIT SYSZULM_991E2094
000010 CFNAME: CF2
000010 COUPLING FACILITY : 002094.IBM.02.00000002991E
000010 PARTITION: 10 CPCID: 00
000010 SITE : N/A
000010 POLICY DUMP SPACE SIZE: 2048 K
000010 ACTUAL DUMP SPACE SIZE: 2048 K
000010 STORAGE INCREMENT SIZE: 512 K
000010 CONNECTED SYSTEMS:
000010 SC63 SC64 SC65 SC70
000010 STRUCTURES:
000010 EJESGGS_UTSCPLX4
000010 SYSZULM_WORKUNIT
F1=HELP F2=SPLIT F3=END F4=RETURN F5=IFIND F6=BOOK
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE
-----
Connected to remote server/host wtsnet.itso.ibm.com using lu/pool SC38TC93 and port HP DeskJet 890C on LPT1:
  
```

Shows us that no new structures will be allocated in this CF

Connector was disconnecting at the same time as the REBUILD was being attempted. Second REALLOCATE command should resolve this

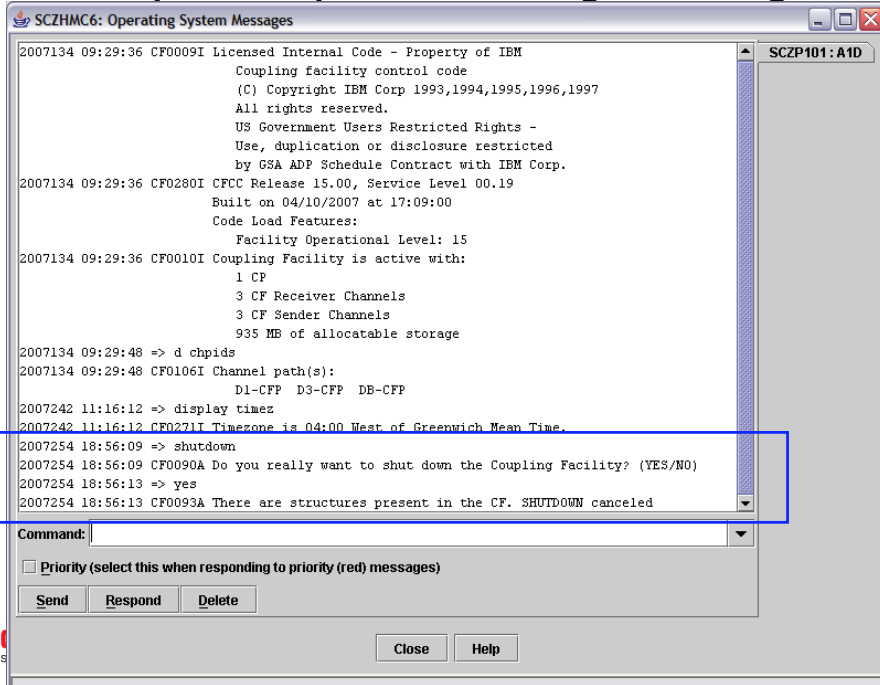
## Managing Planned CF outages

### ▪ Next stage in shutdown procedure is to stop CF LPAR:

- Could use CFDRAIN function in Systems Automation) or equivalent function in other product), if available
  - Recommended action if available
- Can do this by DEACTIVATING the CF LPAR from HMC
  - NOT recommended - too easy to accidentally reset the wrong LPAR
- Can use CF SHUTDOWN command
  - Just as fast as issuing DEACTIVATE from HMC
  - Lot safer as it checks to ensure CF is empty
- If using STP, DO NOT Config CHP all CF links offline from z/OS
  - Configuring links offline will remove timing signals

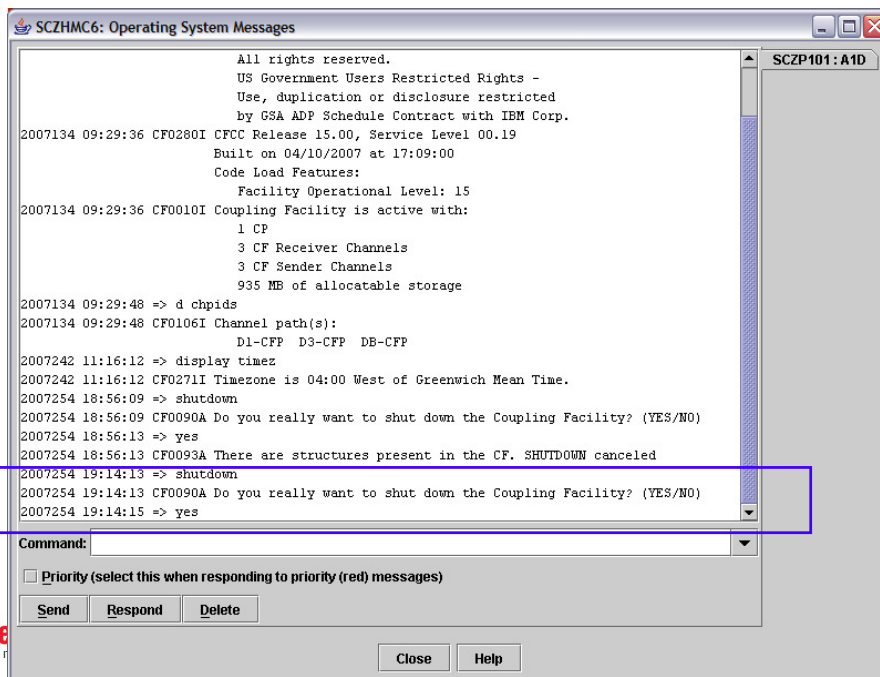
## Managing Planned CF outages

- **SHUTDOWN** protects you from oversights or finger checks



## Managing Planned CF outages

- **After cleaning up remaining structure...**



# Managing Planned CF outages

- And CF LPAR is now down and ready to be serviced

The screenshot shows a 'Sandbox Work Area' window titled 'SCZP101 A1D Details'. The 'Instance Information' tab is active, showing the following details:

- Status: **Not Operating** (circled in red)
- Group: **Sandbox** (circled in red)
- Activation profile: A1D
- Last used profile: A1D
- Operating system: Operating system
- SysPlex name:
- CPU LPAR cluster name:
- Operating system type:
- Operating system level:

Task Information shows 'Task name: Operating System Messages' and 'Task status:'.

Below the window, a terminal window displays system messages:

```

SDSF OPERLOG DATE 09/11/2007 1 WTOR
COMMAND INPUT ==>
000010 IXL157I PATH D2 IS NOW OPERATIONAL TO CUID: FFF6 462
000010 COUPLING FACILITY 002094.IBM.02.00000002991E
000010 PARTITION: 1D CPCID: 00
000010 IXL157I PATH DA IS NOW OPERATIONAL TO CUID: FFF6 463
000010 COUPLING FACILITY 002094.IBM.02.00000002991E
000010 PARTITION: 1D CPCID: 00
000210 IXC507I CLEANUP FOR 464
000210 COUPLING FACILITY 002094.IBM.02.00000002991E
000210 PARTITION: 1D CPCID: 00
000210 HAS STARTED.
000210 TRACE THREAD: 001D67DF.
000210 IXC507I CLEANUP FOR 465
000210 COUPLING FACILITY 002094.IBM.02.00000002991E
000210 PARTITION: 1D CPCID: 00
000210 HAS COMPLETED.
000210 TRACE THREAD: 001D67DF.
000010 IXC517I SYSTEM SC65 ABLE TO USE 466
000010 COUPLING FACILITY 002094.IBM.02.00000002991E
000010 PARTITION: 1D CPCID: 00
000010 NAMED CF2
    
```

# Managing Planned CF outages

- After service is complete, **ACTIVATE** CF LPAR

The screenshot shows a terminal window titled 'Session B - gdps mop whitescreen.ws - [24 x 80]'. The terminal output includes the following messages:

```

SDSF OPERLOG DATE 09/11/2007 1 WTOR
COMMAND INPUT ==>
000010 IXL157I PATH D2 IS NOW OPERATIONAL TO CUID: FFF6 462
000010 COUPLING FACILITY 002094.IBM.02.00000002991E
000010 PARTITION: 1D CPCID: 00
000010 IXL157I PATH DA IS NOW OPERATIONAL TO CUID: FFF6 463
000010 COUPLING FACILITY 002094.IBM.02.00000002991E
000010 PARTITION: 1D CPCID: 00
000210 IXC507I CLEANUP FOR 464
000210 COUPLING FACILITY 002094.IBM.02.00000002991E
000210 PARTITION: 1D CPCID: 00
000210 HAS STARTED.
000210 TRACE THREAD: 001D67DF.
000210 IXC507I CLEANUP FOR 465
000210 COUPLING FACILITY 002094.IBM.02.00000002991E
000210 PARTITION: 1D CPCID: 00
000210 HAS COMPLETED.
000210 TRACE THREAD: 001D67DF.
000010 IXC517I SYSTEM SC65 ABLE TO USE 466
000010 COUPLING FACILITY 002094.IBM.02.00000002991E
000010 PARTITION: 1D CPCID: 00
000010 NAMED CF2
    
```

## Managing Planned CF outages

- Then make it available for selection again using **SETXCF STOP,MAINTMODE** - note that this does NOT trigger re-duplexing or moving of structures

```

Session B - gdps mop whitescreen.ws - [24 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ==> SCROLL ==> CSR
000210 SETXCF STOP,MAINTMODE,CFNM=CF2
000210 IXC569I MAINTENANCE MODE STOPPED FOR 153
000210 COUPLING FACILITY 002094.IBM.02.0000002991E
000210 PARTITION: 1D CPCID: 00
000210 NAMED CF2
000010 IXC369I THE SETXCF STOP MAINTMODE REQUEST FOR COUPLING FACILITY 154
000010 CF2 WAS SUCCESSFUL.
*IMS CONNECT READY* IMSHCUNN
***** BOTTOM OF DATA *****

```

## Managing Planned CF outages

- Then finally repopulate it using **START,REALLOCATE** - note that this moves one structure at a time, so is less disruptive than everything moving and re-duplexing at the same time

```

Session B - gdps mop whitescreen.ws - [24 x 80]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/11/2007 1 WTOR COLUMNS 52- 131
COMMAND INPUT ==> SCROLL ==> CSR
000210 SETXCF START,REALLOCATE
000010 IXC543I THE REQUESTED START,REALLOCATE WAS ACCEPTED. 159
000210 IXC574I EVALUATION INFORMATION FOR REALLOCATE PROCESSING 160
000210 OF STRUCTURE IGWLOCK00
000210 SIMPLEX STRUCTURE ALLOCATED IN COUPLING FACILITY: CF1
000210 ACTIVE POLICY INFORMATION USED.
000210 CFNAME STATUS/FAILURE REASON
000210 -----
000210 CF1 PREFERRED CF 1
000210 INFO110: 00000028 CC000B00 0000000F
000210 CF2 PREFERRED CF ALREADY SELECTED
000210 INFO110: 00000028 CC000B00 0000000F
000010 IXC544I REALLOCATE PROCESSING FOR STRUCTURE IGWLOCK00 161
000010 WAS NOT ATTEMPTED BECAUSE
000010 STRUCTURE IS ALLOCATED IN PREFERRED CF
000210 IXC574I EVALUATION INFORMATION FOR REALLOCATE PROCESSING 162
000210 OF STRUCTURE ISGLOCK
000210 SIMPLEX STRUCTURE ALLOCATED IN COUPLING FACILITY: CF1
000210 ACTIVE POLICY INFORMATION USED.
000210 CFNAME STATUS/FAILURE REASON

```

## Managing Planned CF outages

### ■ Implementation considerations

- New SETXCF MAINTMODE command provided with z/OS 1.9, NOT rolled back to earlier releases
  - However, earlier releases will understand that the CF is "unavailable" for new structure allocations for some reason and will act accordingly - that is, you must do the START and STOP,MAINTMODE commands on an R9 or later system, but earlier releases can be active in the sysplex and won't try to put new structures in the CF after it goes into MAINTMODE. REALLOCATE command can be issued on any member of the sysplex.
  - Recommend installing toleration APAR OA17685 so that back levels will understand why they can't use the CF (and report that correctly in their messages).
- If you happen to fallback from R9 to an earlier release, make sure no CFs are left in MAINTMODE!
  - Sysplex IPL will clear MAINTMODE indicator

## Managing Planned CF outages

### ■ Summary

- Really nice new function, very easy to use, makes procedures a lot simpler for operations
  - Strongly recommend that it is used together with CFCC SHUTDOWN command from CF HMC console
- Can be used as soon as one system in the sysplex moves to z/OS 1.9
- Recommend updating operations procedures (or your home-grown automation) to use this process



## Value summary

### ▪ Customer value:

- Should enable higher availability due to simplifying the process of emptying the CF AND (especially) putting everything back in the right place after the CF comes back online:
  - Should avoid errors in taking the "wrong" CF down
  - It is common to see structures allocated with single point of failure and no one notices
- Less performance impact when CF comes back online
  - Maintmode stops all duplexed structures from trying to reduplex in parallel

### ▪ Ease of implementation:

- 9 out of 10
  - Only complexity is that you need to update your operator procedures

## Changes in Sysplex Failure Management (SFM)

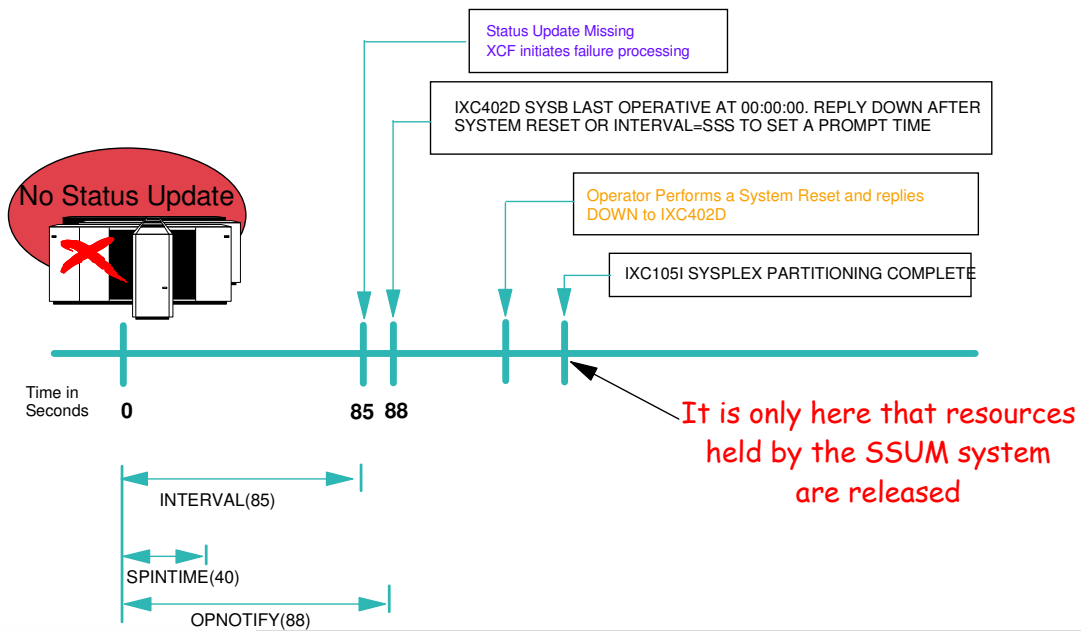
## Sysplex Failure Management (SFM)

- **Sysplex Failure Management (SFM) is a standard component of z/OS. However, you must explicitly enable it.**
- **It can:**
  - Automatically partition a dead system out of the sysplex
  - Automatically complete the removal of a system that was manually removed from the sysplex (V XCF,sysname,OFFLINE)
  - Reconfigure storage from a failed member (rarely used)
  - Control whether structures get rebuilt or not following a CF connectivity failure (REBUILDPERCENT - **NEVER USE THIS!!**)
  - Automatically remove a stalled XCF member (MEMSTALLTIME - new in z/OS 1.8)
  - Automatically remove a system that is not updating its heartbeat, but that is not completely dead either - **SSUMLIMIT - NEW in z/OS 1.9!**

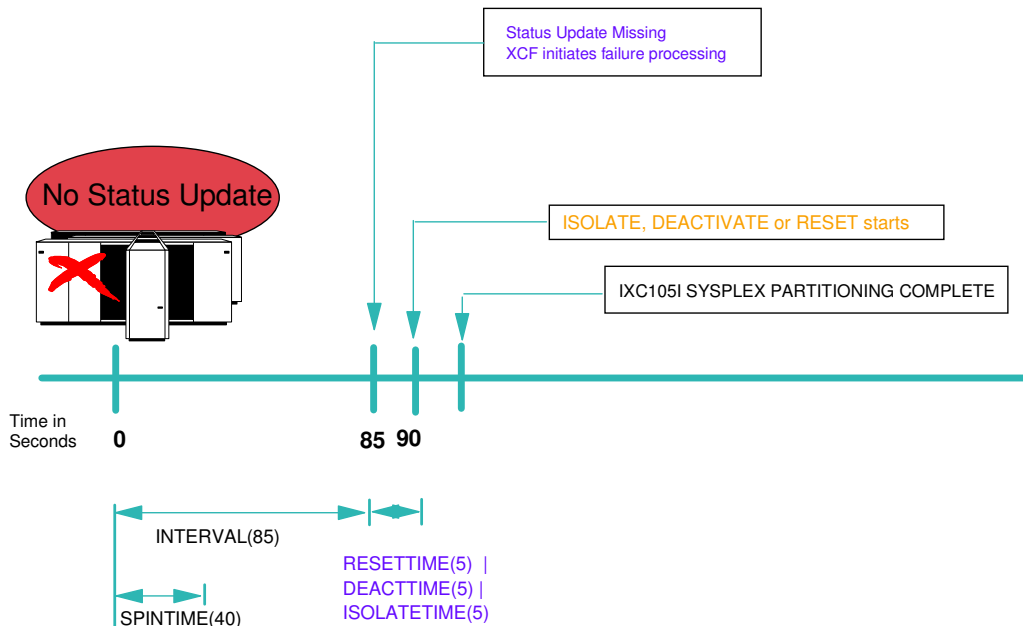
## Importance of speedy response

- **Why is it so important to quickly remove a dead (or dying) system from the sysplex?**
  - System may be holding RESERVEs, but can't release them. As a result, other systems can't update that volume.
  - System may be holding resources (ENQs, Locks) that other systems need. So work on those systems waiting for those resources stalls.
  - The stalled work on those other systems may in turn be holding resources that someone else needs, but those resources can't be released until the work can use the resource held by the dead system
  - The longer this situation goes on, the more work will come to a halt, waiting for a resource held by someone else
  - If allowed to go on for long enough, this can result in sysplex IPL

# Timeline - no status update, no SFM



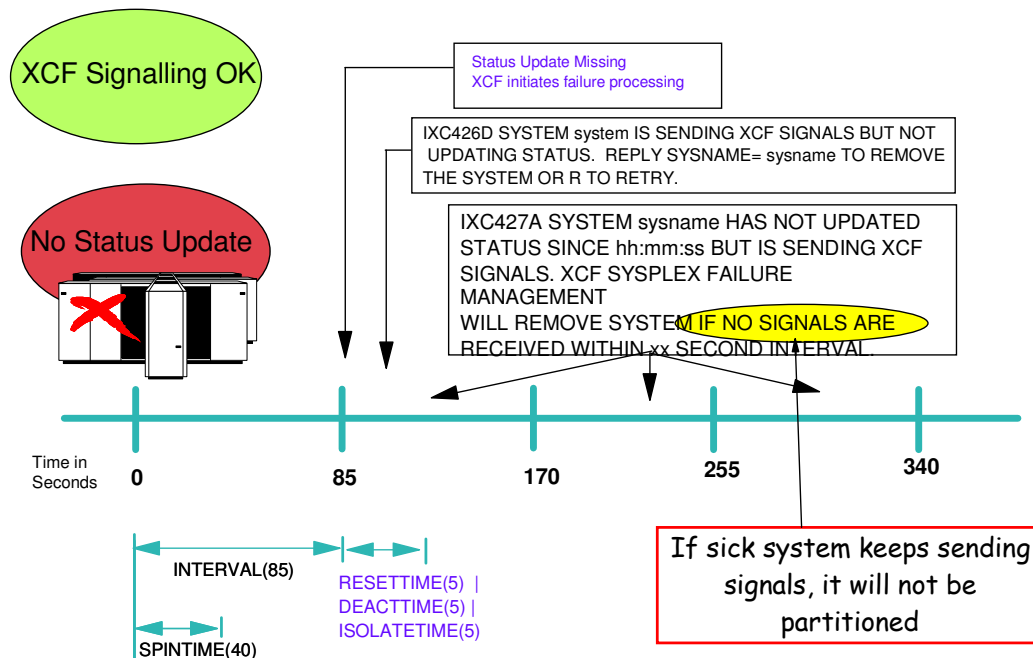
# Timeline - no status update, SFM active



## Enhanced SFM stalled system handling

- Originally, SFM only checked to see if a system had updated its heartbeat in the Sysplex CDS to determine if a system was dead:
  - If the system is dead, recommended action is to immediately partition it out of the sysplex (ISOLATETIME(0) in SFM policy)
- SFM was then enhanced (APAR OW30926) to additionally check to see if a system that has not updated its heartbeat is still issuing XCF signals:
  - A truly dead system will not be issuing XCF signals:
    - If no XCF signalling, partition system from 'plex (depending on ISOLATETIME setting in SFM policy)
    - If XCF signalling IS still active, message IXC426D (instead of IXC402D) issued to Operator

## Timeline - no status update, XCF good, SFM active



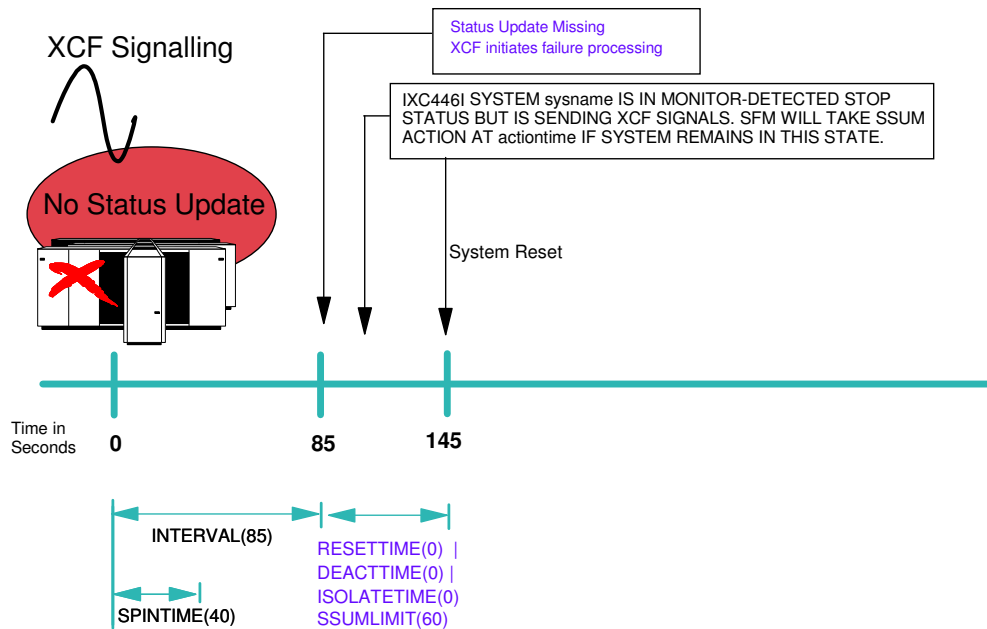
## Enhanced SFM stalled system handling

- A **SSUM** system **CAN** still communicate with the **CF** and **XCF** signals and may still be processing work, but it probably has **I/O** problems. A system should not be allowed stagger along like this indefinitely because it can't see the **sysplex CDS**:
  - Not aware of members leaving or joining **XCF** groups elsewhere in the **sysplex**, for example
  - Members on that system can't join or leave **XCF** groups because **XCF** can't update the **sysplex CDS** to reflect the change
- So..... **z/OS R9** introduces the option to have **SFM** automatically partition such systems out of the 'plex after an installation-specified interval

## Enhanced SFM stalled system handling

- **New SFM Policy keyword: SSUMLIMIT(ssssss)**
  - Can specify default value for **sysplex**, and/or specific values for each system
- If this is exploited, message **IXC446I** is issued instead of message **IXC426D**, stating that the system will be partitioned from the **sysplex** within **xxx** seconds unless the system starts updating the **Sysplex CDS** again.
  - **IXC446I SYSTEM sysname IS IN MONITOR-DETECTED STOP STATUS BUT IS SENDING XCF SIGNALS. SFM WILL TAKE SSUM ACTION AT actiontime IF SYSTEM REMAINS IN THIS STATE.**
- The default is **NOT** to do anything - you must specify a value for **SSUMLIMIT** for any action to be taken.

## Timeline - no status update, XCF good, SSUMLIMIT active



## SSUMLIMIT Implementation

- **Not necessary to wait until all members of the sysplex move to z/OS 1.9:**
  - However, function will only be active on the 1.9 or later systems and only work for systems that are also 1.9:
    - Earlier systems don't understand SSUMLIMIT, so don't give their SSUMLIMIT value to their peers in the sysplex
    - Not rolled back to earlier releases via APAR
- **To activate, change SFM policy to add SSUMLIMIT keyword, update policy from a 1.9 system (or one that is pointing at the 1.9 level of IXCMIAPU), and start new policy on a 1.9 system.**
- **Need to carefully consider SSUMLIMIT to get a balance between killing too soon, and not soon enough...**

## SFM General

- **SFM is still the single best thing you can do to reduce the chance of encountering a sysplex-wide IPL**
  - However we still have customers that don't activate SFM or (just as bad) run SFM with PROMPT rather than ISOLATETIME
  - z/OS Health Checker checks if SFM is active, but doesn't currently check if PROMPT or ISOLATEIME is in use
  - Don't forget the relationship between Failure Detection Interval and your spin recovery actions and SPINTIME
    - Failure Detection Interval is based on the DEFAULT SPINTIME - if you change SPINTIME (or SPINRCVY actions) in EXSPATxx, FDI is NOT changed automatically....

## Value summary

- **Customer value:**
  - Should improve sysplex availability by extending SFM benefits (automating removal from the sysplex) to "sick" systems
- **Ease of implementation:**
  - 8 out of 10
    - Implementation is simple (just update SFM policy), but you need to determine an appropriate value for each system, and make sure that the IXC446I message is immediately brought to the attention of the correct personnel.

# Changes in Recovery Processing for User-Managed Duplexed Structures

## Changes in duplexed structure recovery

- A duplexed structure, from XES' perspective, is one that is in the middle of being rebuilt, but he never got to the point of deleting the "old" instance

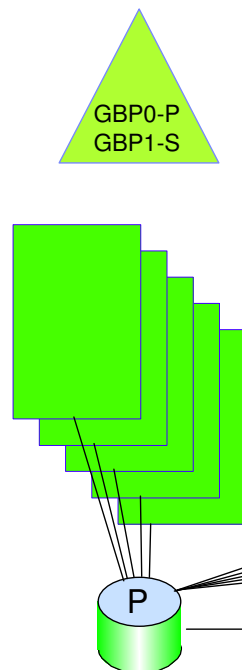
```
D XCF,STR,STRNM=DB8QU_GBP0
IXC360I 11.39.36 DISPLAY XCF 324
STRNAME: DB8QU_GBP0
STATUS: REASON SPECIFIED WITH REBUILD START:
        POLICY-INITIATED
        DUPLExING REBUILD
        METHOD: USER-MANAGED
        PHASE: DUPLEx ESTABLISHED
...
DUPLExING REBUILD NEW STRUCTURE
-----
ALLOCATION TIME: 09/06/2007 11:38:23
CFNAME       : FACIL06
COUPLING FACILITY: 002094.IBM.02.00000002991E
                PARTITION: 1F  CPCID: 00
...
DUPLExING REBUILD OLD STRUCTURE
-----
ALLOCATION TIME: 09/06/2007 11:38:23
CFNAME       : FACIL05
COUPLING FACILITY: 002094.IBM.02.00000002991E
                PARTITION: 1E  CPCID: 00
```



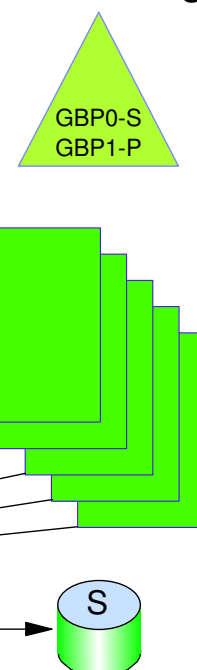
## Changes in duplexed structure recovery

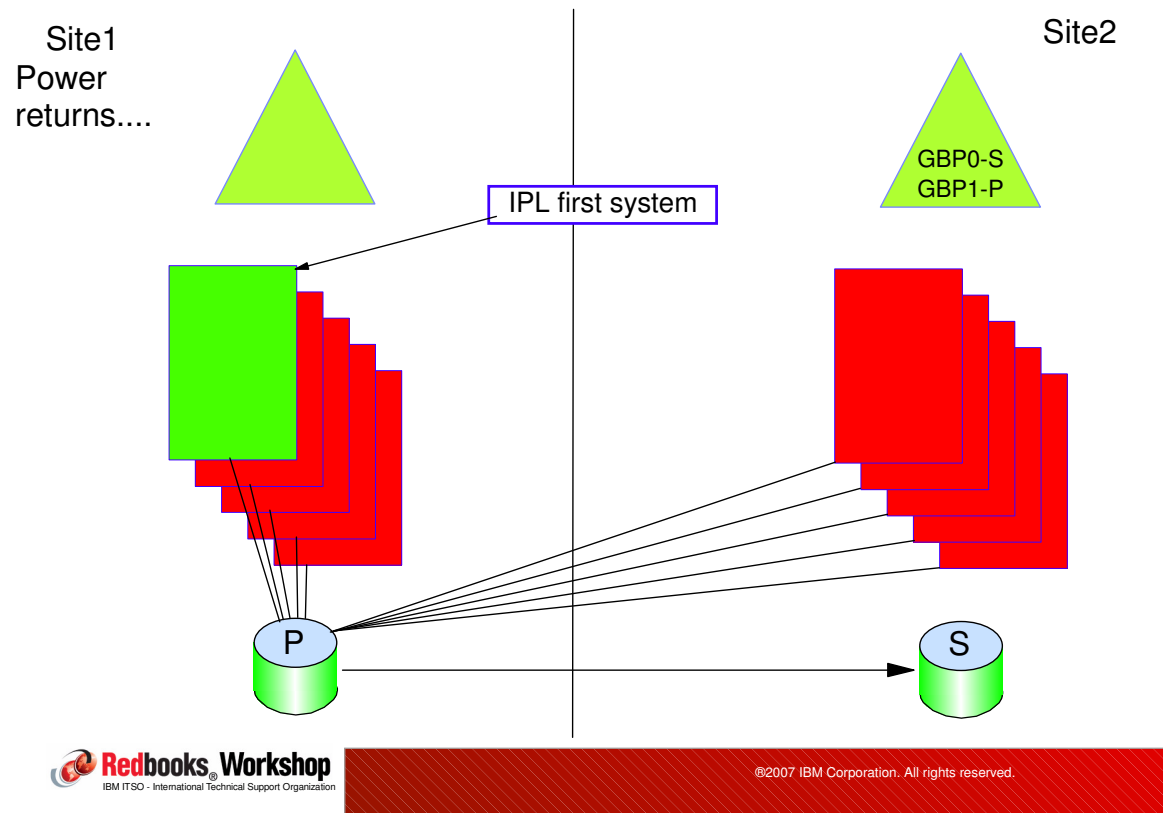
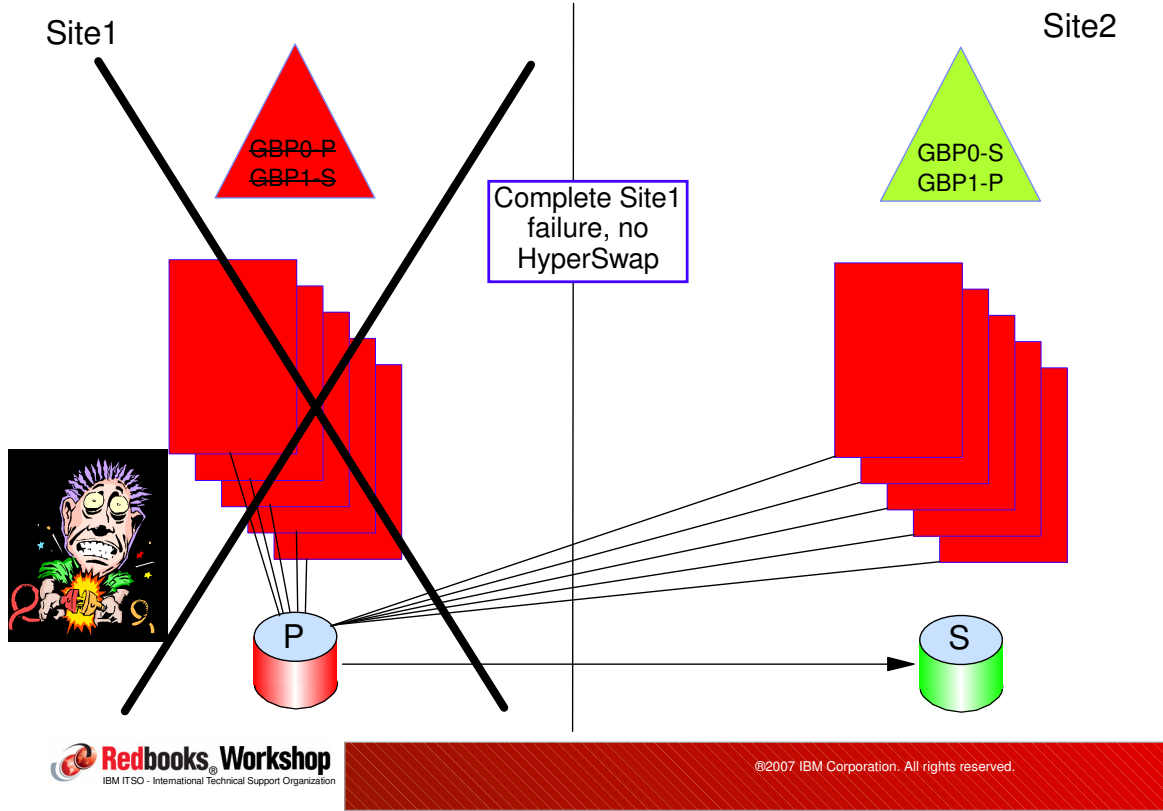
- In the event of an unplanned Sysplex IPL, early in the IPL process, XES used to inspect the CFRM CDS. When he did this, he would find incomplete rebuilds (these are the User-Managed duplex structures) and "complete" the rebuild by removing the information about the "old" instance and make the "new" instance the sole one.
  - Because it is the user's (DB2) responsibility to keep UM Duplexed structures in synch, XES plays safe by reverting to simplex after a sysplex IPL (because he doesn't know if DB2 managed to keep the structures in synch before the sysplex failure).
  - For a System-Managed Duplexed structure, XES is responsible for keeping structures in synch, so he will attempt to keep both instances

Site1



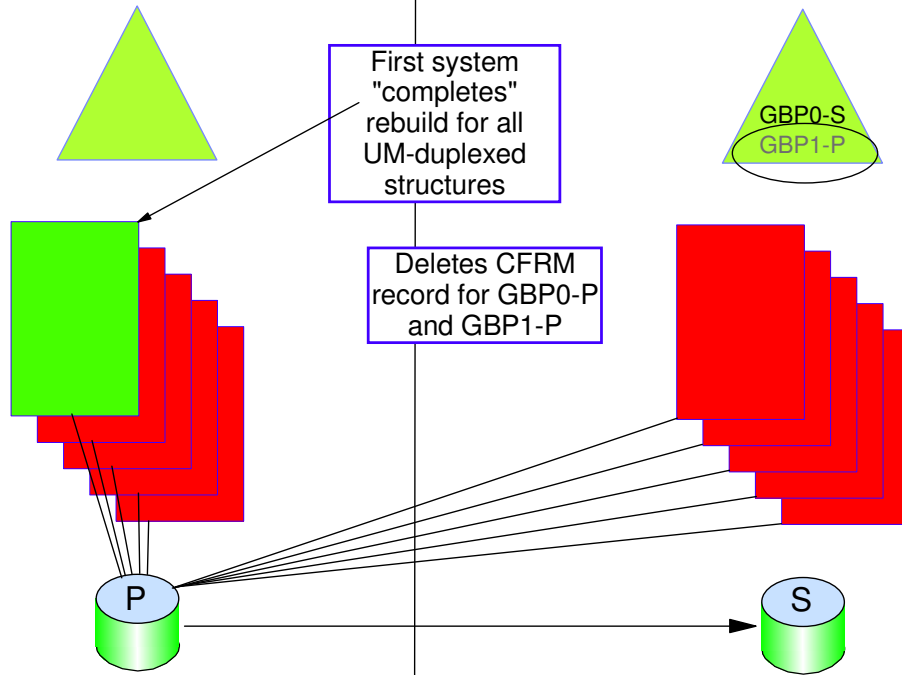
Site2





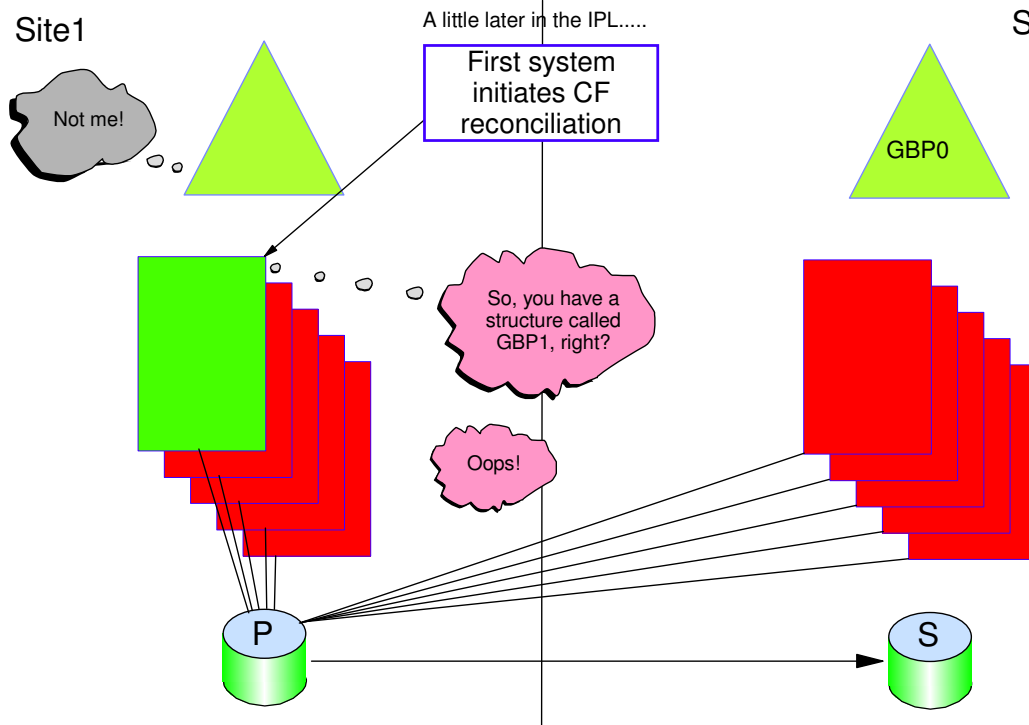
Site1

Site2



Site1

Site2



## Changes in duplexed structure recovery

- Net result of old way of processing was that any User-Managed duplexed structures that no longer have either structure instance will need to be recovered - this can take a LOOOOOOOOOOOONG time.
- To address this, XES APAR OA19151 was opened and delivered earlier in 2007.
  - Changes sequence of operations so that no structure records are removed until AFTER XES talks to each CF to determine which structures actually still exist.

## Implementation

- PTFs for APAR OA19151 available back to z/OS 1.6
- Should be installed by rolling IPL.
  - Until it is installed on all members of the sysplex, you can still get the benefit of it by ensuring that the first system IPLed after a sysplex outage is one that has the PTF applied, AND don't IPL any of the other members until AFTER you get the message for ALL CFs saying that the system is able to use the CF:

