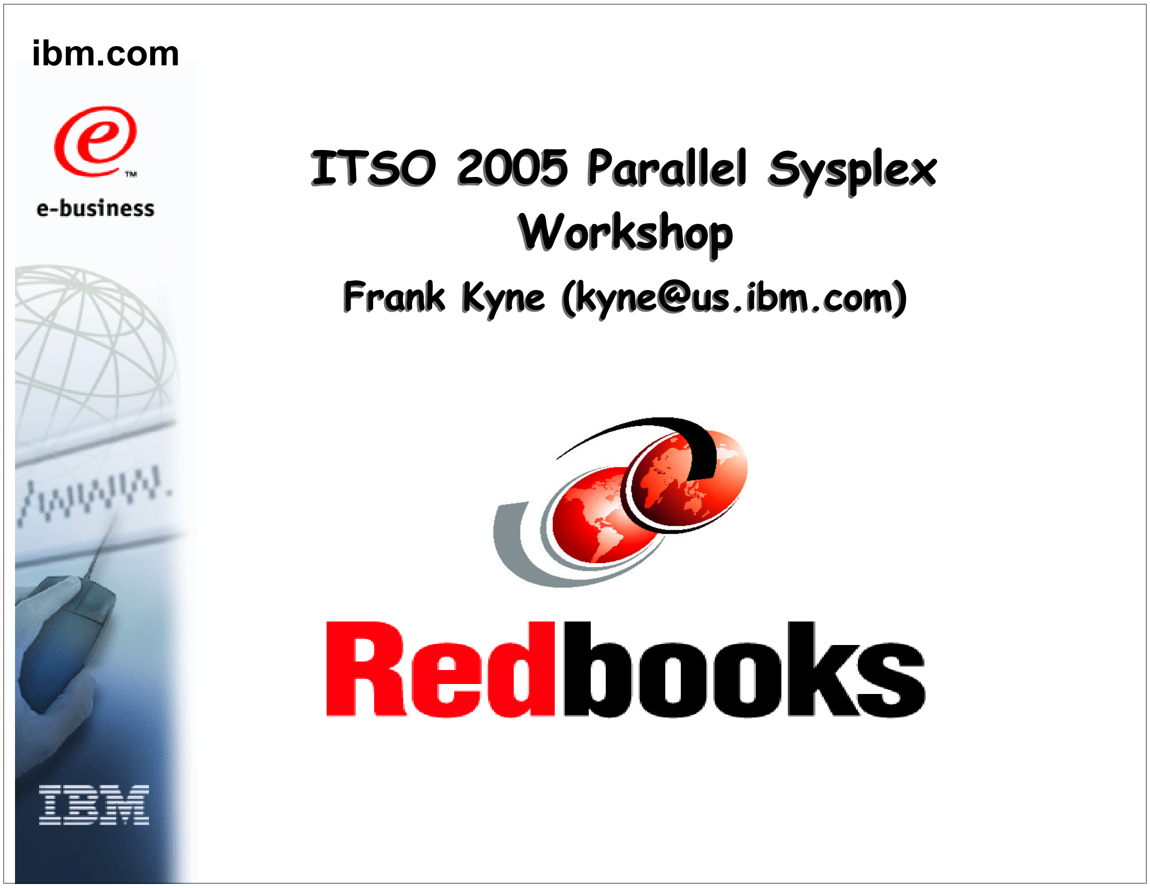
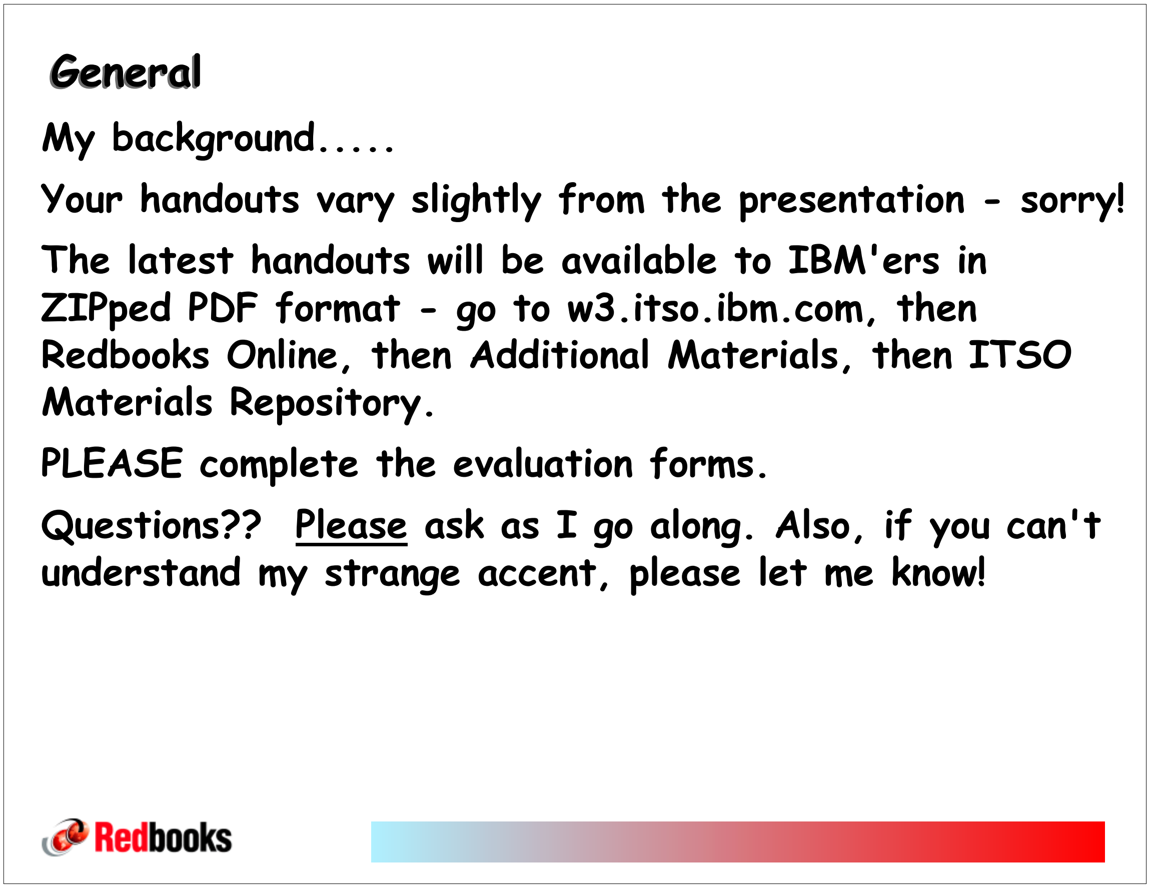


ibm.com

ITSO 2005 Parallel Sysplex Workshop

Frank Kyne (kyne@us.ibm.com)



General

My background.....

Your handouts vary slightly from the presentation - sorry!

The latest handouts will be available to IBM'ers in ZIPped PDF format - go to w3.itso.ibm.com, then Redbooks Online, then Additional Materials, then ITSO Materials Repository.

PLEASE complete the evaluation forms.

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

General

Agenda:

- Start 09:00 (ish!)
- Coffee about 10:30
- Lunch about 12:00 for 2 hours
- Afternoon coffee about 15:30
- Finish about 17:00

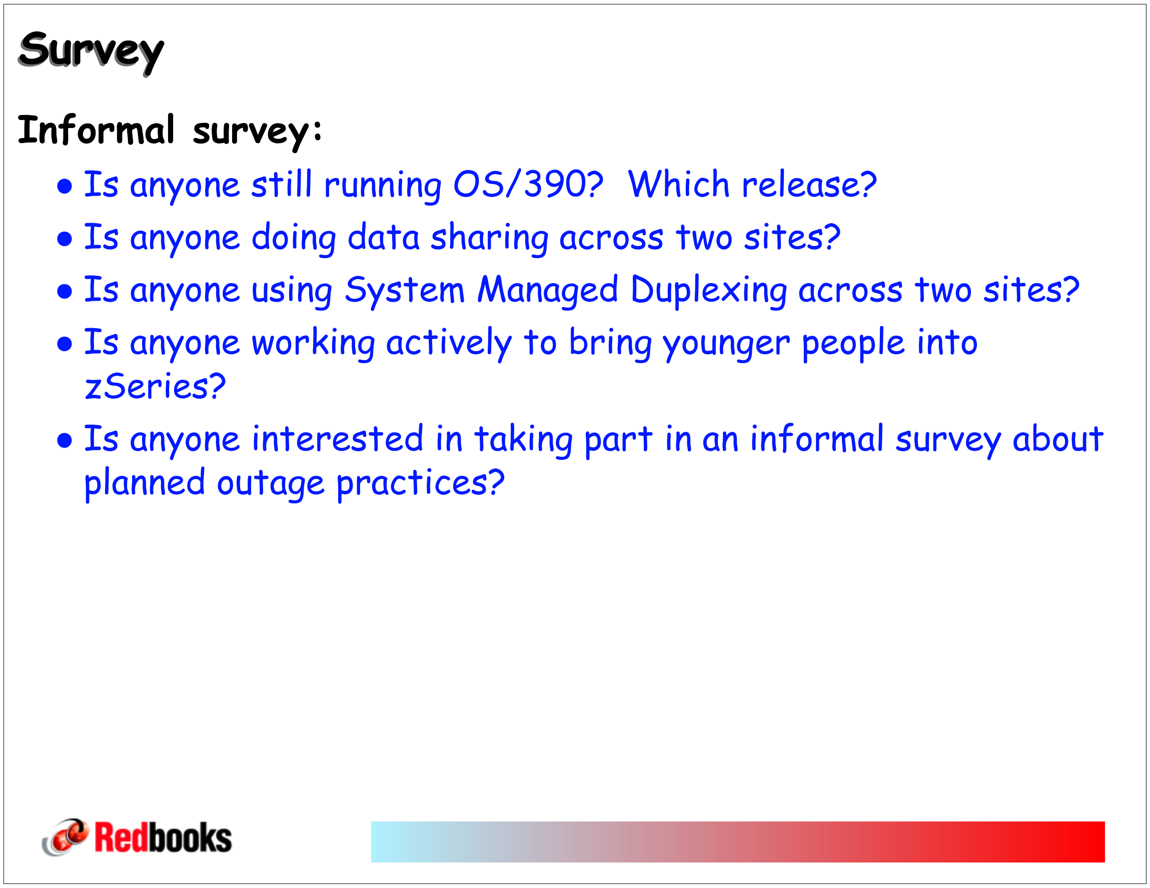
Redbooks

Agenda

Topics:

- Planned outage avoidance
- JES2 scalability considerations
- Sysplex aggregation
- Non-MVS clustering for dinosaurs
- Back to basics - all you never wanted to know about locking in a sysplex
- Bits and bytes
- Automatic Restart Manager (time permitting)

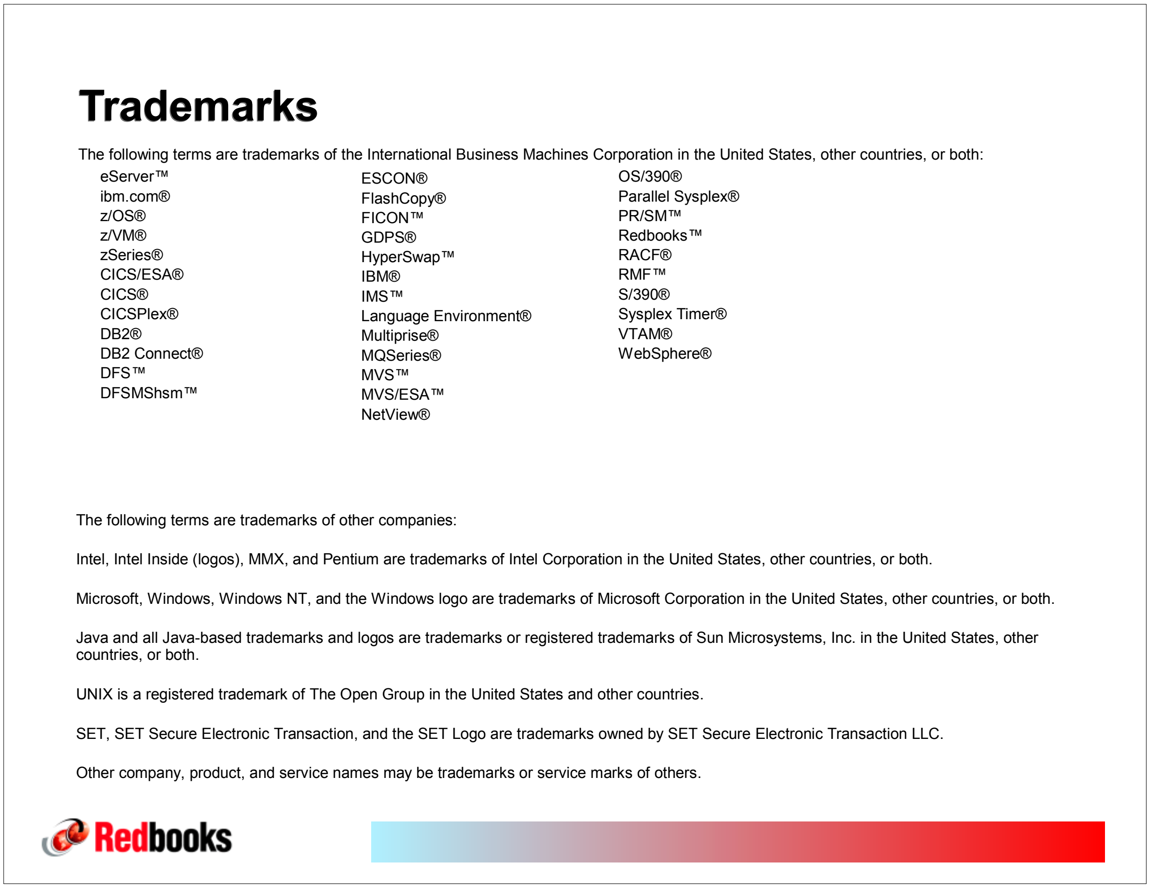
Redbooks



Survey

Informal survey:

- Is anyone still running OS/390? Which release?
- Is anyone doing data sharing across two sites?
- Is anyone using System Managed Duplexing across two sites?
- Is anyone working actively to bring younger people into zSeries?
- Is anyone interested in taking part in an informal survey about planned outage practices?



Trademarks

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

eServer™
ibm.com®
z/OS®
z/VM®
zSeries®
CICS/ESA®
CICS®
CICSPlex®
DB2®
DB2 Connect®
DFS™
DFSMShsm™
ESCON®
FlashCopy®
FICON™
GDPS®
HyperSwap™
IBM®
IMS™
Language Environment®
Multiprise®
MQSeries®
MVS™
MVS/ESA™
NetView®
OS/390®
Parallel Sysplex®
PR/SM™
Redbooks™
RACF®
RMF™
S/390®
Sysplex Timer®
VTAM®
WebSphere®

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.

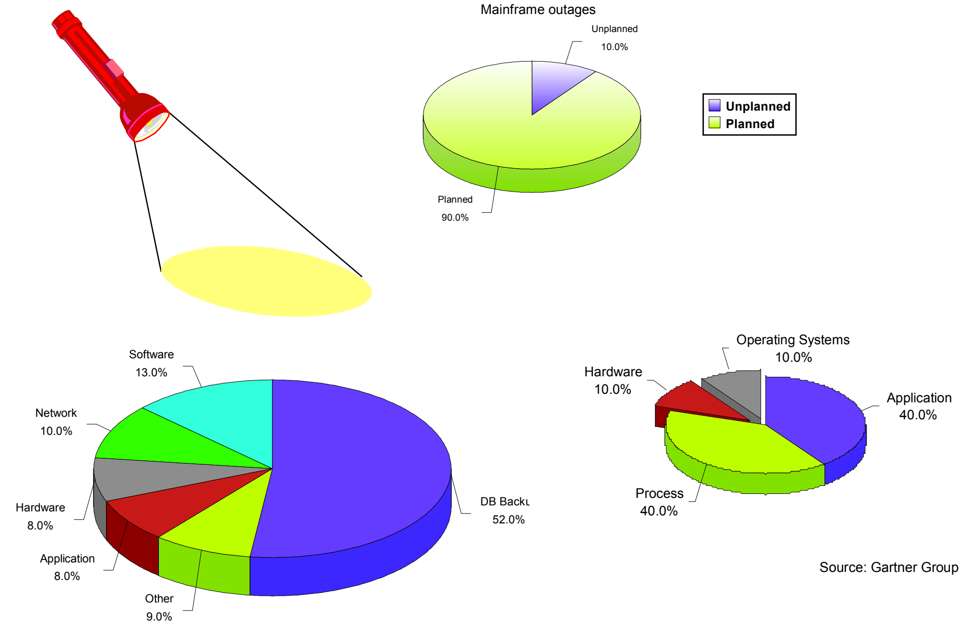
SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

®

Planned Outage Avoidance

Planned outage avoidance



Source: Gartner Group

Planned outage avoidance

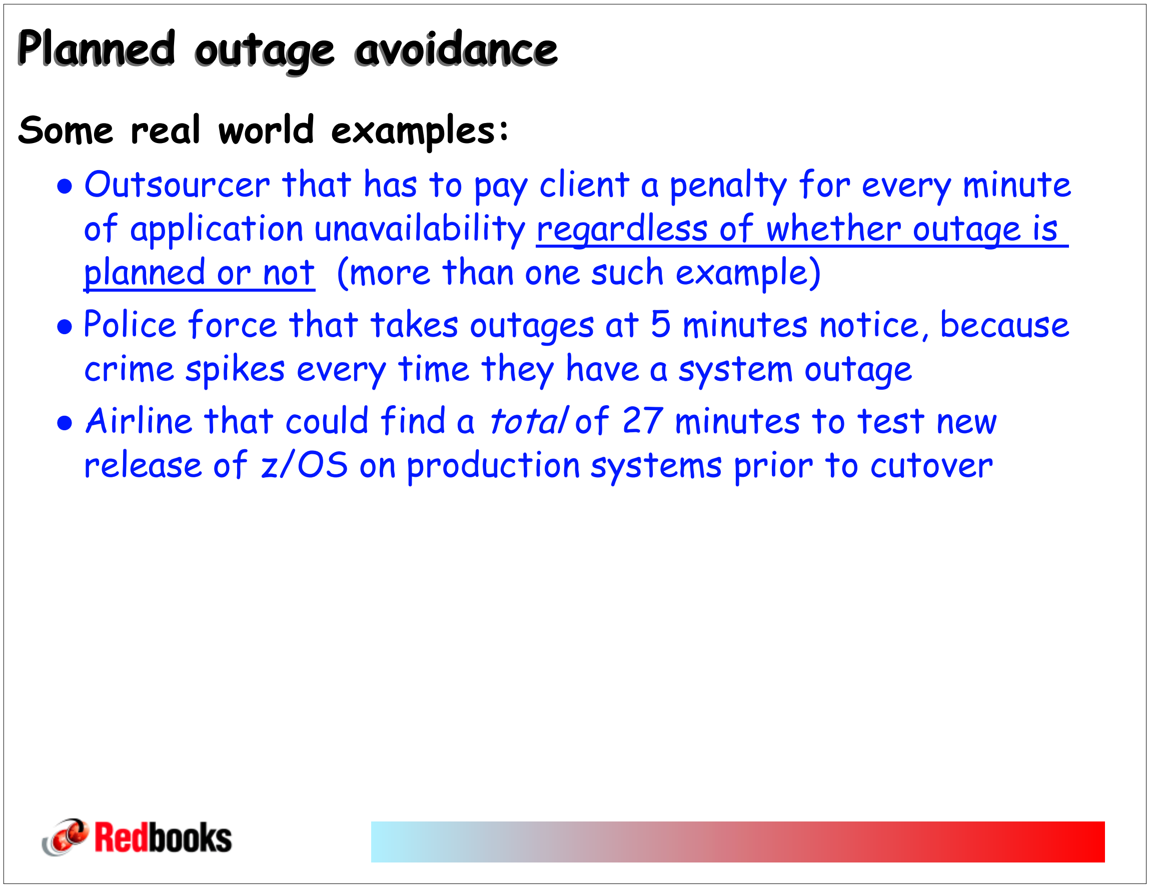
Definition of a "planned outage":

- "An unavailability of service that is known in advance to *all* users and maintainers of that service"
- How do you inform all potential users in advance?
  - Do you even know who all your users are? They are not all 3270 users any more.. What about MQ? DB2 DDF? TSM? LAN Services
  - You may have users on the web, using your system directly (WAS on z/OS) or indirectly (DB2 or IMS as a data server to WAS on another platform)
  - You may have business partners that access your system
  - You may have users in your own company spread out all over the world
- It may not be possible to find a slot when no one wants to use the system
  - A z9 can have nearly 18K+ MIPS and 60 LPARs - will there EVER be a time when <u>someone</u> doesn't want to use it?

Planned outage avoidance

Why is it important to reduce planned outages?

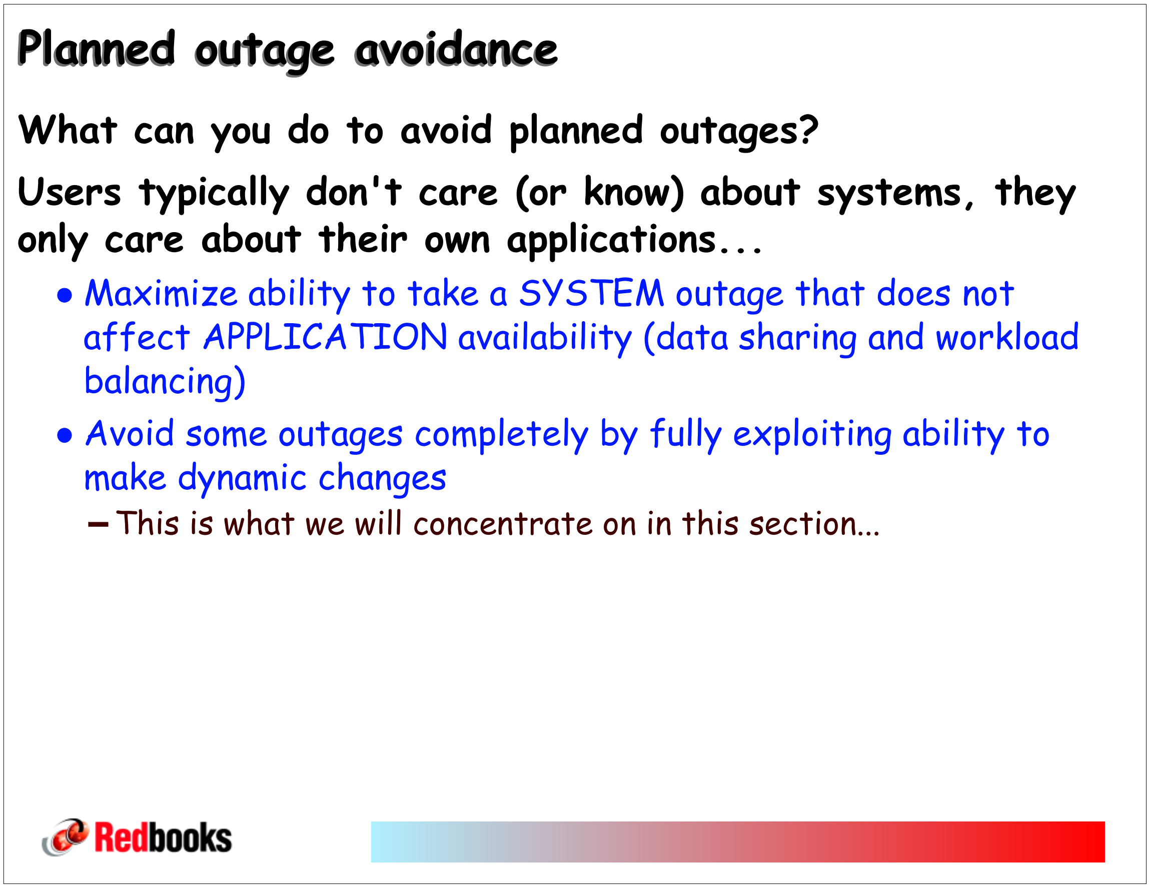
- If you are buying something online and the Web site you are buying from is unresponsive, how long will you wait before taking your service elsewhere?
  - Completely different to traditional in-store buying where you will wait minutes in a queue waiting to be served.
  - Then: Downtime = customer dissatisfaction
  - Now: Downtime = revenue (and maybe a customer) gone to a competitor
- Because the supply of them is very limited, so you need to keep them for when they are REALLY necessary
- Because you are competing against platforms that support a single application and therefore do not have a problem getting planned outages
  - Less users = more outage windows



Planned outage avoidance

Some real world examples:

- Outsourcer that has to pay client a penalty for every minute of application unavailability regardless of whether outage is planned or not (more than one such example)
- Police force that takes outages at 5 minutes notice, because crime spikes every time they have a system outage
- Airline that could find a *total* of 27 minutes to test new release of z/OS on production systems prior to cutover

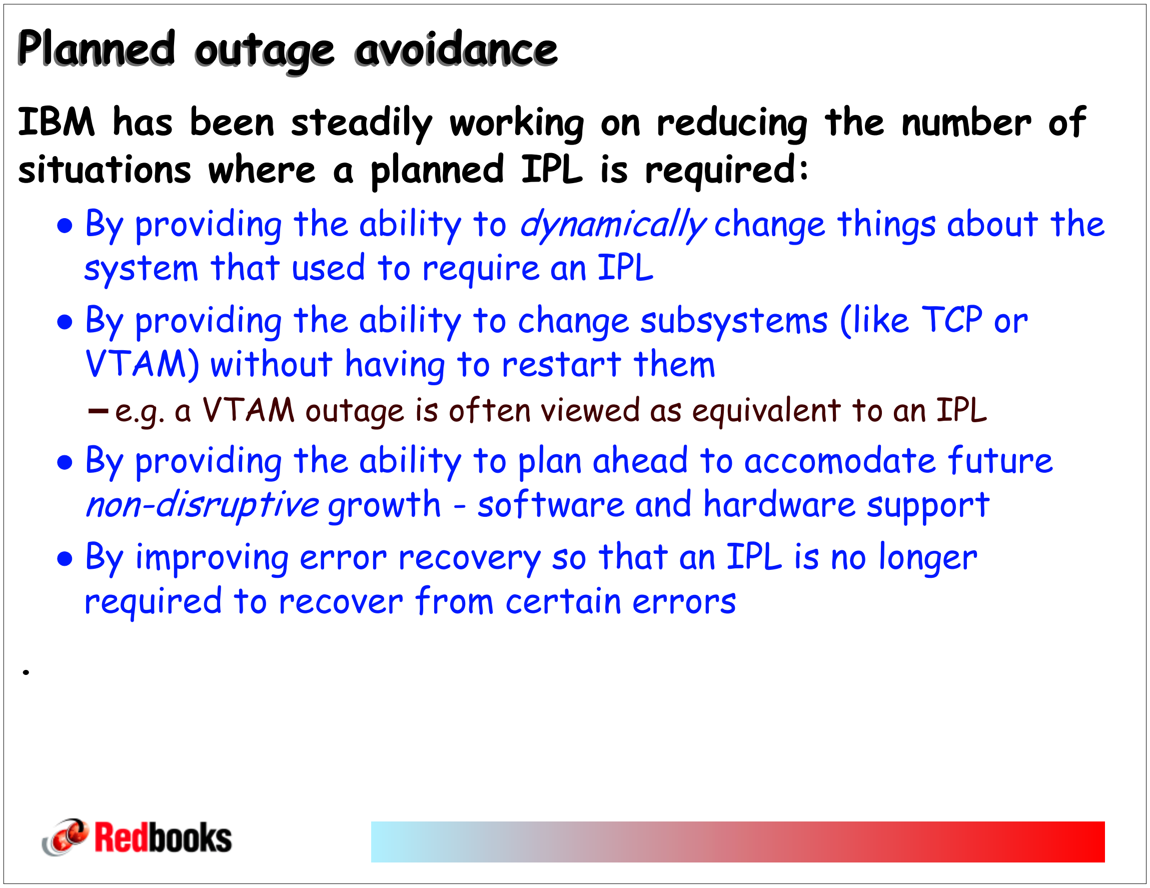


Planned outage avoidance

What can you do to avoid planned outages?

Users typically don't care (or know) about systems, they only care about their own applications...

- Maximize ability to take a SYSTEM outage that does not affect APPLICATION availability (data sharing and workload balancing)
- Avoid some outages completely by fully exploiting ability to make dynamic changes
  - This is what we will concentrate on in this section...

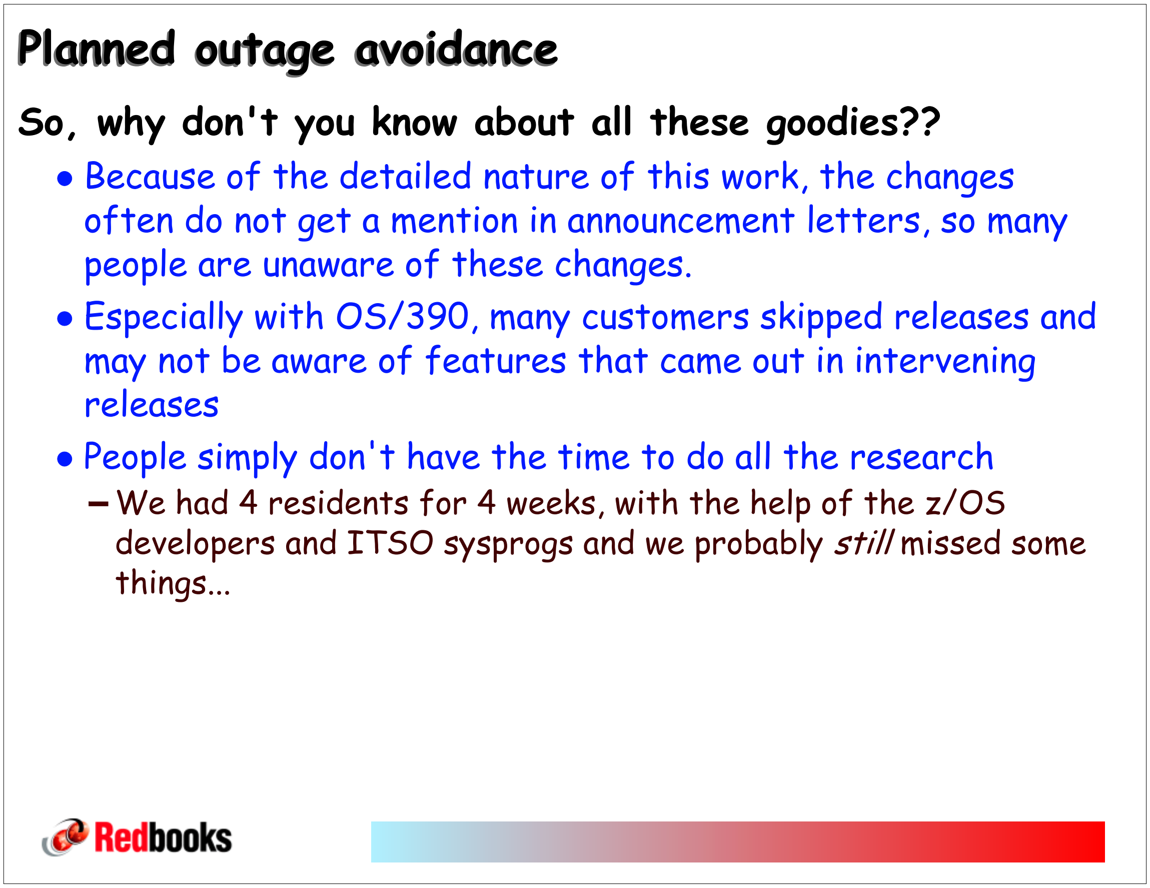


Planned outage avoidance

IBM has been steadily working on reducing the number of situations where a planned IPL is required:

- By providing the ability to *dynamically* change things about the system that used to require an IPL
- By providing the ability to change subsystems (like TCP or VTAM) without having to restart them
  - e.g. a VTAM outage is often viewed as equivalent to an IPL
- By providing the ability to plan ahead to accomodate future *non-disruptive* growth - software and hardware support
- By improving error recovery so that an IPL is no longer required to recover from certain errors

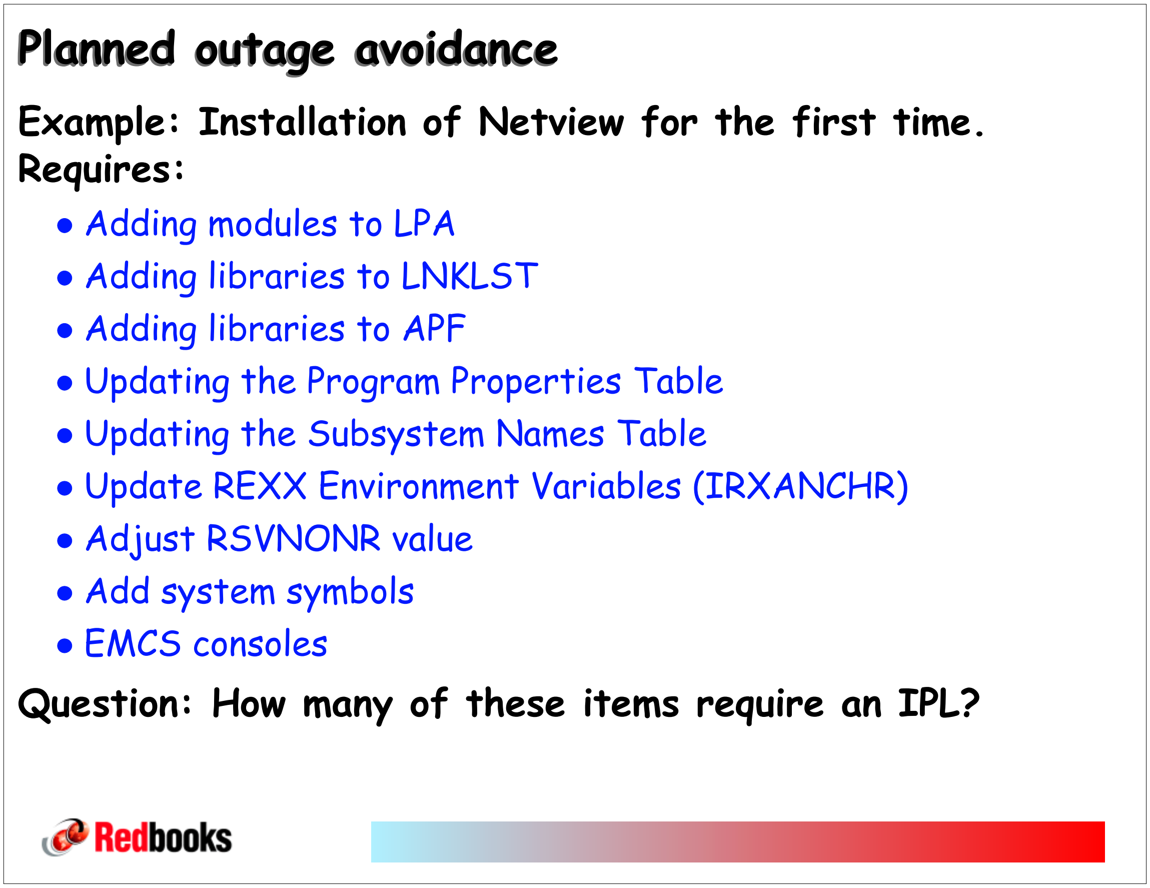
.



Planned outage avoidance

So, why don't you know about all these goodies??

- Because of the detailed nature of this work, the changes often do not get a mention in announcement letters, so many people are unaware of these changes.
- Especially with OS/390, many customers skipped releases and may not be aware of features that came out in intervening releases
- People simply don't have the time to do all the research
  - We had 4 residents for 4 weeks, with the help of the z/OS developers and ITSO sysprogs and we probably *still* missed some things...



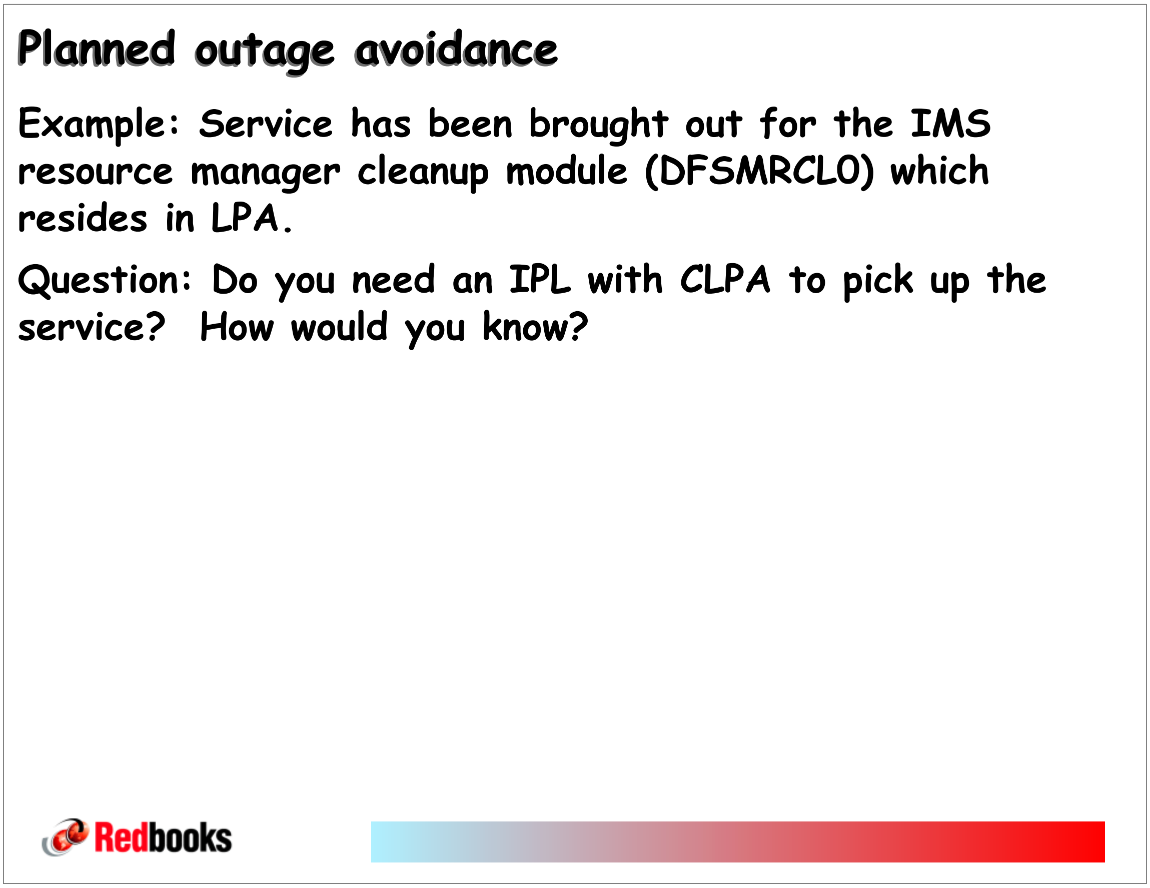
Planned outage avoidance

Example: Installation of Netview for the first time. Requires:

- Adding modules to LPA
- Adding libraries to LNKLST
- Adding libraries to APF
- Updating the Program Properties Table
- Updating the Subsystem Names Table
- Update REXX Environment Variables (IRXANCHR)
- Adjust RSVNONR value
- Add system symbols
- EMCS consoles

Question: How many of these items require an IPL?

Redbooks



Planned outage avoidance

Example: Service has been brought out for the IMS resource manager cleanup module (DFSMRCL0) which resides in LPA.

Question: Do you need an IPL with CLPA to pick up the service? How would you know?

Redbooks

Planned outage avoidance

Example: Service has been brought out for Unix System Services affecting modules that reside in LPA?

Question: To activate this corrective service do you have to wait for the next planned IPL?

Redbooks

Planned outage avoidance

Many customers perform regular, planned IPLs - we call these "Therapeutic IPLs":

- To harden dynamic changes
- To address storage creep/fragmentation
  - You should track and trend and IPL based on actuals
  - If you have a storage creep problem, get it fixed!
- To recover non-reusable ASIDs (IEF352I) or non-reusable LXs
- For operator training (!!)
- To maintain the ability to do an IPL if you really do need it (users get used to the systems being unavailable every nth weekend)
- **Because we have always done it this way!**

Redbooks

Planned outage avoidance

z/OS "things" that can be changed dynamically:

- APF List (SET PROG, SETPROG)
- LNKLST (SET PROG, SETPROG)
- LPA - Add and delete modules
- Exits (SET PROG, SETPROG)
- Subsystems (SETSSI ADD,S=ssn)
- System Symbols (IEASYMUP in z/OS 1.6)
- Number of page data sets (PAGEADD/PAGEDEL)
- PPT (SCHEDxx - SET SCH=xx)
- RACF Started Class, CDT, database templates
- SVCs
- JES2
- Many more - see new Redbook for the exc~~ruc~~i~~a~~ting details!

Redbooks

Planned outage avoidance

This is a two-way session ------ if you are aware of a feature/function/enhancement that I have missed,

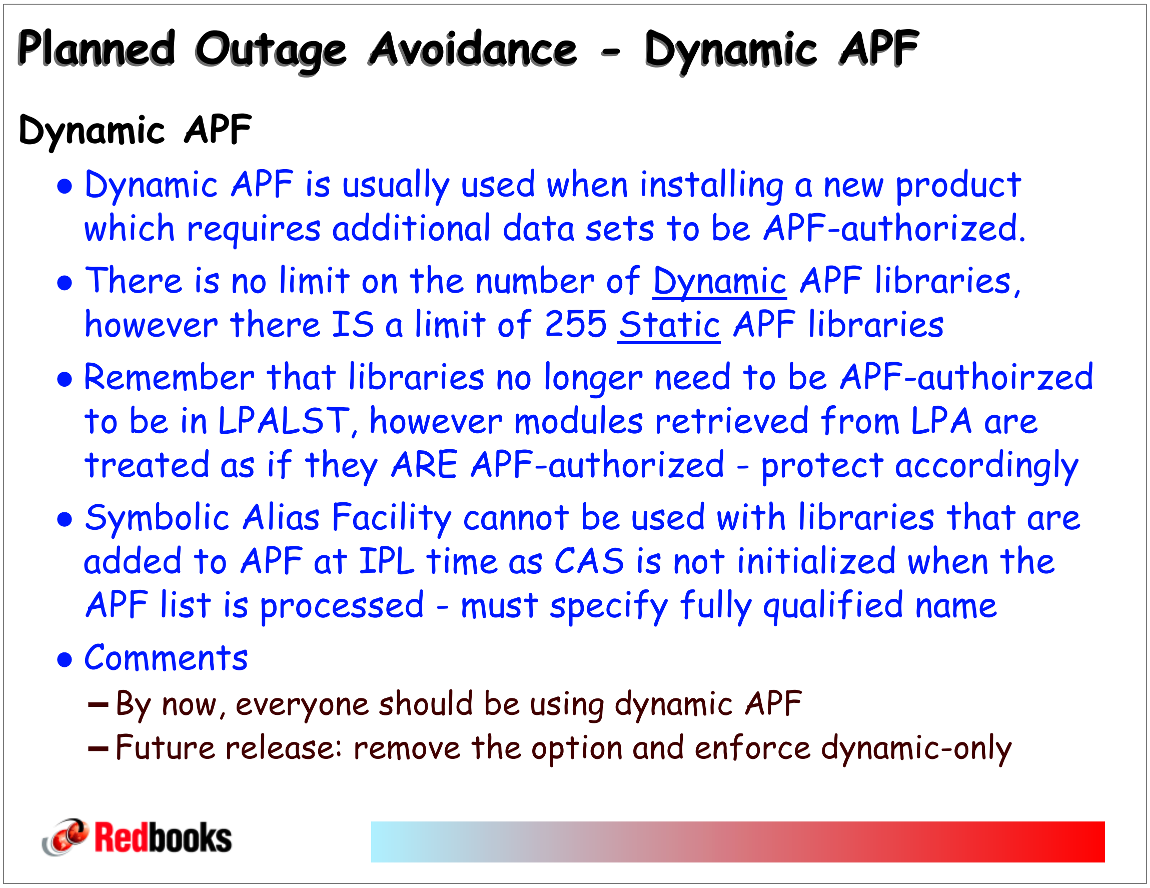
PLEASE

PLEASE

PLEASE

speak up so I can update the presentation and everyone else can benefit...

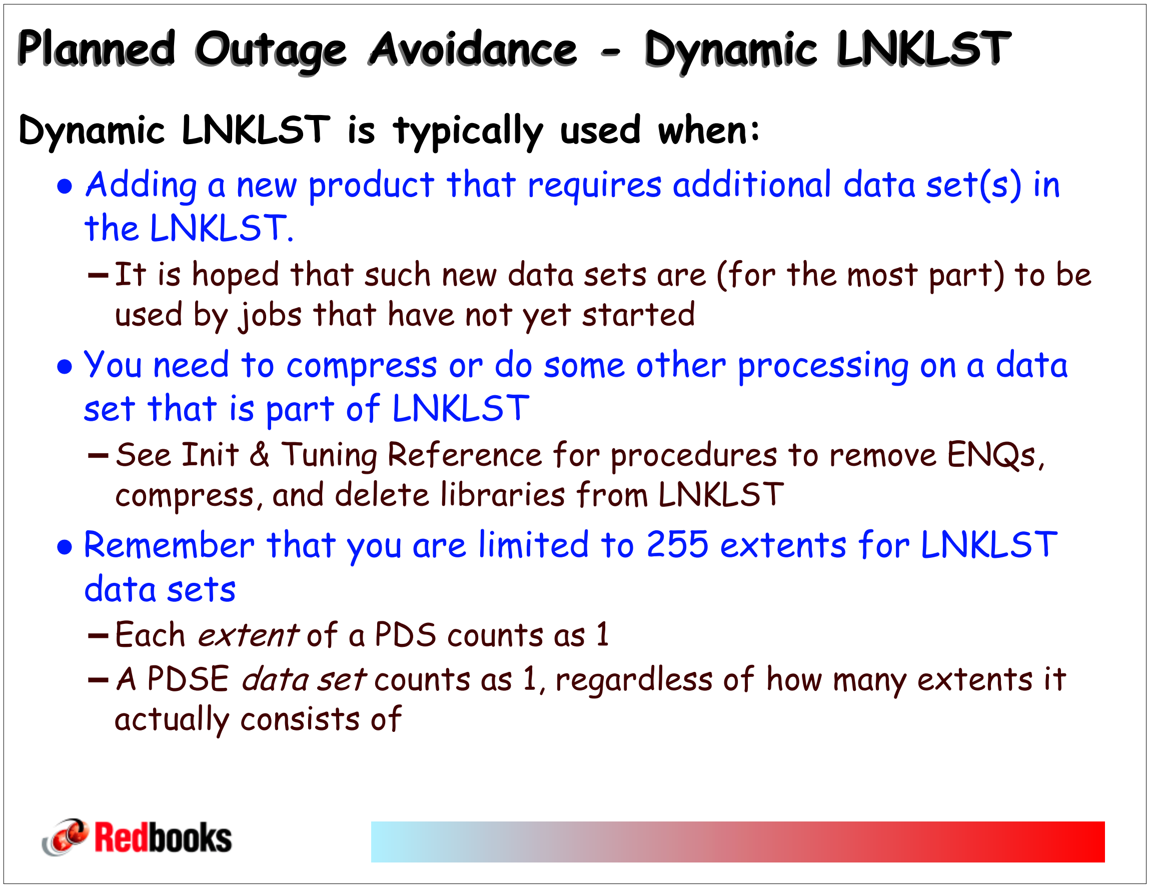
Redbooks



Planned Outage Avoidance - Dynamic APF

Dynamic APF

- Dynamic APF is usually used when installing a new product which requires additional data sets to be APF-authorized.
- There is no limit on the number of Dynamic APF libraries, however there IS a limit of 255 Static APF libraries
- Remember that libraries no longer need to be APF-authoirzed to be in LPALST, however modules retrieved from LPA are treated as if they ARE APF-authorized - protect accordingly
- Symbolic Alias Facility cannot be used with libraries that are added to APF at IPL time as CAS is not initialized when the APF list is processed - must specify fully qualified name
- Comments
  - By now, everyone should be using dynamic APF
  - Future release: remove the option and enforce dynamic-only



Planned Outage Avoidance - Dynamic LNKLST

Dynamic LNKLST is typically used when:

- Adding a new product that requires additional data set(s) in the LNKLST.
  - It is hoped that such new data sets are (for the most part) to be used by jobs that have not yet started
- You need to compress or do some other processing on a data set that is part of LNKLST
  - See Init & Tuning Reference for procedures to remove ENQs, compress, and delete libraries from LNKLST
- Remember that you are limited to 255 extents for LNKLST data sets
  - Each *extent* of a PDS counts as 1
  - A PDSE *data set* counts as 1, regardless of how many extents it actually consists of

Planned Outage Avoidance - Dynamic LNKLST

LNKLST sets:

- Jobs or address spaces continue to use their current LNKLST set until the job ends or the LNKLST set for the job or address space is updated using the UPDATE option.
  - If the new library will only be used by address spaces that have not started yet, this should not be an issue as you do NOT need to use the UPDATE command
    - If you want to use a newly LNKLSTed library with a subsystem added by Dynamic SSI (initrtn=), you MUST do an UPDATE because MASTER must see the new LNKLST in order for the dynamic subsystem add to work - suggest UPDATE,JOB=*MASTER* in this case to minimize the risk
  - There is no situation where it is 100% safe to issue the UPDATE command

Planned Outage Avoidance - Dynamic LPA

Dynamic LPA is typically used when:

- Installing a new product that needs things in LPA
- When a product has items that it needs in common storage that must reside in a PDSE
- Use the SETPROG LPA command to *replace* modules only where the owning product verifies the replacement. Otherwise, replacement could result in partial updates.
  - If the owning product has already saved the module address, the system will NOT conduct an LPA search and will NOT find the updated module.
  - Also, the addresses of all modules that are accessed via a program call (PC) instruction are stored in the PC table. That table is NOT updated by the SETPROG LPA command. Therefore, these modules cannot be replaced using the SETPROG LPA command. You must IPL for the updates to take effect.

Planned Outage Avoidance - Dynamic LPA

```
SETPROG LPA,ADD,MODNAME=(m1,...,mN)|MASK=mmm*,
   DSNAME=d
```

- Add to LPA the named module(s) from the named data set
- Important to add, in same operation, module and **all its aliases**
- Can do an ADD for a module that is already in LPA
  - Modules added to the system by dynamic LPA processing are placed into CSA or ECSA storage. Therefore, it is important to ensure that the system CSA and ECSA sizes are adequately defined to handle the additional consumption of CSA storage resulting from the issuance of the dynamic LPA request. Further protection can be gained through the use of the CSAMIN parameter

Planned Outage Avoidance - Dynamic LPA

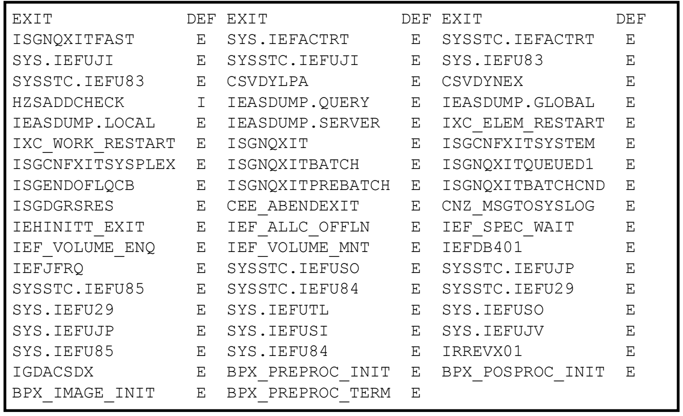
Things not available:

- Safe Delete from LPA (will never be truly safe, but might reject a delete when the use count is not 0)
- Automatically adding ALIASes for LPA ADD
- IPL-time specification of PDSEs for dynamic LPA

Planned Outage Avoidance - Dynamic Exits

Dynamic Exits

- The Dynamic Exits facility (added in MVS 5.1!) lets you associate multiple exit routines with an exit point AND lets you update them dynamically
- Following IBM exits support dynamic exits:



```
EXIT              DEF EXIT              DEF EXIT              DEF
ISGNQXITFAST       E  SYS.IEFACTRT       E  SYSSTC.IEFACTRT    E
SYS.IEFUJI         E  SYSSTC.IEFUJI      E  SYS.IEFU83         E
SYSSTC.IEFU83      E  CSVDYLPA           E  CSVDYNEX           E
HZSADDCHECK        I  IEASDUMP.QUERY     E  IEASDUMP.GLOBAL    E
IEASDUMP.LOCAL     E  IEASDUMP.SERVER    E  IXC_ELEM_RESTART   E
IXC_WORK_RESTART   E  ISGNQXIT           E  ISGCNFXITSYSTEM    E
ISGCNFXITSYSPLEX   E  ISGNQXITBATCH      E  ISGNQXITQUEUED1    E
ISGENDOFLQCB       E  ISGNQXITPREBATCH   E  ISGNQXITBATCHCND   E
ISGDGRSRES         E  CEE_ABENDEXIT      E  CNZ_MSGTOSYSLOG    E
IEHINITT_EXIT      E  IEF_ALLC_OFFLN     E  IEF_SPEC_WAIT      E
IEF_VOLUME_ENQ     E  IEF_VOLUME_MNT     E  IEFDB401           E
IEFJFRQ            E  SYSSTC.IEFUSO      E  SYSSTC.IEFUJP      E
SYSSTC.IEFU85      E  SYSSTC.IEFU84      E  SYSSTC.IEFU29      E
SYS.IEFU29         E  SYS.IEFUTL         E  SYS.IEFUSO         E
SYS.IEFUJP         E  SYS.IEFUSI         E  SYS.IEFUJV         E
SYS.IEFU85         E  SYS.IEFU84         E  IRREVX01           E
IGDACSDX           E  BPX_PREPROC_INIT   E  BPX_POSPROC_INIT   E
BPX_IMAGE_INIT     E  BPX_PREPROC_TERM   E
```

- See MVS/ESA 5.1 Technical Presentation Guide (GG24-4137)

Planned Outage Avoidance - Dynamic SSI

Dynamic SSI:

- Dynamic SSI is used to define new subsystem interfaces (by operator command) without requiring an IPL
- Note that you cannot ADD a subsytem if the initialization routine comes from a library that was dynamically added to LNKLST unless you issue an UPDATE JOB(*MASTER*).
- You can activate and deactivate dynamically added subsystems, but you can't delete them, nor change the attributes you specified on the SETSSI command
- There is no SET SSN command...