```
IXC517I SYSTEM #@3 ABL TO USE 188
          COUPLING FACILITY 002094.IBM.02.00000002991E
                                PARTITION: 1F      CPCID: 00
          NAMED FACIL06
```

- This means that CFRM has completed any cleanup work he needs to do, so the remaining systems can now be started.
- By the way..... this is one of the reasons I recommend both UPS AND Battery Backup for CFs....

## Value summary

### ▪ Customer value:

- Should improve DB2 availability in any configuration where all the systems could go down, but not all the CFs die (for example, multi-site sysplex or config where only some CFs have battery backup)

### ▪ Ease of implementation:

- 10 out of 10
  - Just install the APAR (OA19151) and you are finished. But watch for associated DB2 APAR (APAR number not assigned yet)

## XCF Best Practices

## XCF Best Practices

- Every transport class should have **AT LEAST** two, 1 failure-isolated, paths to every other member of the sysplex
- Use message size, rather than originating XCF group, to assign messages to a transport class 2

- CLASSDEF CLASS (DEFAULT) CLASSLEN (956) MAXMSG (2000)
- CLASSDEF CLASS (DEFMED) CLASSLEN (20412) MAXMSG (2000)
- CLASSDEF CLASS (DEFLARGE) CLASSLEN (62464) MAXMSG (2000)

- Do **NOT** do something like...