Planned outage avoidance   SVCs

Adding/updating SVCs

- There are two aspects to installing a new SVC:
  - Must get the SVC load module into the system
  - Must update the SVC table
- Dynamic LPA should be usable for the former if the module is in LPA (and could be used if it's in the nucleus if the CSECT has no external references, specifying to page-fix the LPA module)
- SVCUPDTE can be used to update the SVC table
- If updating an existing SVC, you must use Dynamic LPA to load the new code, and SVCUPDTE to update the SVC table with the new address.
- Documented in Auth Assembler Services Guide and Auth Assembler Services Reference

Redbooks

Planned outage avoidance   SVCs

Sample program to invoke SVCUPDTE

```
SVCUP    CSECT
SVCUP    AMODE 24
SVCUP    RMODE 24
         USING SVCUP,12
*
         SAVE  (14,12),,&SYSDATE-&SYSTIME-SVCDC
         LR    12,15             BASE.
         LR    14,13             SAVE HI-SAVE PTR.
         LA    13,SAVE           POINT AT LO-SAVE.
         ST    14,4(,13)         CHAIN LO- &..
         ST    13,8(,14)         .HI-SAVE.
         L     2,X'10'(,0)       POINT AT CVT.
         TM    X'74'(2),X'80'    TEST, WHETHER..
         BZ    SVCUPA            .NOT RUNNUNG MVS/XA.
         LA    2,SVCUPA          SET..
         O     2,HBITOW          .31-BIT..
         BSM   0,2               ADDR MODE.
SVCUPA   DS    0H
*
         MODESET KEY=ZERO,MODE=SUP
*
         SVCUPDTE 216,REPLACE,TYPE=3,EPNAME=DFHCSVC
*
         MODESET KEY=NZERO,MODE=PROB
*
         SR    15,15
         L     13,4(,13)
         RETURN (14,12),RC=(15)
         DS    0F
HBITOW   DC    X'80000000'
SAVE     DC    18F'0'
*      THE FOLLOWING MACROS MUST BE INCLUDED IN THE SOURCE PROGRAM   *
*      CVT    - TO MAP THE FIELD CVTNUCLU                            *
*      IHAPSA - TO SUPPLY CVT BASE                                   *
*      NUCLKUP- TO FIND THE SVC UPDATE SERVICE ENTRY POINT (IEAVESTU)*
         PRINT NOGEN
         IHAPSA
         IKJTCB
         CVT   DSECT=YES
         END   SVCUP
```

Redbooks

Planned Outage Avoidance - Page data sets

It is possible to add new LOCAL page data sets to react to an AUX shortage

You can also remove a LOCAL page data set, for example if you need to free up the volume

- Make sure ESQA is large enough - need x'500' bytes per used cylinder in the data set being PAGEDELed during the entire process

Make sure that PAGTOTL is large enough - you can add or delete page data sets dynamically, but an IPL is required to change this value

You CANNOT add or delete COMMON or PLPA data sets

msys for Ops includes automation to PAGEADD on AUX shortage msgs...

Planned Outage Avoidance - PPT

It is possible to add or change Program Properties Table entries dynamically:

- Create or update the SCHEDxx member with your definitions.
- z/OS comes with a number of system entries that are automatically merged with the SCHEDxx entries at IPL or when you issue SET SCH=xx
  - If there is a clash, SCHEDxx specification overrides.

Note that only PPT entries are read from SCHEDxx when you do a SET SCH - other entries, like the size of the master trace table, are NOT processed.

Planned outage avoidance RACF

RACF Started Class Names Table

- How many people still use the old Started Class Names Table to associate Started Tasks with RACF userids?
  - This requires an IPL every time you want to add a new STC
- RACF 1.9 (1989!) introduced the RACF STARTED class that allows you to assign STCs to RACF userids using the panels or RACF commands - no module updates (and no IPLs!) required

Planned outage avoidance RACF

Security Server Dynamic CDT support

- With z/OS V1R6 the Dynamic Class Descriptor Table (CDT) provides the means to add, change, and delete installation-defined classes in the CDT without IPLing the system.
- New RACF class name, CDT, is used to hold the definitions of the installation-defined classes.
- You can use RDEFINE and RALTER commands to define classes.
- Then use command SETROPTS RACLIST(CDT) REFRESH.

Planned outage avoidance RACF

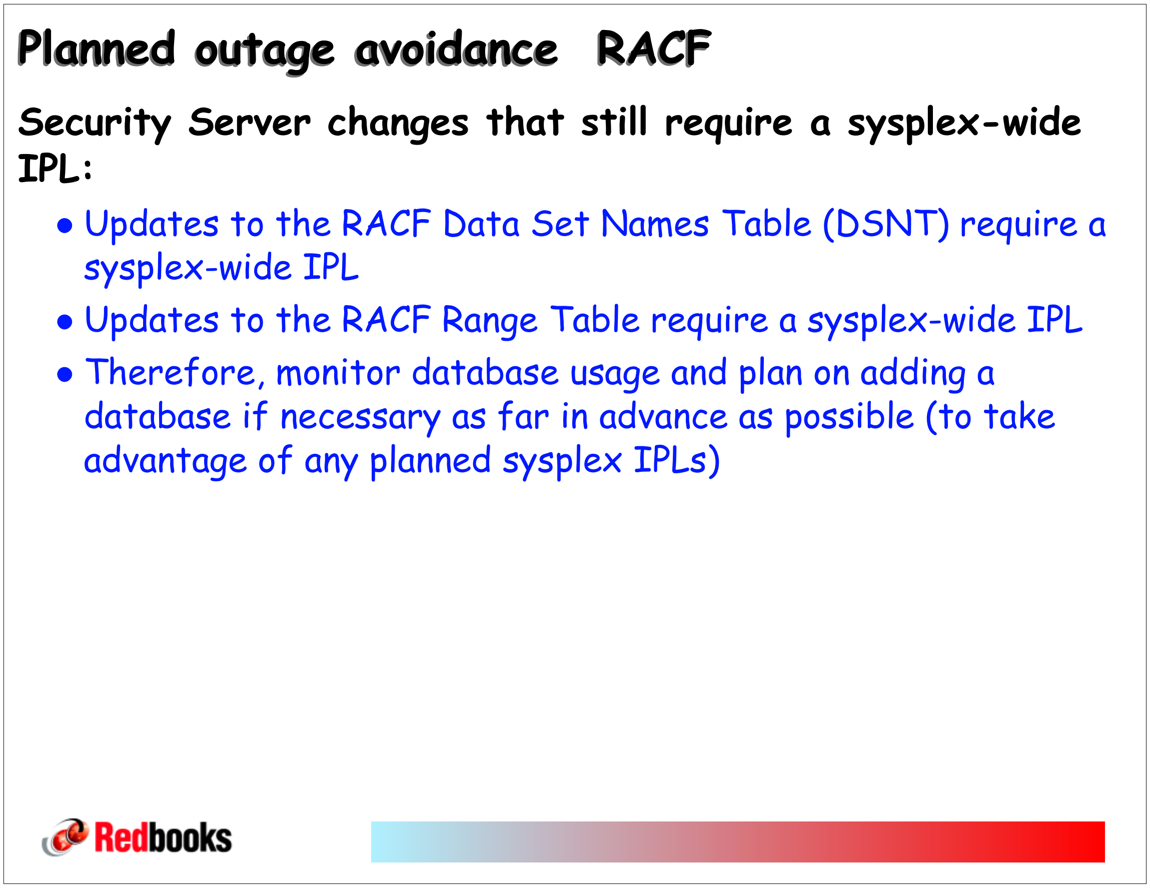
Security Server Removal of the Router Table

- With z/OS V1R6 the IBM-supplied portion of the RACF router table (ICHRFR0X) is removed
  - This eliminates the need to provide an entry in the installation-defined router table (ICHRFR01) for every installation-defined class.
  - Most installation-defined classes will not require any change to ICHRFR01. The most likely candidate for this is DSNR (DB2).
- No longer need to IPL when an installation-defined class is being added.

Planned outage avoidance RACF

Security Server Dynamic Template enhancements

- Templates now have a level indicator which can be used to compare two sets of templates and determine which is the latest. Prevents you from installing a downlevel set of templates onto a RACF database.
- During initialization at IPL, RACF will determine whether the database has the right level of templates and if not, RACF will ignore the templates in the database and use those from IRRTEMP2 automatically.
- When a PTF that changes the templates is applied to a live system, you can run IRRMIN00 and have RACF recognize the new templates without an IPL.
- Prevents complete reinitialization of a RACF database, if that database is "live" on the system where IRRMIN00 is run.

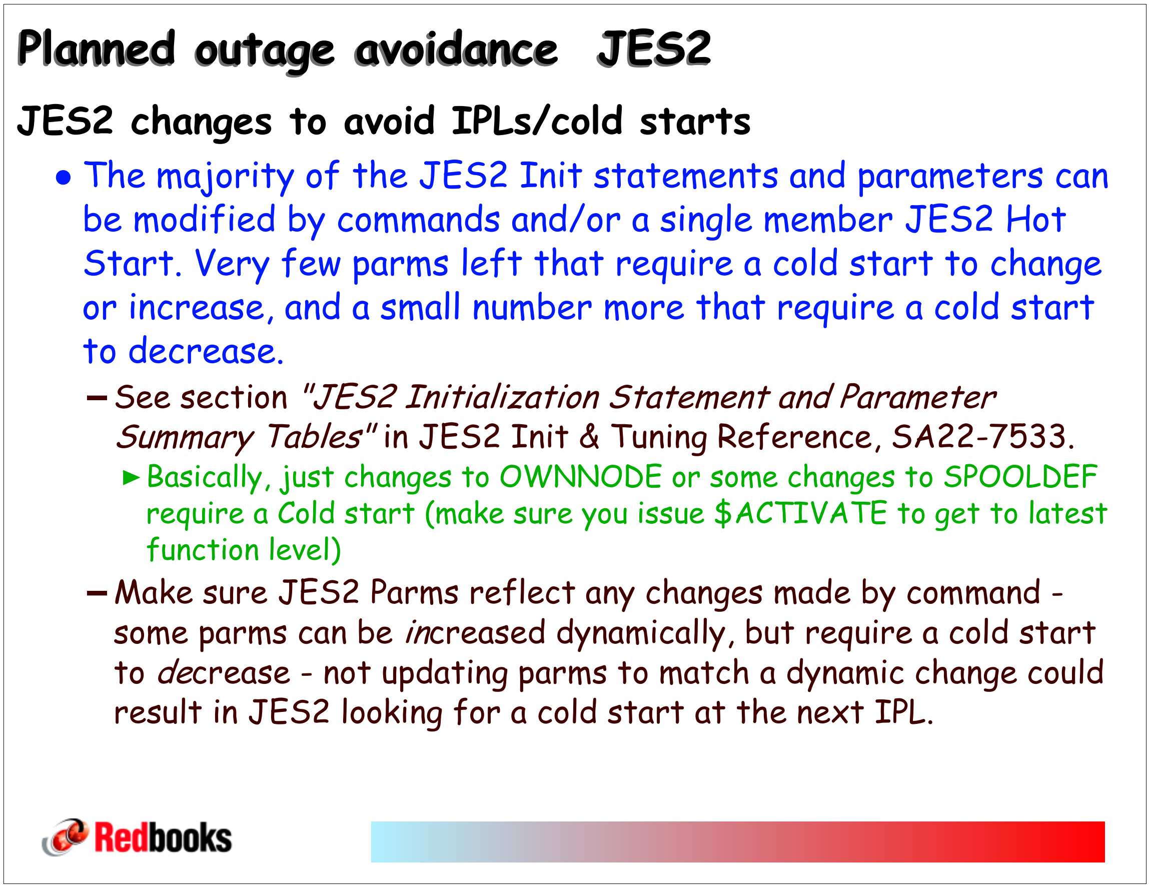


Planned outage avoidance RACF

Security Server changes that still require a sysplex-wide IPL:

- Updates to the RACF Data Set Names Table (DSNT) require a sysplex-wide IPL
- Updates to the RACF Range Table require a sysplex-wide IPL
- Therefore, monitor database usage and plan on adding a database if necessary as far in advance as possible (to take advantage of any planned sysplex IPLs)

Redbooks

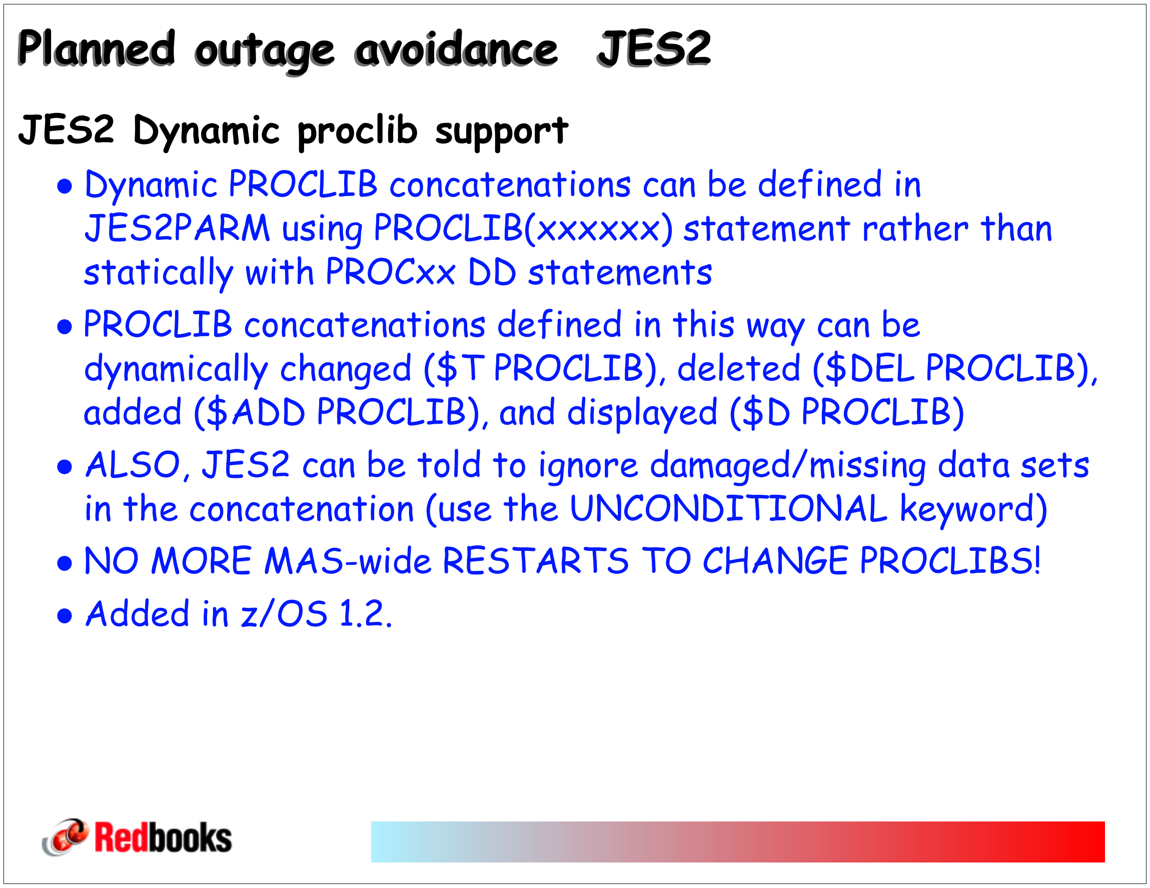


Planned outage avoidance JES2

JES2 changes to avoid IPLs/cold starts

- The majority of the JES2 Init statements and parameters can be modified by commands and/or a single member JES2 Hot Start. Very few parms left that require a cold start to change or increase, and a small number more that require a cold start to decrease.
  - See section *"JES2 Initialization Statement and Parameter Summary Tables"* in JES2 Init & Tuning Reference, SA22-7533.
    - Basically, just changes to OWNNODE or some changes to SPOOLDEF require a Cold start (make sure you issue $ACTIVATE to get to latest function level)
  - Make sure JES2 Parms reflect any changes made by command - some parms can be *in*creased dynamically, but require a cold start to *decrease* - not updating parms to match a dynamic change could result in JES2 looking for a cold start at the next IPL.

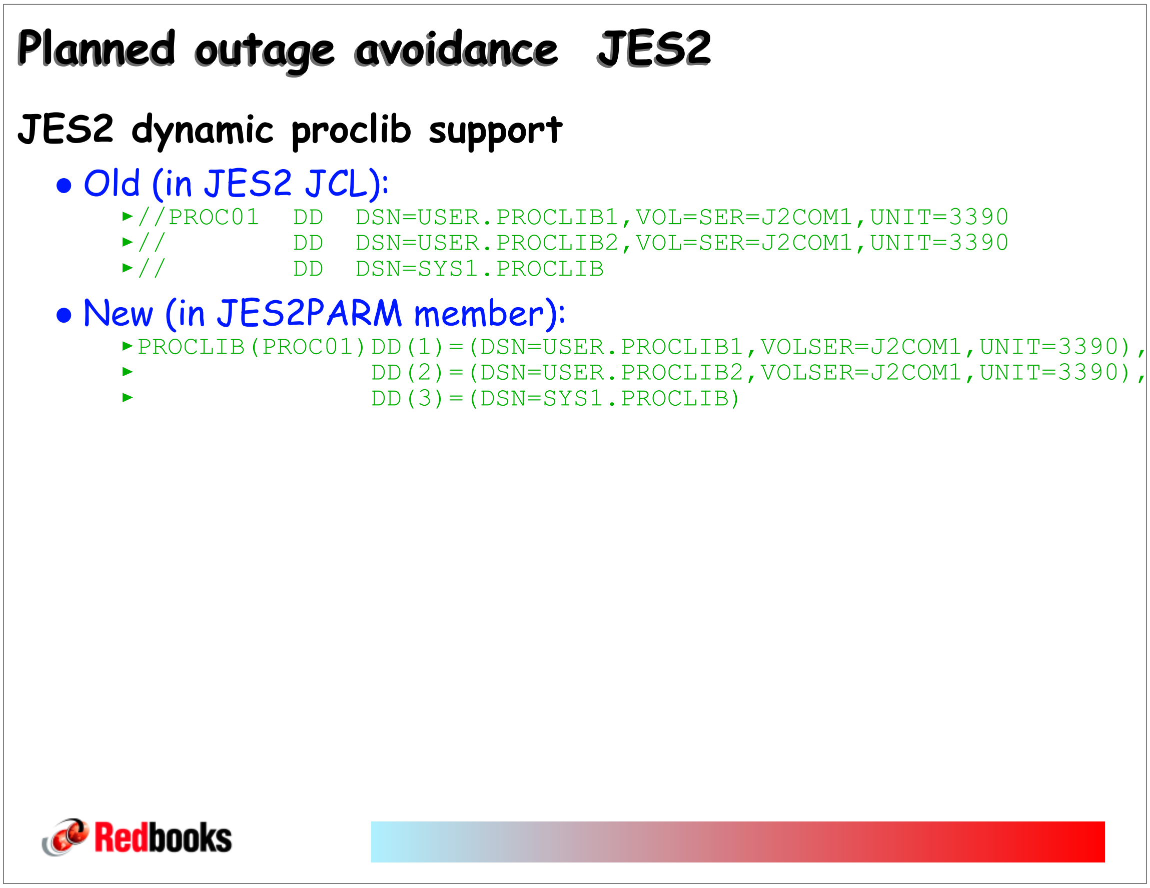
Redbooks



Planned outage avoidance JES2

JES2 Dynamic proclib support

- Dynamic PROCLIB concatenations can be defined in JES2PARM using PROCLIB(xxxxxx) statement rather than statically with PROCxx DD statements
- PROCLIB concatenations defined in this way can be dynamically changed ($T PROCLIB), deleted ($DEL PROCLIB), added ($ADD PROCLIB), and displayed ($D PROCLIB)
- ALSO, JES2 can be told to ignore damaged/missing data sets in the concatenation (use the UNCONDITIONAL keyword)
- NO MORE MAS-wide RESTARTS TO CHANGE PROCLIBS!
- Added in z/OS 1.2.



Planned outage avoidance JES2

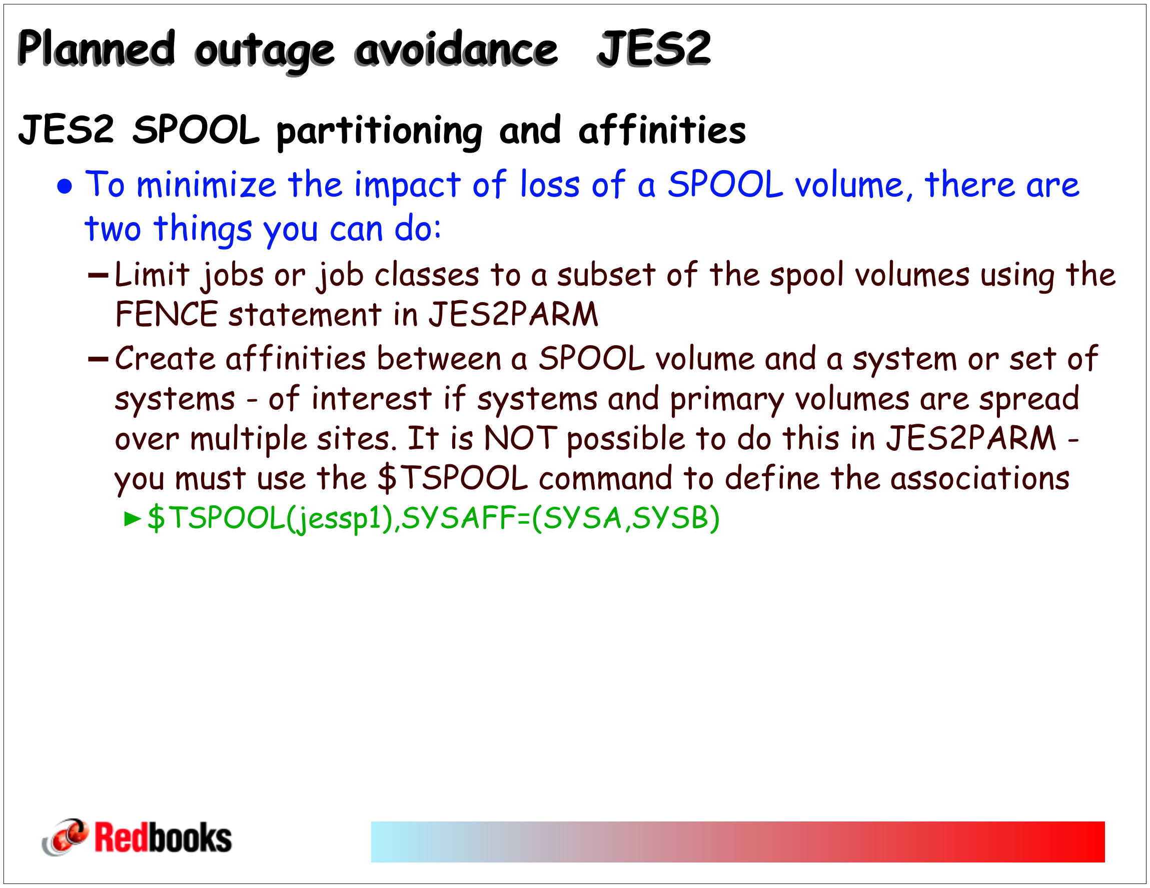
JES2 dynamic proclib support

- Old (in JES2 JCL):

```
▸//PROC01  DD   DSN=USER.PROCLIB1,VOL=SER=J2COM1,UNIT=3390
▸//        DD   DSN=USER.PROCLIB2,VOL=SER=J2COM1,UNIT=3390
▸//        DD   DSN=SYS1.PROCLIB
```

- New (in JES2PARM member):

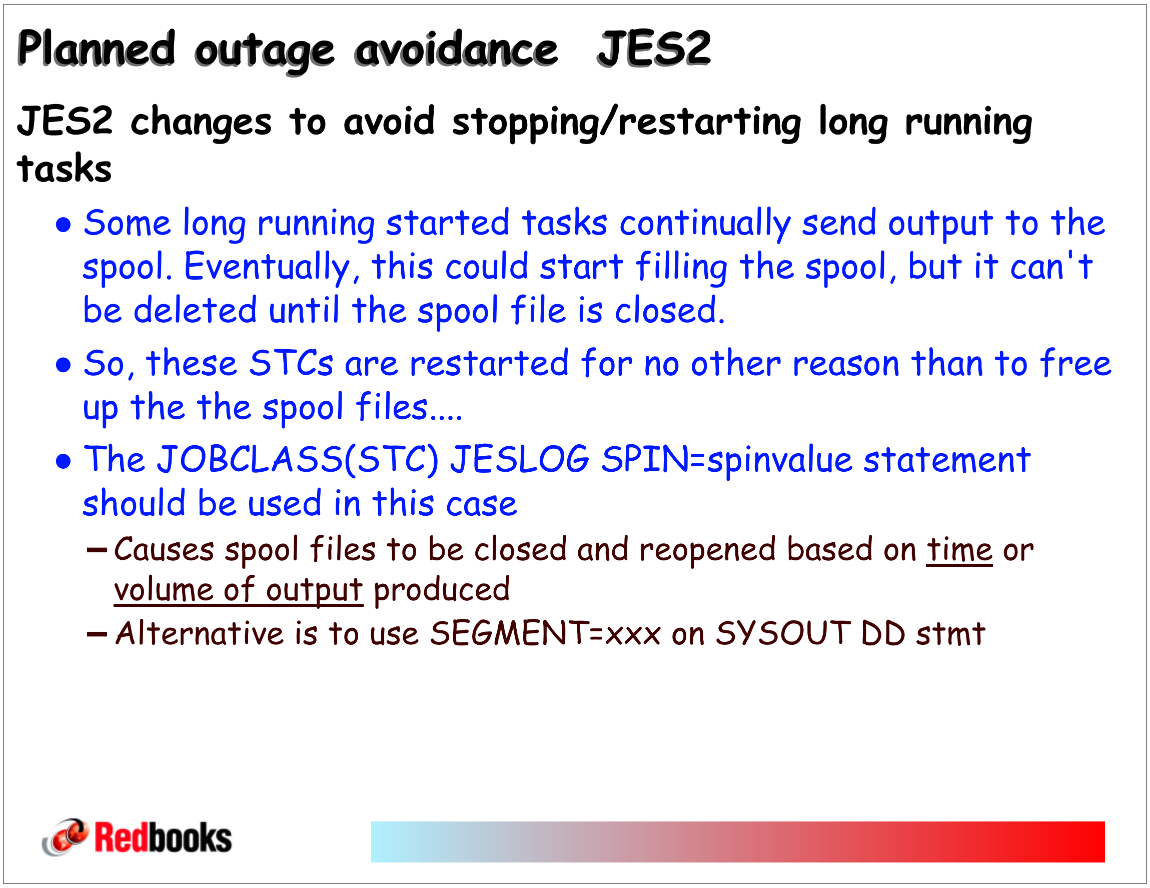
```
▸PROCLIB(PROC01)DD(1)=(DSN=USER.PROCLIB1,VOLSER=J2COM1,UNIT=3390),
▸               DD(2)=(DSN=USER.PROCLIB2,VOLSER=J2COM1,UNIT=3390),
▸               DD(3)=(DSN=SYS1.PROCLIB)
```



Planned outage avoidance JES2

JES2 SPOOL partitioning and affinities

- To minimize the impact of loss of a SPOOL volume, there are two things you can do:
  - Limit jobs or job classes to a subset of the spool volumes using the FENCE statement in JES2PARM
  - Create affinities between a SPOOL volume and a system or set of systems - of interest if systems and primary volumes are spread over multiple sites. It is NOT possible to do this in JES2PARM - you must use the $TSPOOL command to define the associations
    - ►$TSPOOL(jessp1),SYSAFF=(SYSA,SYSB)



Planned outage avoidance JES2

JES2 changes to avoid stopping/restarting long running tasks

- Some long running started tasks continually send output to the spool. Eventually, this could start filling the spool, but it can't be deleted until the spool file is closed.
- So, these STCs are restarted for no other reason than to free up the the spool files....
- The JOBCLASS(STC) JESLOG SPIN=spinvalue statement should be used in this case
  - Causes spool files to be closed and reopened based on time or volume of output produced
  - Alternative is to use SEGMENT=xxx on SYSOUT DD stmt

Planned outage avoidance JES2

JES2 Health Monitor

- Separate address space to monitor JES2
- Starts automatically - no setup or operator intervention required
- Tracks, trends, and compares JES2 critical resources and processes, looking for situations that are different to the norm for this system
  - Monitor keeps 72 hours of data to use as the base for comparison
  - Level of monitoring increases when an out-of-line situation arises
  - Issues HASP9nnn messages when it detects a situation
  - You can display information at any time with $J commands
- SDSF 1.7 enhanced to display information from monitor
  - More information about SDSF interface in Paul Rogers's presentation
- Recommend putting automation in place to raise an alert when a Health Monitor message is issued

Planned outage avoidance JES3

JES3 changes to avoid IPLs

- JES3 LPA modules can be updated using Dynamic LPA (as of OS/390 V2R6)
- JES3 Hot Start Refresh will now pick up majority of changes in JES3 statements
- COMMDEFN, MAINPROC, OUTSERV, STANDARDS, SYSIN, and SYSOUT statements tolerate syntax errors without failing initialization

Planned outage avoidance PDSE

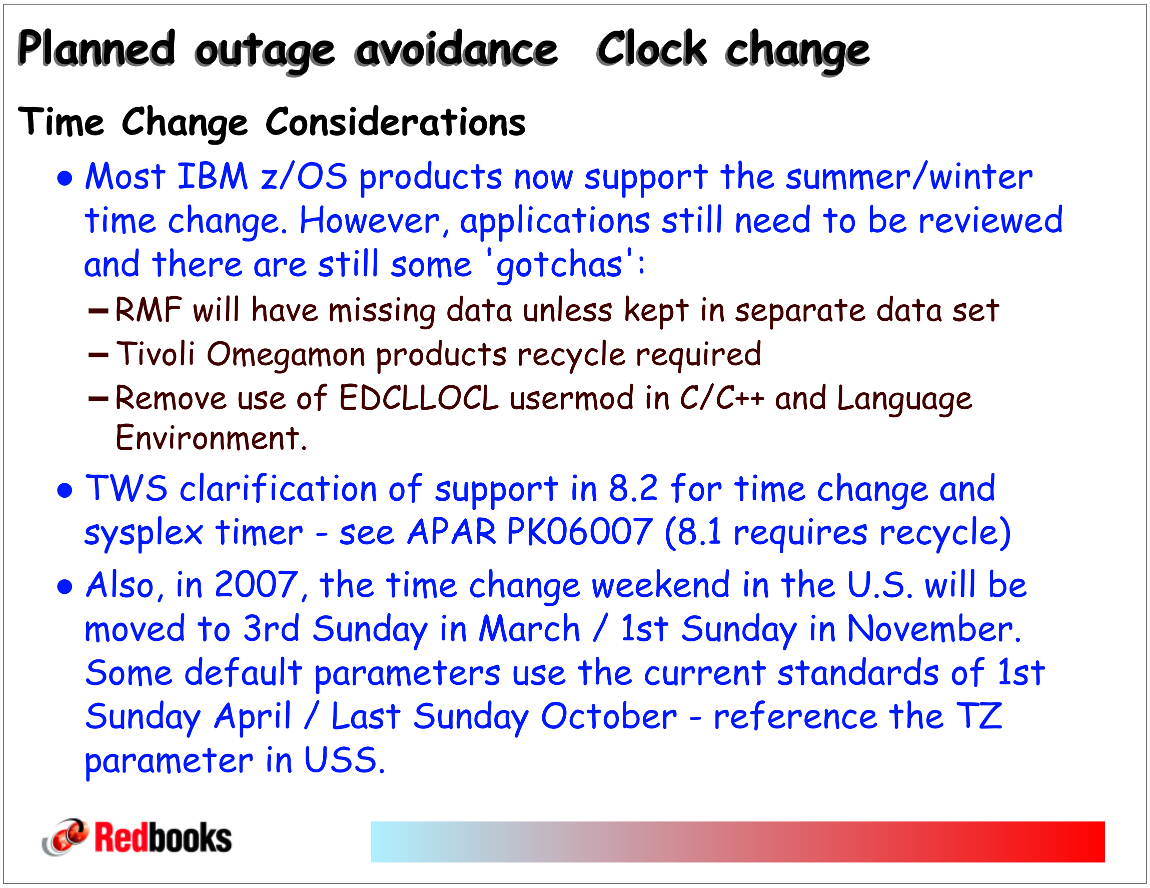
DFSMSdfp recovery enhancements

- In the past, it was necessary to re-IPL a system or systems to resolve a hang condition, deadlock condition, or storage problem in the PDSE address space. With z/OS V1R6, DFSMSdfp will *optionally* use two PDSE address spaces, SMSPDSE and SMSPDSE1. The default will continue to be to only use SMSPDSE.
  - SMSPDSE1 is a restartable address space that provides connections to, and processes requests for, those PDSE data sets that are not part of the global connections associated with SMSPDSE.
  - To create the SMSPDSE1 address space in a sysplex environment, set the IGDSMSxx as follows:
    - PDSESHARING(EXTENDED)
    - PDSE_RESTARTABLE_AS(YES)
  - IPL is recommended for this to be set up initially

Planned outage avoidance Catalog

DFSMSdfp - catalog space monitoring

- Prior to z/OS 1.5, there was no warning when a catalog is about to run out of space
- New enhancement issue message when a specified percentage of maximum extents is reached
  - IEC361I CATALOG catalogname (comptype) HAS REACHED xxx% OF THE MAXIMUM
  - Issued once per catalog per extent and reissued for each subsequent catalog extent
- Invoked by issuing F CATALOG,NOTIFYEXTENT(xxx) where xxx default is 80. 0 or 100 will disable the function
- F CATALOG,ALLOCATED will display the percentage of allocated extents for each catalog in the list in message IEC348I

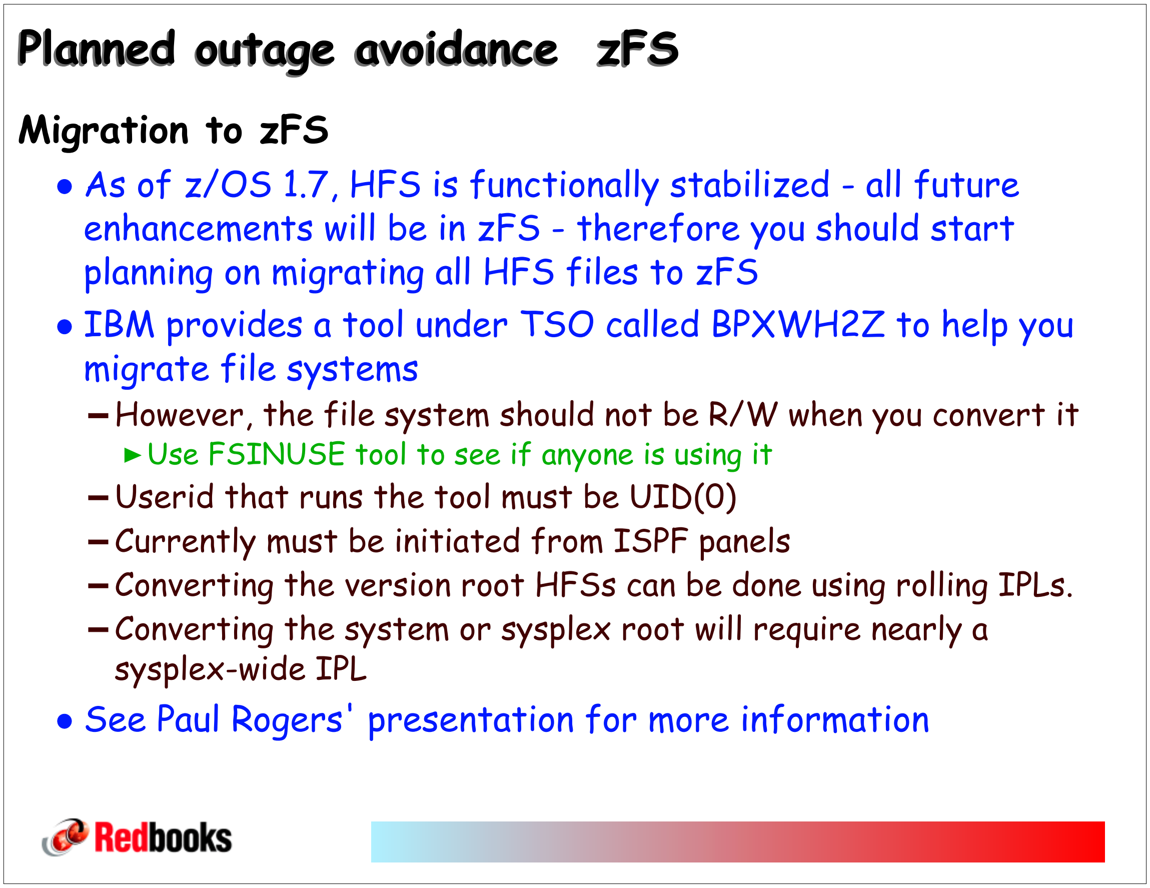


Planned outage avoidance Clock change

Time Change Considerations

- Most IBM z/OS products now support the summer/winter time change. However, applications still need to be reviewed and there are still some 'gotchas':
  - RMF will have missing data unless kept in separate data set
  - Tivoli Omegamon products recycle required
  - Remove use of EDCLLOCL usermod in C/C++ and Language Environment.
- TWS clarification of support in 8.2 for time change and sysplex timer - see APAR PK06007 (8.1 requires recycle)
- Also, in 2007, the time change weekend in the U.S. will be moved to 3rd Sunday in March / 1st Sunday in November. Some default parameters use the current standards of 1st Sunday April / Last Sunday October - reference the TZ parameter in USS.

Redbooks



Planned outage avoidance zFS

Migration to zFS

- As of z/OS 1.7, HFS is functionally stabilized - all future enhancements will be in zFS - therefore you should start planning on migrating all HFS files to zFS
- IBM provides a tool under TSO called BPXWH2Z to help you migrate file systems
  - However, the file system should not be R/W when you convert it
    - Use FSINUSE tool to see if anyone is using it
  - Userid that runs the tool must be UID(0)
  - Currently must be initiated from ISPF panels
  - Converting the version root HFSs can be done using rolling IPLs.
  - Converting the system or sysplex root will require nearly a sysplex-wide IPL
- See Paul Rogers' presentation for more information

Redbooks

Planned outage avoidance Unicode

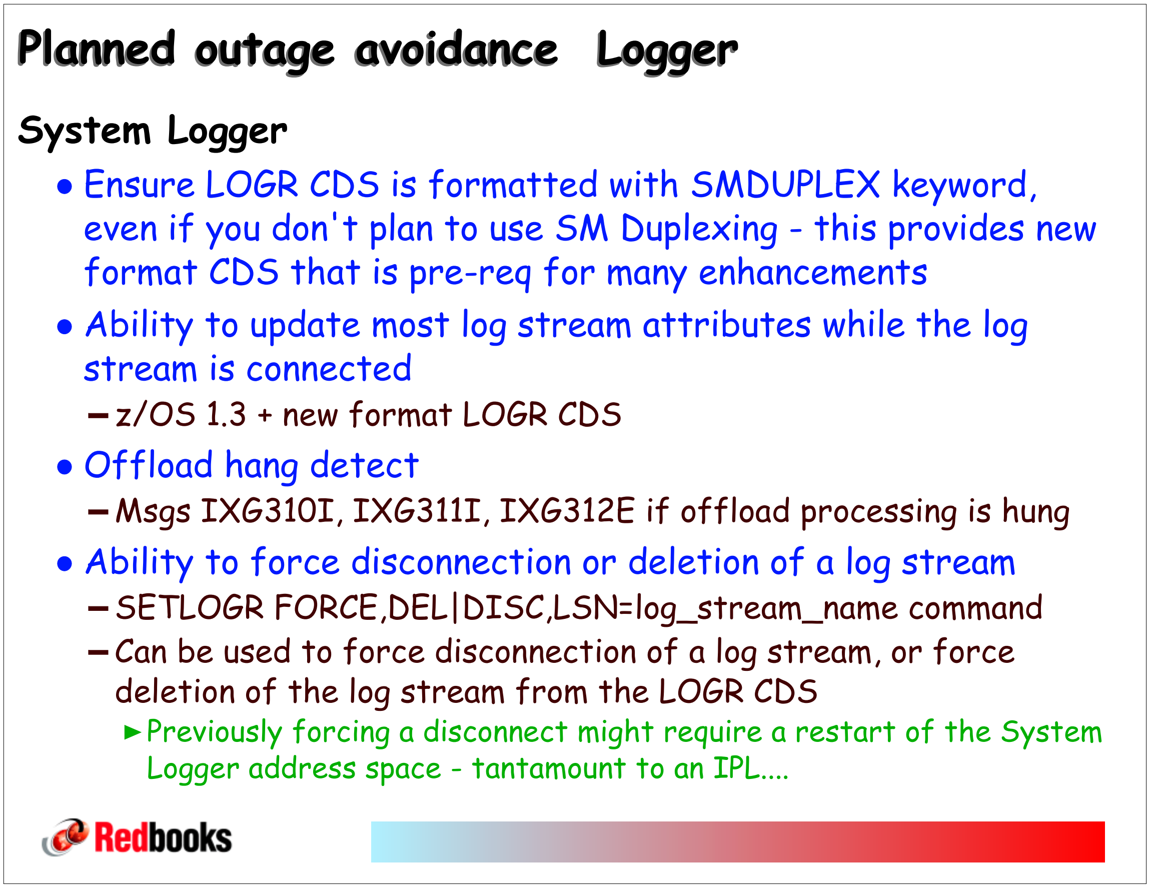
Dynamic UNICODE updates

- Parmlib member CUNINIxx contains statement used to control the conversion environment
  - Pointed to by UNI=xx in IEASYSxx
- In z/OS 1.7, you can update the conversion environment defintions dynamically using SET UNI=xx statement
  - There are some errors in the first release of the 1.7 manual, so make sure all service is applied and all DOC HOLDs reviewed
  - There is also a SETUNI xxx version of the command to make dynamic changes

Planned outage avoidance ASIDs

Non-reuseable address spaces and linkage indexes

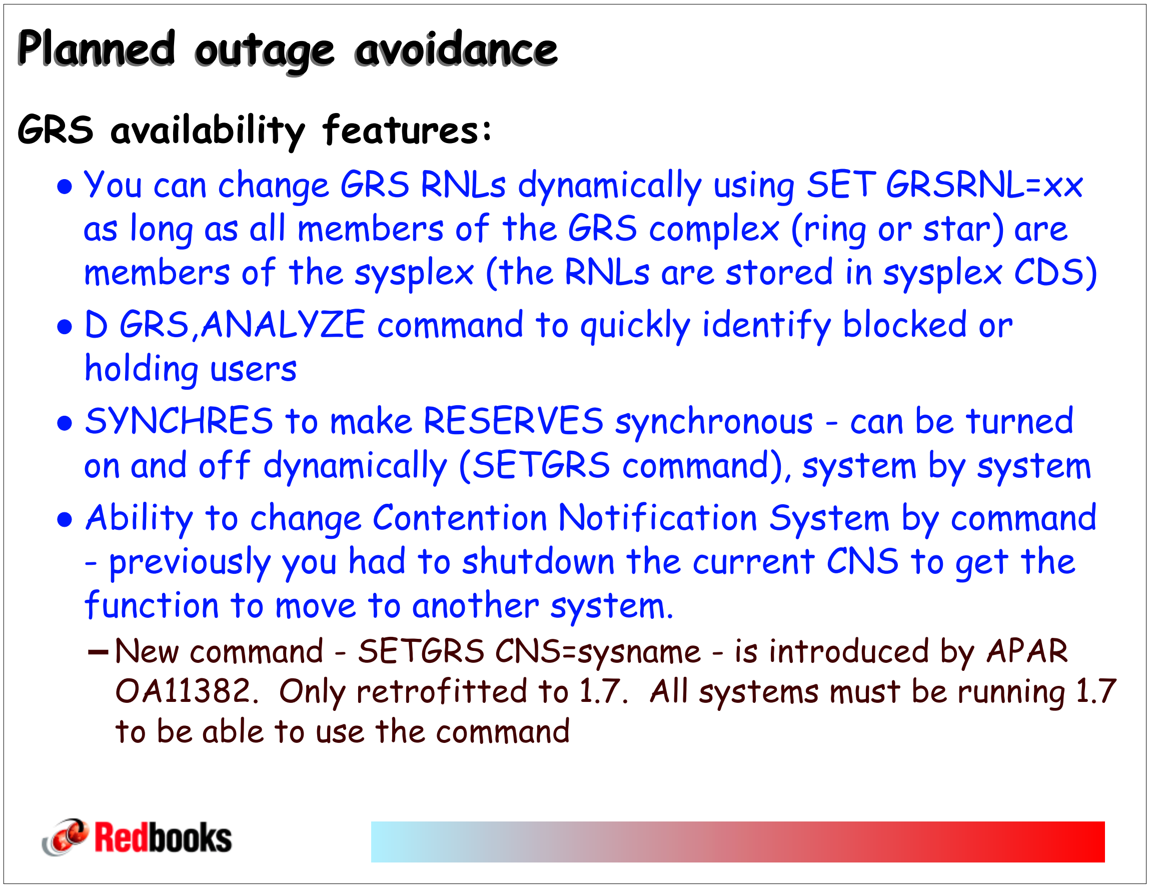
- Prior to z/OS 1.6, system LXs were not reusable. Every time a user of a system LX goes away, the LX becomes dormant. Eventually you run out of LXs and need to re-IPL.
- In z/OS 1.6 (on z990 and later), number of LXs is increased from 2048 to 32K and LXs can be reusable (requires exploitation by the products)
  - NSYSLX max value us 512 prior to 1.6, 512 (12-bit) 8192 (24-bit) after 1.6
- In z/OS 1.7, RRS supports the use of reusable LXs by the products using its services.
- For more information, see http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10273



Planned outage avoidance Logger

System Logger

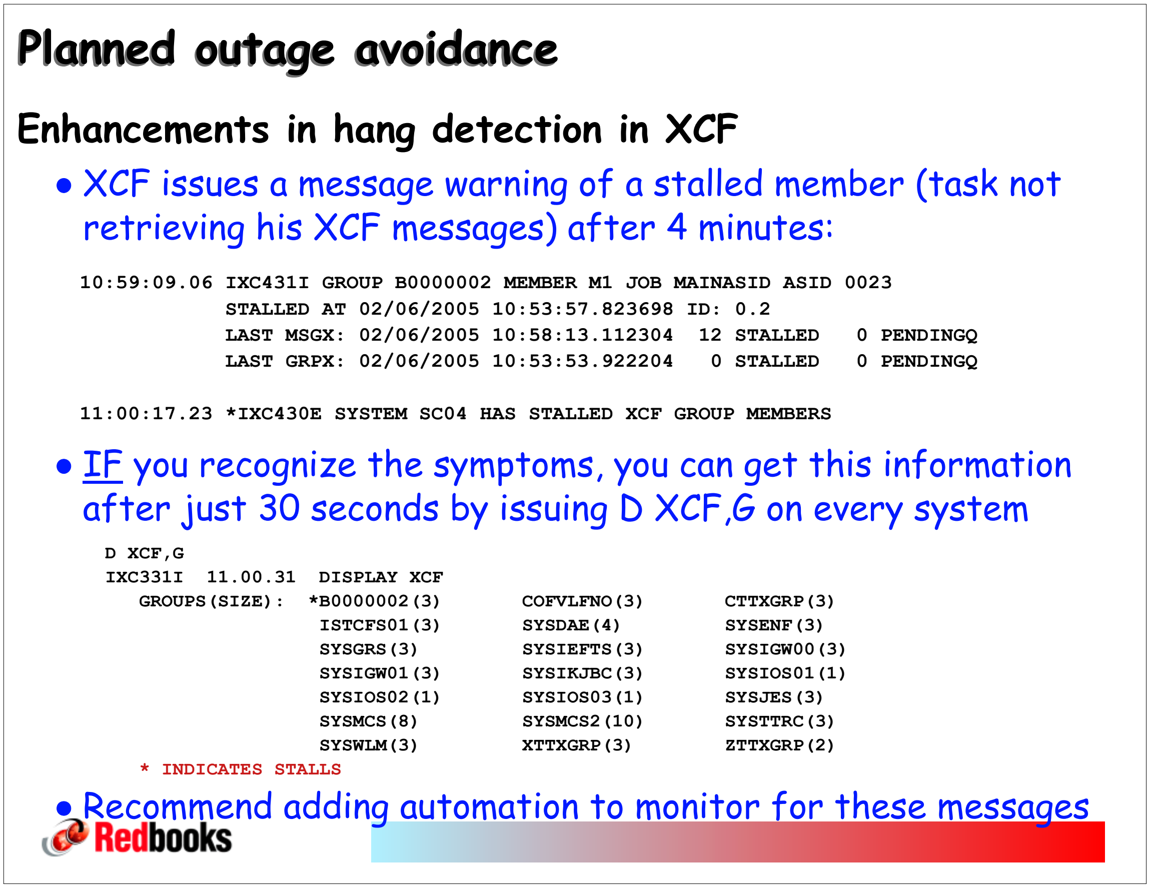
- Ensure LOGR CDS is formatted with SMDUPLEX keyword, even if you don't plan to use SM Duplexing - this provides new format CDS that is pre-req for many enhancements
- Ability to update most log stream attributes while the log stream is connected
  - z/OS 1.3 + new format LOGR CDS
- Offload hang detect
  - Msgs IXG310I, IXG311I, IXG312E if offload processing is hung
- Ability to force disconnection or deletion of a log stream
  - SETLOGR FORCE,DEL|DISC,LSN=log_stream_name command
  - Can be used to force disconnection of a log stream, or force deletion of the log stream from the LOGR CDS
    - ►Previously forcing a disconnect might require a restart of the System Logger address space - tantamount to an IPL....



Planned outage avoidance

GRS availability features:

- You can change GRS RNLs dynamically using SET GRSRNL=xx as long as all members of the GRS complex (ring or star) are members of the sysplex (the RNLs are stored in sysplex CDS)
- D GRS,ANALYZE command to quickly identify blocked or holding users
- SYNCHRES to make RESERVES synchronous - can be turned on and off dynamically (SETGRS command), system by system
- Ability to change Contention Notification System by command - previously you had to shutdown the current CNS to get the function to move to another system.
  - New command - SETGRS CNS=sysname - is introduced by APAR OA11382. Only retrofitted to 1.7. All systems must be running 1.7 to be able to use the command



Planned outage avoidance

Enhancements in hang detection in XCF

- XCF issues a message warning of a stalled member (task not retrieving his XCF messages) after 4 minutes:

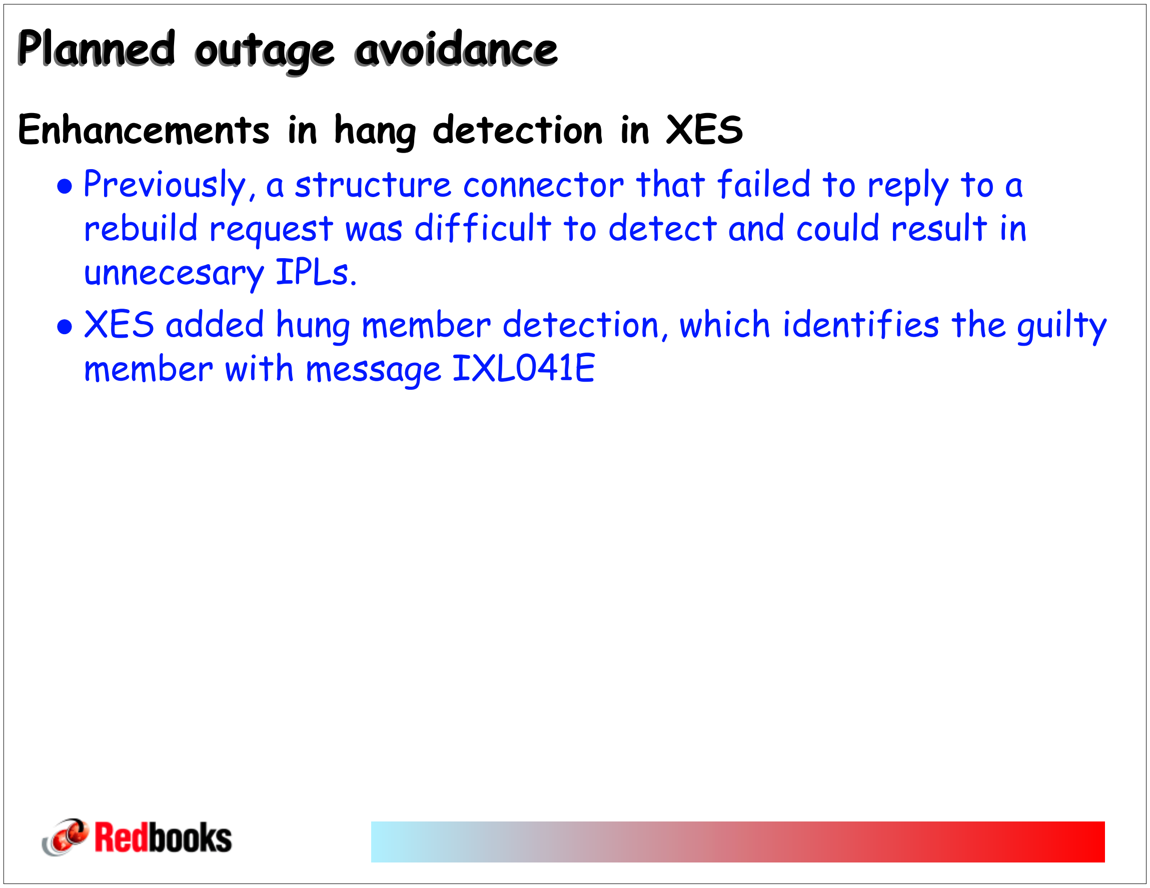
```
10:59:09.06 IXC431I GROUP B0000002 MEMBER M1 JOB MAINASID ASID 0023
            STALLED AT 02/06/2005 10:53:57.823698 ID: 0.2
            LAST MSGX: 02/06/2005 10:58:13.112304  12 STALLED   0 PENDINGQ
            LAST GRPX: 02/06/2005 10:53:53.922204   0 STALLED   0 PENDINGQ

11:00:17.23 *IXC430E SYSTEM SC04 HAS STALLED XCF GROUP MEMBERS
```

- IF you recognize the symptoms, you can get this information after just 30 seconds by issuing D XCF,G on every system

```
D XCF,G
IXC331I  11.00.31  DISPLAY XCF
   GROUPS(SIZE):  *B0000002(3)      COFVLFNO(3)       CTTXGRP(3)
                   ISTCFS01(3)      SYSDAE(4)         SYSENF(3)
                   SYSGRS(3)        SYSIEFTS(3)       SYSIGW00(3)
                   SYSIGW01(3)      SYSIKJBC(3)       SYSIOS01(1)
                   SYSIOS02(1)      SYSIOS03(1)       SYSJES(3)
                   SYSMCS(8)        SYSMCS2(10)       SYSTTRC(3)
                   SYSWLM(3)        XTTXGRP(3)        ZTTXGRP(2)
   * INDICATES STALLS
```