```
CLASSDEF CLASS (GRS) CLASSLEN (956) MAXMSG (2000) GROUP (SYSGRS)
CLASSDEF CLASS (RMF) CLASSLEN (20412) MAXMSG (2000) GROUP (SYSRMF)
CLASSDEF CLASS (VTAMBIG) CLASSLEN (62464) MAXMSG (2000) GROUP (ISTXCF)
```

## XCF Best Practices

- For nearly all customers, 3-4 transport classes should be plenty: 3
  - Generally want a 956-byte class, a large class, and 1 or 2 in between
- Make sure XCF structure sizes match recommendation from CFSizer 4
- Specify MAXMSG of at least 2000 in COUPLExx 5
- Monitor for non-0 REQ REJECT on PATHIN and PATHOUT
  - Get this information from the RMF XCF report
- Consider adding paths if Q Length in RMF XCF report is regularly > 1.0
  - Get this information from the RMF XCF report

## XCF Health Checks

- XCF\_CF\_CONNECTIVITY
  - Change SEV to HIGH
- XCF\_FDI
- XCF\_SFM\_ACTIVE
  - But doesn't check on PROMPT|ISOLATETIME
- XCF\_CLEANUP\_VALUE
- XCF\_CDS\_SEPARATION
- XCF\_SYSPLEX\_CDS\_CAPACITY
- XCF\_TCLASS\_HAS\_UNDESIG ②
- XCF\_TCLASS\_CONNECTIVITY ①
  - But override to say PARM('2')
- XCF\_TCLASS\_CLASSLEN ③
  - But override to say PARM('3')

## XCF Health Checks

- XCF\_SIG\_CONNECTIVITY
  - Change SEV to HIGH
- XCF\_DEFAULT\_MAXMSG ⑤
- XCF\_MAXMSG\_NUMBUF\_RATIO ⑤
- XCF\_SIG\_PATH\_SEPARATION
  - Change SEV to HIGH
- XCF\_SIG\_STR\_SIZE ④
- XCF\_CF\_STR\_PREFLIST
- XCF\_CF\_STR\_EXCLLIST

# RMF Spreadsheet Reporter XCF Support

(with thanks to Harald Bender!)

## RMF Spreadsheet Reporter enhancement

- **RMF Postprocessor provides most of the information you need to set up and tune XCF paths and transport classes**
- **But one of the drawbacks is that the reports all only have a single-system view and are not exactly user-friendly...**
- **To address this, the latest release of the RMF Spreadsheet Reporter (5.2.4) provides information about all the sysplex members in a single spreadsheet**
  - Provides information from the 1st RMF XCF report (XCF Usage by System) and the 3rd report (XCF Path Statistics). Does not currently support the 2nd report (XCF Usage by Member).

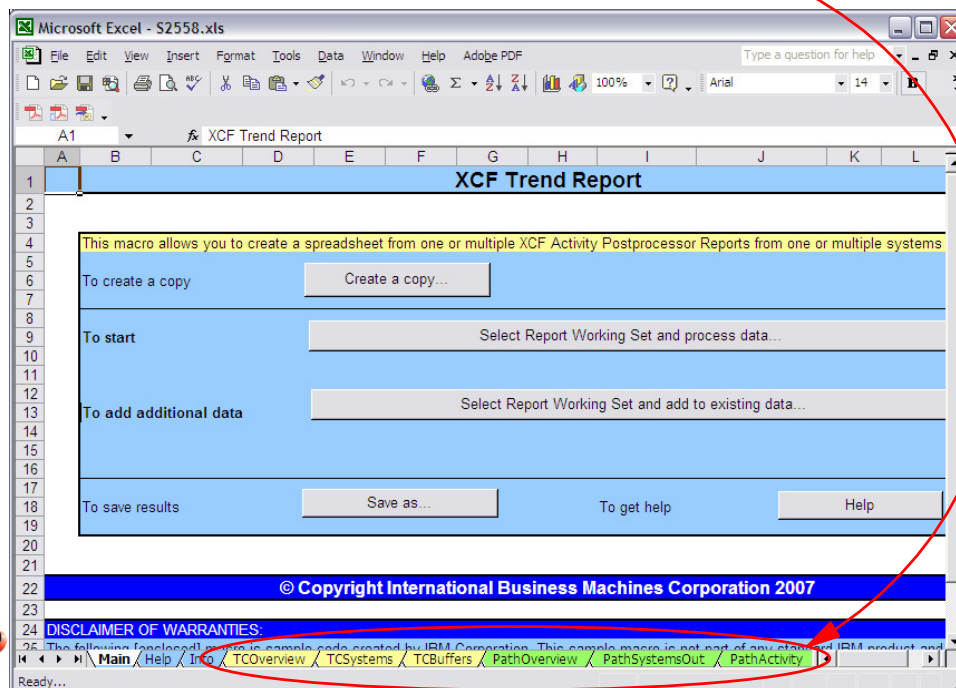


## RMF Spreadsheet Reporter Introduction

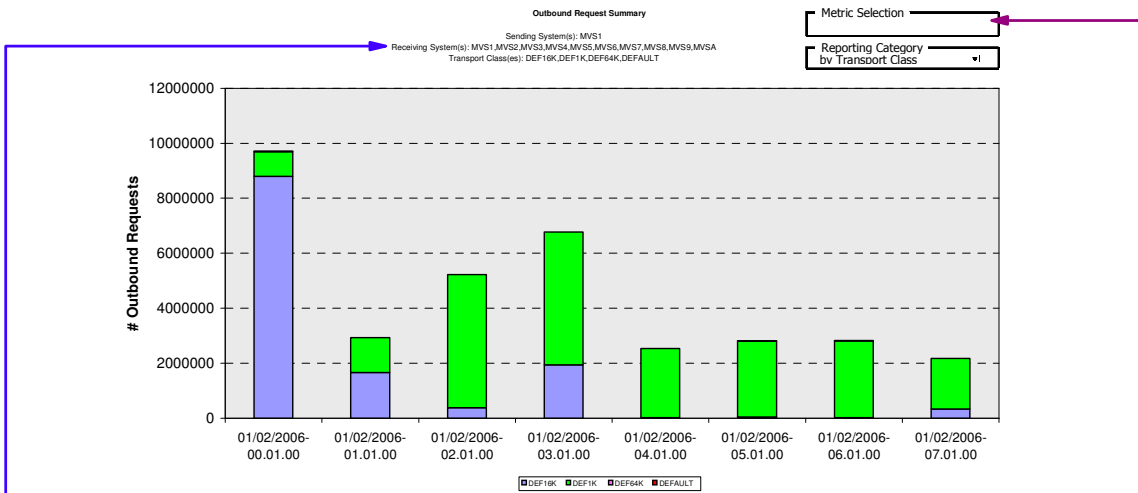
### ▪ I assume you are already familiar with the RMF Spreadsheet Reporter

- If not, you can download it from SYS1.SERBPWS(ERB9R2SW) or get the latest version from the RMF home page
  - <http://www.ibm.com/servers/eserver/zseries/zos/rmf/>
- Information about how to get the Spreadsheet Reporter up and running is available in the RMF Redbooks document, *Effective zSeries Performance Monitoring Using Resource Measurement Facility, SG24-6645*. But to (over)simplify, RMF SR:
  - Submits an RMF Postprocessor job to your MVS system
  - Downloads the resulting reports to workstation
  - Postprocesses the report into a format that can be loaded into a spreadsheet
  - Provides a load of Excel macros to turn these into graphical reports

### RMF SR XCF Macro initial screen - supported reports are listed on tabs across bottom of screen



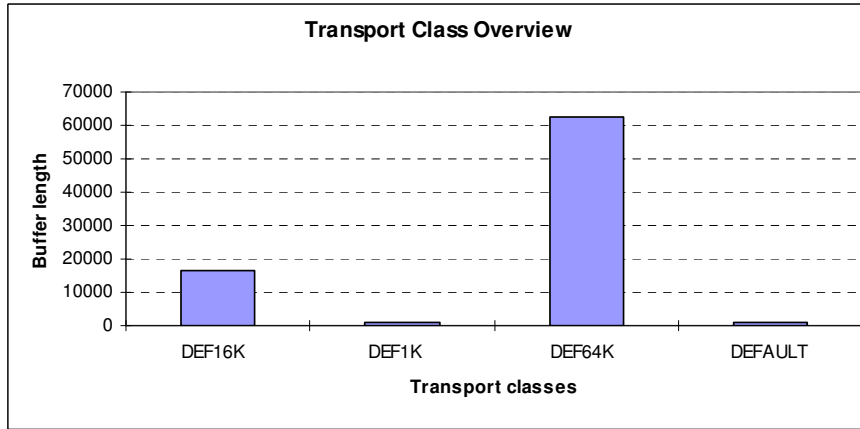
# Let's see which transport classes are the most heavily used



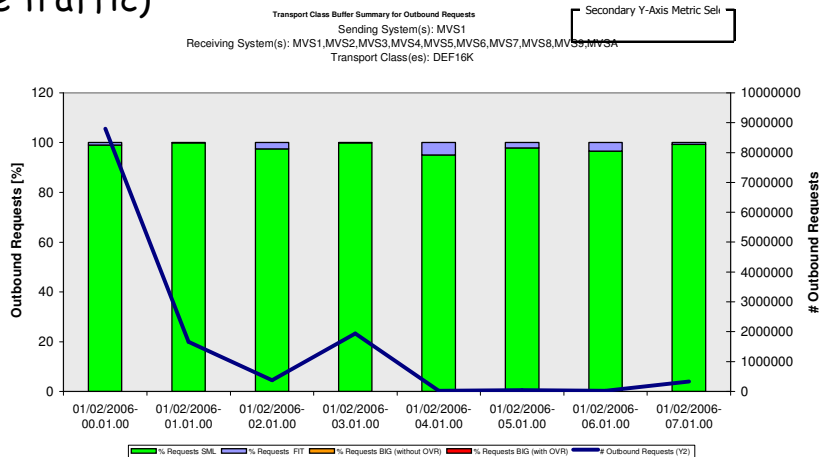
This worksheet shows the XCF signalling activity from 1 system (MVS1) to all the other members of the plex, for all transport classes, for every interval  
You can tailor the sending system, target system, and transport classes  
You can also report on any other field in the XCF Usage By System report

# Example of filtering options

If we want to check how efficient the transport class sizes are, we first need to know WHICH transport classes are defined and what message size their buffers are set up for



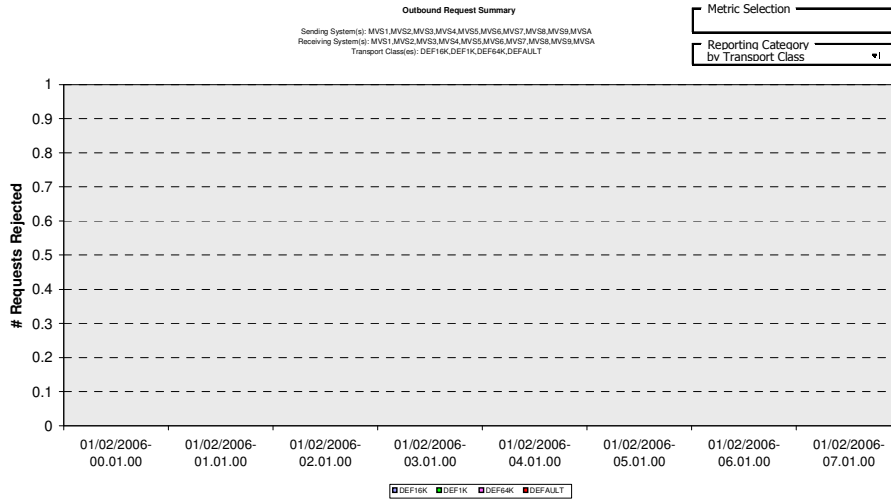
Display %SML, %FIT, and %BIG for each transport class, together with number of requests (so we can ignore ones with little traffic)



This report shows huge %SML for DEF16K transport class from MVS1 - should really check each system to see if pattern is the same

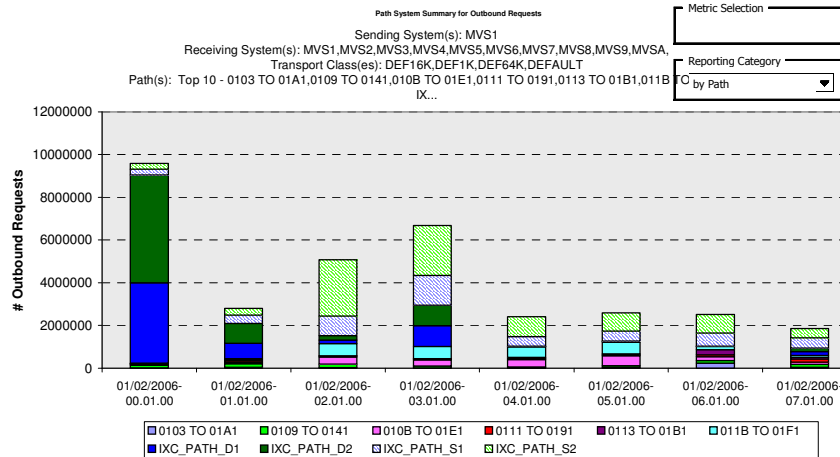
Next step would be to use D XCF,CD commands to break down signal sizes

## Also said that we want to check for Requests Rejected



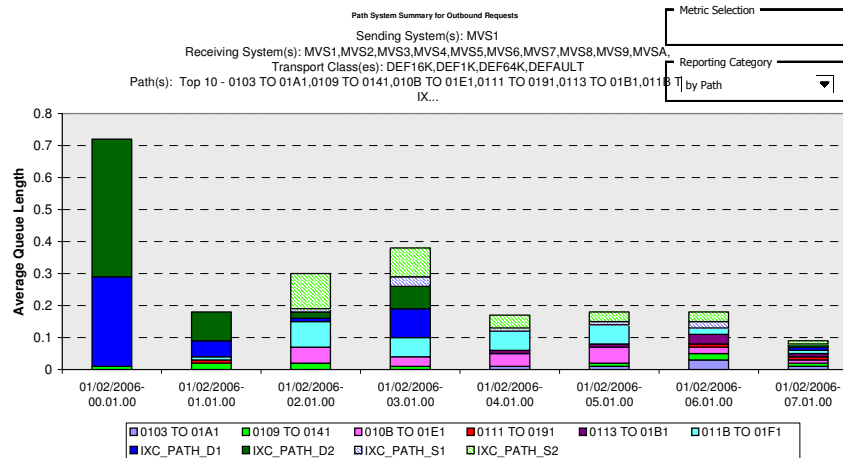
This report shows count of Req Rejects for transport class buffers (that is, on the sending side) for all systems for all transport classes. In this case, there were 0 Req Rejects - good!

## Next, we want to get the big picture on path usage, in preparation for looking at queue lengths



This report shows that the 4 busiest paths are the 4 structures - 2 for DEF1K, and 2 for DEF16K

Finally, we want to look at queue lengths, to see if more paths are required:



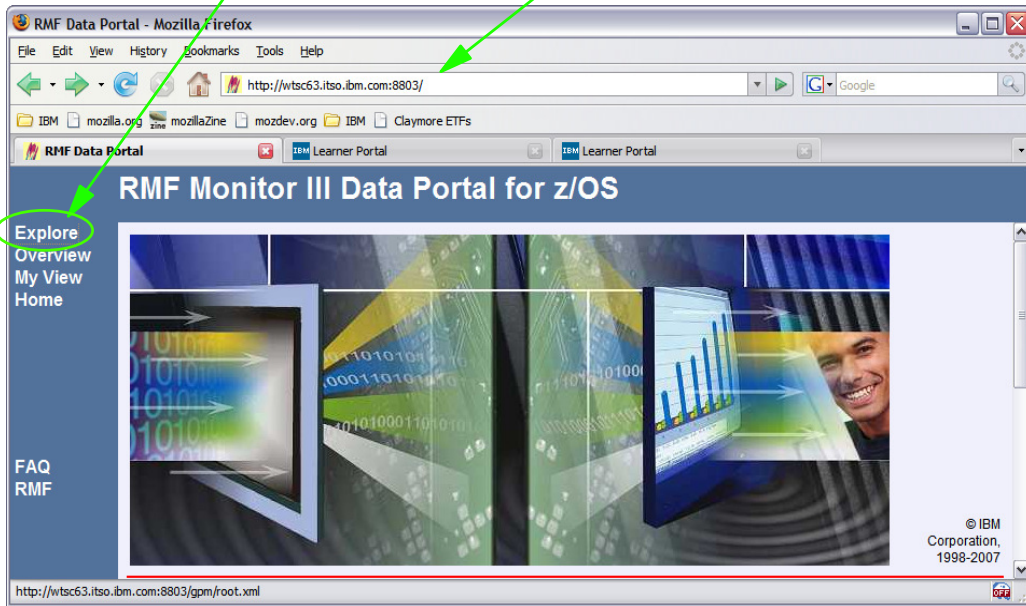
Here we see that the worst case queue length is about .4, for IXC\_PATH\_D2. While this should be monitored, it is not necessary to add any more paths now

## RMF Monitor III Data Portal

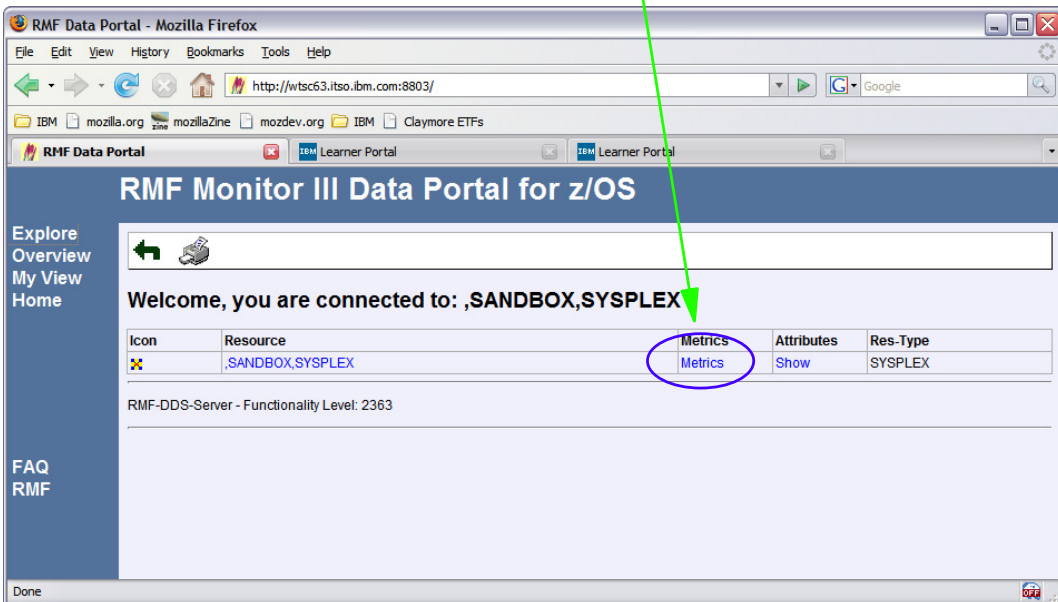
- Whereas the RMF Spreadsheet Reporter reports on historical data, and is based on the RMF Postprocessor, the RMF Monitor III Data Portal feeds off data in the Sysplex Data Server, and is realtime
- The Data Portal uses a standard Web Browser as the front end, and provides access to over 600 metrics from RMF Monitor III
  - It also displays information that is not available in the Monitor III ISPF panels
- z/OS 1.9 adds support for XCF information to the Data Portal

Click on "Explore"

Point Web Browser at port 8803 on your MVS system



Click on "Metrics"



Click on "XCFGROUP"

The screenshot shows the RMF Monitor III Data Portal for z/OS. A green arrow points to the 'XCFGROUP' link in the 'Full RMF Reports:' table. The table lists various report types, with 'XCFGROUP' highlighted under the 'CACHSUM' column.

CACHDET	CACHSUM	CFACT	CFOVER	CFSYS	SPACEG	SPACED
SYSSUM	XCFGROUP	XCFOWW	XCFPATH	XCFSYS		

Available metrics for: ,SANDBOX,SYSPLEX

Metric description	Help	Id
% delay	<a href="#">Explanation</a>	8D0160
% delay for enqueue	<a href="#">Explanation</a>	8D1A20
% delay for i/o	<a href="#">Explanation</a>	8D1A80
% delay for operator	<a href="#">Explanation</a>	8D1AE0
% delay for processor	<a href="#">Explanation</a>	8D1B40
% delay for storage	<a href="#">Explanation</a>	8D1BA0

Presented with information about all the XCF groups in the sysplex, including sysplex-level total (the \*ALL rows)

The screenshot shows the RMF Report [ ,SANDBOX,SYSPLEX ] : XCFGROUP (XCF Group Statistics). The report displays a table of XCF groups with columns for Group Name, Member Name, Status, Status Checking Interval, System Name, Job Name, Outbound Requests, Inbound Requests, and Line Type. The time range is 09/26/2007 19:03:00 - 09/26/2007 19:04:00.

Group Name	Member Name	Status (short)	Status	Status Checking Interval	System Name	Job Name	Outbound Requests	Inbound Requests	Line Type
ATRRRS	*ALL						0	0	G
ATRRRS	SC63	A	Active	0.00	SC63	RRS	0	0	M
ATRRRS	SC64	A	Active	0.00	SC64	RRS	0	0	M
ATRRRS	SC65	A	Active	0.00	SC65	RRS	0	0	M
ATRRRS	SC70	A	Active	0.00	SC70	RRS	0	0	M
COVLFNO	*ALL						0	0	G
COVLFNO	SC63	A	Active	0.00	SC63	VLF	0	0	M

Click on "Outbound Requests" to sort by number of outbound requests

RMF Report [,SANDBOX,SYSPLEX] : XCFGROUP (XCF Group Statistics)  
Time Range: 09/26/2007 19:03:00 - 09/26/2007 19:04:00

Group Name	Member Name	Status (short)	Status	Status Checking Interval	System Name	Job Name	Outbound Requests	Inbound Requests	Line Type
SYSGRS	*ALL						1008	1008	G
XCFJES2A	*ALL						726	724	G
SYSGRS	SC63	A	Active	0.00	SC63	GRS	276	276	M
ISTXCF	*ALL						256	252	G
SYSGRS	SC64	A	Active	0.00	SC64	GRS	244	244	M
SYSGRS	SC65	A	Active	0.00	SC65	GRS	244	244	M
SYSGRS	SC70	A	Active	0.00	SC70	GRS	244	244	M

Back to previous panel. Click on "XCFOVW"

Full RMF Reports:

CACHDET	CACHSUM	CFACT	CFOVER	CFSYS	SPACEG	SPACED
SYSSUM	XCFGROUP	XCFOVW	XCFPATH	XCFSYS		

Available metrics for: ,SANDBOX,SYSPLEX

Metric description	Help	Id
% delay	<a href="#">Explanation</a>	8D0160
% delay for enqueue	<a href="#">Explanation</a>	8D1A20
% delay for i/o	<a href="#">Explanation</a>	8D1A80
% delay for operator	<a href="#">Explanation</a>	8D1AE0
% delay for processor	<a href="#">Explanation</a>	8D1B40
% delay for storage	<a href="#">Explanation</a>	8D1BA0



## Provides summary information about the members of the sysplex

RMF Monitor III Data Portal for z/OS

RMF Report [ ,SANDBOX,SYSPLEX ] : XCF0VW (XCF Systems Overview)

Time Range: 09/26/2007 19:05:00 - 09/26/2007 19:06:00

System Name	SMF Id	Partition Name	System Level	Monitoring Interval	Operator Interval	Status	RMF Master
SC63	SC63	A04	SP7.0.9	8,500.00	8,800.00	Active	Yes
SC64	SC64	A06	SP7.0.9	8,500.00	8,800.00	Active	No
SC65	SC65	A11	SP7.0.9	8,500.00	8,800.00	Active	No
SC70	SC70	A16	SP7.0.9	8,500.00	8,800.00	Active	No

Back to previous panel. Click on "XCFPATH"

RMF Monitor III Data Portal for z/OS

Full RMF Reports:

<a href="#">CACHDET</a>	<a href="#">CACHSUM</a>	<a href="#">CFACT</a>	<a href="#">CFOVER</a>	<a href="#">CFSYS</a>	<a href="#">SPACEG</a>	<a href="#">SPACED</a>
<a href="#">SYSSUM</a>	<a href="#">XCFGROUP</a>	<a href="#">XCF0VW</a>	<a href="#">XCFPATH</a>	<a href="#">XCFSYS</a>		

Available metrics for: ,SANDBOX,SYSPLEX

Metric description	Help	Id
<a href="#">% delay</a>	<a href="#">Explanation</a>	8D0160
<a href="#">% delay for enqueue</a>	<a href="#">Explanation</a>	8D1A20
<a href="#">% delay for i/o</a>	<a href="#">Explanation</a>	8D1A80
<a href="#">% delay for operator</a>	<a href="#">Explanation</a>	8D1AE0
<a href="#">% delay for processor</a>	<a href="#">Explanation</a>	8D1B40
<a href="#">% delay for storage</a>	<a href="#">Explanation</a>	8D1BA0

Presented with information about every path,  
between every pair of systems in the sysplex

RMF Report [.,SANDBOX,SYSPLEX] : XCFPATH (XCF Path Statistics)  
Time Range: 09/26/2007 19:10:00 - 09/26/2007 19:11:00

Systems	Structure or CTC Devices	Path Type	Transport Class	Status	Status (short)	Retry %	Retry Limit	Message Limit	Signals Sent	Times Path Busy	Signals Pending	Storage in Use	Rest Cour
SC63:????	5C51	CTC	BIG	NotOperational	NO	0.0	10	4096	0	0	0	0	0
SC63:????	5C50	CTC	DEFAULT	NotOperational	NO	0.0	10	2000	0	0	0	0	0
SC63:????	5C59	CTC	BIG	NotOperational	NO	0.0	10	4096	0	0	0	0	0
SC63:????	5C58	CTC	DEFAULT	NotOperational	NO	0.0	10	2000	0	0	0	0	0
SC63:SC64	5C71:4C51	CTC	BIG	Working	WR	0.0	10	4096	8	0	0	330	0
SC63:SC64	IXC_DEFAULT_4(0008)	LST	BIG	Working	WR	0.0	10	4096	0	0	0	660	0

To find busiest paths, click on "Signals Sent"

Can move back or forward one interval

RMF Report [.,SANDBOX,SYSPLEX] : XCFPATH (XCF Path Statistics)  
Time Range: 09/26/2007 19:10:00 - 09/26/2007 19:11:00

Systems	Structure or CTC Devices	Path Type	Transport Class	Status	Status (short)	Retry %	Retry Limit	Message Limit	Signals Sent	Times Path Busy	Signals Pending	Storage in Use	Rest Cour
SC64:SC63	5C58:4C78	CTC	DEFAULT	Working	WR	0.0	10	2000	223	0	0	10	0
SC63:SC64	IXC_DEFAULT_2(0010)	LST	DEFAULT	Working	WR	0.0	10	2000	194	0	0	22	0
SC70:SC63	5C58:4D38	CTC	DEFAULT	Working	WR	0.0	10	2000	160	0	0	10	0
SC63:SC70	5D30:4C50	CTC	DEFAULT	Working	WR	0.0	10	2000	157	0	0	10	0
SC70:SC64	5C70:4D30	CTC	DEFAULT	Working	WR	0.0	10	2000	146	0	0	10	0
SC63:SC64	5C78:4C58	CTC	DEFAULT	Working	WR	0.0	10	2000	132	0	0	10	0

Back to previous panel. Click on "XCFSYS"

The screenshot shows the RMF Monitor III Data Portal for z/OS interface. On the left, there is a navigation menu with options: Explore, Overview, My View, Home, FAQ, and RMF. The main content area is titled "Full RMF Reports:" and contains a grid of report links. A green arrow points to the "XCFSYS" link in the grid.

CACHDET	CACHSUM	CFACT	CFOVER	<b>XCFSYS</b>	SPACEG	SPACED
SYSSUM	XCFGROUP	XCFOWW	XCFPATH	XCFSYS		

Below the reports grid, it says "Available metrics for: ,SANDBOX,SYSPLEX" and lists several metrics with descriptions, help links, and IDs.

Presented with information about transport class usage between every pair of systems in the sysplex

Can easily save all information to spreadsheet

The screenshot shows the RMF Report for XCF System Statistics. The report title is "RMF Report [,SANDBOX,SYSPLEX] : XCFSYS (XCF System Statistics)". The time range is "09/26/2007 19:12:00 - 09/26/2007 19:13:00". A yellow callout box points to a save icon in the top right of the report area.

Systems	Transport Class	Signals Sent	Signals Received	Times Path Unavailable	Times Buffer Unavailable	Buffer Length	Fit %	Smaller %	Larger %	Degraded %	System(1)	System(2)	Direction
SC63:SC63	BIG	8	0	0	0	62464	0.0	100.0	0.0	0.0	SC63	SC63	L
SC63:SC63	DEFAULT	168	0	0	0	956	100.0	0.0	0.0	0.0	SC63	SC63	L
SC63:SC64	*ALL	0	351	0	0	0	0.0	0.0	0.0	0.0	SC63	SC64	I
SC63:SC64	BIG	20	0	0	0	62464	0.0	100.0	0.0	0.0	SC63	SC64	O
SC63:SC64	DEFAULT	322	0	0	0	956	100.0	0.0	0.0	0.0	SC63	SC64	O
SC63:SC65	*ALL	0	107	0	0	0	0.0	0.0	0.0	0.0	SC63	SC65	I
SC63:SC65	BIG	19	0	0	0	62464	0.0	100.0	0.0	0.0	SC63	SC65	O

To get information about the meaning of a particular field, move the mouse over the column heading

RMF Monitor III Data Portal for z/OS

RMF Report [,SANDBOX,SYSPLEX] : XCF-SYS (XCF System Statistics)

Time Range: 09/26/2007 20:51:00 - 09/26/2007 20:52:00

Systems	Transport Class	Signals Sent	Signals Received	Fit %	Smaller %	Larger %	Degraded %	System(1)	System(2)	Direction
SC63:SC64	*ALL	0	409	0.0	0.0	0.0	0.0	SC63	SC64	I
SC64:SC63	*ALL	0	393	0.0	0.0	0.0	0.0	SC64	SC63	I
SC63:SC70	*ALL	0	286	0.0	0.0	0.0	0.0	SC63	SC70	I
SC70:SC63	*ALL	0	270	0.0	0.0	0.0	0.0	SC70	SC63	I
SC64:SC70	*ALL	0	256	0.0	0.0	0.0	0.0	SC64	SC70	I
SC70:SC64	*ALL	0	252	0.0	0.0	0.0	0.0	SC70	SC64	I
SC65:SC70	*ALL	0	219	0.0	0.0	0.0	0.0	SC65	SC70	I

XCF signals received

The total number of inbound signals.

This metric is available for the following XCF entities:

- ◆ Group
- ◆ Member
- ◆ Path
- ◆ System

## Enhanced RMF XCF supports

### ■ Summary

- Vast improvement over old system-by-system XCF Postprocessor reports
- Spreadsheet XCF report currently doesn't support Usage by Member part of PP report, but information about paths and transport classes is still very helpful
- Building the SR XCF reports can take a while, especially for a large sysplex, as there is a lot of data to process. But you only need to do this a few times a year, so take the opportunity to grab a coffee!
- The Data Portal is very easy to set up - start GPM SERVE and point Web browser at your system!
  - Data Portal contains loads of other information, including CF information, just used XCF as an example

## Value summary

### ▪ Customer value:

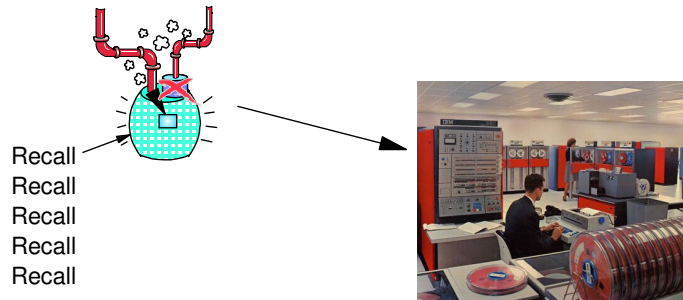
- Delivers improved performance and availability by making it much easier to detect XCF-related performance issues

### ▪ Ease of implementation:

- 7 out of 10
  - Installation of the RMF Spreadsheet Reporter is simple (just download to PC), but setup takes a little getting used-to. But once you are conversant with using it, the XCF support has the same look and feel as all other SR reports.
  - RMF Data Portal is even easier to set up.

## System Logger Migrated data set recall enhancements

- System Logger users sometimes encounter delays in log stream offload processing because of Logger single threaded, synchronous handling of recall requests for migrated log stream data sets:



## System Logger Migrated data set recall enhancements

- This design limitation manifests itself in two ways:
  - If the HSM ML1 or ML2 volume or data set is unavailable for some reason, no recalls will move until that request is somehow cancelled:
    - Problem in one log stream could impact other log streams that don't actually have any recall-related problems, except that they arrived after the request that DOES have a recall problem
    - The delay could result in the log stream filling, causing problems for users of that log stream (CICS, for example)
  - Recalls could simply be processed very slowly because HSM is very busy, or there is contention on the ML2 volumes between different HSM instances
    - This situation could potentially be eased by implementing HSM Common Recall Queue to let all the HSMs in the SMSplex process all RECALL requests from any system, spreading out the workload and providing improved turnaround times for recalls.

## System Logger Migrated data set recall enhancements

- A partial circumvention for this situation was to specify **OFFLOADRECALL(NO)** in the log stream definition:
  - Results in Logger allocating a new offload data set for that log stream if the data set it is trying to *mod on to* is migrated
    - However this can result in using up limit of 168 offload data sets per log stream much faster than expected
  - Also doesn't provide any relief if Logger is trying to *read* from the migrated data set.

## System Logger Migrated data set recall enhancements

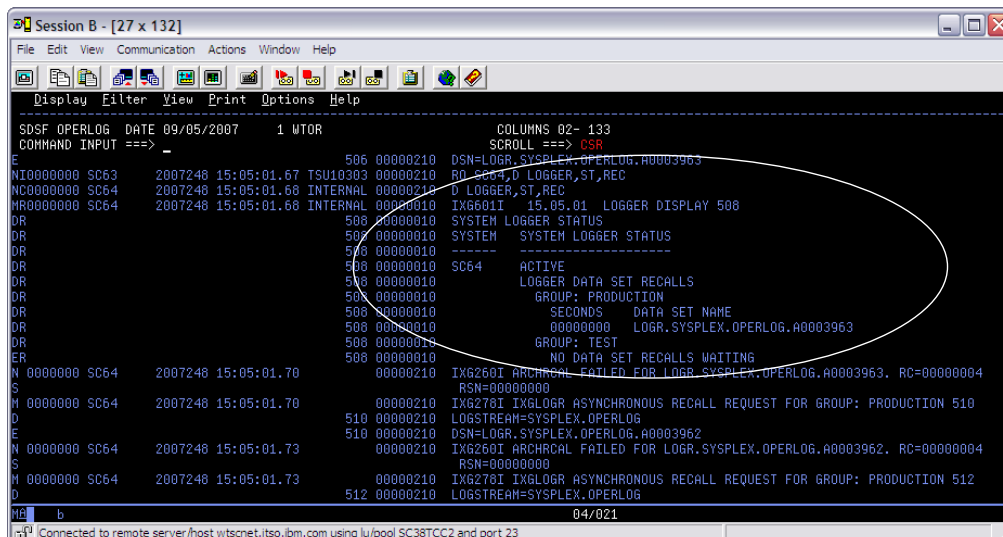
- **z/OS 1.5 delivered new messages to help you identify the offending request and potentially cancel it:**
  - IXG271I Logger Data Set Request Delayed during the last xx seconds for logstream
  - IXG272E LOGGER TASK DELAYED, REPLY "MONITOR", "IGNORE", "FAIL", "EXIT".
  - Doesn't make problem goes away, but assists in more timely identification of cause of the delay
- **z/OS 1.8 addressed this to some extent by adding a second System Logger recall task - one for PRODUCTION log streams, and the second for TEST log streams.**

## System Logger Migrated data set recall enhancements

- **z/OS 1.9 resolves this problem by dramatically increasing the number of concurrent recall requests per System Logger instance:**
  - Up to 24 concurrent requests per System Logger instance for PRODUCTION log streams, plus up to 8 concurrent requests for TEST log streams.
- **Also provides new support to display information about pending recall requests and cancel selected ones:**
  - IXG271/IXG272 only support cancellation of the *oldest* request
    - But you may wish to cancel more than one, without having to wait for Logger to detect that there is a problem with another recall and issue the next IXG271/IXG272 message

## System Logger Migrated data set recall enhancements

- **New Display command (D LOGGER, ST, REC) lets you display information about in-flight recall requests:**



```

Session B - [27 x 132]
File Edit View Communication Actions Window Help
-----
Display Filter View Print Options Help
-----
SDSF OPERLOG DATE 09/05/2007 1 MTOR          COLUMNS 02- 133
COMMAND INPUT ==> -                          SCROLL ==> CSR
E
NI00000000 SC63 2007248 15:05:01.67 TSU10303 00000210 DSN=LOGR.SYSPLEX.OPERLOG.A0003963
NC00000000 SC64 2007248 15:05:01.68 INTERNAL 00000210 RC=0004,D LOGGER,ST,REC
NR00000000 SC64 2007248 15:05:01.68 INTERNAL 00000210 D LOGGER,ST,REC
DR 508 00000010 IXG601I 15.05.01 LOGGER DISPLAY 508
DR 508 00000010 SYSTEM LOGGER STATUS
DR 508 00000010 SYSTEM SYSTEM LOGGER STATUS
DR 508 00000010 SC64 ACTIVE
DR 508 00000010 LOGGER DATA SET RECALLS
DR 508 00000010 GROUP: PRODUCTION
DR 508 00000010 SECONDS DATA SET NAME
DR 508 00000010 00000000 LOGR.SYSPLEX.OPERLOG.A0003963
DR 508 00000010 GROUP: TEST
DR 508 00000010 NO DATA SET RECALLS WAITING
ER 00000000 SC64 2007248 15:05:01.70 00000210 IXG260I ARCHRCAL FAILED FOR LOGR.SYSPLEX.OPERLOG.A0003963, RC=00000004
S 00000000 SC64 2007248 15:05:01.70 00000210 RSN=00000000
N 00000000 SC64 2007248 15:05:01.70 00000210 IXG278I IXGLOGR ASYNCHRONOUS RECALL REQUEST FOR GROUP: PRODUCTION 510
D 510 00000210 LOGSTREAM=SYSPLEX.OPERLOG
E 510 00000210 DSN=LOGR.SYSPLEX.OPERLOG.A0003962
N 00000000 SC64 2007248 15:05:01.73 00000210 IXG260I ARCHRCAL FAILED FOR LOGR.SYSPLEX.OPERLOG.A0003962, RC=00000004
S 00000000 SC64 2007248 15:05:01.73 00000210 RSN=00000000
N 00000000 SC64 2007248 15:05:01.73 00000210 IXG278I IXGLOGR ASYNCHRONOUS RECALL REQUEST FOR GROUP: PRODUCTION 512
D 512 00000210 LOGSTREAM=SYSPLEX.OPERLOG
-----
04/021
Connected to remote server/host wtsnet.itso.ibm.com using lu/pool SC38TCC2 and port 23
  
```



## System Logger Migrated data set recall enhancements

- Having used the Display LOGGER command to identify the offending offload data set, use the new SETLOGR FORCE,NORECALL,DSN=xxxx command to cancel the recall request:

```
setlogr force,norec,dsn=LOGR.SYSplex.OPERLOG.A0003963
IXG601I SETLOGR FORCE NORECALL COMMAND ACCEPTED
DSNAME=LOGR.SYSplex.OPERLOG.A0003963
IXG280I LOGR RECALL REQUEST STOPPED BY SETLOGR COMMAND
DSN=LOGR.SYSplex.OPERLOG.A0003963
IXG661I SETLOGR FORCE NORECALL PROCESSED SUCCESSFULLY
DSNAME=LOGR.SYSplex.OPERLOG.A0003963
```

## System Logger Migrated data set recall enhancements

- New commands (Display and SETLOGR) only work on z/OS 1.9 or later systems and only display information about the system they are issued on:
  - No toleration service required on downlevel systems
  - Downlevel systems reject the commands if issued on that system
- However you don't have to wait for all systems to move to z/OS 1.9 before you can use it - as soon as each system moves to 1.9, you can use these commands

## Value summary

### ▪ Customer value:

- Can improve availability, especially for CICS and IMS (less likely for log stream to fill because offload is delayed).
- Can provide improved performance for any task trying to browse a log stream with migrated offload data sets.

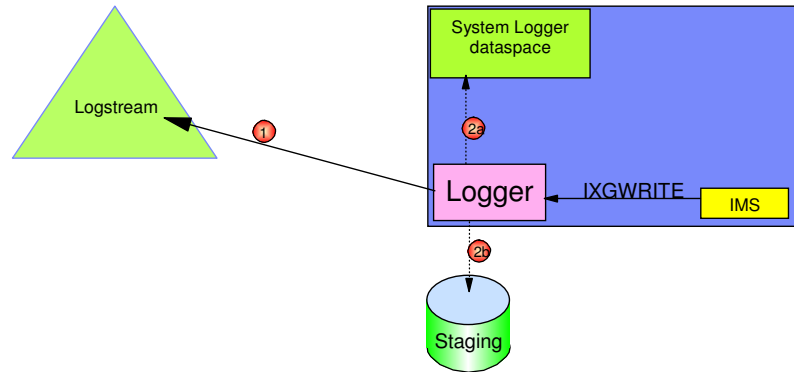
### ▪ Ease of implementation:

- 10 out of 10
  - Nothing to do - once you move to z/OS 1.9, function is automatically active
  - May want to update operator procedures for handling Logger recall problems

## Performance considerations for use of System Logger staging data sets and remote copy

## System Logger Staging data sets and remote copy

- When using a (non-System-Managed Duplexed) CF log stream, you have 2 choices about the location of the second copy of the data that is currently residing in the CF: <sup>1</sup>
  - 1) Place it in a data space in z/OS storage, or <sup>2a</sup>
  - 2) Place it in a Staging Data Set on DASD <sup>2b</sup>



## System Logger Staging data sets and remote copy

- The first option (data space) provides the best performance, but you also need to consider recoverability. Data space and CF should have no single point of failure (consideration for ICFs) in order to ensure log data is not lost.
- However, if you are remote copying your DASD, you will probably want to use the second option (staging data sets), so the log data corresponding to the remote-copied data sets is also available in the remote site:
  - The write to the Staging Data Set happens after the CF write but delays the response to the IXGWRITE (write to log stream) request, so the longer the DASD response time, the larger the impact on the user of the log stream.

## System Logger Staging data sets and remote copy

- As a result of the response time impact on IXGWRITE requests, IBM used to recommend that System Logger users with high performance requirements (IMS Shared Message Queue, for example) should avoid the use of Staging Data Sets
  - Old IMS Redbooks document, "IMS Version 7 Performance Monitoring and Tuning Update", SG24-6404 (2002), recommends not to use Staging Data Sets with IMS SMQ
- However, both DASD technology and remote copy technology have improved a lot since that book was written, so is that recommendation still valid??

## System Logger ITSO measurement

- The objective of the measurement was to see what impact the use of Staging data sets, combined with PPRC over 20km, would have on transaction response times when using IMS Shared Message Queue, and if the old recommendation still stands
- But first, let's take a step back to understand the interaction between Shared Message Queue, System Logger, and transaction response times....

## System Logger ITSO measurement

- **IMS provides two fundamental exploitations of Parallel Sysplex:**
  - Database sharing, where multiple database manager instances can update the same set of databases
    - The implementation of this introduces additional processing ("overhead") for every request to a shared database
      - The more database calls the transaction issues, the higher the additional cost
  - Shared Message Queue, where incoming transactions are placed on a shared queue structure in the CF, and the outgoing responses are again placed on the shared queue
    - This results in some time being added before the transaction is selected for processing, and some more time after it completes. Therefore the impact is relatively fixed, and is not at all related to the current elapsed time or CPU- or database-intensiveness of the transaction

## System Logger ITSO measurement

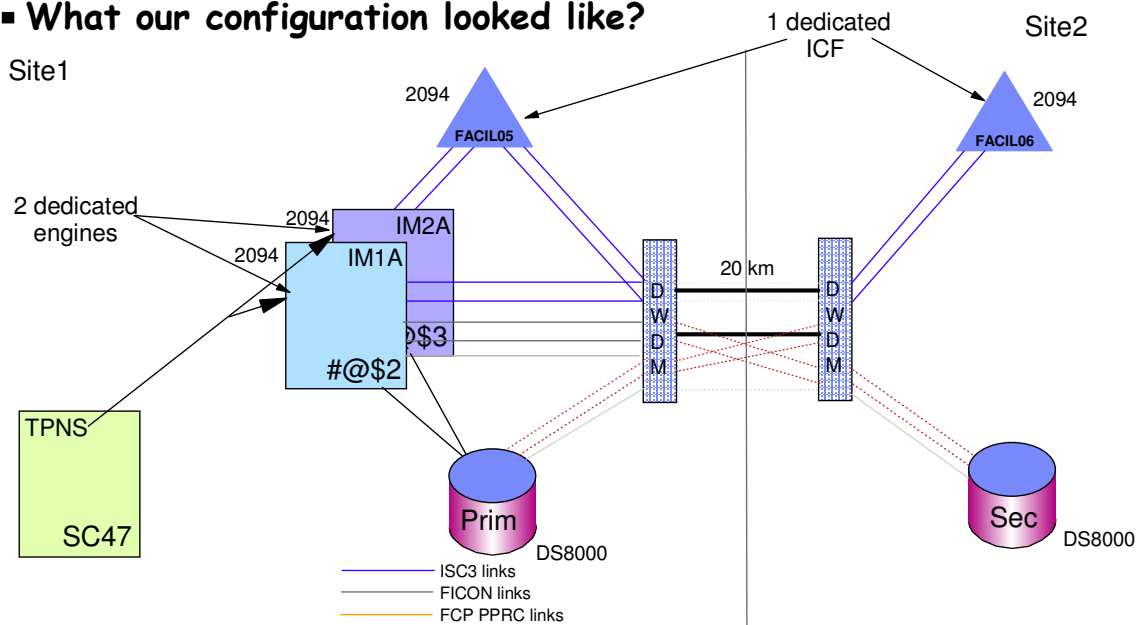
- **So what does System Logger have to do with all this?**
  - Every time a message is placed on, or retrieved from, the shared queue, or a response is placed on or retrieved from the shared queue, IMS writes a log record to System Logger:
    - This permits IMS to recover from a failure affecting the shared queue structure
  - The number of log records created for each IMS transaction is fairly stable - just over 4 log records per transaction on average.
  - Note that the only use IMS makes of System Logger is in relation to shared message queues - IMS continues to do his own traditional logging for all his other processing, so this discussion about System Logger and IMS only applies to IMS customers that are using SMQ. However, the general discussion about the impact of staging data sets and PPRC applies to all users of Logger staging data sets.

## System Logger Staging data sets and remote copy

- In response to a customer request, ITSO ran a measurement comparing various IMS SMQ configurations:
  - CF-only log stream, with no Staging Data Set, to set the baseline best-case measurement (not a valid configuration for DR purposes)
  - CF log stream with Staging Data Sets PPRCed at 20km and various numbers of stripes (1, 2, 4) for the Staging Data Sets
  - CF log stream with simplex Staging Data Sets (no PPRC) and 4 stripes
  - System-Managed Duplexed CF log stream and no Staging Data Sets
    - Log data duplexed between 2 CFs, but not to DASD
    - Only run for comparison
    - Not a valid DR config
    - Used CF-Level 14 level of System Managed Duplexing

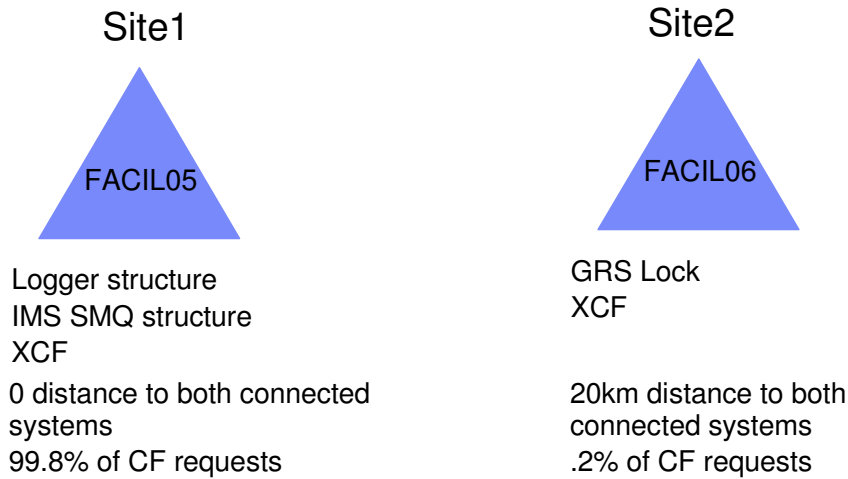
## System Logger Staging data sets and remote copy

### What our configuration looked like?



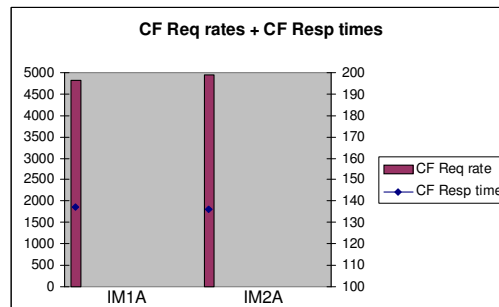
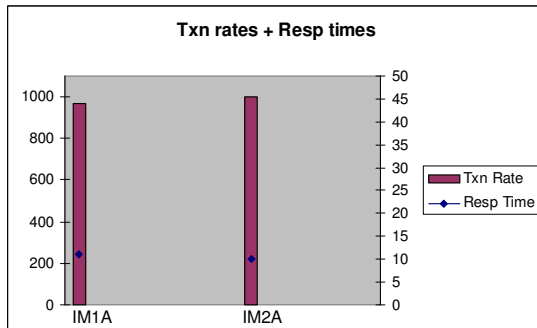
## CF Contents

### ▪ A little more detail on CF contents



## System Logger Staging data sets and remote copy

### ▪ Base measurement - CF-only log streams



## System Logger ITSO measurement

### ■ Observations:

- The point of this measurement was to identify the best possible performance that could be achieved from the configuration:
  - Note that this is *not* a valid configuration for a data mirroring environment as the log data would not be mirrored until it reaches the offload data sets
  - Even in a single-site configuration, the use of data spaces for the second copy of interim data should only be considered if the CF is failure-isolated from all connected systems AND the CF is non-volatile.
- Txn rate nearly 1000 per second per IMS subsystem
  - Only constraint was in how IMS was set up - neither CPU nor CF saturated, DASD performance not relevant in this test

## System Logger ITSO measurement

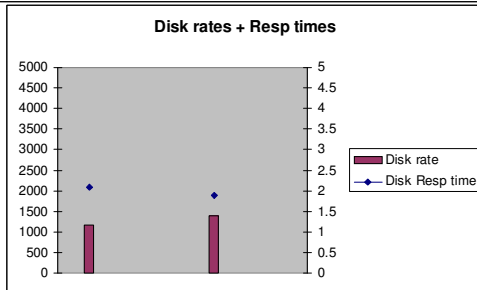
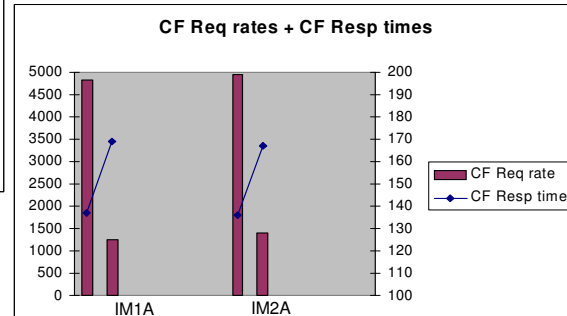
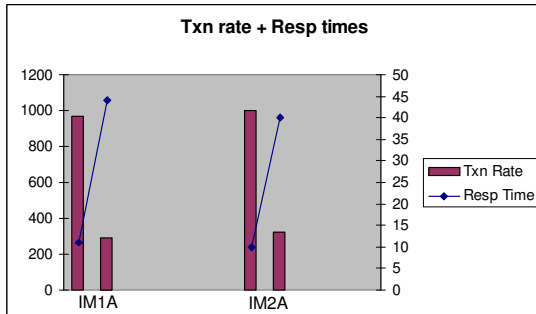
### ■ Observations:

- High CF response times
  - Need to understand how Logger uses these types of log streams:
    - Lots of "small" writes (typically about 2KB)
    - Long reads (64KB each) every time there is an offload (and we were constantly offloading)
    - Long delete commands, to remove log blocks from the structure after they have been successfully moved to the offload data sets
  - The relatively large number of long-running commands, coupled with a single-engine CF AND ISC links (relatively low bandwidth) resulted in high average CF response times



## System Logger Staging data sets and remote copy

### ■ First staging measurement - CF + Staging + PPRC @ 20 km



## System Logger Staging data sets and remote copy

### ■ Observations:

- Look at that DASD I/O rate! When the referenced IMS Redbook was written, best I/O rate to a SIMPLEX staging data set was about 400-450 I/Os per second, equivalent to about 100 txns/sec. Now we are achieving nearly 1200 I/Os per second (300 txns/sec)- to a volume that is PPRCed over 20km. This has a fundamental impact on the old recommendation not to use staging data sets with IMQ SMQ.

# System Logger Staging data sets and remote copy

## Observations:

- IMS transaction rate dropped due to each transaction taking longer to complete
  - IMS was set up to handle a given number of concurrent transactions, so as each txn took longer to complete, IMS was able to process fewer txns per second.
  - Also, remember that just about the only thing these transactions were doing was related to Logger. All txns were hanging around waiting to talk to Logger - this does NOT represent a normal transaction profile. Normally transactions would be using CPU and doing DB I/O, so the relative impact of slower IXGWrites would be a lot smaller

Rate of requests to CF dropped in line with decreased number of transactions. So why did CF response time INcrease?

STRUCTURE NAME = IM0A_LOGM TYPE = LIST STATUS = ACTIVE													
SYSTEM NAME	# REQ	AVG/SEC	REQUESTS				REASON	# REQ	DELAYED REQUESTS				
			# REQ	% OF ALL	-SERV TIME (MIC)- AVG	STD_DEV			# REQ	% OF REQ	--- /DEL	AVG TIME (MIC)	STD_DEV
TOTAL	5862K	9770	33K	0.6	114.8	175.3	NO SCH	1	0.0	21.0	0.0	0.0	
			ASYNC	5829K	99.4	136.2	191.4	PR WT	0	0.0	0.0	0.0	0.0
			CHNGD	1	0.0			PR CMP	0	0.0	0.0	0.0	0.0
							DUMP	0	0.0	0.0	0.0	0.0	

Bulk of requests were async due to sync/async heuristic algorithm, so overall "average" response time really is avg async time

CF-only

Sync time was nearly identical, Async time (which includes time waiting for XCF to get dispatched again) was one that increased

CF + Staging

STRUCTURE NAME = IM0A_LOGM TYPE = LIST STATUS = ACTIVE													
SYSTEM NAME	# REQ	AVG/SEC	REQUESTS				REASON	# REQ	DELAYED REQUESTS				
			# REQ	% OF ALL	-SERV TIME (MIC)- AVG	STD_DEV			# REQ	% OF REQ	--- /DEL	AVG TIME (MIC)	STD_DEV
TOTAL	1583K	2638	7150	0.5	115.0	188.4	NO SCH	0	0.0	0.0	0.0	0.0	
			ASYNC	1576K	100	168.3	175.9	PR WT	0	0.0	0.0	0.0	0.0
			CHNGD	0	0.0			PR CMP	0	0.0	0.0	0.0	0.0
							DUMP	0	0.0	0.0	0.0	0.0	

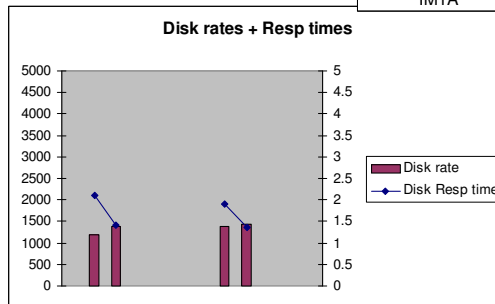
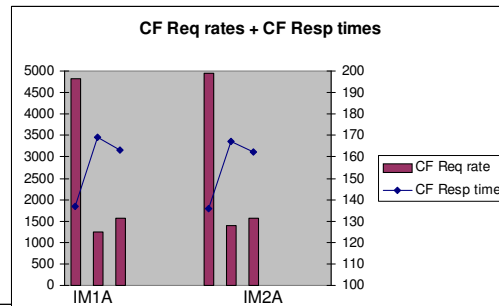
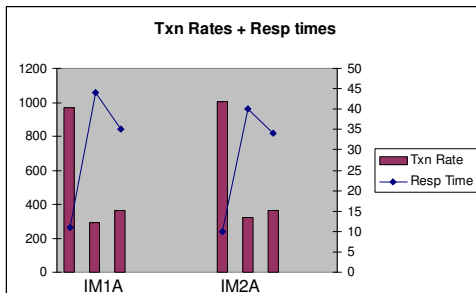
## System Logger ITSO measurement

### ■ Observations:

- With no changes other than the introduction of PPRCed staging data sets, txn resp time increased by about 30 milliseconds (from 10 ms.):
  - Average of 4 staging data set I/Os per transaction, at 2 ms per I/O = increase of 8 ms per txn purely due to waiting for disk I/Os to complete
  - Remainder of the increase was due to increased queueing within IMS.
    - The arrival rate of transactions into IMS remained constant, so as throughput decreased, the time spent waiting in IMS increased
- Disk response times
  - Because the staging disk I/Os are writes, they all included the cost of PPRCing over 20km (about .6 ms per I/O - .4 for PPRC + .1 per 10km.)
  - I/Os to the offload data sets happen after the transaction has ended and have no impact on transaction response time
    - Data for the offload is read from the CF, not from staging data sets

## System Logger Staging data sets and remote copy

### ■ Measurement - CF + Staging + PPRC 20 km + 2 stripes



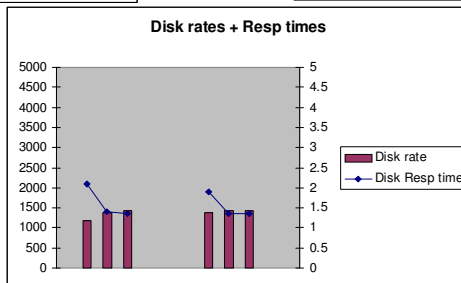
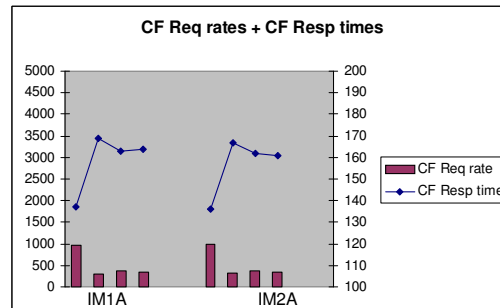
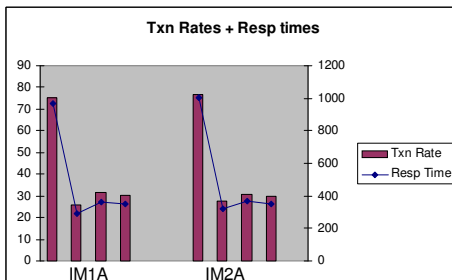
# System Logger ITSO measurement

## Observations:

- Increasing the number of stripes in the staging data sets to 2 resulted in improved disk response time:
  - Resp time dropped from 2.1 ms to average of 1.4 ms
  - This means that each IXGWRITE request took less time - 4 x .7 ms per txn
  - Meaning that the logging of each IMS message (into and out of IMS) took less time:
    - Combined with resulting reduction in queue time within IMS, this resulted in reduction of 6-9 ms per txn
- Net - Moving from 1 to 2 stripes showed a measureable benefit in IMS txn resp times and throughput
  - For more information on striping, see "VSAM Demystified", SG24-6105

# System Logger Staging data sets and remote copy

## Next measurement - CF + Staging + PPRC 20 km + 4 stripes



## System Logger ITSO measurement

### Observations:

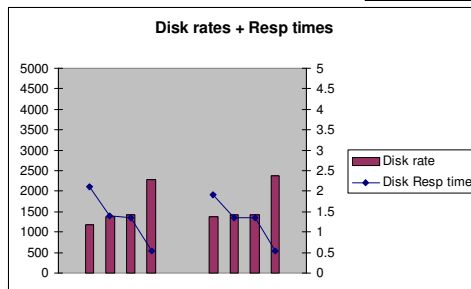
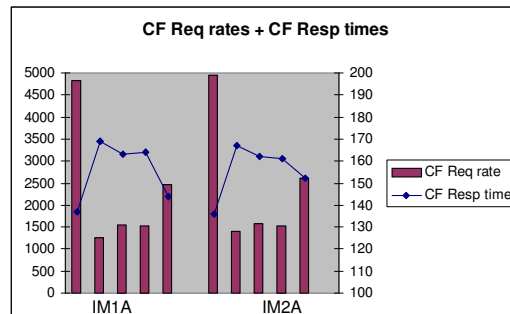
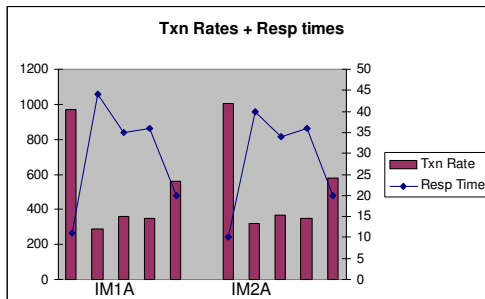
- Increasing the number of stripes in the staging data sets from 2 to 4 resulted in no significant change:
  - Change in txn resp time and transaction rate was within margin of error
- All metrics basically unchanged

### Net - Moving from 2 to 4 stripes for the IMS log streams showed no additional benefit

- And remember that each additional stripe requires one additional volume PER SYSPLEX MEMBER, so no point in having more stripes than you get benefit from
- Other Logger users were not measured - it is possible that other users may get more benefit from >2 stripes than this workload did
  - Depends on size of log blocks (greater than 4KB?), and number of concurrent IXGWrites to the log stream

## System Logger Staging data sets and remote copy

### Measurement - CF + Staging + NO PPRC + 4 stripes



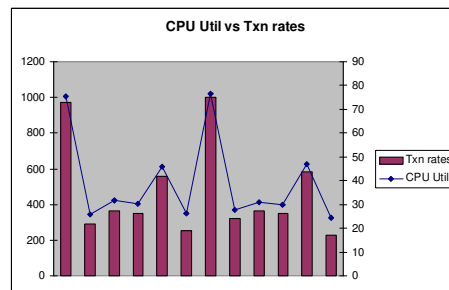
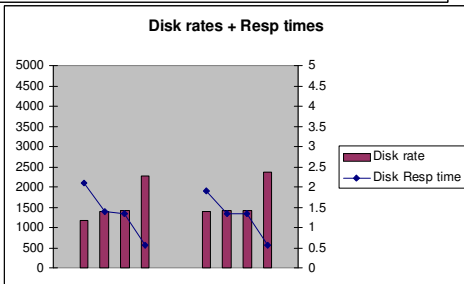
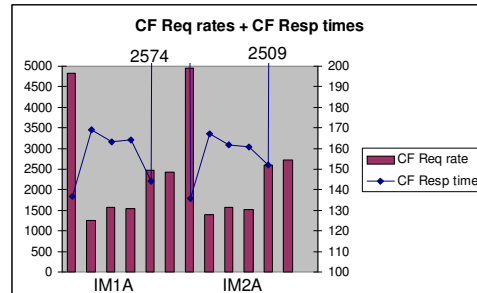
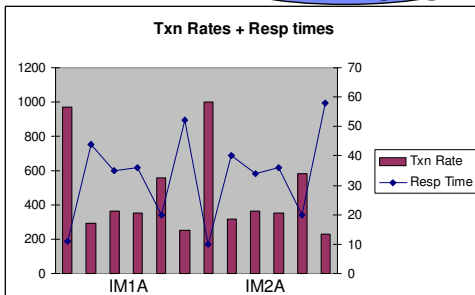
# System Logger ITSO measurement

## Observations:

- As expected, turning PPRC off resulted in improved disk response times (down from 1.35 to .55 ms in this case).
- Note the disk I/O rate - now nearly 2400 I/O per second across 4 stripes (up from a little over 1400/sec when using PPRC). This corresponds to an IMS transaction rate of nearly 600 txns/sec per subsystem - up from just 100/sec when the book was written back in 2002
- Improved disk times, combined with attendant reduction in queueing within IMS resulted in 16 ms reduction in response times, back down to just 20 ms. We are now seeing an impact of just 10ms per txn for using staging data sets compared to using CF-only log stream.

# System Logger Staging data sets and remote copy

## Last measurement - CF + System-Managed Duplexing of Logger Structure + NO Staging



## System Logger Staging data sets and remote copy

### ▪ Observations:

- This run was only done for comparison purposes. **System-Managed Duplexing is NOT the CF equivalent of PPRC. In case of a connectivity failure (and without GDPS CF Hint support) there is no way to know which instance of a duplexed structure will be kept.**
- The type of CF requests that Logger issues (multiple-entry requests) result in many CF-to-CF signals per request. Even when Logger writes a single block, that resulted in at least 4 CF-to-CF exchanges, compared to 1 site-to-site exchange for PPRC.
  - As the distance between the CFs increases, the higher number of exchanges wipes out any inherent performance benefit for the CF.
  - The result is higher response times for the CF structure than the staging data sets. This means longer IMS txn response times (up by 8-18 ms per txn) compared to single stripe PPRC case.

## System Logger Staging data sets and remote copy

### ▪ Summary of findings:

- As expected, use of staging data sets elongates response time for transactions using shared message queue.
- However, the worst case increase when using staging data sets and PPRC in this configuration was about 30 ms. This increase would be roughly the same regardless of whether the previous IMS Txn response time was 10 ms (as in our case) or 300 ms (a more realistic scenario). By striping the staging data sets over 2 volumes each, increase was reduced to about 24 ms.
- In an all-ICF environment, where you have a choice between simplex staging data sets or duplexing the logstream structure, it would probably be better to use staging data sets. In our measurement, the impact of turning on simplex staging data sets was only 10 ms. per transaction

## System Logger Staging data sets and remote copy

### ▪ Summary:

- The net is that the old recommendation to never use staging data sets with IMS Shared Message Queue has changed to say that this *may* be a reasonable option, depending on a number of things:
  - DASD technology. We used DS8K. Expect that the impact would have been significantly higher with older technology
  - The distance between the sites - obviously the higher the distance, the bigger the impact (and don't forget about the knock-on queuing within IMS that results from the longer response times).
  - Your current SMQ logging rates for each IMS subsystem (remember that staging data sets are one per z/OS image, whereas the Logger structure is shared between all connected IMS subsystems)
  - Can your applications withstand the response time impact of turning on staging, and the resulting possible impact to throughput?

## Value summary

### ▪ Customer value:

- These results may enable the combination of IMS Shared Message Queue, DASD mirroring, and full recovery that was not feasible previously

### ▪ Ease of implementation:

- ? out of 10
  - Implementing the use of staging data sets for the SMQ log stream is a trivial task. Determining the impact of the resulting response time increase on your applications is more complex.



## Important recent Logger APARs

- **OA20055** - Address log stream damage if log stream is rebuilt while structure is transitioning from duplex to simplex
- **OA20281** - Expand meaning of OFFLOADRECALL(NO). If this is set, Logger will allocate a new offload data set if:
  - The data set he wants to MOD on to is migrated, OR
  - The data set he wants to MOD on to already has an ENQ against it
- There is also important new text about the use of LOGR CDS in R9 level of Setting Up a Sysplex - see section entitled "Logger couple data set use considerations" - applies to ALL levels of z/OS, not just R9....
- Also recommend shutting down all Logger applications prior to IPL if possible, especially RRS

## SMF use of System Logger

- Limitations and drawbacks of current approach
- How SMF uses System Logger
- Comparison of SMF post-processing before and after 1.9
- Deciding whether you want to move to the new paradigm
- Implementation steps

## SMF Challenges

- **Performance challenges of current SMF mechanisms:**
  - The volume of SMF records that can be saved is gated by the performance of the DASD containing the SYS1.MANx data sets:
    - SMF doesn't support extended format data sets, so no striping support
    - Only one active SYS1.MANx data set per system
    - If SMF data is consistently created faster than it can be saved to DASD, you will eventually start losing it (when the buffers fill)
      - Many installations have to give up collecting useful data in order to ensure vital SMF data isn't discarded
  - The amount of buffer space available for SMF records is gated by the fact that SMF buffers are stored in the SMF Address Space above-the-line private, limiting you to 1GB in z/OS 1.8
  - There is only a single SRB, so only 1 I/O can be driven at a time, AND that represents a single point of failure

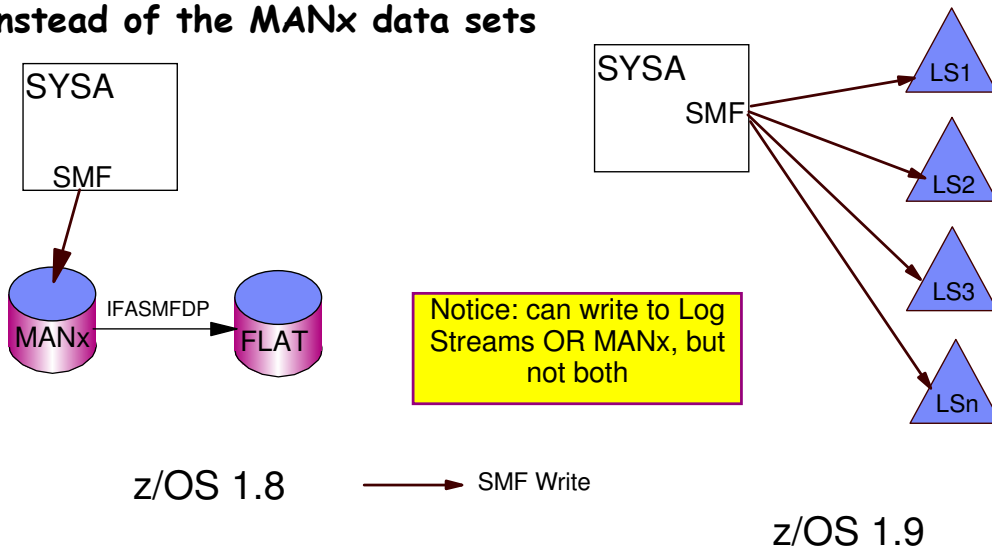
## SMF Challenges

### ▪ Data Management challenges:

- IFASMFDP only supports a single output file, so the same SMF data ends up being processed over and over to create various GDGs
- The way SMF data is collected (every record type, for just one system, all in the same data set) does not reflect how the data is used by the installation, increasing the processing that must be done to get the records grouped as you would like
- "Losing" SMF data is not uncommon, due to the number of data sets, GDG generations, and days/weeks/months to be managed

## SMF enhancement in z/OS 1.9

- To address these constraints, z/OS 1.9 introduces the ability to optionally write SMF data to System Logger log streams instead of the MANx data sets



## How does this help me?

### ▪ Performance:

- Most importantly, SMF can write to **MANY** log streams concurrently, eliminating the bottleneck of only being able to write to **ONE** data set
- A side effect of the use of log streams is that instead of writing data to "slow" DASD, SMF can write to a "fast" Coupling Facility
- Rather than having just a single 1GB buffer, when using log streams you have a 2GB data space *per log stream*
- Rather than just 1 SRB to write all records, there is 1 write task *per log stream*

## How does this help me?

### ▪ Flexibility in how the SMF data is *created*:

- Because you can handle the generation of more data, you can collect all the data you *want*, instead of only the data that you can safely handle
- Instead of **ALL** SMF data being grouped by system (in the MANx data sets), you can now group it by use (sysplex-wide) in a log stream :
  - You can use DASDONLY log streams. Don't get the merging or performance benefits of CF, but you can have many of them, each with its own 2GB buffer, so still much faster than MANx
  - You can use CF log stream(s) dedicated to a given system. Performance and buffer benefits, but not the merging benefits
  - You can have SMF subplexes. So a subset of systems write to one set of log streams, and another subset writes to another
  - You can have sysplex-wide, function-specific, CF log streams
  - You can have any mix of the above

## How does this help me?

- **Flexibility in how/where the data is kept until it expires:**
  - If you wish, you have the possibility to process data (semi-)directly out of the log stream, rather than having to create and manage GDGs
    - Must use IFASMFDP to retrieve data from the log stream
  - If you want to continue the existing model with GDGs, the new IFASMFDP can create multiple output files on a single pass of the SMF log stream(s), eliminating a lot of the inefficiencies of the current process with IFASMFDP

## SMF Archiving process considerations

- **What your process looks like now...**

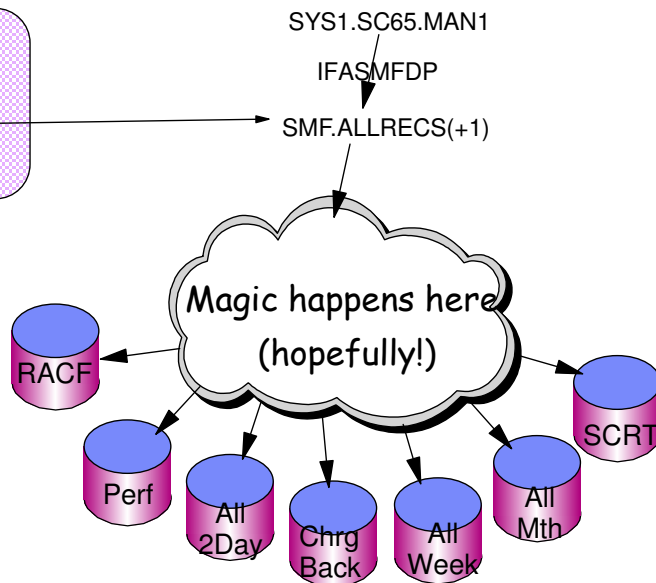
```
Run
Run
Run
IEE391A SMF ENTER DUMP FOR DATA SET ON VOLSER SBOX2D,
      DSN=SYS1.SC65.MAN1
IEFU29 =====> S SMFDUMP,MAN=MAN1
IEE388I SMF NOW RECORDING ON VOLSER SBOX2D,
      DSN=SYS1.SC65.MAN2 TIME=07.00.00
```

SYS1.SC65.MAN1

IFASMFDP

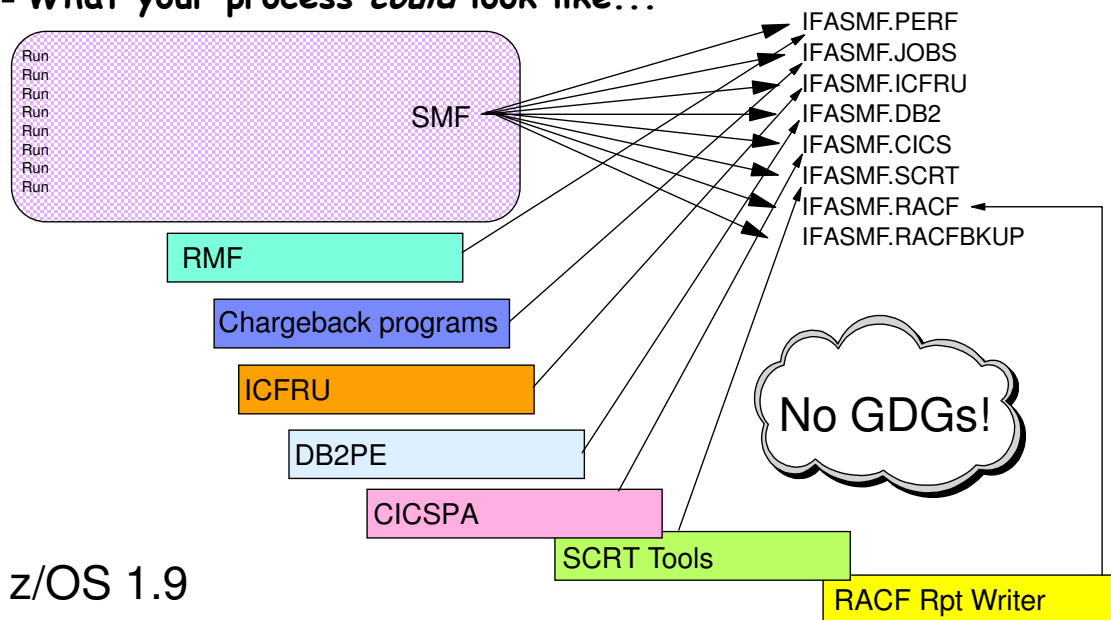
SMF.ALLRECS(+1)

- Considerations:**
- Process initiated by MANx switch
  - MANx read once then cleared
  - Have to cater for:
    - New day
    - New week
    - New month
    - Merging



## SMF Archiving process considerations

### ▪ What your process *could* look like...



## SMF log stream - is it for you?

### ▪ The first decisions that must be made are:

- Is the additional bandwidth this function provides *required*?
- If you don't need it for the performance, would the flexibility to create task-oriented log streams provide value to your company?

### ▪ IF you decide to switch to writing the SMF records to log streams, what about the GDGs?

- Do you continue with the GDG model for archived SMF records, or move to just keeping the SMF data in the log stream until it is no longer required by the enterprise (and is deleted using the Logger retention period function) or a mix of the two?
  - What is the SMF data used for today? Must identify all processes and users that use SMF data and how long each record type is retained for

## SMF log stream implementation considerations

### ▪ GDG model benefits:

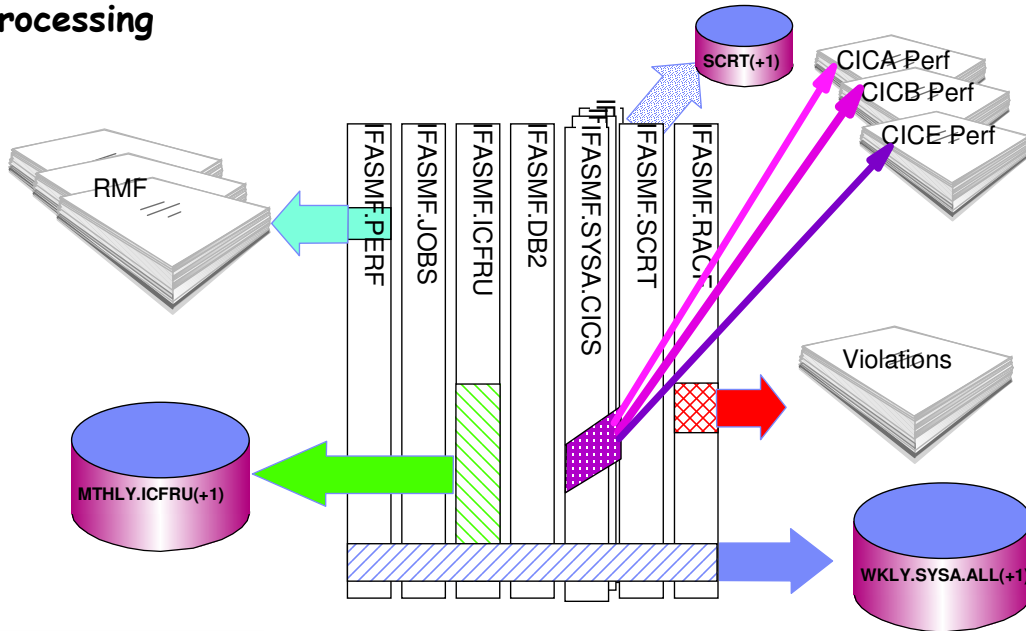
- No changes required to the downstream processes that use the SMF data
- It is a familiar and tried and proven process that has evolved and been fine-tuned over the years
- No transition considerations
- Less work to implement than a complete transition to keeping all data in log streams

## SMF log stream implementation considerations

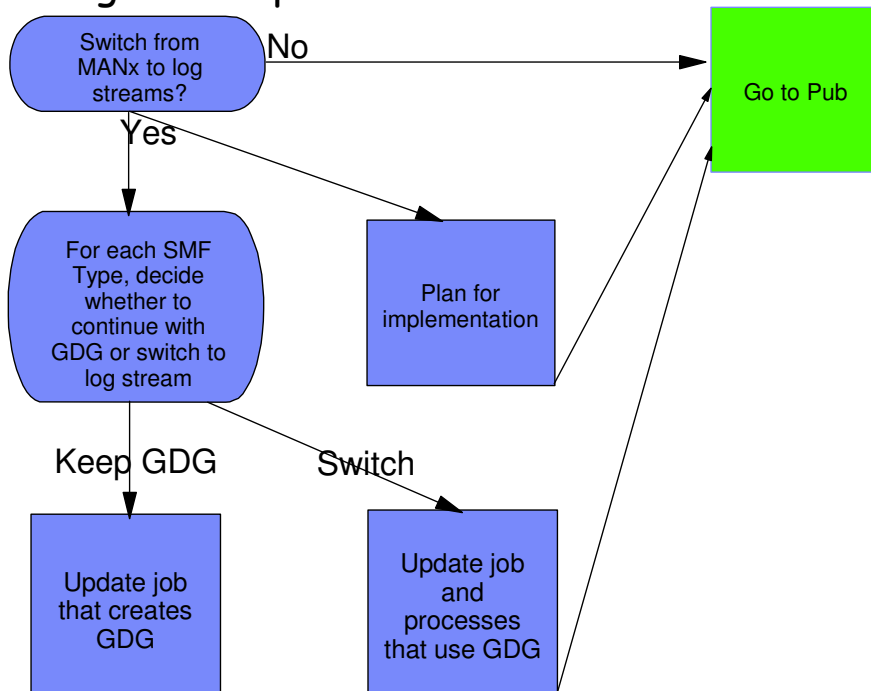
### ▪ Log stream model benefits:

- Simplicity - data is written to log stream by SMF and resides there until it expires
- More efficient - required data is accessed directly (using log block IDs) rather than reading sequentially through the GDG
- Don't need to work out relative generations to find a particular date - just specify start and end dates
- Easier management - all the rules relating to retention periods for SMF data are in a single place - the Logger policy
- Save the CPU time associated with creating the GDGs
- Not exposed to GDG-related problems

- If continuing with the GDG model, replace IFASMFDP with IFASMF DL and streamline process to eliminate unnecessary processing



## Deciding how to proceed





## Implementation summary

1. **Decide what log streams you want, and scope (system or plex)**
  - Even if you will continue to use GDGs, it makes sense to set up log streams based on use
  - Place record types with similar retention periods in the same log stream
2. **Size each log stream using GDG data set sizes or IFASMFDP output**
  - Consider volume of data and retention period (need to size both offload datasets and CF structure)
  - Remember limit of 168 offload extents (or increase DSEXTENT) and aim for not more than 1 offload/minute per log stream
3. **Decide on log stream-to-structure mapping (ideally each log stream should be connected to by 2 CPUs)**

## Implementation summary

4. **Define structures and log streams in CFRM and Logger structures**
5. **Set up replacement for SMFDUMP process**
  - Remember that no switch will automatically occur to kick off the process
6. **Update SMFPRMxx and activate**
7. **Issue SETSMS RECORDING(LOGSTREAM)**
8. **Remove MANx data sets from SMFPRM when happy**

## SMF writing to log streams

### ■ Implementation:

- Define structures in CFRM policy.
- Define structures and log streams in LOGR policy:
  - Recommend MAXBUFSIZE of 65532 on DEFINE STRUCTURE
- Update SMFPRMxx member to add LSNAME definitions
  - DEFAULTLSNAME(IFASMF.SMFALL)
  - LSNAME(IFASMF.SMF210,TYPE(210))
  - LSNAME(IFASMF.SMF211,TYPE(211))
  - LSNAME(IFASMF.SMF212,TYPE(212))
  - .....
- Recommend retaining MANx definitions and NOT specifying RECORDING(LOGSTREAM) until you are ready to permanently cutover:
  - But be aware that a SET SMF=xx will revert back to DATASET mode....

## Implementation

- Issue SET SMF=xx to pick up new definitions