- Recommend adding automation to monitor for these messages

Redbooks



Planned outage avoidance

Enhancements in hang detection in XES

- Previously, a structure connector that failed to reply to a rebuild request was difficult to detect and could result in unnecesary IPLs.
- XES added hung member detection, which identifies the guilty member with message IXL041E

Redbooks

Planned outage avoidance

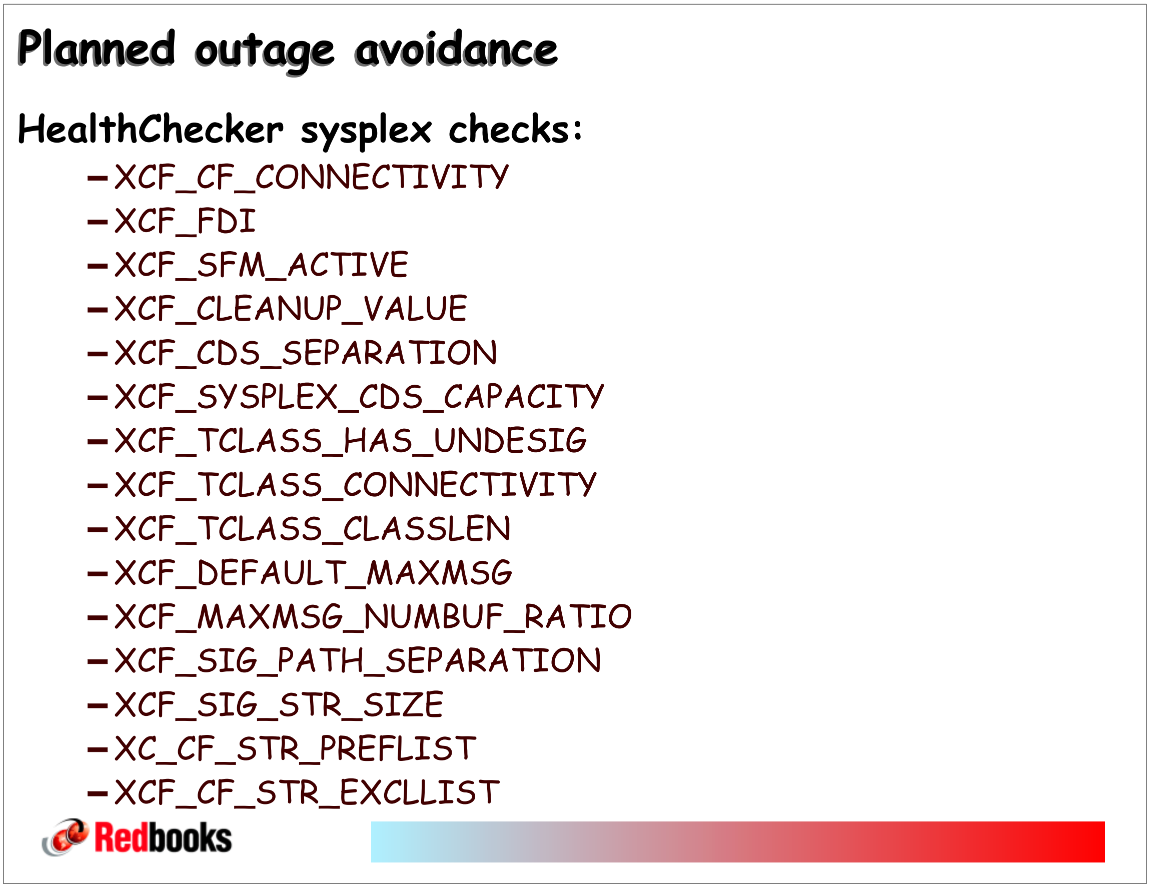
VTAM Generic Resources

- When an application registers as a generic resource, information about the GR name is stored in various places in VTAM (and in the ISTGENERIC structure).
- If you want to reuse that name for something else, all that information must be deleted.
- Prior to 1.7, this meant defining a new GR structure, stopping ALL VTAMs, and bring them up with the new structure - similar impact to a sysplex IPL
- In 1.7, there is a new VTAM command to delete GR information dynamically:
  - F NET,GR,GRNAME=neta.grappl,OPTION=DELETE
- For more information, refer to section 6.1.3.7 "Removing a Generic Resource", in *SNA Network Implementation Guide*

Planned outage avoidance

z/OS HealthChecker:

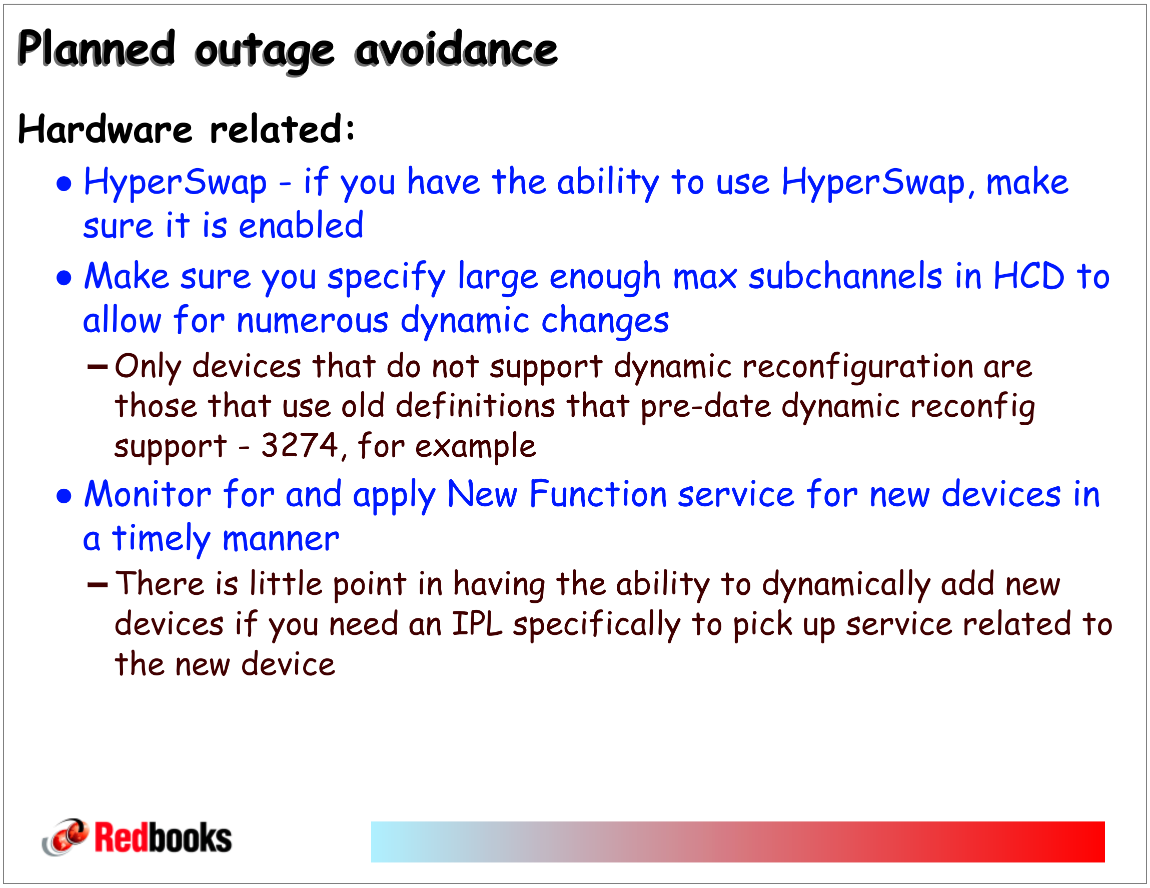
- VERY popular with customers - may warn you of problems before they become critical, thereby avoiding an IPL
- New version of HealthChecker included in z/OS 1.7
  - The same code is available as a Web download for z/OS 1.4 to 1.6
  - Available for download from Sept 30 from:
    - http://www.ibm.com/servers/eserver/zseries/zos/downloads/#asis
- Everyone should run every check and either address any "issues" that are raised, or adjust the HealthChecker parms so it understands this situation is normal in your shop. Then run it continually.
  - Note that the message numbers now reflect the component that owns the check (IXCHCxxx) rather than HealthChecker (HZS)
- Paul Rogers covers the new HealthChecker in detail.



Planned outage avoidance

HealthChecker sysplex checks:

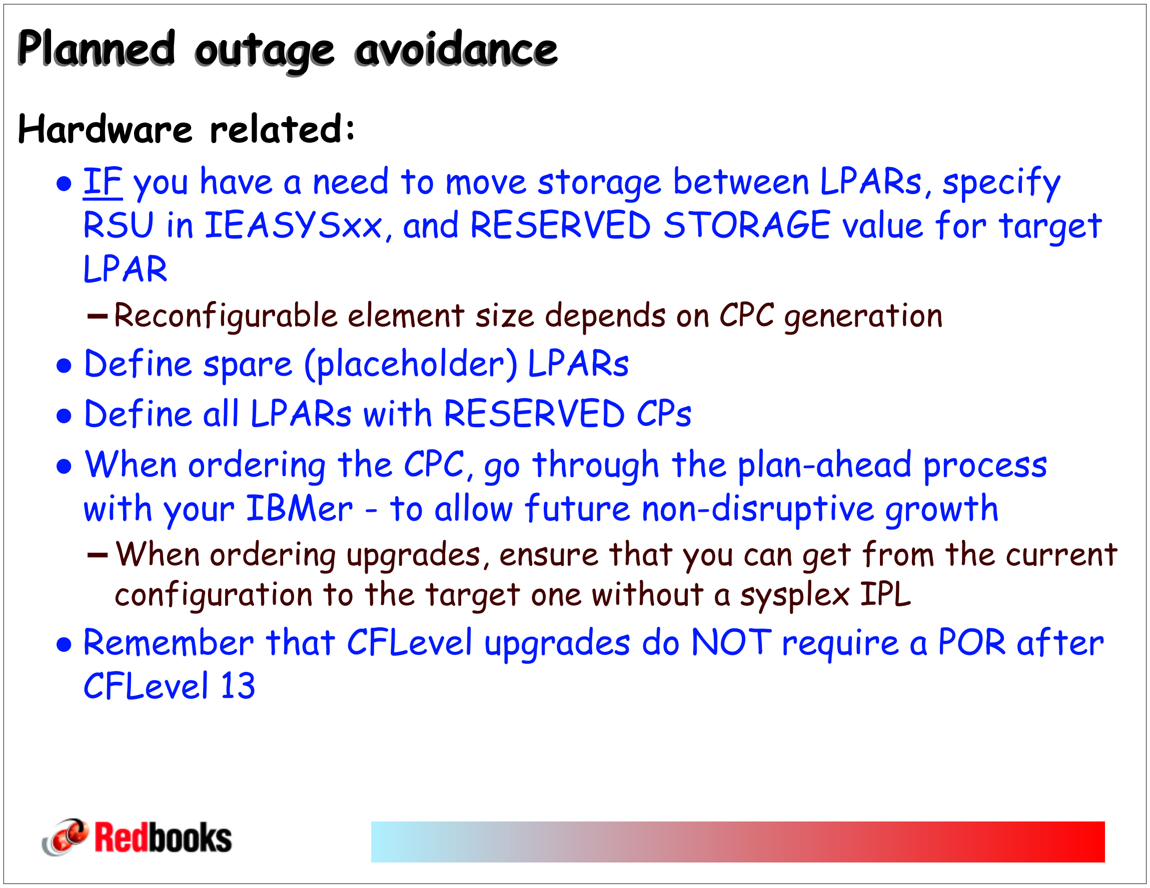
- XCF_CF_CONNECTIVITY
- XCF_FDI
- XCF_SFM_ACTIVE
- XCF_CLEANUP_VALUE
- XCF_CDS_SEPARATION
- XCF_SYSPLEX_CDS_CAPACITY
- XCF_TCLASS_HAS_UNDESIG
- XCF_TCLASS_CONNECTIVITY
- XCF_TCLASS_CLASSLEN
- XCF_DEFAULT_MAXMSG
- XCF_MAXMSG_NUMBUF_RATIO
- XCF_SIG_PATH_SEPARATION
- XCF_SIG_STR_SIZE
- XC_CF_STR_PREFLIST
- XCF_CF_STR_EXCLLIST



Planned outage avoidance

Hardware related:

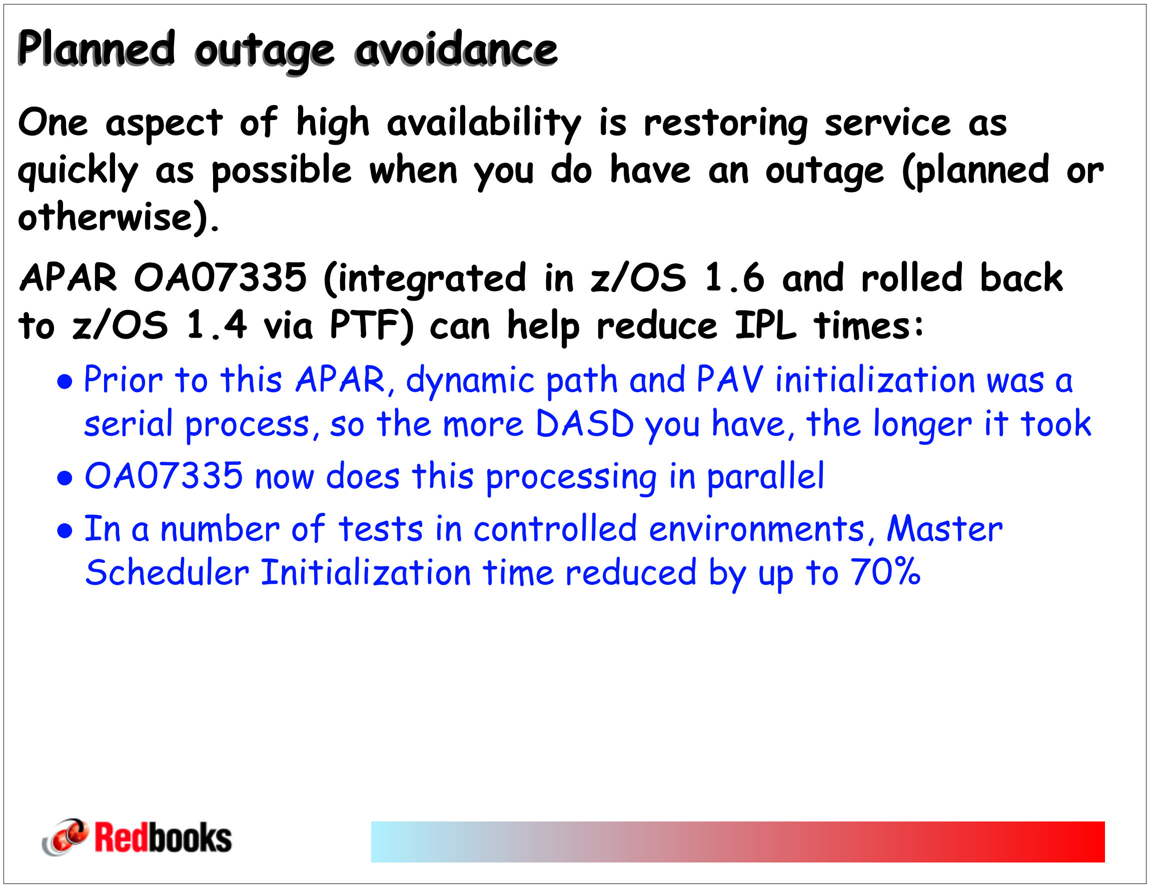
- HyperSwap - if you have the ability to use HyperSwap, make sure it is enabled
- Make sure you specify large enough max subchannels in HCD to allow for numerous dynamic changes
  - Only devices that do not support dynamic reconfiguration are those that use old definitions that pre-date dynamic reconfig support - 3274, for example
- Monitor for and apply New Function service for new devices in a timely manner
  - There is little point in having the ability to dynamically add new devices if you need an IPL specifically to pick up service related to the new device



Planned outage avoidance

Hardware related:

- IF you have a need to move storage between LPARs, specify RSU in IEASYSxx, and RESERVED STORAGE value for target LPAR
  - Reconfigurable element size depends on CPC generation
- Define spare (placeholder) LPARs
- Define all LPARs with RESERVED CPs
- When ordering the CPC, go through the plan-ahead process with your IBMer - to allow future non-disruptive growth
  - When ordering upgrades, ensure that you can get from the current configuration to the target one without a sysplex IPL
- Remember that CFLevel upgrades do NOT require a POR after CFLevel 13

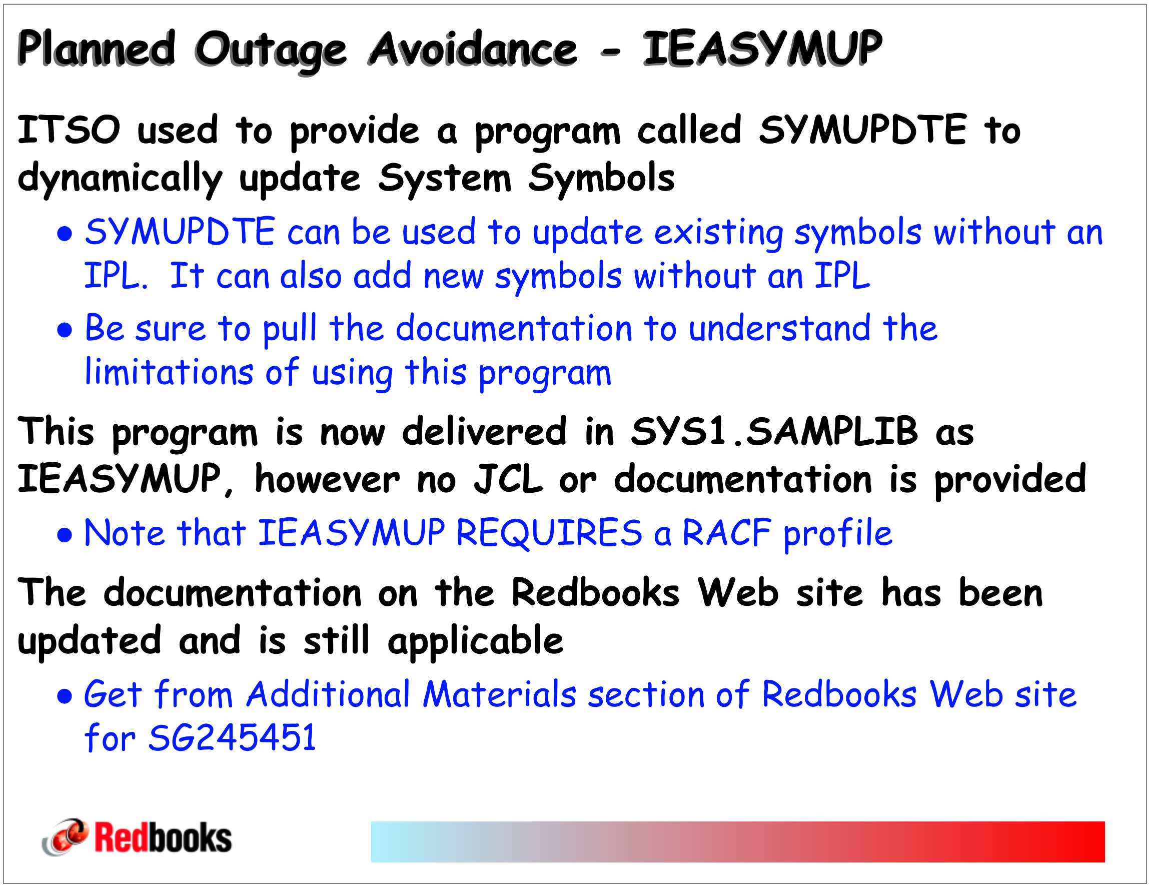


Planned outage avoidance

One aspect of high availability is restoring service as quickly as possible when you do have an outage (planned or otherwise).

APAR OA07335 (integrated in z/OS 1.6 and rolled back to z/OS 1.4 via PTF) can help reduce IPL times:

- Prior to this APAR, dynamic path and PAV initialization was a serial process, so the more DASD you have, the longer it took
- OA07335 now does this processing in parallel
- In a number of tests in controlled environments, Master Scheduler Initialization time reduced by up to 70%



Planned Outage Avoidance - IEASYMUP

ITSO used to provide a program called SYMUPDTE to dynamically update System Symbols

- SYMUPDTE can be used to update existing symbols without an IPL. It can also add new symbols without an IPL
- Be sure to pull the documentation to understand the limitations of using this program

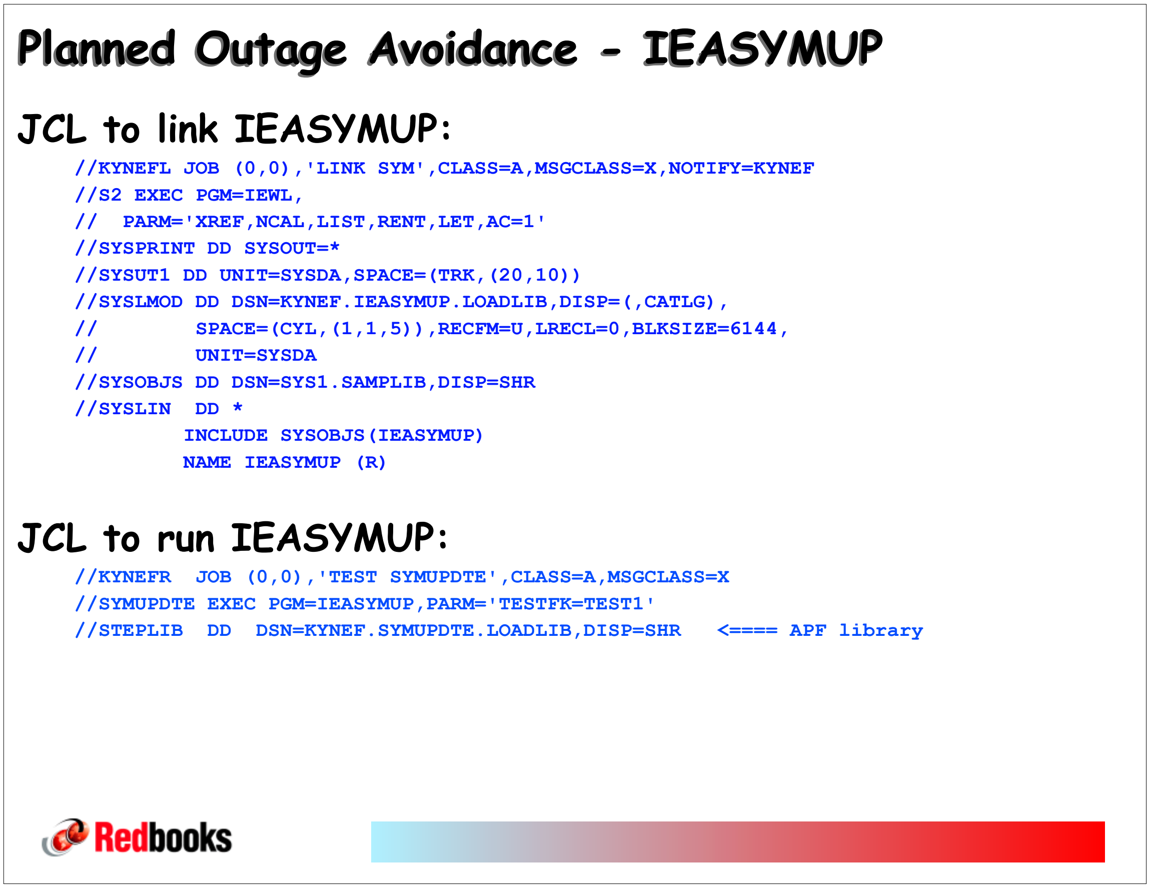
This program is now delivered in SYS1.SAMPLIB as IEASYMUP, however no JCL or documentation is provided

- Note that IEASYMUP REQUIRES a RACF profile

The documentation on the Redbooks Web site has been updated and is still applicable

- Get from Additional Materials section of Redbooks Web site for SG245451

Redbooks



Planned Outage Avoidance - IEASYMUP

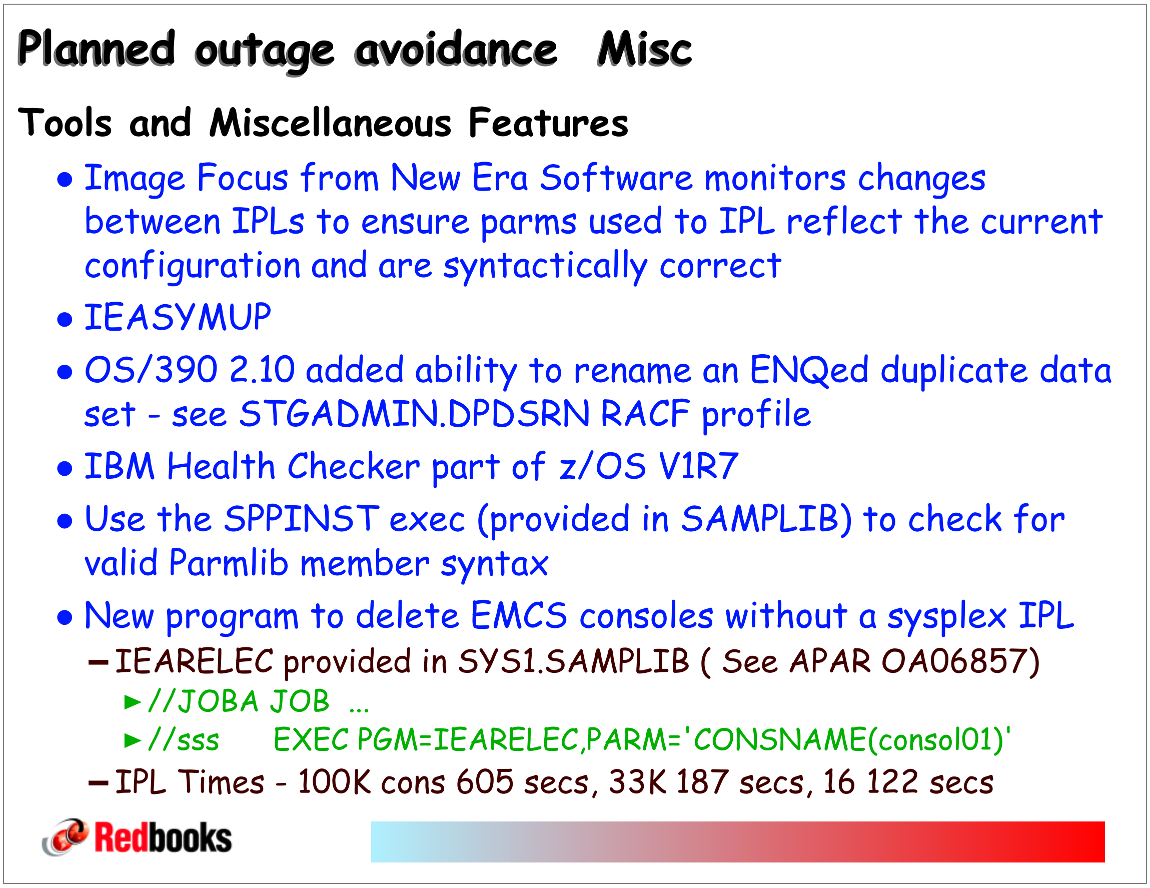
JCL to link IEASYMUP:

```
//KYNEFL JOB (0,0),'LINK SYM',CLASS=A,MSGCLASS=X,NOTIFY=KYNEF
//S2 EXEC PGM=IEWL,
//  PARM='XREF,NCAL,LIST,RENT,LET,AC=1'
//SYSPRINT DD SYSOUT=*
//SYSUT1 DD UNIT=SYSDA,SPACE=(TRK,(20,10))
//SYSLMOD DD DSN=KYNEF.IEASYMUP.LOADLIB,DISP=(,CATLG),
//         SPACE=(CYL,(1,1,5)),RECFM=U,LRECL=0,BLKSIZE=6144,
//         UNIT=SYSDA
//SYSOBJS DD DSN=SYS1.SAMPLIB,DISP=SHR
//SYSLIN  DD *
         INCLUDE SYSOBJS(IEASYMUP)
         NAME IEASYMUP (R)
```

JCL to run IEASYMUP:

```
//KYNEFR  JOB (0,0),'TEST SYMUPDTE',CLASS=A,MSGCLASS=X
//SYMUPDTE EXEC PGM=IEASYMUP,PARM='TESTFK=TEST1'
//STEPLIB  DD  DSN=KYNEF.SYMUPDTE.LOADLIB,DISP=SHR   <==== APF library
```

Redbooks

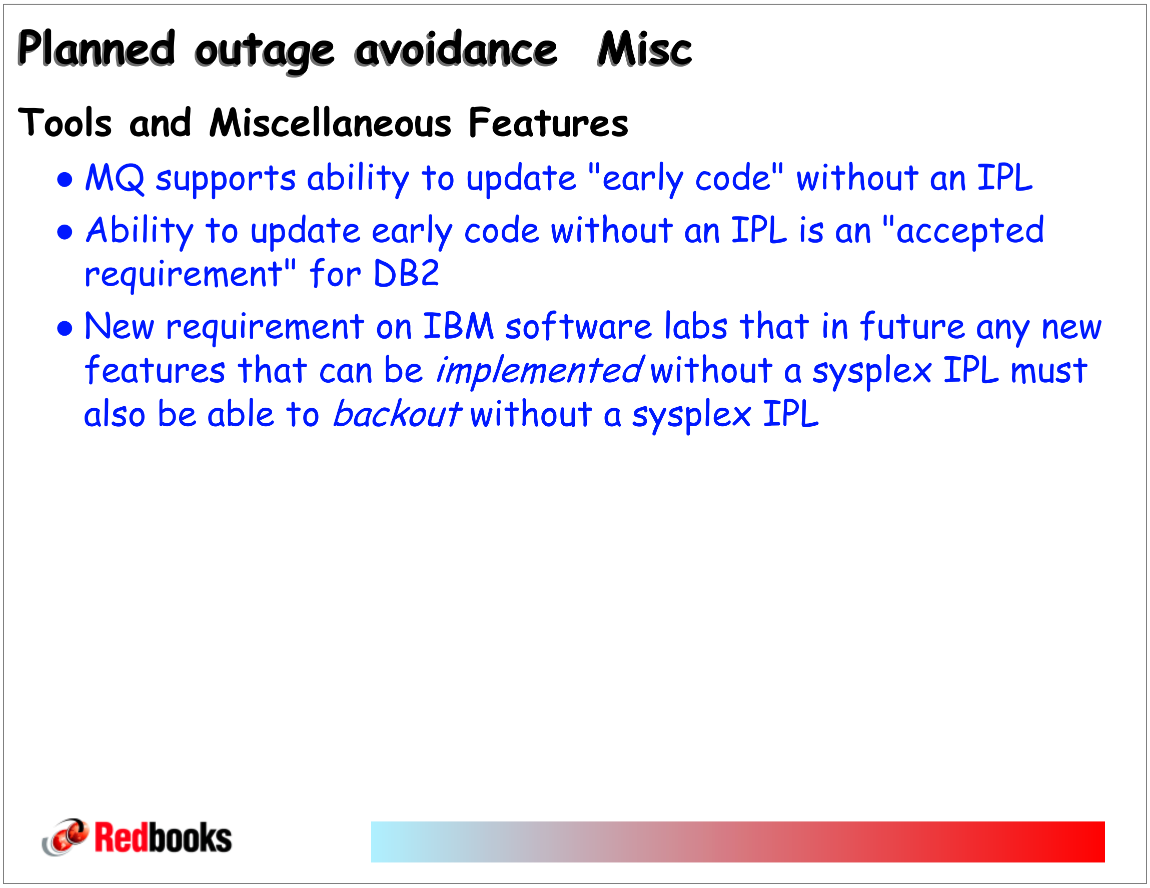


Planned outage avoidance Misc

Tools and Miscellaneous Features

- Image Focus from New Era Software monitors changes between IPLs to ensure parms used to IPL reflect the current configuration and are syntactically correct
- IEASYMUP
- OS/390 2.10 added ability to rename an ENQed duplicate data set - see STGADMIN.DPDSRN RACF profile
- IBM Health Checker part of z/OS V1R7
- Use the SPPINST exec (provided in SAMPLIB) to check for valid Parmlib member syntax
- New program to delete EMCS consoles without a sysplex IPL
  - IEARELEC provided in SYS1.SAMPLIB (See APAR OA06857)
    - //JOBA JOB ...
    - //sss EXEC PGM=IEARELEC,PARM='CONSNAME(consol01)'
  - IPL Times - 100K cons 605 secs, 33K 187 secs, 16 122 secs

Redbooks



Planned outage avoidance Misc

Tools and Miscellaneous Features

- MQ supports ability to update "early code" without an IPL
- Ability to update early code without an IPL is an "accepted requirement" for DB2
- New requirement on IBM software labs that in future any new features that can be *implemented* without a sysplex IPL must also be able to *backout* without a sysplex IPL

Redbooks

Planned outage avoidance

What does still need an IPL ?

- New PLPA and COMMON page datasets
- RACF Dataset name table, RACF range table changes (SYSPLEX IPL)
- Adding JES3 Spool volumes
- Adding MCS and SMCS consoles
- Backouts for GRS STAR, BRLM and PDSE Sharing (SYSPLEX)
- Changing MAXUSER, RSVSTRT and RSVNONR values
- Service:
  - Mass PTF apply
  - Updates to NUCLEUS or LPA (unless specifically supported by vendor)
- Changes to the USS SWA() and SYSPLEX() parameters
- LOAD member changes e.g. new MCAT
- System Symbols - can use IEASYMUP but be careful!
- Parmlib members - ALLOCxx, BLSCECT, BLSCUSER, CNIDTRxx, CONFIGxx, DEVSUP, EPHWP00, IEAAPP00, IEAFIX00, IEAPAKxx, IOEPRMxx, LOADxx, MSTJCLxx, NUCLSTxx, and IEASYSxx still require an IPL to modify.

Planned outage avoidance

What are the considerations for fewer IPLs?

- When you DO have one, many more changes are squeezed into one outage
- If you changed something dynamically and forgot to reflect the change back into Parmlib, there is less chance you will remember the change the longer it is until the next IPL

Planned outage avoidance

Dynamic changes

- Good, because they help you avoid an IPL
- BUT, they require foolproof system management and processes to avoid regressing them in the future

Should you use `SET aaa=xx`, or `SETaaa parm.....?`

- Using SET aaa=xx has the benefit of ensuring that you update the Parm member AND it syntax checks the change
- But... what if many people changed that member - do you activate changes that you are not ready for yet?
- There is no "right" answer - it varies from installation to installation

Redbooks

Planned outage avoidance

To do dynamic or not dynamic??

- Any dynamic changes made should also be reflected in the appropriate Parmlib members at the same time to retain the integrity of your systems. If the changes are not made statically to the corresponding members of Parmlib, they will be lost with the next IPL.
- This is particulary important with the SETxxx (i.e. SETAPPC, SETCEE) commands which in general do not read the corresponding Parmlib members for system information.
- Use of source management products should be seriously considered along with the regular use of syntax checkers from IBM and third parties.

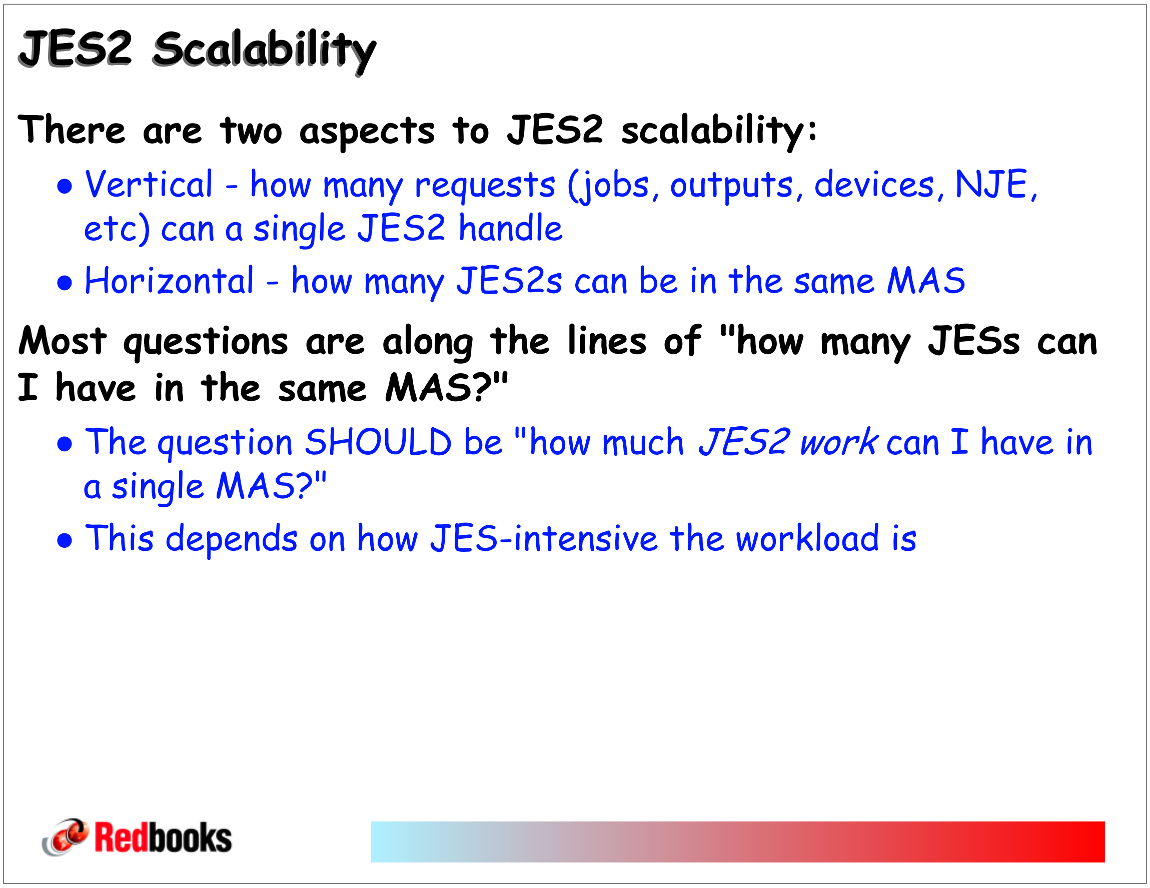
Redbooks

Planned outage avoidance

Further information:

- ITSO RedPaper *Planned Outage Avoidance Checklist,* REDP-4069 (in progress)
- *z/OS Installation and Tuning Reference*, SA22-7592
- *z/OS Introduction and Release Guide*, GA22-7504
- *z/OS Migration*, GA22-7499
- *z/OS System Commands*, SA22-7627

JES2 Scalability considerations



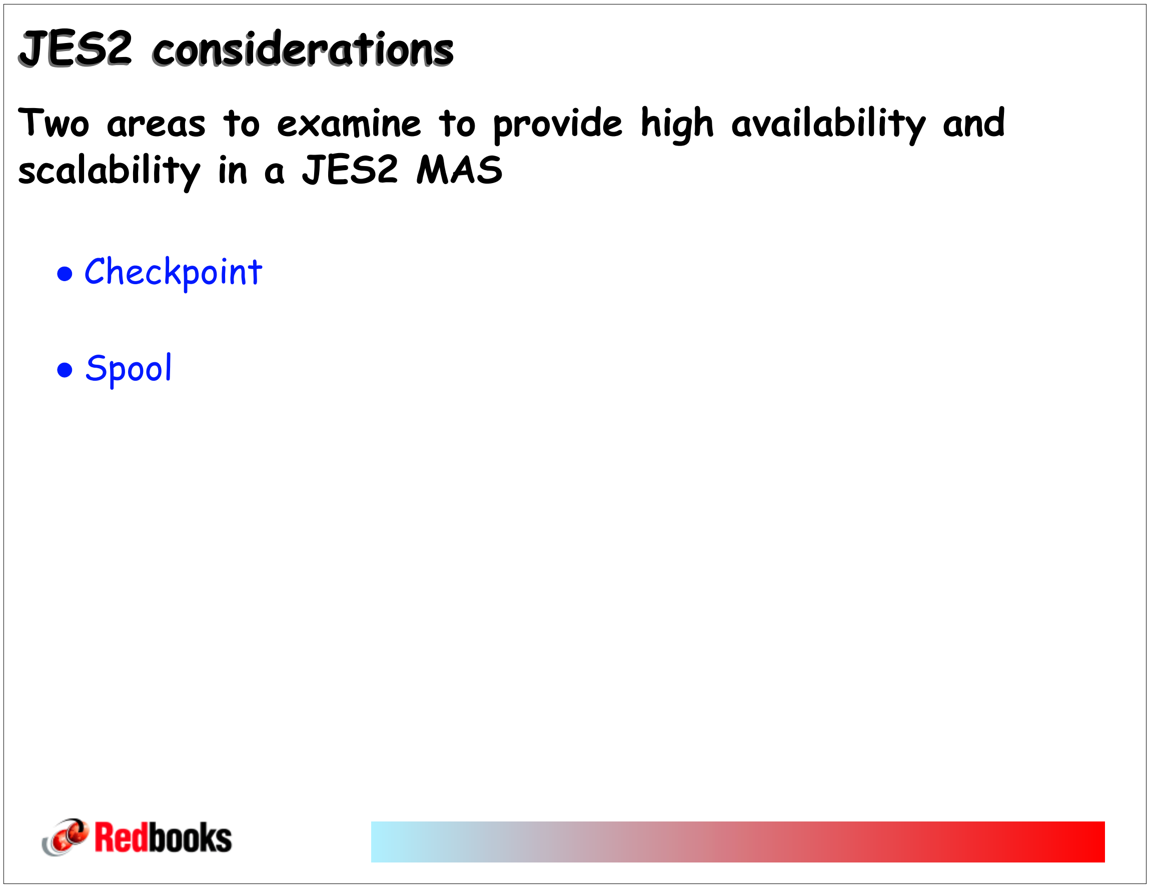
JES2 Scalability

There are two aspects to JES2 scalability:

- Vertical - how many requests (jobs, outputs, devices, NJE, etc) can a single JES2 handle
- Horizontal - how many JES2s can be in the same MAS

Most questions are along the lines of "how many JESs can I have in the same MAS?"

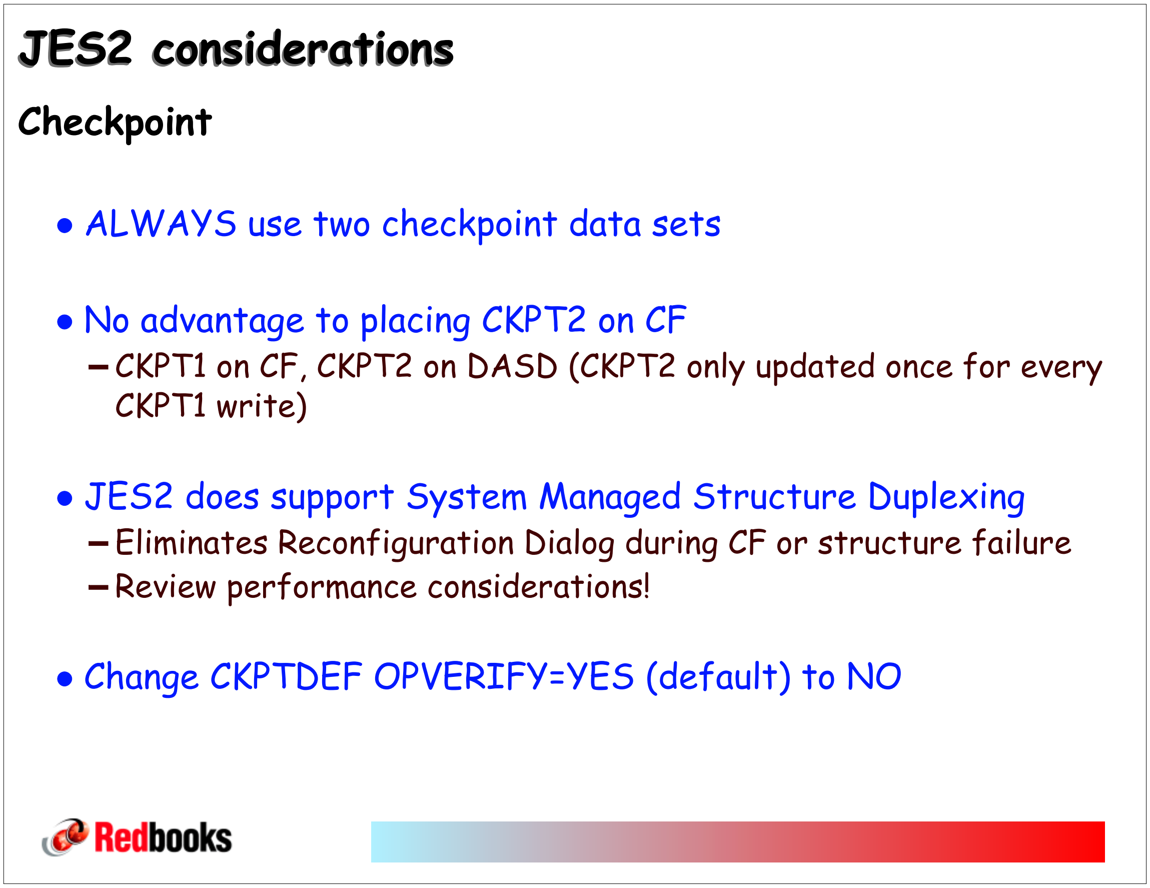
- The question SHOULD be "how much *JES2 work* can I have in a single MAS?"
- This depends on how JES-intensive the workload is



JES2 considerations

Two areas to examine to provide high availability and scalability in a JES2 MAS

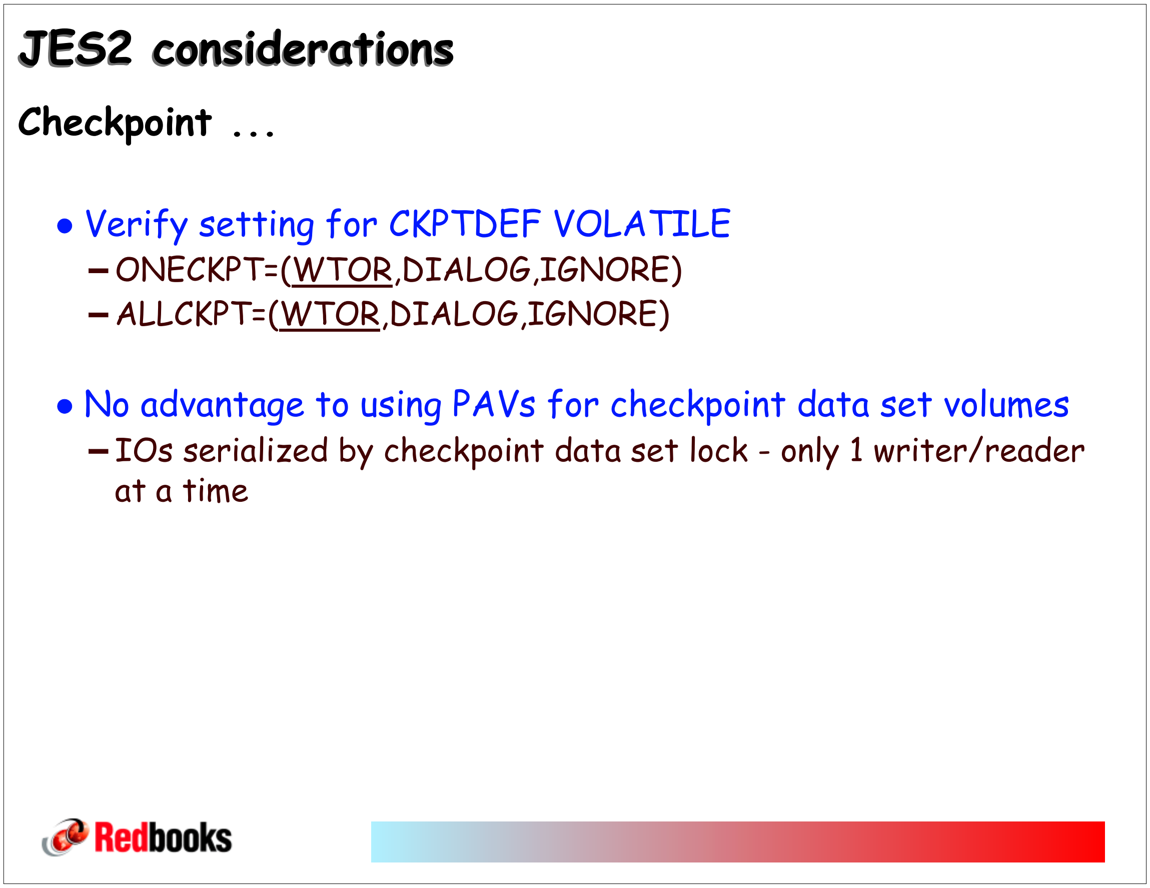
- Checkpoint
- Spool



JES2 considerations

Checkpoint

- ALWAYS use two checkpoint data sets
- No advantage to placing CKPT2 on CF
  - CKPT1 on CF, CKPT2 on DASD (CKPT2 only updated once for every CKPT1 write)
- JES2 does support System Managed Structure Duplexing
  - Eliminates Reconfiguration Dialog during CF or structure failure
  - Review performance considerations!
- Change CKPTDEF OPVERIFY=YES (default) to NO



JES2 considerations

Checkpoint ...

- Verify setting for CKPTDEF VOLATILE
  - ONECKPT=(WTOR,DIALOG,IGNORE)
  - ALLCKPT=(WTOR,DIALOG,IGNORE)
- No advantage to using PAVs for checkpoint data set volumes
  - IOs serialized by checkpoint data set lock - only 1 writer/reader at a time

JES2 considerations

SPOOL

- PAVs can be beneficial
  - I/Os done can performed by multiple address spaces and on multiple systems at the same time
- Consider SPOOL fencing
  - Without fencing, data spread across all available spool volumes
    - Loss of a single SPOOL volume disrupts many jobs
    - But better performance by spreading I/O load
  - FENCE=YES isolates a job to as few SPOOL volumes as possible
    - Loss of a single SPOOL volume disrupts subset of jobs
- Larger TGSIZE
  - Fewer I/Os for large jobs
    - 15-30% improvement in batch performance
  - Less efficient use of space

Redbooks

®

Sysplex Aggrevgation

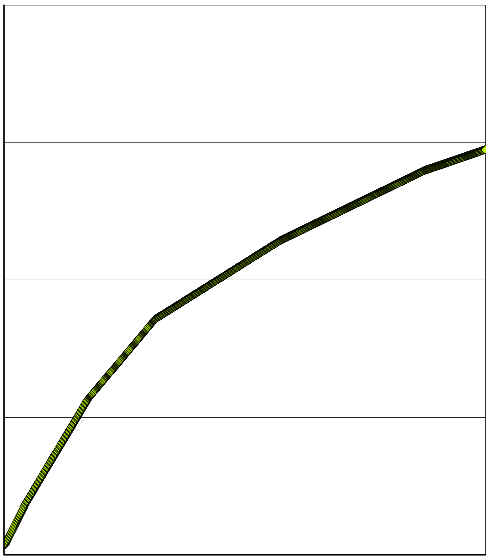
Redbooks

Sysplex aggregation

The objective of sysplex aggregation is to reduce the software cost of incremental growth on zSeries

- The cost for each additional MSU DEcreases as the number of MSUs INcreases
- Sysplex aggregation lets you pay for software on 2 or more CPCs as if they were 1 large CPC
- Why is this good? Because the savings can be 10Ks to 100Ks per month

Software pricing curve



Redbooks

Sysplex aggregation

So, what are the requirements in order to be able to aggregate CPCs?

- Must be connected to the same Common Time Source and at least one common CF
- Must be running MVS 5.2.2 or later
- All systems in the qualifying systems must be running at least one common "systems enablement function"
- **The "PrimaryPlex" must account for over 50% of the used MVS-based MSUs on each CPC, averaged over the 40 hours of prime shift in the week**

"PricingPlex" is the term used to describe the group of CPCs that are aggregated

Redbooks

Sysplex aggregation

Anything else of interest?

- Only affects PSLC or VWLC products - OTC or Fixed Workload products not affected
- Capacity consumed on zAAPs is not included in the calculations
- You MUST include SMF 70s for every system running on the PricingPlex CPCs
- What if the systems in the 'plex have different GMT offsets?
- You must submit a Sysplex Verification Package (available on the Web from http://ibm.com/zseries/swprice/sysplex/pdf/svp.pdf)
  - When you add a new CPC to the PricingPlex
  - When you create a new PricingPlex
  - You have an anniversary for an overlay agreement (OIO or ELA or similar)
- Make sure you pull the latest version of the PLEXCALC tool:
  - http://ibm.com/servers/eserver/zseries/swprice/sysplex/sysplex_calc.html

Sysplex aggregation

Anything else of interest?