```

SET SMF=FK
IEE252I MEMBER SMFPRMFK FOUND IN SYS1.PARMLIB
IEE967I 16.10.50 SMF PARAMETERS 254
  MEMBER = SMFPRMFK
  MULCFUNC -- DEFAULT
  SYNCVAL(00) -- DEFAULT
  DUMPABND(RETRY) -- DEFAULT
  BUFSIZMAX(0128M) -- PARMLIB
  BUFUSEWARN(10) -- PARMLIB
  LSNAME(IFASMF.SMF214,TYPE(214)) -- PARMLIB
  LSNAME(IFASMF.SMF213,TYPE(213)) -- PARMLIB
  LSNAME(IFASMF.SMF212,TYPE(212)) -- PARMLIB
  LSNAME(IFASMF.SMF211,TYPE(211)) -- PARMLIB
  LSNAME(IFASMF.SMF210,TYPE(210)) -- PARMLIB
  LSNAME(IFASMF.SMF220,TYPE(220)) -- PARMLIB
  DEFAULTLSNAME(IFASMF.SMFALL) -- PARMLIB
  PROMPT(LIST) -- PARMLIB
  DSNAME(SYS1.SC64.MAN4) -- PARMLIB
  DSNAME(SYS1.SC64.MAN3) -- PARMLIB
  
```

## Implementation

- Displaying SMF information at this point only shows the MANx data sets (because we haven't switched to log stream mode yet):

```

D SMF
IEE974I 16.11.05 SMF DATA SETS 272
      NAME                VOLSER  SIZE (BLKS)  %FULL  STATUS
P-SYS1.SC64.MAN3        SBOX2J     3000      99  DUMP REQUIRED
S-SYS1.SC64.MAN4        SBOX2J    90000      4  ACTIVE
  
```

## Implementation

- Issue **SETSMF RECORDING(LOGSTREAM)** to turn on recording to log stream

```

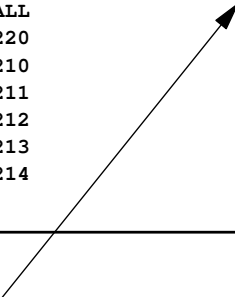
SETSMF RECORDING (LOGSTREAM)
IXC582I STRUCTURE SMF_TYPE21X ALLOCATED BY SIZE/RATIOS. 274
  PHYSICAL STRUCTURE VERSION: C13BB292 AC86F601
  STRUCTURE TYPE:                LIST
  CFNAME:                          CF1
  ALLOCATION SIZE:                  100352 K
  ...
IXC582I STRUCTURE SMF_TYPE200 ALLOCATED BY SIZE/RATIOS. 277
  PHYSICAL STRUCTURE VERSION: C13BB293 0E3BCB09
  STRUCTURE TYPE:                LIST
  CFNAME:                          CF2
  ALLOCATION SIZE:                  100352 K
  POLICY SIZE:                     120000 K
  ...
IXC582I STRUCTURE SMF_TYPEALL ALLOCATED BY SIZE/RATIOS. 280
  PHYSICAL STRUCTURE VERSION: C13BB293 B098D8C9
  STRUCTURE TYPE:                LIST
  CFNAME:                          CF3
  ALLOCATION SIZE:                  15360 K
  ...
IFA711I LOGSTREAM PARAMETERS ARE IN EFFECT
IEE712I SETSMF PROCESSING COMPLETE
  
```

## Implementation

- Now we check SMF again (D SMF)...

```

D SMF
IFA714I 16.11.32 SMF STATUS 293
      LOGSTREAM NAME      BUFFERS      STATUS
A-IFASMF.SMFALL          7925      CONNECTED
A-IFASMF.SMF220           0      CONNECTED
A-IFASMF.SMF210           0      CONNECTED
A-IFASMF.SMF211           0      CONNECTED
A-IFASMF.SMF212           0      CONNECTED
A-IFASMF.SMF213           0      CONNECTED
A-IFASMF.SMF214           0      CONNECTED
  
```



Number of bytes currently sitting in SMF buffers, waiting to be written to the corresponding log stream

## Implementation

- Now start writing some data to the log streams...

```

16:31:56.22 KYNEF 00000210 D SMF
16:31:56.23 KYNEF 00000010 IFA714I 16.31.56 SMF STATUS 349
      LOGSTREAM NAME      BUFFERS      STATUS
349 00000010      A-IFASMF.SMFALL          56317      CONNECTED
349 00000010      A-IFASMF.SMF220          41252      CONNECTED
349 00000010      A-IFASMF.SMF210          41252      CONNECTED
349 00000010      A-IFASMF.SMF211          61876      CONNECTED
349 00000010      A-IFASMF.SMF212          20628      CONNECTED
349 00000010      A-IFASMF.SMF213          30940      CONNECTED
349 00000010      A-IFASMF.SMF214          30940      CONNECTED

16:32:01.33 KYNEF 00000210 D SMF
16:32:01.35 KYNEF 00000010 IFA714I 16.32.01 SMF STATUS 354
      LOGSTREAM NAME      BUFFERS      STATUS
354 00000010      A-IFASMF.SMFALL           6294      CONNECTED
354 00000010      A-IFASMF.SMF220          61876      CONNECTED
354 00000010      A-IFASMF.SMF210          61876      CONNECTED
354 00000010      A-IFASMF.SMF211          10316      CONNECTED
354 00000010      A-IFASMF.SMF212          20628      CONNECTED
354 00000010      A-IFASMF.SMF213          41252      CONNECTED
354 00000010      A-IFASMF.SMF214          61876      CONNECTED
  
```

## Implementation

- As interim storage (CF structures) start to fill up, we see him allocating offload data sets....