- You must submit a Sysplex Verification Package (available on the Web from http://ibm.com/zseries/swprice/sysplex/pdf/svp.pdf)
  - When you add a new CPC to the PricingPlex
  - When you create a new PricingPlex
  - You have an anniversary for an overlay agreement (OIO or ELA or similar)
- Make sure you pull the latest version of the PLEXCALC tool (dated at least Sept 19, 2005):
  - http://ibm.com/servers/eserver/zseries/swprice/sysplex/sysplex_calc.html

Sysplex aggregation

```
=============== SYSPLEX CALCULATOR ===============

Release Date        7/31/2005
Customer Name       CUSTOMER NAME

Machine             Serial     MSUs     LPARs
CPC1                   11111      248 FK15D, FK15A(1), FK15B
CPC2                   22222      492 FKI7A(1), FKI7C, FKI7B
CPC3                   33333      402 FKI1B(1), FKI1A
CPC4                   44444      410 FKI14A
CPC5                   55555      350 FKI4D, FKI4A, FKI4B(1), FKI4C
CPC6                   66666      392 FKI5A(1), FKI5B
CPC7                   77777      330 FKI12A(1), FKI12B
CPC8                   88888      410 FKI17A
CPC9                   99999      187 FKI10A(1), FKI10B, FKI10C, FKI10I

This Sysplex Calculator is designed to enable you to analyze your sysplex environment for compliance with the LPAR usage criterion
of IBM's Parallel Sysplex Aggregation rules.  The tool assumes, but does not verify, your compliance with the other criteria for Parallel Sysplex aggregation.
For a complete list of the Parallel Sysplex Aggregation criteria, please visit http://ibm.com/zseries/swprice/sysplex

This Sysplex Calculator is not guaranteed to be error-free and is provided to you on an 'AS-IS' basis, with no warranties of any kind, express
or implied, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.  Use of the Sysplex
Calculator is not designed to arrive at a conclusive determiniation of your eligibility for Parallel Sysplex Aggregation, which determination
may only be made by IBM.  Please see your IBM representative for details.


Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC1 is a member of FPKD
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC2 is a member of FPKE
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC3 is a member of FPKU
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC4 is a member of FPKE
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC5 is a member of FPKU
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC6 is a member of FPKW
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC7 is a member of FPKW
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC8 is a member of FPKW
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC9 is a member of FPKD

Machine =>          CPC1      CPC1      CPC1            CPC2      CPC2      CPC2            CPC3      CPC3            CPC4            CPC5      CPC5
LPAR =>             FK15D     FK15A     FK15B           FKI7A     FKI7C     FKI7B           FKI1B     FKI1A           FKI14A          FKI4D     FKI4A
Sysid =>            DIPJ      DIPN      NWRD            DIPH      NWRA      PMEA            DIPS      DIPU            DIPL            DIPQ      DIPT
SysName =>          DIPJ      DIPN      NWRD            DIPH      NWRA      PMEA            DIPS      DIPU            DIPL            DIPQ      DIPT
Plex =>             FPKJ      FPKD      FPKT            FPKE      FPKN      FPKE            FPKP      FPKU            FPKE            AMXQ      FPKP
Contribution % =>     32.40%    65.50%     2.00%          77.00%     8.00%    15.00%          43.30%    56.70%         100.00%          15.70%    30.70%
Contribution MSU =>       68       138         4             248        26        48             124       162             319              37        73


Interval
01 Jun 05 - 09:00         24       158                       352        36        52             131       211             347              38        68
01 Jun 05 - 10:00         38       155                       373        38        59             114       211             373              47        54
01 Jun 05 - 11:00         44       142                       377        38        59             109       211             377              34        44
01 Jun 05 - 12:00         66       146                       353        36        50             121       205             344              44        55
01 Jun 05 - 13:00         43       138                       338        34        60             172       182             343              34        65
01 Jun 05 - 14:00         62       135                       334        33        48             163       177             325              36        61
01 Jun 05 - 15:00        110       118                       307        32        47             142       169             271              45       106

(1) Times of other LPARs were reset to match this LPAR.
```

Sysplex aggregation

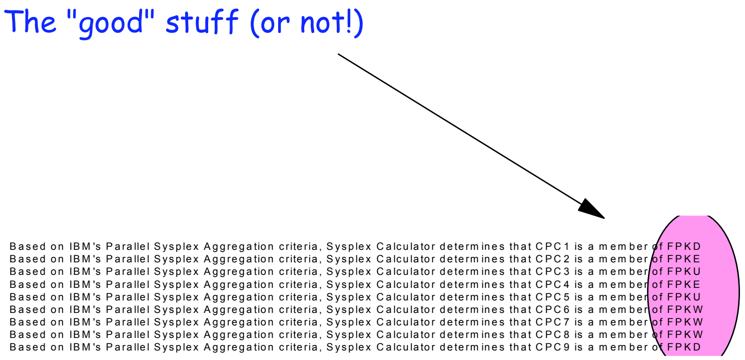
Indicates PLEXCALC version

```
=============== SYSPLEX CALCULATOR ===============

Release Date        7/31/2005
Customer Name       CUSTOMER NAME

Machine             Serial     MSUs     LPARs
CPC1                   11111      248 FK15D, FK15A(1), FK15B
CPC2                   22222      492 FKI7A(1), FKI7C, FKI7B
CPC3                   33333      402 FKI1B(1), FKI1A
CPC4                   44444      410 FKI14A
CPC5                   55555      350 FKI4D, FKI4A, FKI4B(1), FKI4C
CPC6                   66666      392 FKI5A(1), FKI5B
CPC7                   77777      330 FKI12A(1), FKI12B
CPC8                   88888      410 FKI17A
CPC9                   99999      187 FKI10A(1), FKI10B, FKI10C, FKI10I
```

One line per CPC, lists serial, Capacity, LPARs (both those that you provide SMF 70s for and others)

```
This Sysplex Calculator is designed to enable you to analyze your sysplex environment for compliance with the LPAR usage criterion
of IBM's Parallel Sysplex Aggregation rules.  The tool assumes, but does not verify, your compliance with the other criteria for Parallel Sysplex aggregation.
For a complete list of the Parallel Sysplex Aggregation criteria, please visit http://ibm.com/zseries/swprice/sysplex

This Sysplex Calculator is not guaranteed to be error-free and is provided to you on an 'AS-IS' basis, with no warranties of any kind, express
or implied, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.  Use of the Sysplex
Calculator is not designed to arrive at a conclusive determiniation of your eligibility for Parallel Sysplex Aggregation, which determination
may only be made by IBM.  Please see your IBM representative for details.
```

Sysplex aggregation

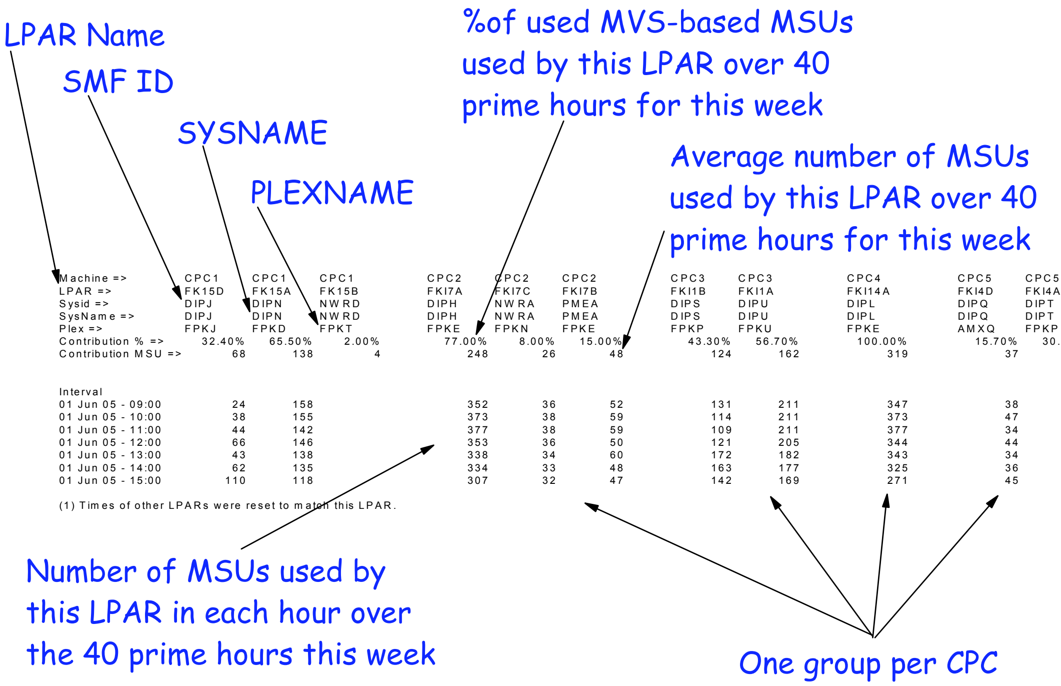


The "good" stuff (or not!)

```
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC1 is a member of FPKD
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC2 is a member of FPKE
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC3 is a member of FPKU
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC4 is a member of FPKE
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC5 is a member of FPKU
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC6 is a member of FPKW
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC7 is a member of FPKW
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC8 is a member of FPKW
Based on IBM's Parallel Sysplex Aggregation criteria, Sysplex Calculator determines that CPC9 is a member of FPKD
```

(1) Times of other LPARs were reset to match this LPAR.

Sysplex aggregation



LPAR Name

SMF ID

SYSNAME

PLEXNAME

%of used MVS-based MSUs used by this LPAR over 40 prime hours for this week

Average number of MSUs used by this LPAR over 40 prime hours for this week

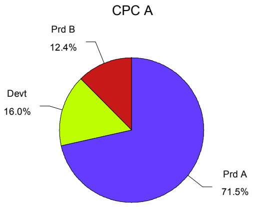
```
Machine =>           CPC1     CPC1     CPC1         CPC2     CPC2     CPC2        CPC3     CPC3        CPC4          CPC5     CPC5
LPAR =>              FK15D    FK15A    FK15B        FKI7A    FKI7C    FKI7B       FKI1B    FKI1A       FKI14A        FKI4D    FKI4A
Sysid =>             DIPJ     DIPN     NWRD         DIPH     NWRA     PMEA        DIPS     DIPU        DIPL          DIPQ     DIPT
SysName =>           DIPJ     DIPN     NWRD         DIPH     NWRA     PMEA        DIPS     DIPU        DIPL          DIPQ     DIPT
Plex =>              FPKJ     FPKD     FPKT         FPKE     FPKN     FPKE        FPKP     FPKU        FPKE          AMXQ     FPKP
Contribution % =>     32.40%   65.50%    2.00%        77.00%    8.00%   15.00%      43.30%   56.70%      100.00%        15.70%   30.70%
Contribution MSU =>       68      138        4           248       26       48         124      162          319            37       73


Interval
01 Jun 05 - 09:00         24      158                    352       36       52         131      211          347            38       68
01 Jun 05 - 10:00         38      155                    373       38       59         114      211          373            47       54
01 Jun 05 - 11:00         44      142                    377       38       59         109      211          377            34       44
01 Jun 05 - 12:00         66      146                    353       36       50         121      205          344            44       55
01 Jun 05 - 13:00         43      138                    338       34       60         172      182          343            34       65
01 Jun 05 - 14:00         62      135                    334       33       48         163      177          325            36       61
01 Jun 05 - 15:00        110      118                    307       32       47         142      169          271            45      106

(1) Times of other LPARs were reset to match this LPAR.
```

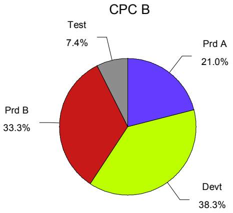
Number of MSUs used by this LPAR in each hour over the 40 prime hours this week

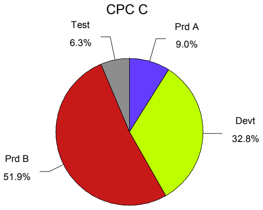
One group per CPC

Sysplex aggregation

Can we aggregate this configuration?



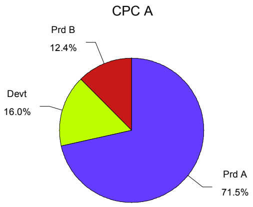


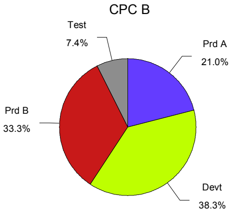


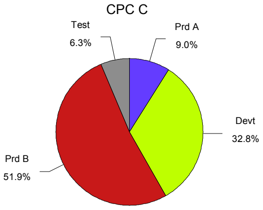
Could we aggregate if there was workload rebalancing?

Sysplex aggregation

Can we aggregate this configuration?







1000 MSUs

200 MSUs

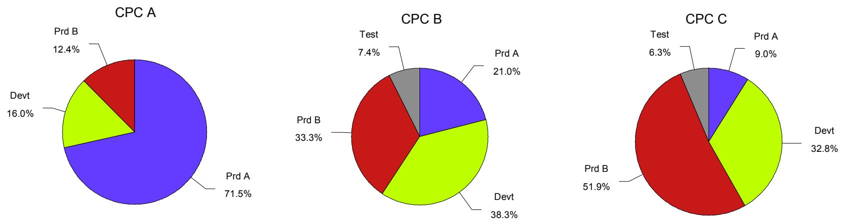
150 MSUs

Could we aggregate if there was workload rebalancing?

	Prd A	Prd B	Devt	Test
CPC A	715	124	160	0
CPC B	42	66	77	15
CPC C	14	78	50	10
Total	771	268	287	25

Sysplex aggregation

Can we aggregate this configuration?

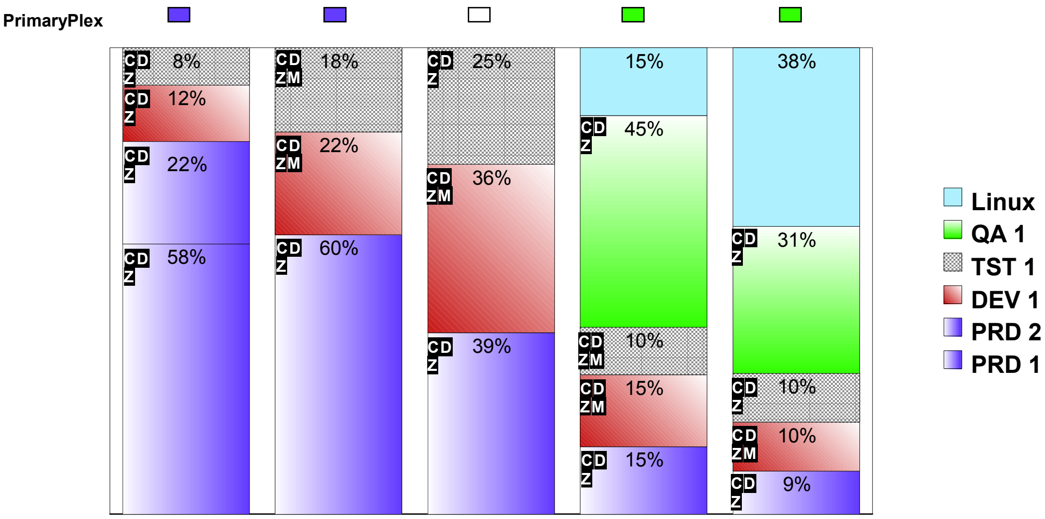


100 MSUs 200 MSUs 1000 MSUs

Could we aggregate if there was workload rebalancing?

	Prd A	Prd B	Devt	Test
CPC A	72	12	16	0
CPC B	42	66	77	15
CPC C	90	519	328	63
Total	204	593	421	78

Sysplex aggregation



	CPC A	CPC B	CPC C	CPC D	CPC E
MSUs	1500	1000	1000	1250	1250

Key
C CICS
D DB2
M WebSphere MQ
Z z/OS

Sysplex aggregation

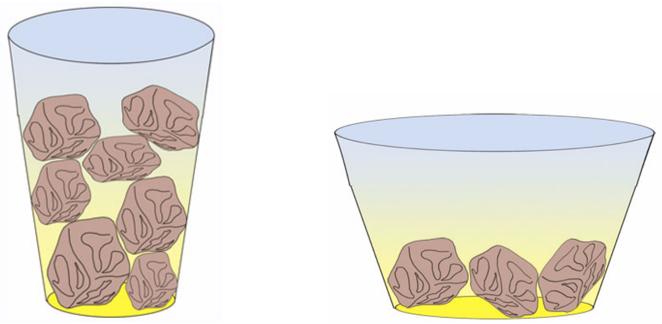
The good news:

Configuration	Monthly cost in "Redbits"	Savings due to aggregation
No aggregation	1,897K	N/A
CPCs A & B aggregated, CPCs D and E aggregated, CPC C not aggregated	1,535K	19%
All CPCs aggregated (if you could achieve this)	1,139K	40%

Sysplex aggregation

The not-so good news:

Q: What happens when you upgrade?

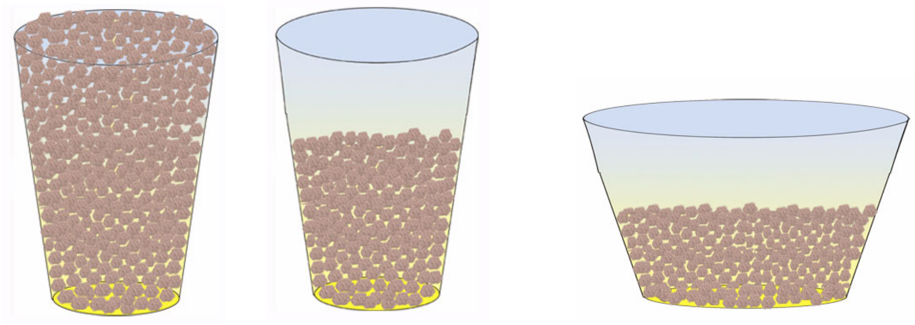


A: Maintaining qualification is not so easy

Sysplex aggregation

So how do we address this?

Q: What would happen if the "stones" were "pebbles"?



A: Work will flow (at the txn level) more evenly around the available capacity

Redbooks

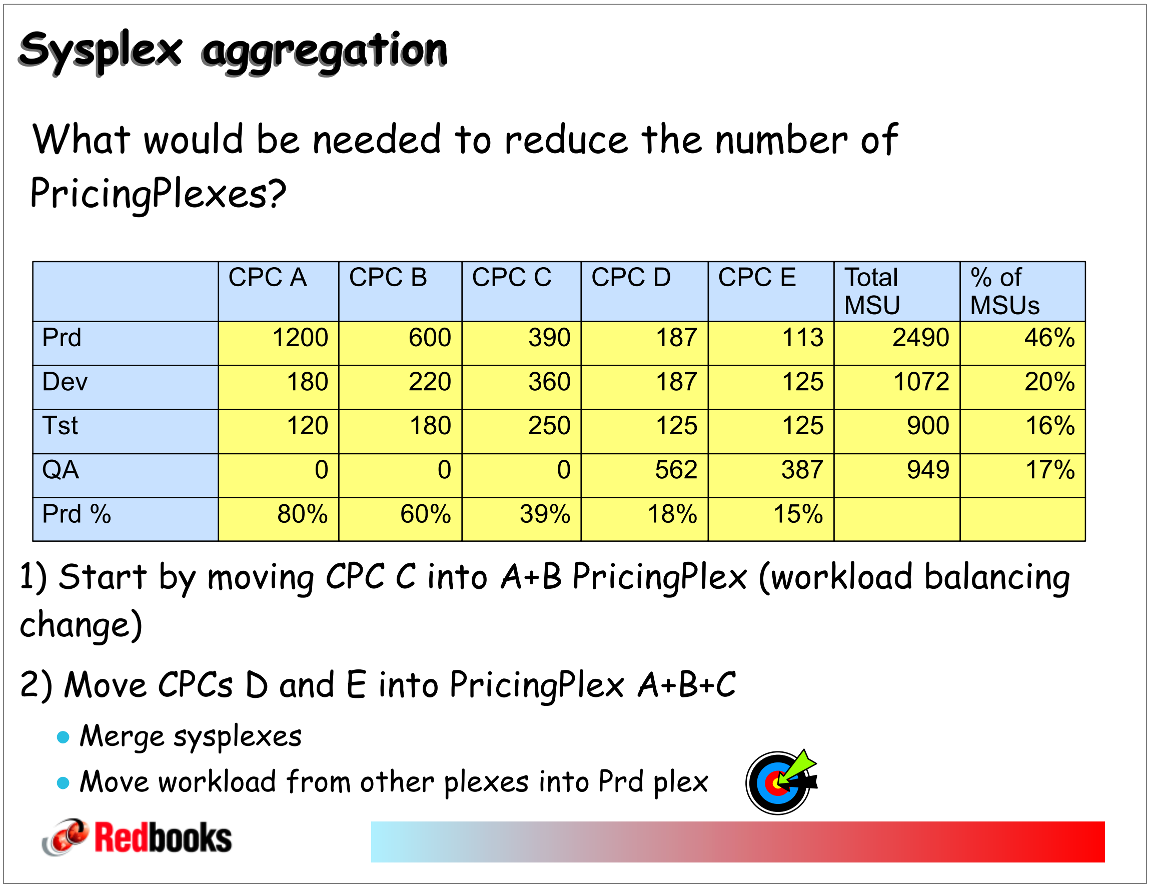
Sysplex aggregation

Stones and pebbles?

- "Stones" are monolithic workloads that must be moved as a single unit - a DB2 subsystem, all connected CICS regions, batch jobs, remote users, and so on (that is, non-data sharing)
- "Pebbles" are independent units of work in a data sharing environment - CICS transactions, individual batch jobs, single users, and so on.

To get the benefit of balanced pebbles, you need dynamic workload balancing in addition to data sharing....

Redbooks



Sysplex aggregation

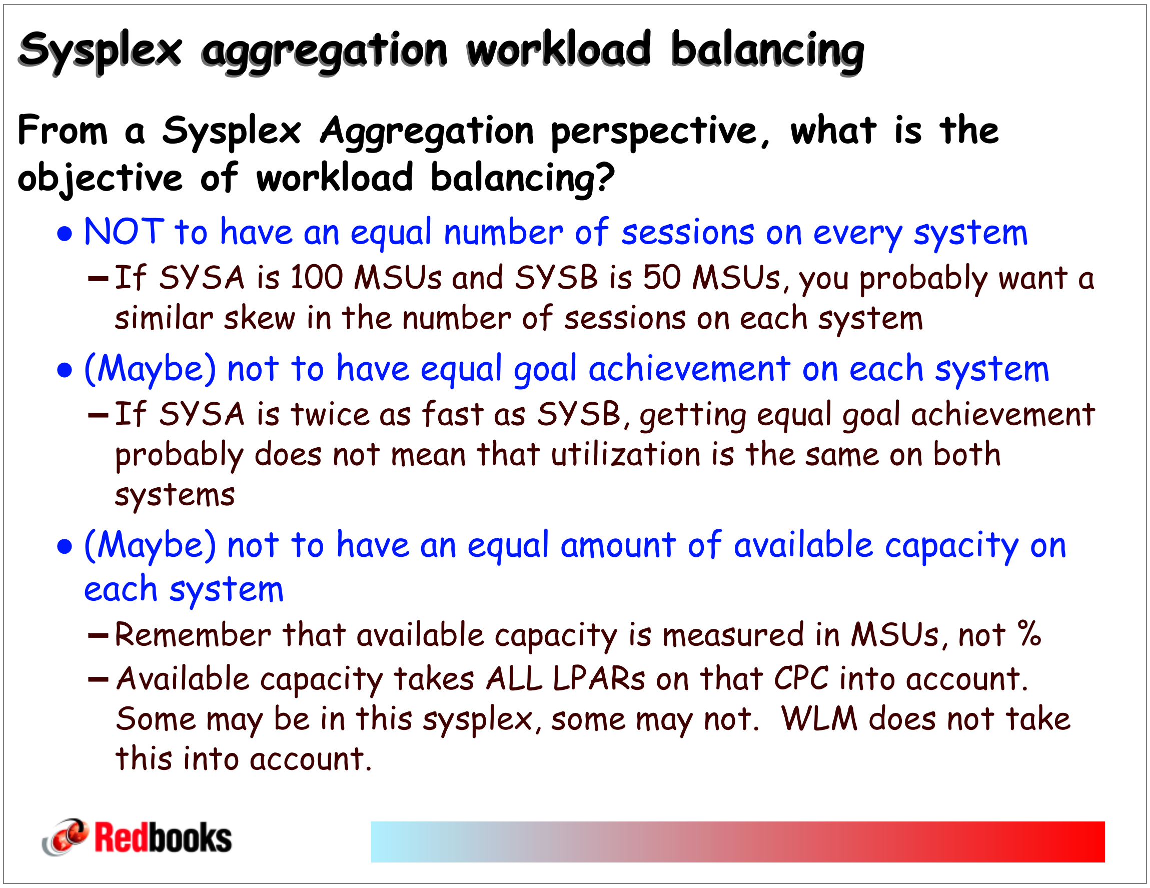
What would be needed to reduce the number of PricingPlexes?

	CPC A	CPC B	CPC C	CPC D	CPC E	Total MSU	% of MSUs
Prd	1200	600	390	187	113	2490	46%
Dev	180	220	360	187	125	1072	20%
Tst	120	180	250	125	125	900	16%
QA	0	0	0	562	387	949	17%
Prd %	80%	60%	39%	18%	15%		

1) Start by moving CPC C into A+B PricingPlex (workload balancing change)

2) Move CPCs D and E into PricingPlex A+B+C

- Merge sysplexes
- Move workload from other plexes into Prd plex



Sysplex aggregation workload balancing

From a Sysplex Aggregation perspective, what is the objective of workload balancing?

- NOT to have an equal number of sessions on every system
  - If SYSA is 100 MSUs and SYSB is 50 MSUs, you probably want a similar skew in the number of sessions on each system
- (Maybe) not to have equal goal achievement on each system
  - If SYSA is twice as fast as SYSB, getting equal goal achievement probably does not mean that utilization is the same on both systems
- (Maybe) not to have an equal amount of available capacity on each system
  - Remember that available capacity is measured in MSUs, not %
  - Available capacity takes ALL LPARs on that CPC into account. Some may be in this sysplex, some may not. WLM does not take this into account.

Sysplex aggregation workload balancing

So, what IS the objective of workload balancing?

- To let the system *automatically* get you closer to a balanced utilization
- Rather than having to make large, disruptive, configuration changes, you can tweak the balance using system controls

Why doesn't workload balancing / WLM do this automatically?

- Not all components that play a role in workload balancing communicate with WLM
- Sysplex aggregation is a pricing mechanism, not a technical one....

Sysplex aggregation workload balancing

What are the stated objectives of MVS Workload Manager?

- To manage the resources of the sysplex so that workloads achieve their performance and throughput objectives, based on their importance as specified by the installation,
- To attempt to ensure that resources are not over-committed to the extent that performance is impacted.
- To optimize and maximize the use of sysplex resources in order to get the work through the system/sysplex as quickly as possible.

Don't see anything in here about trying to balance utilization across systems or CPCs, or about trying to achieve the 50% rule....