```

16:32:05.43      00000210  IEF196I  IGD100I  6D16  ALLOCATED TO DDNAME  SYS00050  DATACLAS  (LOGR24K)
16:32:13.86      00000210  IEF196I  IGD100I  C530  ALLOCATED TO DDNAME  SYS00052  DATACLAS  (LOGR24K)
16:32:16.22      00000210  IEF196I  IGD100I  6D16  ALLOCATED TO DDNAME  SYS00054  DATACLAS  (LOGR24K)
16:32:26.79      00000210  IEF196I  IGD100I  6D16  ALLOCATED TO DDNAME  SYS00056  DATACLAS  (LOGR24K)
16:32:27.34      00000210  IEF196I  IGD100I  C530  ALLOCATED TO DDNAME  SYS00058  DATACLAS  (LOGR24K)
16:32:27.69      00000210  IEF196I  IGD100I  6D16  ALLOCATED TO DDNAME  SYS00060  DATACLAS  (LOGR24K)
16:32:58.90      00000210  IEF196I  IGD100I  6D16  ALLOCATED TO DDNAME  SYS00062  DATACLAS  (LOGR24K)
16:33:00.45      00000210  IEF196I  IGD100I  C530  ALLOCATED TO DDNAME  SYS00064  DATACLAS  (LOGR24K)

```

- Remember that data is available for processing as soon as it written to the log stream - don't need to wait for it to be offloaded

## SMF use of System Logger

### ■ Summary

- Excellent concept that delivers improved performance, greater flexibility, and improved efficiencies:
  - Over time, makes sense that this will become the defacto standard way of handling SMF data
- However, it will require some re-engineering (maybe significant) of how you post-process the MANx data sets:
  - Less impact if you continue to use GDGs
  - If you transition to just log streams, need to address how to handle when some data is in GDGs and more recent data is in log stream
- Expect that the customers most likely to take this up first are those that are limited to collecting less SMF record types than they would like because of the volume of SMF data they produce

## Value summary

### ▪ Customer value:

- Less likely to lose business-critical SMF records
- Use fewer processor cycles for post-processing of SMF data
- Ability to collect more data to help you fine-tune use of system, if you wish

### ▪ Ease of implementation:

- 4 out of 10
  - Enabling function should take less than 1 hour
  - Complex part is re-engineering all your SMF processes to handle the new repository of SMF data AND handling coexistence during the transition process

## SMF

### ▪ Some general SMF-related tips while we are on the subject of SMF...

- Make sure MANx data sets are allocated with 26KB CI Size for optimum write performance
- Turn off SMF Type 19s if you don't use them:
  - Especially if your z/OS system has access to non-z/OS DASD volumes
    - On ITSO system, I SMF switch took 9 seconds to write out IEE498I messages about LSPACE errors on non-z/OS volumes - SMF had to buffer all records created during this time...
- If you have enough real storage in z/OS, increase SMF buffer space with BUFSIZMAX, to allow you to handle larger spikes in SMF write rates
- Use ERBSCAN and ERBSHOW (provided with RMF) to easily browse SMF records from ISPF 3.4



International Technical Support Organization

[ibm.com/redbooks](http://ibm.com/redbooks)

*Sysplex-related hardware considerations*



© 2007 IBM Corporation. All rights reserved.

[ibm.com/redbooks](http://ibm.com/redbooks)

International Technical Support Organization



## Understanding the impact of z/OS processor changes on CF response times

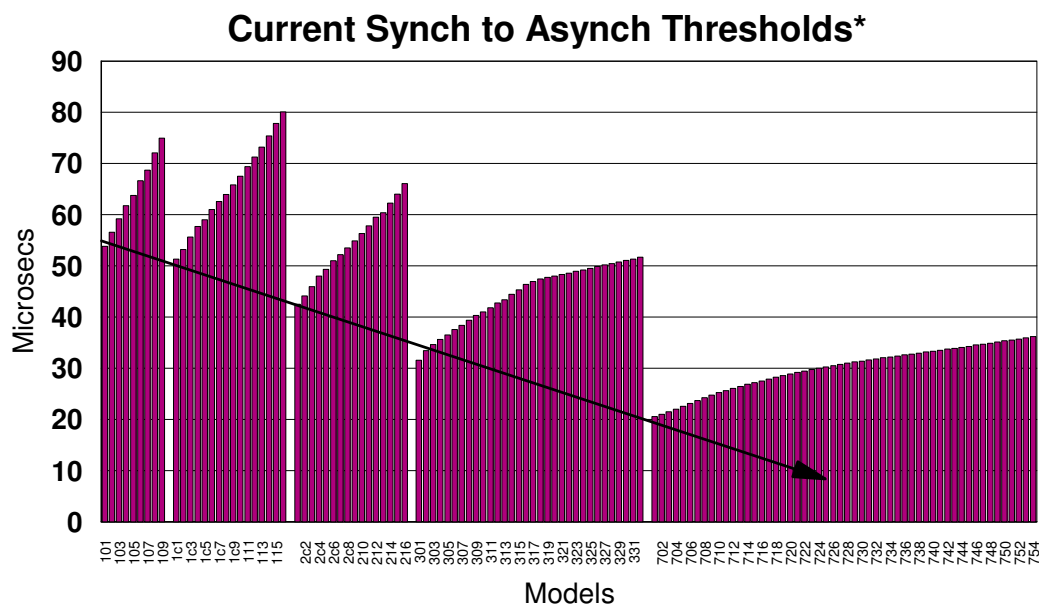


©2007 IBM Corporation. All rights reserved.

## z/OS Heuristic Algorithm

- Introduced in z/OS 1.2
- Intent is to handle CF requests as efficiently as possible, from a z/OS perspective
  - For short resp times, sync requests are more efficient
  - There is a machine-speed-dependent threshold where it becomes more efficient to send requests async and do some other work while we wait for CF to respond
- But important to remember there are TWO perspectives:
  - Each request wants to be sync, because that is always faster than async
  - Sysprog looks at *overall/average* resp times, and balance between sync and async
    - Turning a slow sync request into async improves average sync response time

## Thresholds for simplex CF requests



\* Note: The thresholds are higher for SM Duplexed structures - Thresholds for simplex requests changed by APAR 21635.

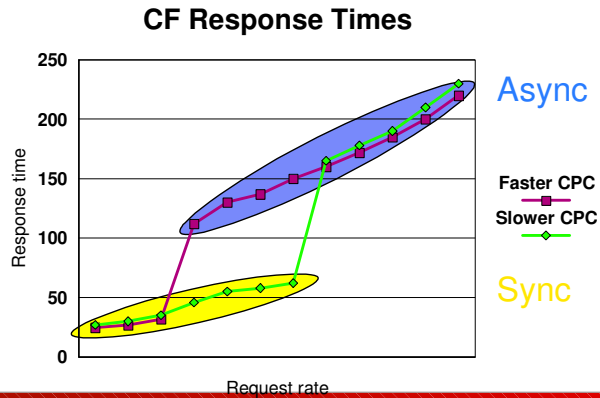


## z/OS Heuristic algorithm

### ▪ Considerations:

- If you upgrade just the CPC containing the operating system to a faster processor type, you may find that response times for some CF requests *increase*, because the threshold is lower on a faster CPC
  - Will probably find a larger percentage of requests getting converted to asynchronous

- May see a *decrease* in overall *average* synch response time as a result of the longer-running synch requests getting converted to asynch

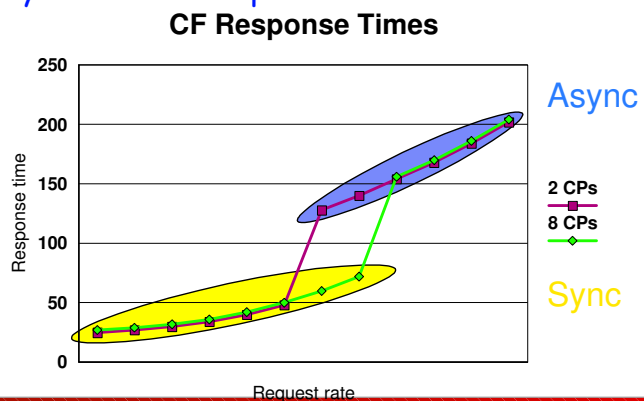


## z/OS Heuristic algorithm

### ▪ Considerations:

- What about if you upgrade an existing CPC by adding more engines?
- The per-CP speed will *DEcrease*, *INcreasing* the threshold, probably meaning that more requests will go synchronous. This movement of longer running requests into the synchronous bucket may have the effect of *INcreasing* average synchronous response times.

- Probably will not have as noticeable effect on average asynch response times



## z/OS Heuristic algorithm

### ▪ Any other considerations?

- CPU time for *synch* requests is all charged back to the requestor
- CPU time for *asynch* requests gets charged:
  - Some to the requestor
  - Some to the XCF Address Space
  - Some to whoever happens to be running when the response arrives back
- So, a slower engine speed (because you added more engines) *may* result in an increase in reported CPU time for certain address spaces (such as DB2) because a higher percent of requests go *synch*
  - There should be a corresponding decrease for the other address spaces, but this is probably less noticeable.

## Thresholds for simplex CF requests

- Based on trend for *synch/asynch* threshold to be getting very low, together with a review of the impact of code changes and performance of different generations of System z processors, thresholds for simplex CF requests will be adjusted upwards by XES APAR OA21635. PTFs available now, going back to z/OS 1.6.
  - The effect of this APAR should be that the z/OS CPU required to drive a given set of CF requests should be the same or less than without the APAR
- For more information, refer to excellent new description of this topic in the z/OS 1.9 level of *Setting Up a Sysplex*

## CF Configuration Options: Understanding the impact of Dynamic CF Dispatching and shared CF engines

### Sharing CF engines

- **The issues:**
  - Current System z engines are so big (and getting bigger with every new generation), I don't want to dedicate a whole one to my CF. Or,
  - For the last 20 years I have used shared engines for z/OS with acceptable results, why can't I do the same for my CFs?
  - How does my CF response time impact z/OS CPU utilization (=)?
- **There is no absolute answer on shared engines that applies across every customer - here we will describe the considerations, so each site can make the best decision for their configuration**

## Sharing CF engines

- **Expectations for a CF are different to those for z/OS:**
  - z/OS response times are measured in seconds (online txn. resp times) or maybe milliseconds (DASD resp times). CF response times are measured in microseconds (1/1000th of a millisecond).
    - Parallel Sysplex was designed on assumption that CF would be MUCH faster than DASD in order to minimize the cost of data sharing.
  - In order to deliver the response times required to minimize the overhead of using the CF, the design mode of operation (which assumes production data sharing) is for CFCC to run in a polling loop, either processing requests, or looping looking for new requests, but not releasing the engine until PR/SM takes it away. This ensures very good response times (no interrupt code to go through). However, it will consume all the cycles available to that LPAR.

## Considerations for picking the "right" CF...

- **For a production CF:**
  - Does it contain anything that is response-time critical, like data sharing structures?
  - What is the request rate?
  - Is there a failure-independence requirement? Is it using System-Managed Duplexing?
- **Is it a test or development or sandbox CF?**
  - Generally speaking, non-production CFs are used for function testing, so performance is not critical
  - Recommend that failure-independence (or lack of) should be similar to production CF, for more representative recovery testing (irrespective of whether it is using shared engines or not).

## Financial considerations

- How do shared engines impact you?
  - Increased CF response times, compared to using dedicated engines
- So, the cost of a dedicated engine is the \$ to buy the engine
- But shared engines also have a cost - the cost of higher overhead in the connected z/OS systems because the shared engine results in longer CF response times (= higher software bills)
  - The more requests you send to the CF, the higher the cumulative impact

## The cost of shared engines

- Another way to look at this table is that the boxes on the left represent CF response times

Faster z/OS systems →

Host CF	G5	G6	z800	z900 1xx	z900 2xx	z890	z990	z9-109
R06-HL	12%	14%	16%	17%	19%	22%	26%	---
R06-ICB	9%	10%	---	13%	14%	17%	20%	---
G5/6-IC	8%	8%	---	---	---	---	---	---
z800 ISC	11%	12%	11%	12%	13%	15%	18%	20%
z800 ICB/IC	---	---	9%	10%	11%	12%	14%	16%
z900 ISC	11%	12%	10%	11%	12%	14%	16%	18%
z900 ICB/IC	8%	9%	8%	9%	10%	11%	12%	13%
z890 ISC	8%	9%	9%	10%	11%	13%	15%	17%
z890 ICB/IC	8%	8%	7%	8%	8%	9%	10%	11%
z990 ISC	8%	9%	9%	10%	11%	13%	14%	15%
z990 ICB/IC	8%	8%	7%	8%	8%	9%	9%	10%
z9-109 ISC	---	---	8%	9%	10%	12%	13%	14%
z9-109 ICB/IC	---	---	7%	7%	7%	8%	8%	9%

↓ Better CF response times

Based on roughly 9 CF Requests/1 MIPS/second and dedicated engines for CFs  
 With z/OS 1.2 and later, CF Overhead is capped at about 18%

## What are "good" CF response times?

"Good" is very relative. Today (4Q 2007), the following are considered good CF response times. As the technology changes, the definition of "good" also changes

	9672 R64 to C04	2064 114 to 1xx	2084 3xx to 3xx	2094 109-7xx
SYNC	100-175	15-30	10-25	5-20
ASYNCR	500-1500	150-450	100-350	60-300

Fastest time is for lock request with no data (GRS)

Assumes 0 distance between CF and CPC

Longest time is for largest allowed data transfer (64KB)

Assumes fastest link type supported on that generation of CPC (usually ICB)

Add 15 (lock request) to 30 mics (list/cache) if using Fiber (ISC) links rather than ICB

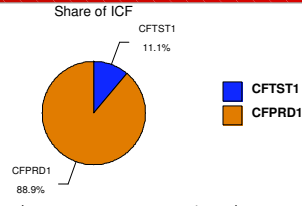
## Relationship between resp times and shared engines

### ■ How does sharing engines impact CF response times?

- First, let's understand what we are comparing against - lock resp times of **5 microseconds with a dedicated engine**
- If a CF is using a shared engine, any requests that arrive immediately before, or while the CF has lost the engine, will need to wait for the CF to get dispatched again. How long that is depends on:
  - Relative LPAR weights (which specify the LP's "fair share" of the engine)
  - The *actual* engine usage of each sharing LPAR, compared to its fair share
  - The Dynamic CF Dispatching setting of all CF LPARs sharing the engine
  - The duration of the PR/SM timeslice (default is 12,500 microseconds)
  - Special PR/SM override that ensures a ready CF LPAR will get dispatched at least once..... **every 100,000 microseconds!**
  - **PR/SM will not take CP away from a z/OS that is spinning, waiting for a response to a sync CF response**

Let's look at some examples...

- 2 CF LPARs, both with DCFD OFF
- Weights:
  - CFTST1 10
  - CFPRD1 90

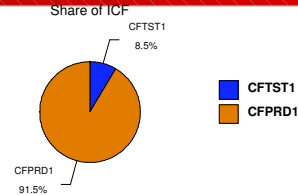


Time (us)	Action	Reason	CFTST1 Time	CFTST1 %	CFPRD1 Time	CFPRD1 %
.000000	CFTST1 gets engine	Due 10%	0		0	
.012500	Timeslice runs out		12500	100%	0	0%
.012500	CFPRD1 gets engine	Due 90%	12500		0	
.025000	Timeslice runs out		12500	50%	12500	50%
.025000	CFPRD1 gets engine	Due 90%	12500		12500	
.037500	Timeslice runs out		12500	33%	25000	66%
.037500	CFPRD1 gets engine	Due 90%	12500		25000	
.050000	Timeslice runs out		12500	25%	37500	75%
.050000	CFPRD1 gets engine	Due 90%	12500		37500	
.062500	Timeslice runs out		12500	20%	50000	80%
.062500	CFPRD1 gets engine	Due 90%	12500		50000	
.075000	Timeslice runs out		12500	16%	62500	84%
.075000	CFPRD1 gets engine	Due 90%	12500		62500	
.087500	Timeslice runs out		12500	14%	75000	86%
.087500	CFPRD1 gets engine	Due 90%	12500		75000	
.100000	Timeslice runs out		12500	12.5%	87500	87.5%
.100000	CFPRD1 gets engine	Due 90%	12500		87500	
.112500	Timeslice runs out		12500	11.1%	100000	88.9%
.112500	CFTST1 gets engine	Waiting 100 ms.	12500		100000	



Next example...

- 2 CF LPARs
- Weights:
  - CFTST1 10 DCFD ON
  - CFPRD1 90 DCFD OFF



Time (sec)	Action	Reason	CFTST1 Time	CFTST1 %	CFPRD1 Time	CFPRD1 %
.000000	CFTST1 gets engine	Due 10%	0		0	
.001500	Release engine	Sleep for 10ms	1500	100%	0	0%
.001500	CFPRD1 gets engine	Due 90%	1500		0	
.014000	Timeslice runs out		1500	10.7%	12500	89.3%
.014000	CFPRD1 gets engine	Due 90%	1500		12500	
.026500	Timeslice runs out		1500	5.6%	25000	94.4%
.026500	CFTST1 gets engine	Behind share	1500		25000	
.028000	Release engine	Sleep for 10ms	3000	10.7%	25000	89.3%
.028000	CFPRD1 gets engine	Due 90%	3000		25000	
.040500	Timeslice runs out		3000	7.4%	37500	92.6%
.040500	CFTST1 gets engine	Behind share	3000		37500	
.044500	Release engine	Sleep for 5ms	7000	15.7%	37500	84.3%
.044500	CFPRD1 gets engine	Due 90%	7000		37500	
.057000	Timeslice runs out		7000	12.3%	50000	87.7%
.057000	CFPRD1 gets engine	Due 90%	7000		50000	
.069500	Timeslice runs out		7000	10.1%	62500	89.9%
.069500	CFPRD1 gets engine	Due 90%	7000		62500	
.082000	Timeslice runs out		7000	8.5%	75000	91.5%
.082000	CFTST1 gets engine	Due 10%	7000		75000	



## Shared engines and the heuristic algorithm

### ▪ Let's come back to how shared engines impact z/OS utilization...

- How will the heuristic algorithm react to the changing response times delivered by CFPRD1? What ARE the response times?
  - Fast, Fast, Fast, REALLY Slow, Fast, REALLY Slow....
- What is the result of that?
  - z/OS decides to send the requests asynchronously.
  - The CPU time to process an asynch request is much higher than the time to process a well-performing synch request. ==>>
    - The use of a shared engine for the production CF results in more z/OS CPU time per request than if the CF had a dedicated engine.
    - How *much* more depends on the difference between the synch time and the threshold (the threshold represents the amount of CPU time required to process an async request).

## Translating into actual results

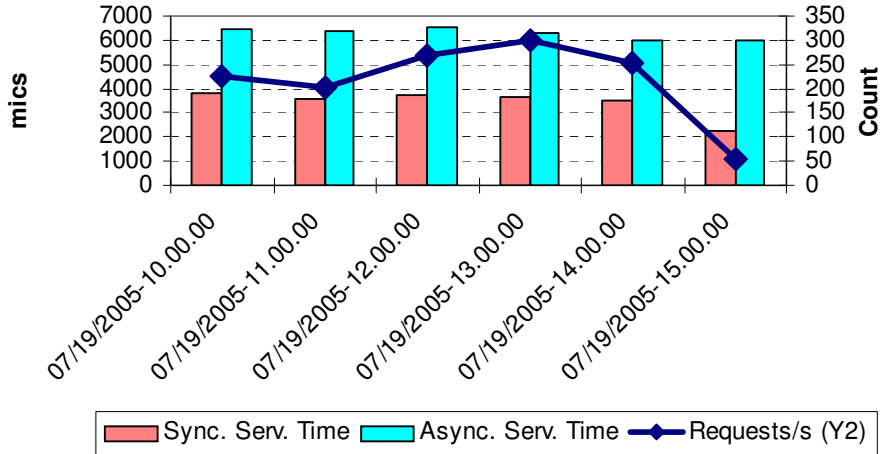
- This was just an example - let's look at an actual customer situation using shared engines
  - Customer had 4 CF LPs sharing a single ICF
  - All CFs were production, but 3 had very light CF loads
  - Customer followed IBM guidance and turned DCFD OFF for all CFs (because they were all production)



## Sharing CF engines - how NOT to do it!

### Single ICF, 4 CF LPARs, DYNDISP OFF for all LPs

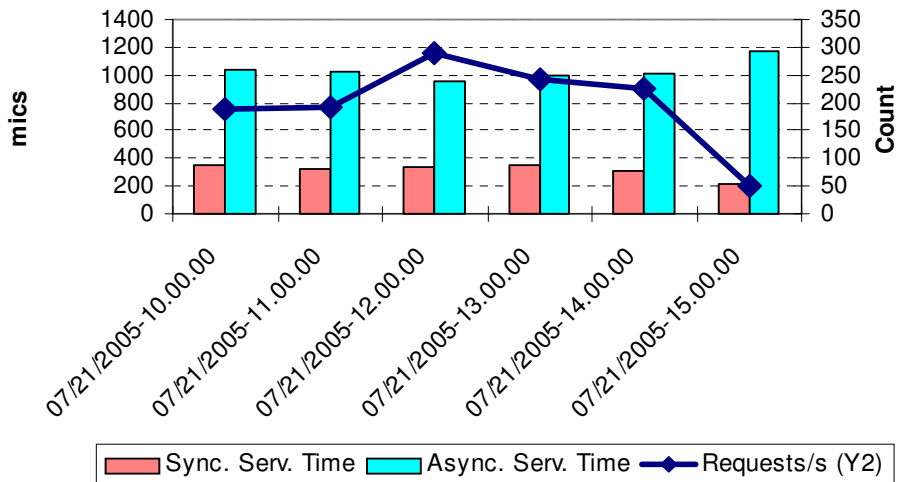
Report for Coupling Facility: FPKF07, System: FPKA



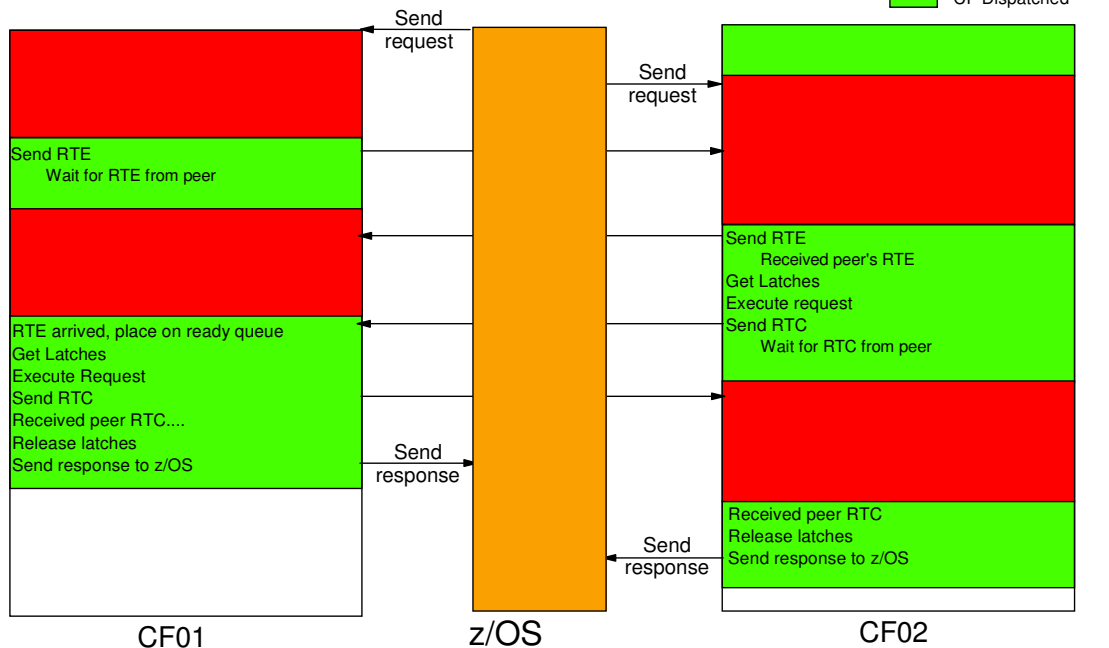
## Sharing CF engines - how NOT to do it!

### Single ICF, 4 CF LPARs, DYNDISP ON for all LPs

Report for Coupling Facility: FPKF07, System: FPKA



## Mixing shared engines and SM Duplexing



## What this means for you

- If doing production data sharing, preference is dedicated engine
- Don't have >1 CF with DCFD OFF sharing a given engine
- Consider the cost of higher response times on z/OS CPU consumption
- If using shared engines on a CF-only box, consider reducing max LPAR timeslice ("Processor Running Time) from 12.5 milliseconds to something lower in Options tab in CPC Activation Profile on HMC
- If doing System-Managed Duplexing, both CFs must have dedicated engines or VERY high weight



## Considerations for running Parallel Sysplex under z/VM

- **Increasing number of customers using z/VM for their test sysplexes**
  - However if not set up correctly, availability of these sysplexes may be impaired
    - 1) Ensure that VM directory entries contain DASD WRKAL statements immediately after the MDISK statement for the CDS volumes. Without this, CDS corruption is nearly guaranteed
    - 2) CF virtual machines must deliver acceptable response times (that is, they should be defined with high "relative weights" compared to connect z/OS guests.
      - Specify QUICKDISP in OPTIONS stmt in directory - keeps CF virtual machine in dispatch list
      - Specify SHARE REL xxx in directory entry for CF virtual machines. Default is 100, so make CFVMs much higher than that

## Interesting Web sites

- **Redbooks (and workshops and residencies..):**
  - <http://www.redbooks.ibm.com>
- **z/OS Test Reports:**
  - <http://www.ibm.com/servers/eserver/zseries/zos/integtst/>
- **Red Alerts:**
  - <https://www14.software.ibm.com/webapp/set2/sas/f/redAlerts/home.html>
- **Statements of direction:**
  - [http://www.ibm.com/servers/eserver/zseries/zos/zos\\_sods.html](http://www.ibm.com/servers/eserver/zseries/zos/zos_sods.html)
- **Parallel Sysplex home page:**
  - <http://www.ibm.com/servers/eserver/zseries/ps/>
- **WSC Documents:**
  - <http://www.ibm.com/support/techdocs>
- **IBM Tech Journals:** <http://www.research.ibm.com/journal>

## LookAT

- **Provides online IBM message explanations:**
  - Available on the Web at:  
[www.ibm.com/servers/s390/os390/bkserv/lookat/lookat.html](http://www.ibm.com/servers/s390/os390/bkserv/lookat/lookat.html)
  - Can also be installed and run under Windows
  - Can also be installed and run on TSO
  - Can also be installed on Palm VIX (if you want to REALLY impress your friends and those young UNIX people!)
    - All these can be downloaded from:  
<ftp://ftp.software.ibm.com/ps/products/ibmreader/tools/lookat>
    - They are also available on the IBM Online Library Collection for z/OS
  - Also provides access to latest DOC APARs (updated weekly)

## Interesting WSC documents

### URL:

- <http://www.ibm.com/support/techdocs/atmastr.nsf/Web/TechDocs>

### Documents:

- FLASH10451 - Withdrawal of z/OS Functions
- TD103286 - z/OS Best Practices: Large Stand-Alone Dump Handling - Version 2
  - Also Feb 2007 article in Hot Topics newsletter
- FLASH10598 - z/OS Performance: Performance Enhancements for Large Real Storage Environments
- FLASH10572 - CF Level 15 changes to structure sizes
- FLASH10593 - Testing STP Recovery
- TD104031 - STP External Time Source : AUTHORITIES RESPONSIBLE FOR THE TIME DISSEMINATION SERVICES
- WP100743 - XCF Performance Considerations V3.1
- WP100258 - Performance Considerations When Moving to Fewer, Faster CPUs

## Interesting Hot Topics articles

There is no easy way to search across all Hot Topics newsletters for a given keyword, but....

Go to

[http://publibz.boulder.ibm.com/cgi-bin/bookmgr\\_OS390/Shelves/HOTOPICS](http://publibz.boulder.ibm.com/cgi-bin/bookmgr_OS390/Shelves/HOTOPICS)

Select "Search Documents" button

Enter search word and select "Search" button

You will be presented with a list of the issues that contain your keyword.

- If you select a book, you can see the article name, but you can't view it.
- Need to go to the Hot Topics site and open the PDF....

## End-of-support dates

Following dates are for US. May be different in other countries:

- **z/OS 1.4 and 1.5. March 31, 2007**
  - z/OS 1.6. September 30, 2007
  - Last date to order 1.8 ServerPac - Oct 9, 2007
  - Last date to order 1.8 SystemPac - June 23, 2008
- **CICS TS 2.2. April 30, 2008**
  - CICS TS 2.1. April 30, 2006
- **DB2 V7. June 30, 2008**
  - DB2 V6. June 30, 2005
- **IMS V8. November 5, 2008**
  - IMS V7. November 8, 2005
- **MQ 5.3. October 31, 2005**

See <http://www-306.ibm.com/software/support/lifecycle> or  
<http://www.ibm.com/services/sl/products/java.html>



## Automation considerations - sysplex-related messages

- IXC101, IXC105 - sysplex partitioning started/completed
- IXC102, IXC402, IXC409 - manual actions need to be taken during sysplex partitioning
- IXC244 - cannot use couple dataset, could not be opened
- IXC246 - couple dataset I/O delays
- IXC255 - unable to use couple dataset, inconsistent
- IXC256 - couple dataset switch cannot complete, waiting for systems to participate
- IXC259 - I/O error on couple dataset
- IXC267 - now processing without an alternate couple dataset (you need to make a new one)
- IXC406 - inconsistent time source
- IXC418 - system is now active in the sysplex
- IXC426, IXC427, IXC446 - system is half-dead (not updating system status, but still sending signals)



## Automation considerations - sysplex-related messages

- IXC430, IXC431, IXC432 - stalled XCF members on system
- IXC440 - critical stalled XCF members on system
- IXC458, IXC459 - stopped signalling path
- IXC467 - XCF signalling path recovery actions taken
- IXC501 - confirm request to use CF
- IXC512 - CFRM policy changes are pending
- IXC517 - system ABLE to use CF
- IXC518 - system NOT ABLE to use CF
- IXC519 - CF failure (damaged CF)
- IXC522 - rebuild being stopped (usually a rebuild failure of some kind)
- IXC538 - structure could not be duplexed
- IXC552, IXC553 - lack of duplex failure isolation between duplexed structure instances
- IXC573 - error during system-managed processing

## Automation considerations - sysplex-related messages

- IXC585 - structure has exceeded structure full threshold
- IXC615 - SFM taking action against stalled member
- IXC631, IXC640 - stalled member having a sysplex impact
- IXC700 - couple dataset capacity problems
- IXC800 - ARM restarts not processed

## Automation considerations - sysplex-related messages

- IXL010 - critical message reported from the CF (usually the CF reporting an error of some kind)
- IXL013 - IXLCONN connection failure
- IXL040, IXL041 - XES external hang detect is pointing the finger at a stalled connector
- IXL044 - coupling facility experiencing IFCC errors
- IXL045 - connector may be experiencing delays due to SRB throttling
- IXL157 - CF path operational
- IXL158 - CF path not operational
- IXL159 - CF support facility hardware error detected
- IXL160, IXL162 - request time ordering required but not enabled, cannot use CF

## APARs of interest

**OA14593 (open) - Deadlock when recovering from double failure (see APAR for workaround)**

**OA14465 - Problem partitioning a system from the sysplex**

**OA17708 - System unable to join the plex after multiple PSWITCH and ACOUPLE commands**

**OA19302 - Problem with GRS Ring system trying to rejoin the sysplex**

**OA20749 - Support for page data sets > 4GB**

**CFMON - XCF monitoring tool previously available from ITSO - PLEASE DO NOT USE IT - has been known to cause sysplex outages**



## Help please?

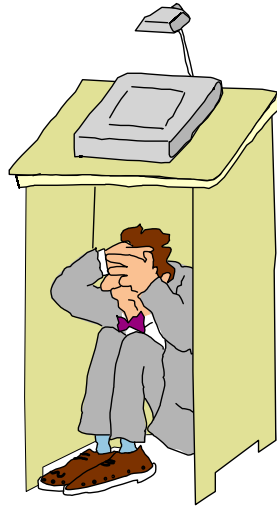
- **APAR OA22414 adds information to the SMF Type 23 record to help us better characterize customer environments relative to the LSPR workload suite.**
  - Intent is to help us more accurately predict the actual capacity customers should expect when migrating to a new processor type
    - Currently the LSPR numbers have broad categories, and it is not always easy to correctly categorize a particular set of workloads.
  - We are looking for customers willing to send us these records, together with the SMF Type 70-77 for the corresponding interval
    - If you would like to contribute, please send an email to [kyne@us.ibm.com](mailto:kyne@us.ibm.com)
- **What would you think of additional sections on:**
  - TCO considerations?
  - Runaway transaction protections?

## Blatant advertising.....

- **The value of IBM Redbooks depends on a steady supply of skilled, experienced, and enthusiastic residents.**
  - Books are written by IBMers, customers, and Business Partners
- **Residencies are generally run beside the related IBM lab and provide an opportunity for the residents to work with Development and develop/extend their network of technical experts within IBM**
  - It also doesn't hurt your resume to be able to say that you co-authored an IBM RedBooks publication!
- **All residencies are listed on redbooks Web site, and you can subscribe to topics you are interested in**
- **IBM covers all travel, hotel, and meal expenses**

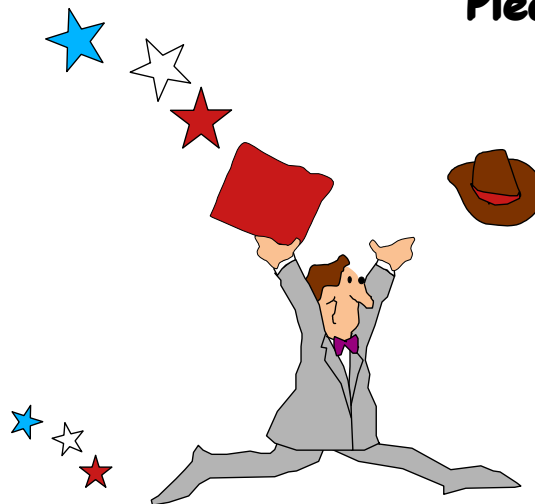


# Questions?



# Thanks!!

**Please come again**





International Technical Support Organization

[ibm.com/redbooks](http://ibm.com/redbooks)

## Miscellaneous enhancements in R9



© 2007 IBM Corporation. All rights reserved.

[ibm.com/redbooks](http://ibm.com/redbooks)

International Technical Support Organization



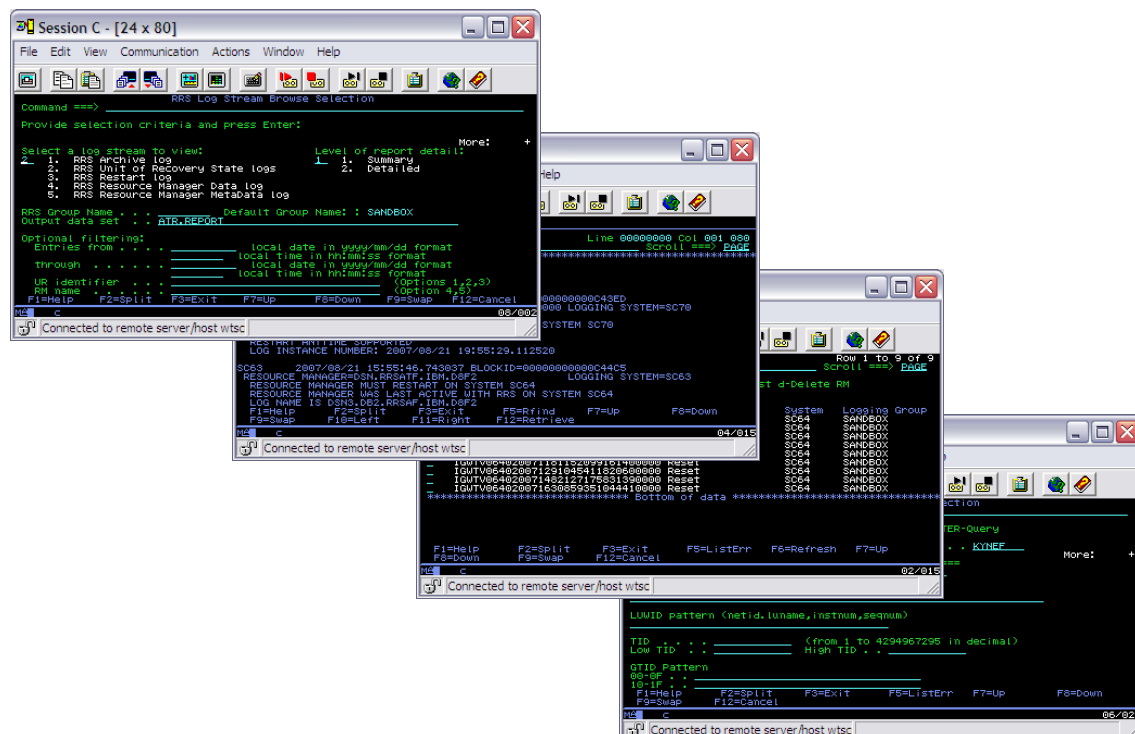
## RRS enhancements



©2007 IBM Corporation. All rights reserved.

## RRS Problem Documentation tool

- RRS provides powerful reporting functions via its ISPF panel interface:
  - Browse an RRS log stream
  - Display/Update RRS related Resource Manager information
  - Display/Update RRS Unit of Recovery information
  - Display/Update RRS related Work Manager information
  - Display/Update RRS UR selection criteria profiles
  - Display RRS-related system information
- If an RRS-related problem is encountered, valuable diagnostic information can be extracted from the panels - IF you know what you are looking for



The screenshot displays the RRS Log stream Browse Selection tool interface, showing multiple overlapping panels. The main panel displays the following information:

```

Command ==>
Provide selection criteria and press Enter:
Select a log stream to view:
1. RRS Archive log
2. RRS Unit of Recovery State logs
3. RRS Restart log
4. RRS Resource Manager Data log
5. RRS Resource Manager Metadata log
Level of report detail:
1. Summary
2. Detailed
RRS Group Name: . . . Default Group Name: : SANDBOX
Output data set : . . . ATB,REPORT
Optional filtering:
Entries from : . . . local date in yyyy/mm/dd format
Through : . . . local time in hh:mm:ss format
UR identifier : . . . local date in yyyy/mm/dd format
RM Name : . . . local time in hh:mm:ss format
Options 1,2,3
F1=Help F2=Split F3=Exit F4=Up F5=Down F6=Swap F7=Cancel
  
```

Other panels show log data and system information:

```

Line 00000000 Col 001 000
*****
00000000C43EC
000 LOGGING SYSTEM=SC70
SYSTEM SC70
  
```

```

RESTRICTIONLINE SUPPORTED
LOG INSTANCE NUMBER: 2007/08/21 19:55:29.112520
SC63 2007/08/21 15:55:46.743837 BLOCKID=0000000000C44C5
RESOURCE MANAGER=DRS,RSMT,IBM.DRP2 LOGGING SYSTEM=SC63
RESOURCE MANAGER WAS RESTART ON SYSTEM SC64
RESOURCE MANAGER WAS LAST ACTIVE WITH RRS ON SYSTEM SC64
LOG NAME IS D3S-D32-RSMP-IBM.DRP2
F1=Help F2=Split F3=Exit F4=Up F5=Down F6=Swap F7=Cancel
  
```

```

System Logging Group
SC64 SANDBOX
SC64 SANDBOX
SC64 SANDBOX
SC64 SANDBOX
SC64 SANDBOX
SC64 SANDBOX
SC64 SANDBOX
  
```

```

LUUID pattern (netid,uname,instnum,seqnum)
TID : : (from 1 to 4294967295 in decimal)
Low Tid : :
High Tid : :
GTID Pattern
955
ID-IF :
F1=Help F2=Split F3=Exit F4=ListErr F7=Up F8=Down
F6=Swap F12=Cancel
  
```

## RRS Diagnostics gathering

- **To make it easier to automatically gather important information in an automated way, z/OS 1.9 provides a batch equivalent (ATRQSRV) of the ISPF functions:**
  - Described in Appendix in *MVS Programming: Resource Recovery*
  - Description of all the "STATEMENTS" subparameters provided in manual after sample JCL.
    - Will probably want automation to build control statements dynamically based on content of error message
- **Recommend adding some automation to submit appropriate job following RRS error**
  - This will be based on the experiences from YOUR installation. Currently there is no IBM list of suggested messages to trap on.

## RRS ATRQSRV utility

- **ATRQSRV must be run on a z/OS R9 or later system**
  - But can extract information about other systems using the *SYSNAME* parameter
- **Also supports new ability to unregister a Resource Manager**
  - Using *UNREGRM* keyword
  - Only works for Resource Managers running on z/OS 1.9 or later systems
    - If RM is on a back-level system, get "ATR538I - The ATRSRV request was processed on a downlevel RRS system that could not honor the request."

## RRS Availability enhancement

- Sometimes, unregister processing for a resource manager (RM) may not complete successfully
- This leaves the RM in limbo:
  - "Unregistered" with Registration Services
  - But still "set" with RRS
- As a result, attempts to restart the RM will fail because RRS thinks it is already up
  - Only discover this state when you try to restart the RM and it fails with a status indicating that it is already active
- Prior to z/OS 1.9, the only way to resolve this was to recycle RRS, impacting all other users of RRS on that system

## Unregistering RMs

- New ISPF (and batch) interface to let you unregister an RM that has already unregistered from Registration Services but not unset from RRS:
- 
- 
- Only supports RMs still listed with status "Run" - trying to unregister an active RM will fail

```

Session C - [24 x 80]
RRS Resource Manager List
Command: v-View Details u-View URs r-Remove Interest d-Delete RM
n-Unregister RM
S RM Name State System Logging Group
DSN,RRSPAS,IBM,D8F2 Run SC64 SANDBOX
DSN,RRSPAS,IBM,D8F2 Run SC64 SANDBOX
DSN,RRSPAS,IBM,D8F2 Run SC64 SANDBOX
IGUTV864922871523478251490000 Reset SC64 SANDBOX
IGUTV86492287118115209916140000 Reset SC64 SANDBOX
IGUTV8649228715219451132080000 Reset SC64 SANDBOX
IGUTV864922871482127175331390000 Reset SC64 SANDBOX
IGUTV8649228718388935194441890000 Reset SC64 SANDBOX
***** BOTTOM of data *****
F1=Help F2=Split F3=Exit F5=ListErr F6=Refresh F7=Up
F8=Down F9=Swap F12=Cancel

```

```

Session C - [24 x 80]
RRS Resource Manager List
Command: v-View Details u-View URs r-Remove Interest d-Delete RM
n-Unregister RM
S RM Name State System Logging Group
DSN,RRSPAS,IBM,D8F2 Run SC64 SANDBOX
IGUTV86492287118115209916140000 Reset SC64 SANDBOX
IGUTV8649228715219451132080000 Reset SC64 SANDBOX
IGUTV864922871482127175331390000 Reset SC64 SANDBOX
IGUTV8649228718388935194441890000 Reset SC64 SANDBOX
***** BOTTOM of data *****
ATR5361 RM DSN,RRSPAS,IBM,D8F2 is still registered with Registration
Services and cannot be unregistered with RRS.
F8=Down F9=Swap F12=Cancel

```

## Unregistering RMs

- **Support delivered with z/OS 1.9**
  - Documented in *MVS Programming: Resource Recovery*
- **RRS ISPF panels show sysplex-wide view of activity.**
  - Unregister command can be issued against any valid RM, however RM must be running on z/OS 1.9 or later system to be accepted.

## RRS Health Checks

- **While we are on the topic of RRS, is everyone running the RRS HealthChecks?**
  - Check that RMDATA log stream data is protected from failure
    - Medium severity, z/OS 1.4 & later
    - Availability related (loss of RMDATA data requires RRS cold start)
  - Check size of RRS offload data sets (4 checks)
    - Low severity, z/OS 1.4 & later
    - Performance related
  - Check that archive log stream is in its own CF structure
    - Low severity, z/OS 1.8 & later
    - Performance related

## RRS Survey

- **Is there anyone here that understands RRS as well as you would like to?**
  - RRS is used by *APPC, CICS, DB2, IMS, MQ, and WebSphere*
- **Any interest in a class on RRS, aimed at MVS people? Or is this seen as a subsystem issue?**

## VTAM Generic Resources enhancements

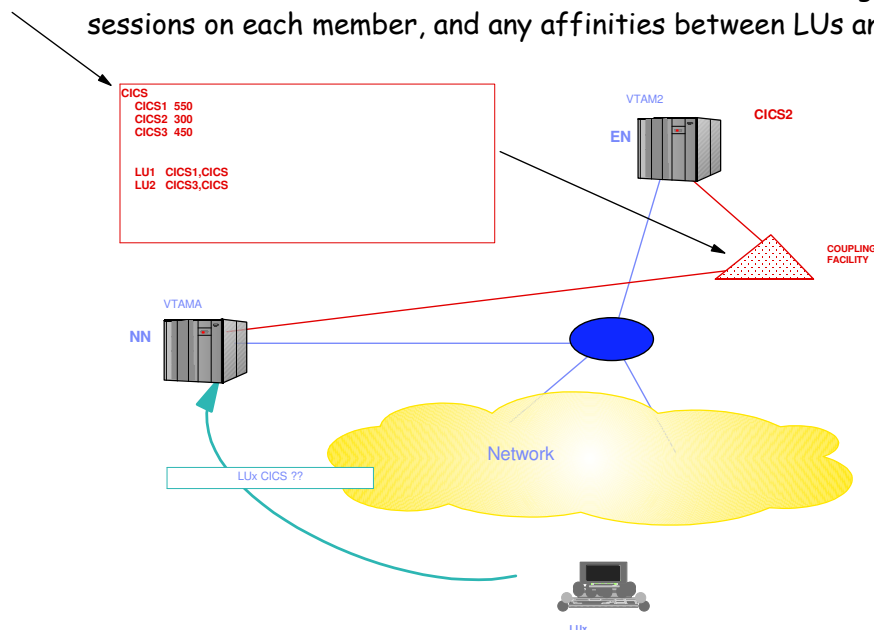


## VTAM Generic Resources

### ▪ VTAM GR can deliver significant benefits:

- Masks outages from application users - as long as at least one member of the GR group is active, users can logon as normal, unaware that most of the members are down
- Provides user with single logical image of the application - doesn't need to worry about using an alternate logon in case of problems
- Balances users across the systems, so loss of one system only affects a subset of users
- Provides degree of load balancing

ISTGENERIC structure maintains info about members of each GR group, number of sessions on each member, and any affinities between LUs and Apps



## VTAM Generic Resources

- **Prior to z/OS 1.9, VTAM used a combination of WLM and a user-written exit (ISTEXCGR) to implement the following (default) rules when deciding where to route a logon request:**
  - 1) If there is an existing affinity between LU and app instance, use it
  - 2) Select local instance of target application if one is available
  - 3) Select app instance with best goal achievement (as per WLM)
  - 4) If all have same goal achievement, select instance with smallest number of sessions

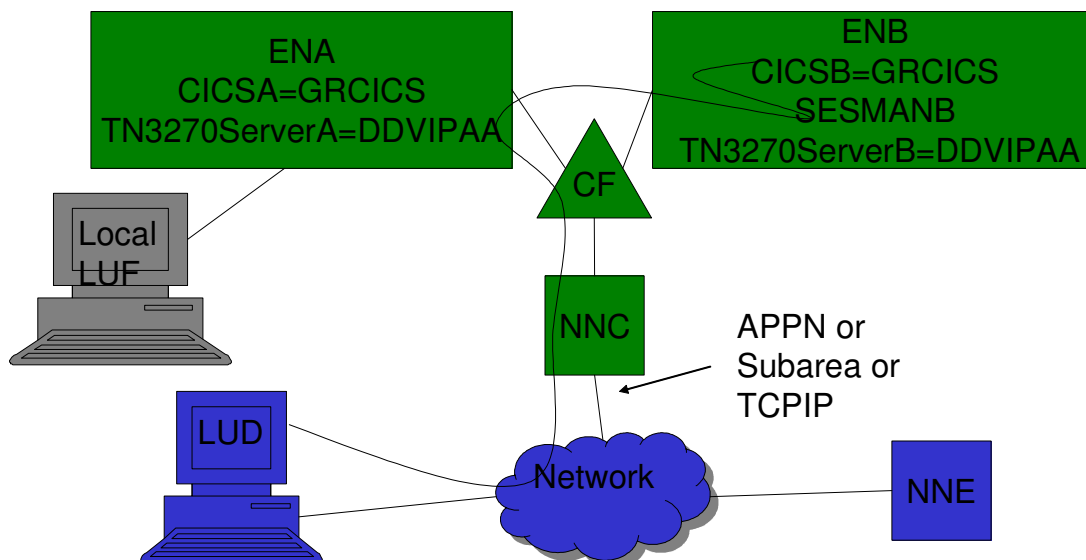
## VTAM Generic Resources

- **While these rules can be overridden using VTAM ISTEXCGR exit, many installations don't want to get into writing exits, so end up not running GR as they would really like to.**
  - Also, the ISTEXCGR exit was complex in that it had TWO functions:
    - 1) Decide which application instance to select from the list of candidates presented to the exit
    - 2) Potentially change the settings that controlled the list of candidates that would be presented THE NEXT time the exit is called
  - By moving control of the candidate list outside the exit, if you DO want to use the exit, the logic can be simpler.
    - Also, the exit can be more effective because you can decide which VTAM applications you DO or DO NOT want to use it for

## VTAM GR - R9 changes

- **z/OS R9 adds ability to have similar function to the ISTEYCGR exit, but without having to do any Assembler coding, PLUS it adds some new capabilities:**
  - New VTAMLST member to define session resolution preferences (equivalent to bit settings in ISTEYCGR) - this determines the candidate list that VTAM will build
  - Can now specify different resolution preferences for different VTAM applications (based on the generic resource name)
  - New information about Originating Logical Unit available to the decision-making process for applications that do a CLSDST PASS to a generic resource, enabling more efficient decisions when multiple workload balancing points

## Duplicate load balancing - DVIPA and GR



## VTAM GR - R9 changes

- A new VBUILD type GRPREFS has been created to identify the generic resource preferences table.
- A new definition statement GRPREF has been defined within the GRPREFS table to define GR resolution preferences. A GRPREF statement can be defined for each GR name. A nameless GRPREF statement can be used to define default GR preferences.

## VTAM GR - R9 changes

- Five operands can be defined on the GRPREF definition statement:
  - GREXIT=YES|NO (DEFAULT=NO)
  - LOCAPPL=YES|NO (DEFAULT=YES)
  - LOCLU=YES|NO (DEFAULT=YES)
  - PASSOLU=YES|NO (DEFAULT=NO)
  - WLM=YES|NO (DEFAULT=YES)
- Except for the new function of PASSOLU, these operands default to the same behavior as the corresponding GR EXIT flags.

## VTAM GR - R9 changes

### ▪ Let's look at a sample....

```
GRHOST01  VBUILD  TYPE=GRPREF
          GRPREF  GREXIT=NO, WLM=YES, LOCAPPL=YES, LOCLU=YES, PASSOLU=NO
GRCICS    GRPREF  GREXIT=NO, WLM=NO, LOCAPPL=YES, LOCLU=YES, PASSOLU=YES
GRTSO     GRPREF  GREXIT=YES, WLM=YES, LOCAPPL=YES, LOCLU=YES, PASSOLU=NO
```

### ▪ Specifies that all VTAM GR applications (other than GRCICS and GRTSO):

- Should NOT call the ISTEXCGR exit,
- WILL call WLM to see which application instance is currently performing best against its goal,
- WILL prefer Local Instances for requests coming from a local LU or local application,
- And the originating LU will NOT be used in the routing decision.

## VTAM GR - R9 changes

```
GRHOST01  VBUILD  TYPE=GRPREF
          GRPREF  GREXIT=NO, WLM=YES, LOCAPPL=YES, LOCLU=YES, PASSOLU=NO
GRCICS    GRPREF  GREXIT=NO, WLM=NO, LOCAPPL=YES, LOCLU=YES, PASSOLU=YES
GRTSO     GRPREF  GREXIT=YES, WLM=YES, LOCAPPL=YES, LOCLU=YES, PASSOLU=NO
```

### ▪ For all requests for CICS regions using a GR name of GRCICS:

- The ISTEXCGR exit will NOT be called,
- WLM will NOT be called (meaning that the GR member with the smallest number of sessions will be selected),
- Local Instances ARE preferred for requests coming from a local LU or local application,
- And the originating LU WILL be used in the routing decision when CLSDST PASS is used

## VTAM GR - R9 changes

```
GRHOST01  VBUILD  TYPE=GRPREF
          GRPREF  GREXIT=NO, WLM=YES, LOCAPPL=YES, LOCLU=YES, PASSOLU=NO
GRCICS    GRPREF  GREXIT=NO, WLM=NO, LOCAPPL=YES, LOCLU=YES, PASSOLU=YES
GRTSO     GRPREF  GREXIT=YES, WLM=YES, LOCAPPL=YES, LOCLU=YES, PASSOLU=NO
```

- **For all requests for TSOs using a GR name of GRTSO:**
  - The ISTEXCGR exit **WILL** be called (presumably in this case, the installation has some function in their ISTEXCGR that selects a specific TSO instance),
  - WLM IS called (and his recommendation will be available to the ISTEXCGR exit),
  - Local Instances **ARE** preferred for requests coming from a local LU or local application,
  - And the originating LU will **NOT** be used in the routing decision.

## VTAM GR - R9 changes

- **Once the new GR preferences table is built, it can be activated (and replaced) using the V NET, ID=mbrname, ACT command**
  - You can't delete the table, but you could replace it with one that provides the same preferences as the default GR rules
    - Use a member with one, "nameless" entry and no parameters

## VTAM GR - R9 changes

- Can also display the currently active settings:

```
D NET, GRPREFS
IST097I DISPLAY ACCEPTED
IST350I DISPLAY TYPE = GR PREFERENCES TABLE
IST075I NAME = GRHOST01, TYPE = GR PREFERENCES
IST924I -----
IST2210I GR PREFERENCE TABLE ENTRY = **NAMELESS**
IST2202I GREXIT = NO WLM = YES LOCLU = YES
IST2204I LOCAPPL = YES PASSOLU = NO
IST924I -----
IST2210I GR PREFERENCE TABLE ENTRY = GRCICS
IST2202I GREXIT = NO WLM = NO LOCLU = YES
IST2204I LOCAPPL = YES PASSOLU = YES
IST924I -----
IST2210I GR PREFERENCE TABLE ENTRY = GRTSO
IST2202I GREXIT = YES WLM = YES LOCLU = YES
IST2204I LOCAPPL = YES PASSOLU = NO
IST314I END
```

Using  
sample  
statements:

If no table  
defined:

```
D NET, GRPREFS
IST097I DISPLAY ACCEPTED
IST350I DISPLAY TYPE = GR PREFERENCES TABLE 178
IST075I NAME = NONE, TYPE = GR PREFERENCES
IST924I -----
IST2210I GR PREFERENCE TABLE ENTRY = **DEFAULT**
IST2202I GREXIT = NO WLM = YES LOCLU = YES
IST2204I LOCAPPL = YES PASSOLU = NO
IST314I END
```

## VTAM GR Changes - Implementation

- New GR preferences table capability only applies to z/OS 1.9 and later.
  - Function is provided as part of z/OS 1.9
  - Default table (if you don't specify any parameters) results in same behaviour as previous releases using defaults (that is, you weren't using the ISTEXCGR exit to change or override anything).
- Not necessary to have all systems using R9 - can be implemented on system-by-system basis and only affects that system

## VTAM GR Changes - Documentation

### ▪ More information:

- For further information on using this new capability, refer to the section entitled "Initiating sessions using the generic resource name" in *z/OS Communications Server: SNA Network Implementation Guide* and the section "Generic resources preference table" in *z/OS Communications Server: SNA Resource Definition Reference*
- See Doris' Comms Server session later/earlier in this week.

## Sysplex-related IBM service offerings

**Redbooks Workshop**

IBM ITSO - International Technical Support Organization



## Parallel Sysplex Training Environment

### Background

- As part of an availability review, an IBM customer requested that IBM provide some mechanism for providing their operators and Systems Programmers with the skills required to manage a Parallel Sysplex - a "Flight Simulator" for z/OS!
- To determine the interest in such an offering, IBM surveyed a number of large customers. 75% of respondents said that IBM should provide something. All said they currently have no dedicated training environment.

## Background

- **The survey participants suggested several possibilities, but the common thread was that people wanted:**
  - An environment where operators and system programmers could do **destructive testing**.
  - IBM-provided **workloads** to create a realistic environment.
  - **Documentation** specifically for the students, to guide them through the tasks they should be familiar with.
  - An environment that could be **easily upgraded and maintained**.

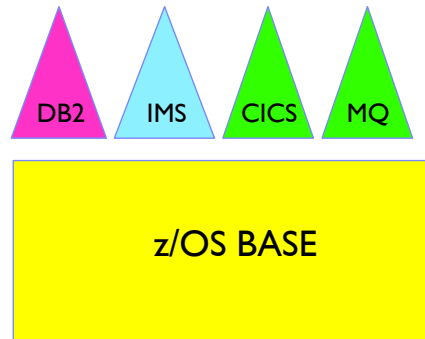
## The result

- **An offering called the Parallel Sysplex Training Environment which consists of:**
  - A load-and-go Parallel Sysplex that can be installed on customers' existing hardware, in LPARs or under z/VM
  - Batch-driven workloads to create a realistic environment.
  - An Installation Guide that helps plan for and install the Trainer.
  - An Exercise Guide covering planned and unplanned scenarios for all CF exploiters.
  - Packaged with 1-day on-site initial education.
  - A subscription service that provides annual updates. New levels include the latest levels of z/OS, CICS, DB2, IMS, and MQ, updated to add workloads for any new CF exploiters.

## Software platform

### Includes:

- z/OS Base
- CICS TS Base \*
- IMS/TM Base \*
- DB2 Base \*
- MQ Series\*
- \* Optional and Self contained



#### Exploiters:

- ▶ XCF Signaling
- ▶ GRS Star
- ▶ JES2 Checkpoint
- ▶ OPERLOG
- ▶ LOGREC
- ▶ Enhanced Catalog Sharing
- ▶ RACF
- ▶ VTAM GR
- ▶ TSO/E GR
- ▶ Unix File System Sharing

#### Planned configuration changes:

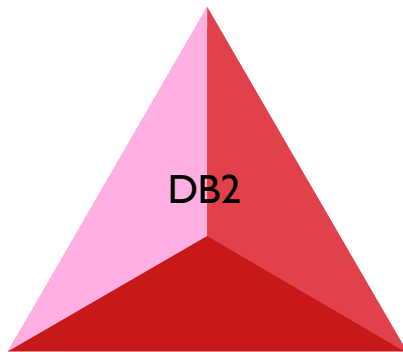
- ▶ Starting use of the structure
- ▶ Disabling the exploiter
- ▶ Rebuild structure
- ▶ Resize structure

#### Unplanned configuration changes:

- ▶ Loss of a CF link
- ▶ Loss of a CF
- ▶ Loss of a system
- ▶ SFM partitioning

#### System capabilities exercised:

- ▶ Sysplex Failure Management
- ▶ Automatic Restart Manager



For each of the Lock, SCA, and Group Buffer Pool structures:

- ▶ Starting the use of the structure
- ▶ Stopping the use of the structure
- ▶ Changing the structure size
- ▶ Moving the structure to another CF using Rebuild
- ▶ Recovering from a CF failure
- ▶ Recovering from a CF Link failure with SFM active
- ▶ Recovering from a CF Link failure without SFM active
- ▶ Recovering from a system failure

In addition, for the Group Buffer Pools:

- ▶ Starting the use of User-managed duplexing
- ▶ Stopping the use of User-managed duplexing

## Workloads

- **To provide a realistic environment for testing, workloads are provided to drive activity to each CF structure**
  - Workloads are generally batch-driven, for simplicity
  - Workloads are self contained - for example, the CICS workload only requires CICS - no need for DB2 or IMS. Similarly for the other workloads.

## Documentation

### ■ Installation Guide:

- Introduction to the Trainer
- Basic instructions on how to install, download, and refresh the environment. Contains information about the IODF, hardware configuration and other prerequisites
- Sample restore and initialization jobs, sample VM definitions

## Documentation

### ■ Exercise Guide:

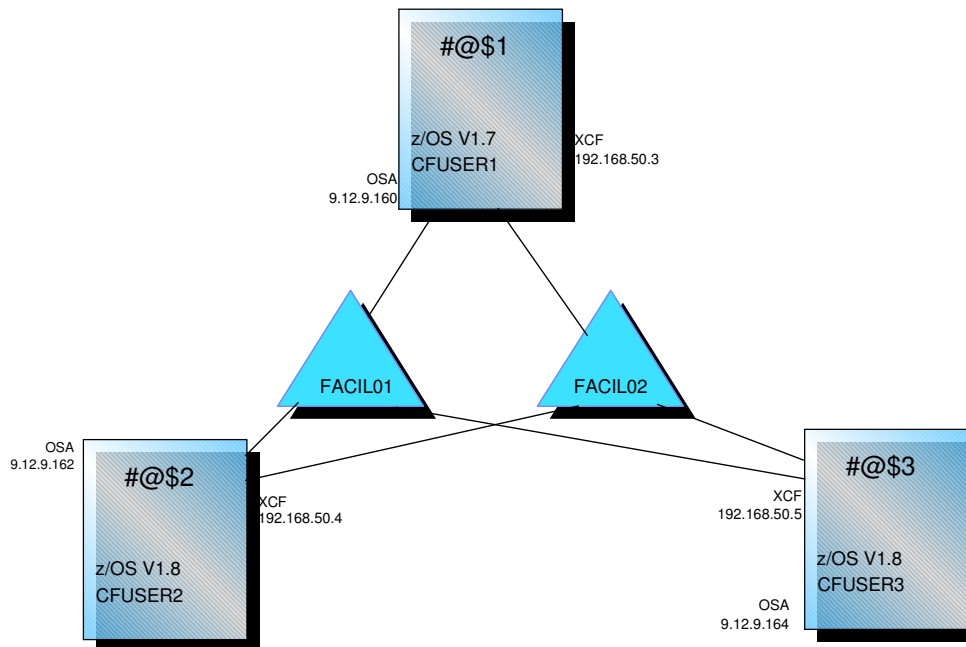
- Starts with introduction to sysplex concepts and components
- Remainder is broken in many parts, one part per CF structure
- Each part is broken into an introduction to how the exploiter uses the CF, followed by a series of exercises

## System environment

- System comes with 50 operator and system programmer ids already defined. Initial passwords are provided, but must be changed the first time they are used.
- National characters are used in volume serial numbers, userids, system names, subsystem names, and data set names.
  - Designed to avoid clashes with existing volumes
  - Designed to avoid mistakes by inadvertently entering a command at the wrong system.

## Hardware requirements

- **DASD:** Max of 25 volumes\*
- **Processor Storage:** c. 1 GB
- **Processor MIPS:** Minimal when not in use
- **Tape:** For install only
- **Console controller:** For LPAR install only (not needed when installing under VM)
- 
- **Option to use z/VM.....**
- **\* DASD space depends on options selected**



## Products provided

- Latest two releases of the operating system - currently z/OS V1.7 and 1.8
- "Current" level of IMS - V9
- "Current" level of DB2 - V8
- "Current" level of CICS - TS 3.1
- "Current" level of MQ - V6
- New release delivery aimed at matching releases installed in leading customers
  - No point delivering on GA date as no one has that level yet
- SMP/E environment provided, to facilitate installation of other products if you wish

## Installation

- **Installation Guide describes installation and customization tasks**
- **Install under VM requires NO further customization and can be achieved in about 4 hours (including coffee break time!)**
  - Install in LPARs takes longer, but should still be possible in one day

## Customization flexibility

- **The delivered environment is a standard z/OS system which can be customized just like any z/OS system.**
- **The SMP/E environment is supplied to allow you to install other IBM or non-IBM products if you wish.**
- **However...**
  - The more customization you do, the longer it takes to install a new release.
  - It is up to each customer to customize as you wish.
  - Customizing may impact the supplied workloads.
  - Each install will completely replace the system, replacing all your customization.



## Software licencing considerations

- **The Parallel Sysplex Training Environment provides z/OS base with all components.**
  - Enable/Disable based on your licenses as you do for any ServerPac install
- **Other products (CICS, DB2, IMS, MQ) optional - you just select the products you are licensed for.**
- **Licensing options:**
  - If run on an existing CPC where all products are already licensed, no additional licences required
    - But remember z/VM license

## Delivery

- **Delivery is through IBM IT Education Services**
- **Packaged with ITES education**
  - Once-off, onsite to ensure the offering is installed and working OK and to work through some sample exercises
- **Sample installation material available - VM directory entries, IPL execs, DSS restore jobs, and so on**

## Is this a solution or a tool?

- **The direction for Parallel Sysplex management is twofold:**
  - Automate as much as possible, as close as possible (autonomics).
  - Provide highly skilled staff to handle the 1% of exceptions - **the need for trained humans to control the system/sysplex will NEVER disappear completely.**
- **We can provide the infrastructure, the education, and even the measurement mechanism (Certification), but only the customer can provide staff with the incentive to use these.**
- **However, the benefits can be significant:**
  - Better availability
  - Better motivated Operators
  - Reduced System Programmer workload/interruptions
  - Provide an in-house breeding ground for future System Programmers

## Other benefits

- **Gives System Programmer access to latest releases much faster than would normally be the case, and with minimal effort on their behalf.**
- **Provides a real working Parallel Sysplex that can be used to identify new ways the installation can extend their sysplex exploitation.**
- **Provides actual working examples and the jobs that were used to set up the environment.**
- **Can help identify potential problems before they hit the production environment.**
- **Provides working examples of new technology, for example, DVIPA, UNIX File System Sharing, and many others**

## Further information

- **If you would like more information, contact your local ITES representative, or send an email to [kyne@us.ibm.com](mailto:kyne@us.ibm.com)**

## IBM GTS Offerings

- **IBM Global Technology Services have a range of new sysplex offerings**
- **Aimed at both experienced sysplex customers, as well as those new to sysplex**
- **Eight modules, each separately orderable**
- **Fixed price for a defined piece of work**
- **Exploit IBM's best practices, experiences, expertise, and methodologies developed from working with many sysplex customers all over the world**

## IBM GTS Offerings

### ▪ **Module I : z/OS & Sysplex Infrastructure Review**

- Understand z/OS and SYSPLEX infrastructure requirements
- Identify required infrastructure configuration data to be collected for analysis
- Perform high level assessment of current z/OS and SYSPLEX infrastructure
  - z/OS Software Currency
  - z/OS Maintenance Strategy
  - Identify level of z/OS component functional exploitation
  - z/OS and SYSPLEX availability single point of failure
- Document observations and recommendations
- Conduct assessment planning and review meetings

## IBM GTS Offerings

### ▪ **Module 2: z/OS & SysplexIBM GTS Offerings Infrastructure Planning, Design and Implementation**

- Understand z/OS and SYSPLEX infrastructure requirements
- Identify or develop high level z/OS and SYSPLEX system standards
- Identify required customization data to be collected for system build
- System design based on requirements with optimal system layout
- Conduct SystemPac order and customization data collection review meeting
- SystemPac build and test off-site
- SystemPac install at customer's site with best practices samples
- Conduct planning and design review meetings

## IBM GTS Offerings

### ▪ **Module 3: Detailed z/OS Sysplex Availability and Performance Assessment**

- Understand customer's z/OS and SYSPLEX infrastructure availability and performance requirements
- Identify required infrastructure configuration data to be collected for analysis
- On-site data collection
- Perform detailed assessment of current infrastructure (off-site)
- Document observations and recommendations
- Conduct planning and review meetings

## IBM GTS Offerings

### ▪ **Module 4: z/OS & Sysplex Configuration & Migration**

- Understand z/OS Parallel SYSPLEX configuration requirements
- Review customer's SYSPLEX naming standards
- On-site Parallel Sysplex Customization assistance
  - Design COUPLExx member
  - Set up environment for HFS sharing
  - Infrastructure enablement for data sharing (optional)
  - HFS to zFS migration (optional)
  - JES2 exit migration (optional)
- Document all configuration changes
- Conduct planning and review meetings

## IBM GTS Offerings

### ▪ **Module 5: z/OS Sysplex Operator Training**

- Assess current experience level of operator staff (Basic, intermediate, Advanced)
- Identify areas of concern, or focus. (IPL, Shutdown, Problem determination and error recovery)
- Develop presentation material based on customer's requirements and system
- Conduct on-site Sysplex Operator training session(s)

## IBM GTS Offerings

### ▪ **Module 6: z/OS Sysplex Maintenance Strategy**

- Identify customer's business requirements that may influence software management strategy
- Develop proposed z/OS maintenance strategy based on
- Business Requirement "influencers"
- IBM recommended maintenance strategy
- Document proposed maintenance strategy
- Conduct planning and review meetings

## IBM GTS Offerings

### ■ **Module 7: Sysplex maintenance environment Design and Implementation**

- Conduct planning meeting to review z/OS and SYSPLEX naming system standards and maintenance strategy
- Design maintenance environment
- Implement maintenance environment
- Set up SMPE environment to utilize "RECEIVE ORDER" functions
- Set up EPSPT (Enhanced Preventive Service Planning Tool)
- Set up automated process to run SMPE ERRORSYSMOD reports
- Assist with roll out of first z/OS cloned image
- Document process and procedures

## IBM GTS Offerings

### ■ **Module 8: z/OS Sysplex production cutover and Deployment Planning**

- Review z/OS SYSPLEX configuration
- Review Production Cut-over & Deployment requirements
- Review Applications Testing requirements
- Identify constraints and available resources
- Review / Develop cut-over and deployment strategy (including SYSPLEX rolling IPLs) and back out considerations
- Review Testing Plan Strategy
- Develop project plan
- Conduct planning and review meetings
- Document plan and processes

## System z Hands-on facilities for IBMers

- **In order to maintain their value to customers, it is important that IBMers keep their System z skills up to date.**
  - There is also a need for an environment where working examples of products can be accessed
- **One option is the Demo facility in Dallas. Contains most major subsystems - CICS, DB2, IMS, MQ, WAS, Tivoli**
  - For complete list see <http://w3.demopkg.ibm.com/LPage/DCnew091207>
  - Systems also contain scripted demos
- **IBMers can get a userid at no charge and access system via Intranet**
  - Go to <http://w3.demopkg.ibm.com/LPage/DNDEMOMVSACCESS> to request a TSO ID

## System z Personal Development Tool

- **An alternative to DemoMVS is to run z/OS on your laptop, desktop, or pSeries**
  - IBM announced zPDT in April 2007
  - Provides about 25 MIPS, supports up to 2GB of z/OS memory
  - Needs fast PC, at least 3 GB of memory
  - Emulated I/O
- **Runs z/OS under Linux on selected platforms**
- **Only available to IBMers**
  - For personal education and hands-on experience
  - For demoing products and solutions to customers



## System z Personal Development Tool

- **Runs current z/OS, z/VM, z/VSE, Linux for System z**
  - Prebuilt packages available for z/OS, z/VM
  - Compilers, CICS, DB2, TSO, TCP/IP, MQ, etc, etc
  - No WAS at this time
  - OSA, QDIO, full System z instruction set
  - Coming soon: zIIPs, zAAPs, ...
- **For more information, see:**
  - <http://w3.demopkg.ibm.com/LPage/DPSYSTEMZPDT>
- **For ordering instructions, see:**
  - <http://w3-03.ibm.com/support/techdocs/atmastr.nsf/WebIndex/TD104025>

## Considerations for System Logger and DFSMSHsm

## System Logger and DFSMSHsm

- **Some customers have encountered intermittent contention between HSM and System Logger.**
  - If HSM is backing up offload data sets, it will issue ENQ against each data set as it backs it up.
  - If Logger tries to browse or mod-on to the data set during the backup, you get:

```
IEF196I IEC161I 052 (015,DFHSM)-084, IEESYSAS, IXGLOGR, SYS00047,,,
IEF196I IEC161I IXGLOGR.SYSPLEX.LOGREC.ALLRECS.A0000000,
IEF196I IEC161I IXGLOGR.SYSPLEX.LOGREC.ALLRECS.A0000000.D,LOGCAT.USERCAT
IEC161I 052 (015,DFHSM)-084, IEESYSAS, IXGLOGR, SYS00047,,,
IEC161I IXGLOGR.SYSPLEX.LOGREC.ALLRECS.A0000000,
IEC161I IXGLOGR.SYSPLEX.LOGREC.ALLRECS.A0000000.D,LOGCAT.USERCAT
```

```
IXG268I LOGSTREAM DATASET IXGLOGR.SYSPLEX.LOGREC.ALLRECS.A0000000
CAN NOT BE OPENED FOR JOB READLOG DUE TO
INCORRECT VSAM SHAREOPTIONS OR OTHER ERROR, REQUESTED DATA MAY NOT BE
AVAILABLE.
```

```
READLOG Logstream BROWSE Failed RetCode= 00000008 RsnCode=0000084A
```

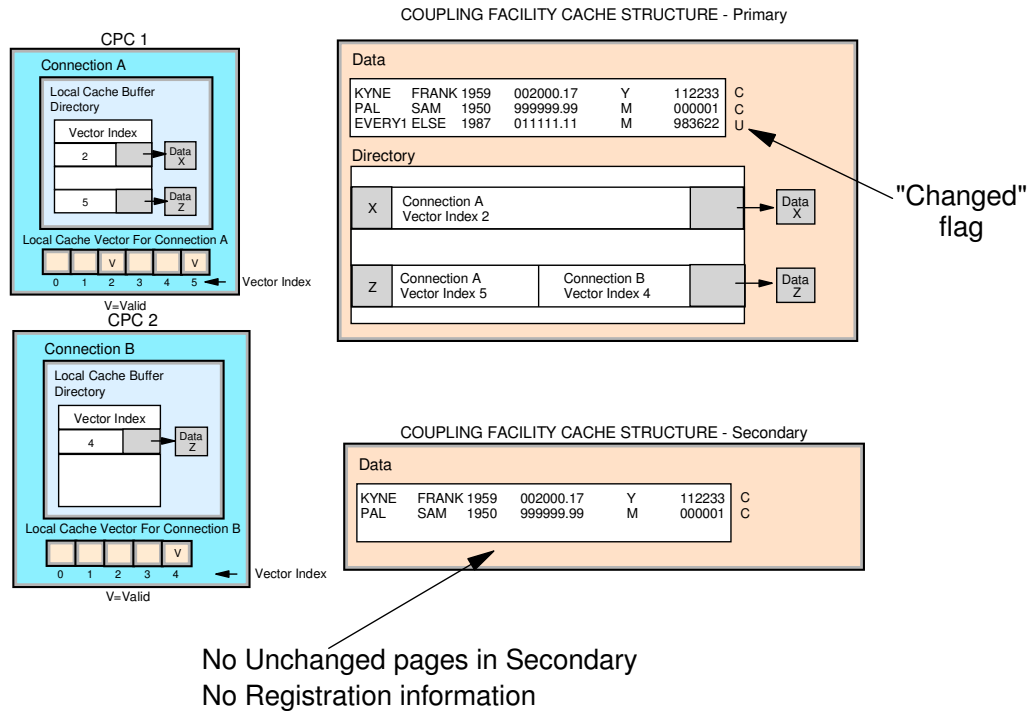
## System Logger and DFSMSHsm

- **If this is an issue for you, you can create a HSM ARCBDEXT exit that will stop HSM issuing ENQs against System Logger offload data sets (but it should still issue the ENQ against staging data sets):**
  - Logger only mods on to the end of Offload data sets, so the HSM backup will still be valid (if old)
  - But for the Staging data sets, Logger can write anywhere within the data set, so backing up without an ENQ could result in an inconsistent backup.
    - Staging data sets are ENQed at log stream connect time and ENQ is held until disconnect, so less likely to encounter Logger processing failing because of the ENQ
      - But the HSM backup might fail because it can't get the ENQ

## Change in GBP cross-invalidation process in DB2 V9

And on the topic of DB2 and duplexed GBPs.....

- **The design of DB2 use of Duplexed GBPs structures is that information about which DB2 has a copy of which pages from the shared databases (known as registration) is only kept in the primary GBP structure**



## Duplexed GBPs in DB2 V8 and V9

- **Registration information is used to drive Cross-Invalidates (XIs) when one DB2 has a local copy of a page that was updated by another DB2 instance.**
- **Changed pages (those that have not yet been hardened to DASD) are kept in both structures, so if either structure is lost, no recovery should be required.**
- **The registration information is only intended to be kept in one instance (the Primary) ...**
  - If the secondary instance is lost, there is no impact on the ability of the remaining CF to know who has an in-storage copy of what data
  - If the primary structure is lost, XES invalidates all local DB2 buffers affected by the failure, but only impact is reduced performance until buffers are re-populated with active data.

Tells us this is "old" instance of duplexed structure

STRUCTURE NAME = DSNREP0_GBP1 TYPE = CACHE STATUS = ACTIVE PRIMARY												
SYSTEM NAME	# REQ	TOTAL AVG/SEC	# REQ	REQUESTS			REASON	# REQ	DELAYED REQUESTS			AVG TIME (MIC) /ALL
				% OF ALL	-SERV TIME (MIC) AVG	STD_DEV			# REQ	% OF REQ	/DEL	
FK0A	716K	SYNC	3372	0.1	97.3	44.7	NO SCH	830	0.1	4287	9657	5.0
	397.7	ASYNC	712K	15.2	177.6	780.9	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	825	0.0	INCLUDED IN ASYNC		PR CMP	0	0.0	0.0	0.0	0.0
FK0C	300K	SYNC	8475	0.2	27.5	13.9	NO SCH	45	0.0	2321	1525	0.3
	166.8	ASYNC	292K	6.2	130.1	654.7	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	45	0.0	INCLUDED IN ASYNC		PR CMP	0	0.0	0.0	0.0	0.0
FK0K	1238K	SYNC	5975	0.1	102.1	45.2	NO SCH	351	0.0	12205	38819	3.5
	687.6	ASYNC	1231K	26.3	190.9	1081.8	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	351	0.0	INCLUDED IN ASYNC		PR CMP	0	0.0	0.0	0.0	0.0
...												
TOTAL												
	4690K	SYNC	759K	16.2	39.2	16.6	NO SCH	1660	0.0	7475	32924	2.6
	2605	ASYNC	3929K	83.8	162.8	840.9	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	1655	0.0			PR CMP	0	0.0	0.0	0.0	0.0
											--- DATA ACCESS ---	
											READS 955678	
											WRITES 445357	
											CASSTOOTS 519937	
											XI'S 193633	

How many times a local buffer was invalidated

Tells us this is "new" instance of duplexed structure

STRUCTURE NAME = DSNREP0_GBP1 TYPE = CACHE STATUS = ACTIVE SECONDARY												
SYSTEM NAME	# REQ	TOTAL AVG/SEC	# REQ	REQUESTS			REASON	# REQ	DELAYED REQUESTS			AVG TIME (MIC) /ALL
				% OF ALL	-SERV TIME (MIC) AVG	STD_DEV			# REQ	% OF REQ	/DEL	
MVSA	80117	SYNC	0	0.0	0.0	0.0	NO SCH	24	0.0	830.1	1912	0.2
	44.51	ASYNC	80K	22.6	103.6	155.2	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	0	0.0	INCLUDED IN ASYNC		PR CMP	0	0.0	0.0	0.0	0.0
MVSC	12880	SYNC	0	0.0	0.0	0.0	NO SCH	0	0.0	0.0	0.0	0.0
	7.16	ASYNC	13K	3.6	154.7	169.3	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	0	0.0	INCLUDED IN ASYNC		PR CMP	0	0.0	0.0	0.0	0.0
MVSK	144K	SYNC	0	0.0	0.0	0.0	NO SCH	47	0.0	5454	20772	1.8
	80.23	ASYNC	144K	40.8	117.6	231.3	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	0	0.0	INCLUDED IN ASYNC		PR CMP	0	0.0	0.0	0.0	0.0
...												
TOTAL												
	354K	SYNC	90	0.0	26.6	1.5	NO SCH	74	0.0	5146	20314	1.1
	196.6	ASYNC	354K	100	126.0	604.8	PR WT	0	0.0	0.0	0.0	0.0
		CHNGD	0	0.0			PR CMP	0	0.0	0.0	0.0	0.0
											--- DATA ACCESS ---	
											READS 0	
											WRITES 445304	
											CASSTOOTS 0	
											XI'S 87130	

- This field would normally be 0
- Secondary GBP should only report WRITES - all other fields should be 0s

## Duplexed GBPs in DB2 V8 and V9

- **This change was introduced in DB2 V8 and is addressed in DB2 V9:**
  - Does not have any integrity considerations
  - Is most likely to have a noticeable effect in cross-site data sharing configurations:
    - When updated pages are written to a GBP that contains registration information, the CF write request does not complete until any cross-invalidates have completed
  - Fix not retrofitted to V8

## Value summary

- **Customer value:**
  - Should improve performance in environments with a lot of GBP writes and/or cross-site data sharing configurations.
- **Ease of implementation:**
  - 10 out of 10
    - Improvement comes automatically with DB2 V9. No migration or co-existence considerations