WLM basics

Important WLM concepts

- "Available capacity":
  - Capacity guaranteed by weight(1) less actual consumed capacity
  - + "Fair" share of unused capacity, based on relative weight
  - Guaranteed capacity is lesser of relative share of capacity or the capacity that can be delivered based on the number of currently online Logical CPs or the capped capacity
  - Ensure APAR OA10006 is applied
  - IWMWSYSQ service available to return avail capacity information
- Goal achievement (Performance Index)
  - Indicator of whether specified goal is being achieved
    - > 1 indicates goal is being missed
    - <= 1 indicates goal is being achieved or exceeded
- Sysplex vs system
  - WLM's primary objective is to achieve goals at the sysplex level
  - Many of the controls within WLM are at the sysplex or MAS level

Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
WLM Managed Inits	Yes	Yes	No	Yes	Can only control # of WLM Inits at the MAS, not the System lvl
WLM Scheduling Environments	Yes	No	Under user control	No	N/A
WLM Resource Groups	Yes	Yes, but at the sysplex, not the system level	N/A	N/A	No
VTAM Generic Resources	Yes		Attempts to balance number of sessions if all PIs are =	Yes	ISTEXCGR VTAM user exit
Sysplex Distributor	Yes	BASEWLM	ROUNDROBIN	SERVERWLM	Sys Dist Routing Policy
Load Balancing Advisor	Yes	As for Sysplex Distributor	No	As for Sysplex Distributor	No

Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
CICSPlex Systems Manager	Only to obtain Txn goals	No	No, but can select AOR based on queue length	Yes, gets response times from AORs directly	Can provide your own EYU9WRAM exit
CICS Data Tables in CF, Temp Storage in CF, Named Counter Server, Global ENQ/DEQ	These are all workload balancing enablers	N/A	N/A	N/A	N/A
CICS MRO structure	No, but CICS supports VTAM GR	N/A	N/A	N/A	Can provide your own Dynamic Transaction Routing Exit

Workload balancing players

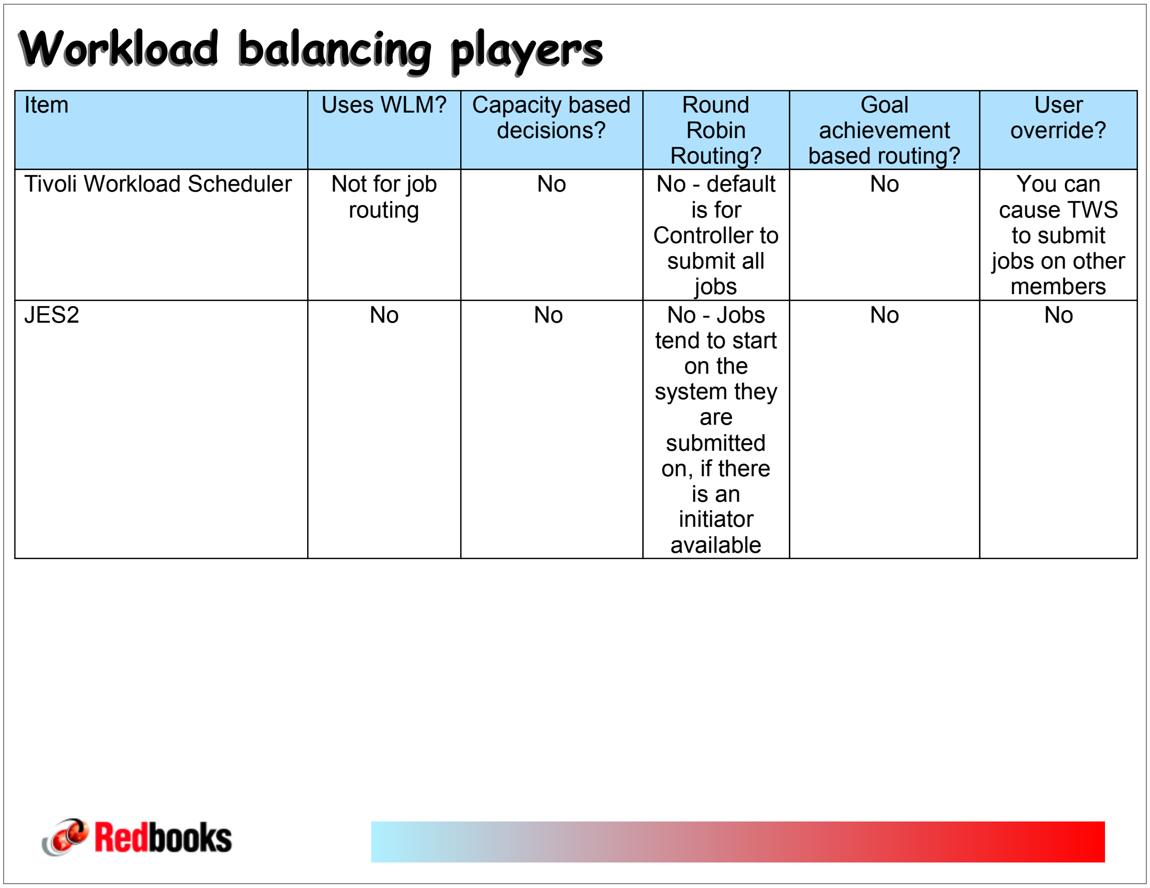
Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
DB2 Connect	Optionally	Yes	No	No	No
DB2 Group Attach Facility - permits batch jobs and CICS regions to connect to any member of the data sharing group	No	No	No - but if there are > 1 members per system can balance randomly across members	No	No
DB2 DDF	Used to classify txns, but not for routing - uses Sys Dist	N/A	N/A	N/A	
DB2 Sysplex Query Parallelism	No	Yes - distributes request based on number of CPs in each image	No	No	No

Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
IMS Connect	No, but supports Sysplex Distributor	No	Controlled through user exit	No	Yes User Exit or use IMS Connect Extensions
IMS Group Connect (IMSGROUP) - permits batch jobs and CICS regions to connect to any member of the data sharing group	No	No	No - but if there are > 1 members per system can balance randomly across members	No	No
IMS Shared Message Queue	No But IMS does support VTAM GR and Sys Dist	Subsystem with most capacity will tend to pull more messages	N/A	N/A	
IMS Workload Router	No	No	Yes, based on user specified rules	No	User defined

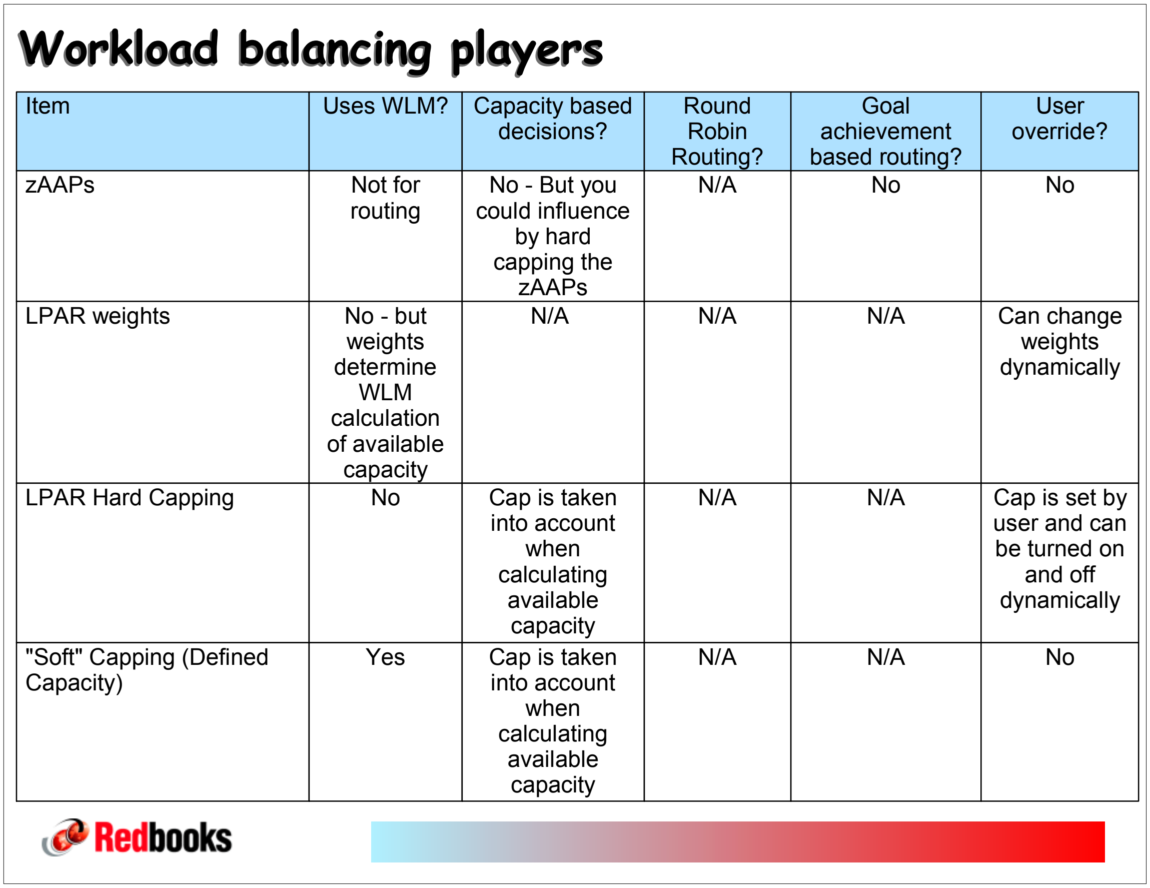
Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
MQ Shared Queue	No, but supports VTAM GR and Sysplex Distributor	Subsystem with most capacity will tend to pull more messages	N/A	N/A	No
WebSphere Application Server	Yes, but only within a single system. Can use Sys Dist to balance requests across WAS regions	No	No	Indirectly. WLM controls number of servant regions to achieve goals. A WAS that is meeting its goals may be given more requests by Sys Dist.	No



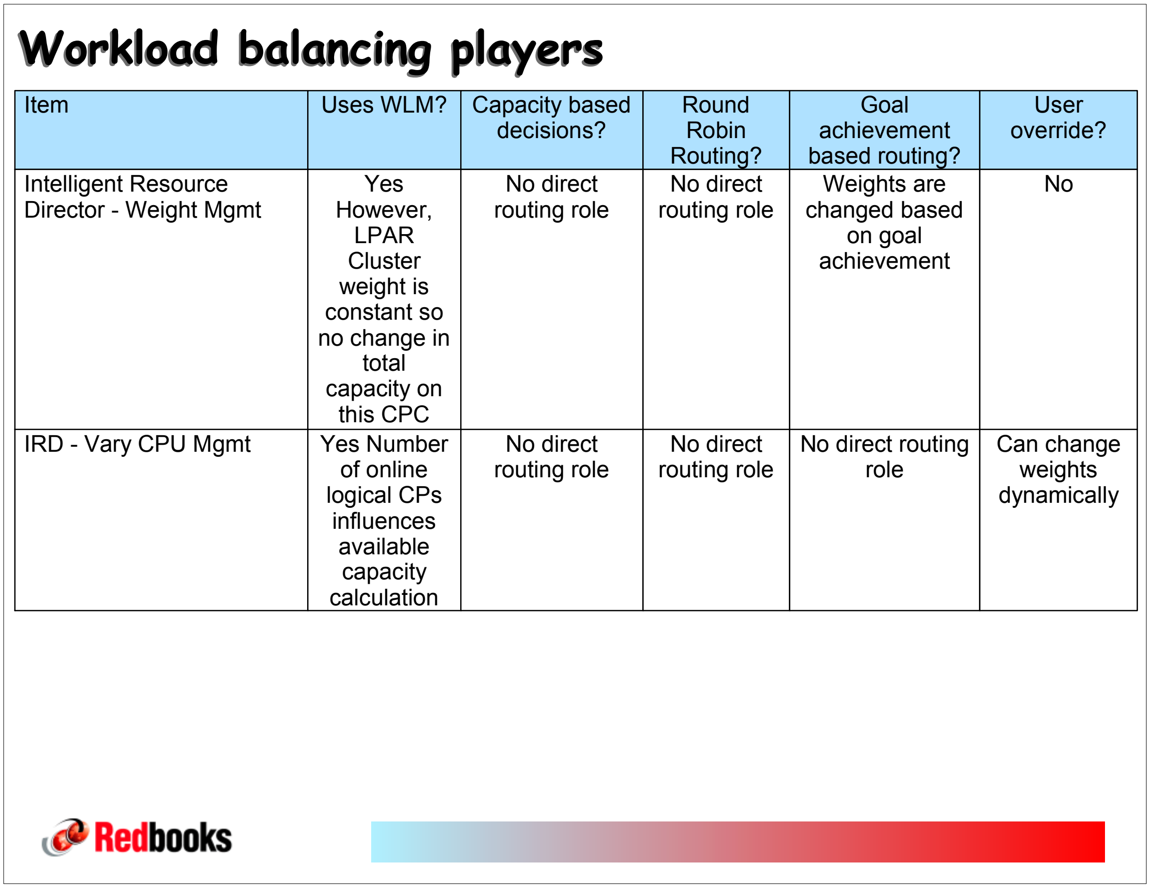
Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
Tivoli Workload Scheduler	Not for job routing	No	No - default is for Controller to submit all jobs	No	You can cause TWS to submit jobs on other members
JES2	No	No	No - Jobs tend to start on the system they are submitted on, if there is an initiator available	No	No



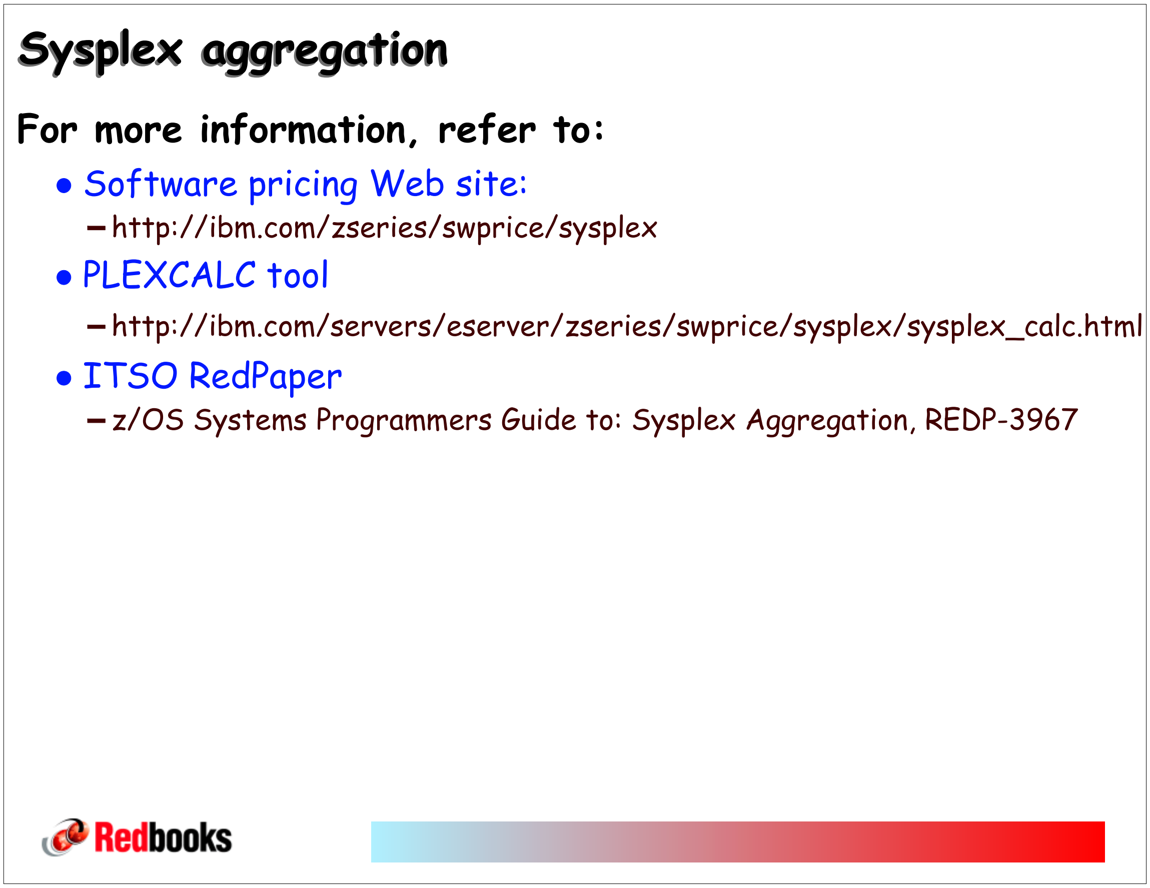
Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
zAAPs	Not for routing	No - But you could influence by hard capping the zAAPs	N/A	No	No
LPAR weights	No - but weights determine WLM calculation of available capacity	N/A	N/A	N/A	Can change weights dynamically
LPAR Hard Capping	No	Cap is taken into account when calculating available capacity	N/A	N/A	Cap is set by user and can be turned on and off dynamically
"Soft" Capping (Defined Capacity)	Yes	Cap is taken into account when calculating available capacity	N/A	N/A	No



Workload balancing players

Item	Uses WLM?	Capacity based decisions?	Round Robin Routing?	Goal achievement based routing?	User override?
Intelligent Resource Director - Weight Mgmt	Yes However, LPAR Cluster weight is constant so no change in total capacity on this CPC	No direct routing role	No direct routing role	Weights are changed based on goal achievement	No
IRD - Vary CPU Mgmt	Yes Number of online logical CPs influences available capacity calculation	No direct routing role	No direct routing role	No direct routing role	Can change weights dynamically



Sysplex aggregation

For more information, refer to:

- Software pricing Web site:
  - http://ibm.com/zseries/swprice/sysplex
- PLEXCALC tool
  - http://ibm.com/servers/eserver/zseries/swprice/sysplex/sysplex_calc.html
- ITSO RedPaper
  - z/OS Systems Programmers Guide to: Sysplex Aggregation, REDP-3967

®

Introduction to "clustering"

Introduction

As Open systems develop, there are many claims made about highly available and scalable clustering solutions that provide 'mainframe-like' functionality(?)/levels of service.

Many zSeries people have very little exposure to Open Systems clustering and are unfamiliar with the terminology of clustering

This presentation aims to give zSeries people an overview of clustering concepts and technologies in open systems and how they relate to zSeries and z/OS functionality.

What is clustering ?

Formal Definition:

- A cluster is a collection of interconnected computers, called nodes or cluster members, working together as a single system.

Real world:

- It depends on who you ask !
- A cluster may be anything from a very simple active/standby cluster on two nodes to a massively parallel high performance computing environment with hundreds of nodes

Redbooks

Clustering

Background:

- Remember the history of the different platforms:
- MVS:
  - 40 years of (compatible) development
  - Predominant users are the business community
    - Data integrity is number 1 priority
    - Stability/reliability is next
  - DASD has always been external
  - Hardware was expensive, so users wanted to maximize utilization and minimize number of boxes (many apps per MVS image)
- UNIX/PCs:
  - Single user background
  - Users were education/science/resear h community
    - Low cost and flexibility top priorities
  - Internal/integrated DASD was the model
  - Hardware was cheap - availability delivered by redundancy (typically 1 app per UNIX system)

Redbooks

Why cluster

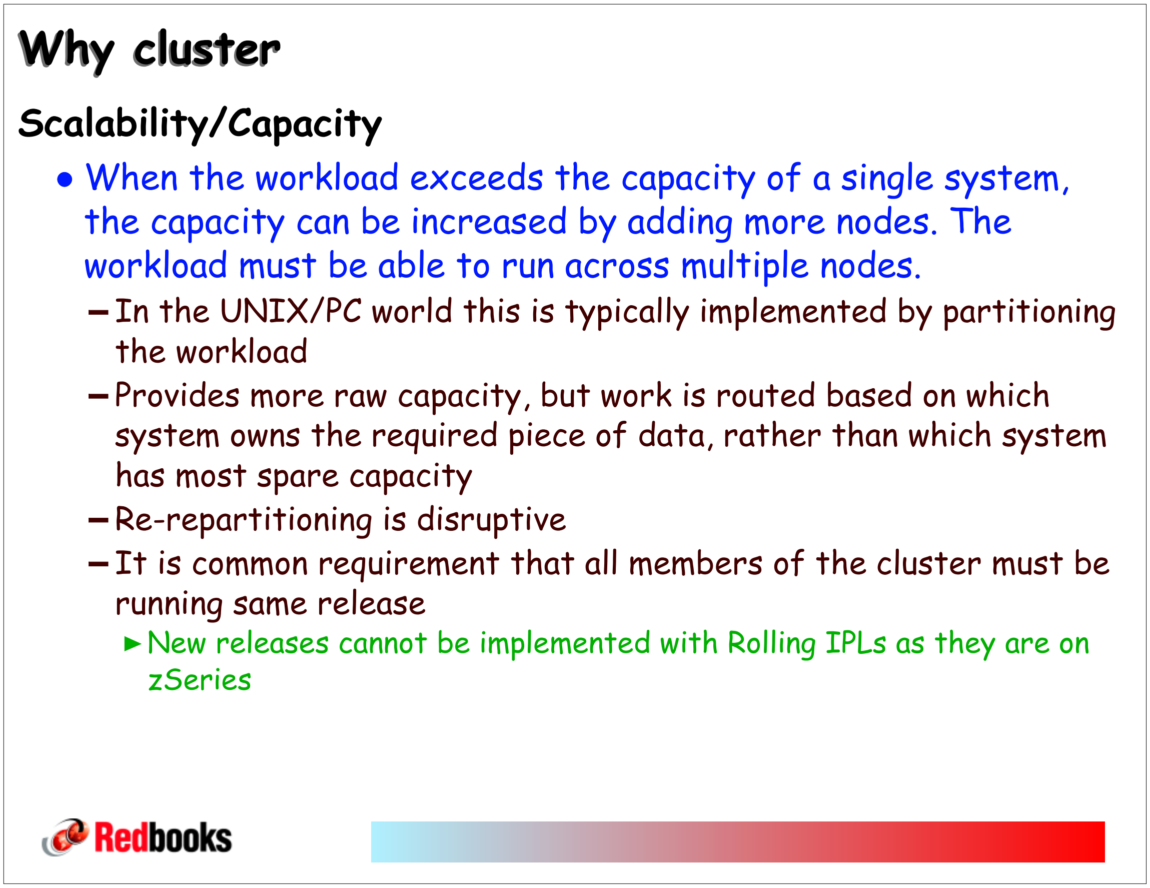
High Availability

- Open systems H/W and Operating systems have traditionally been less resilient than mainfame and MVS. As a result, most critical systems running on Unix or Windows are clustered to some extent to provide high(er) availability.
- This type of cluster has traditionally been a simple active/standby two node cluster.

Why cluster

Manageability

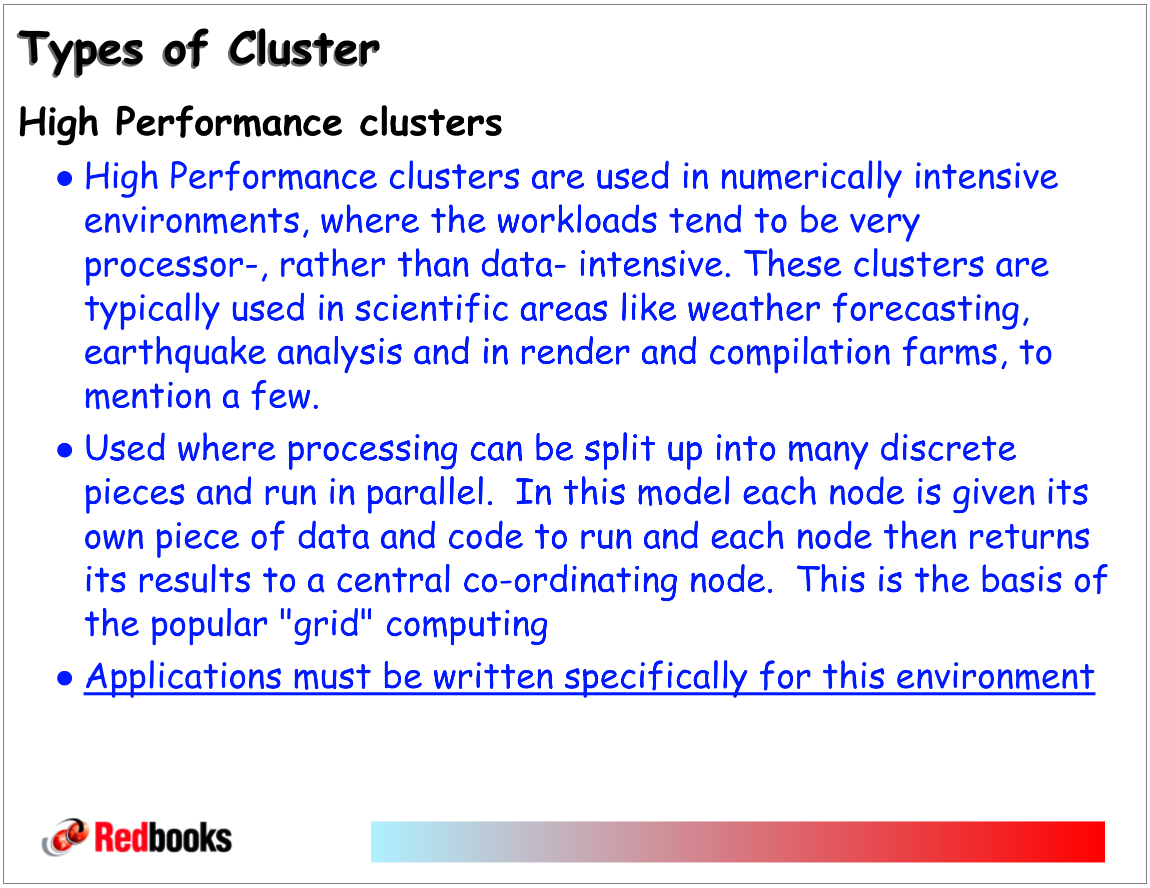
- A cluster may also make systems maintainance easier by allowing nodes to be taken down for upgrades etc while still allowing access to applications. Even in a two node active/standby cluster there are benefits.
  - UNIX systems typically do not allow the same degree of dynamic changes that z/OS does, resulting in more frequent IPLs than would be the norm in z/OS



Why cluster

Scalability/Capacity

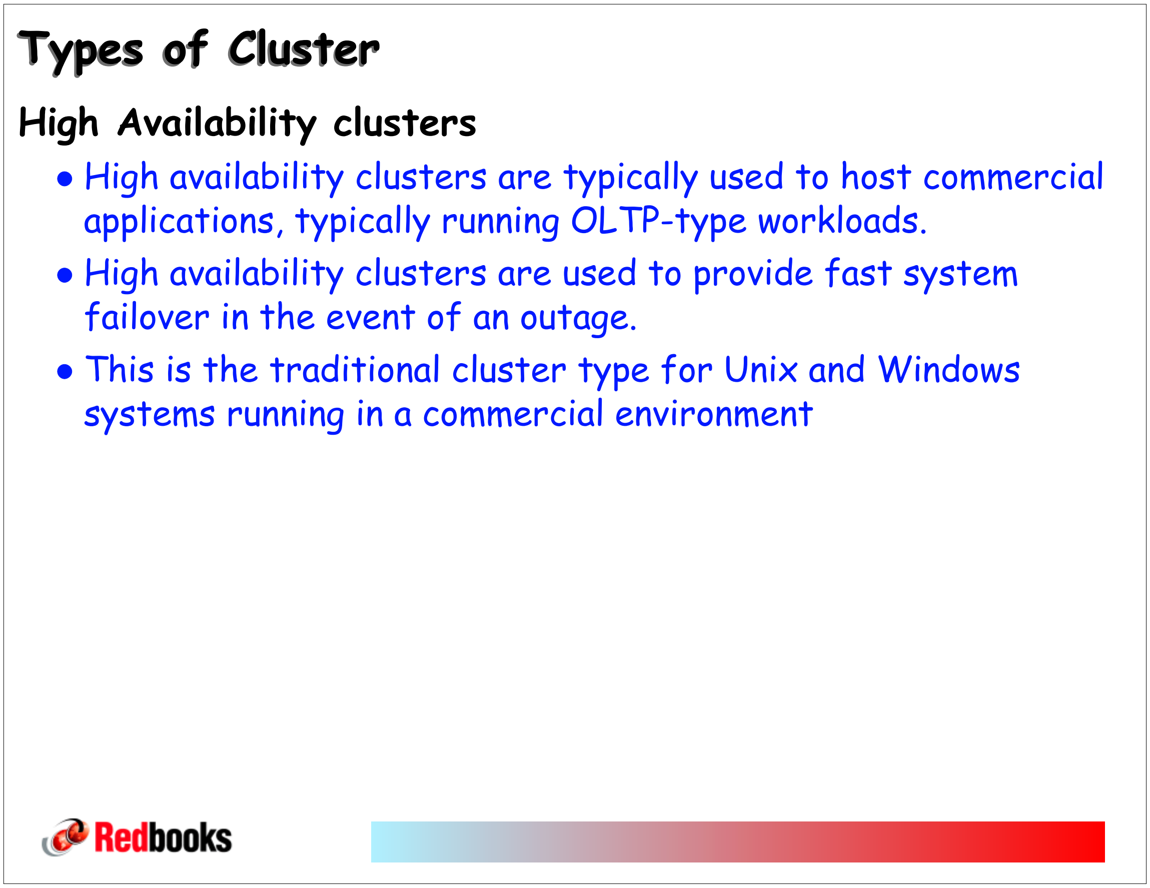
- When the workload exceeds the capacity of a single system, the capacity can be increased by adding more nodes. The workload must be able to run across multiple nodes.
  - In the UNIX/PC world this is typically implemented by partitioning the workload
  - Provides more raw capacity, but work is routed based on which system owns the required piece of data, rather than which system has most spare capacity
  - Re-repartitioning is disruptive
  - It is common requirement that all members of the cluster must be running same release
    - New releases cannot be implemented with Rolling IPLs as they are on zSeries



Types of Cluster

High Performance clusters

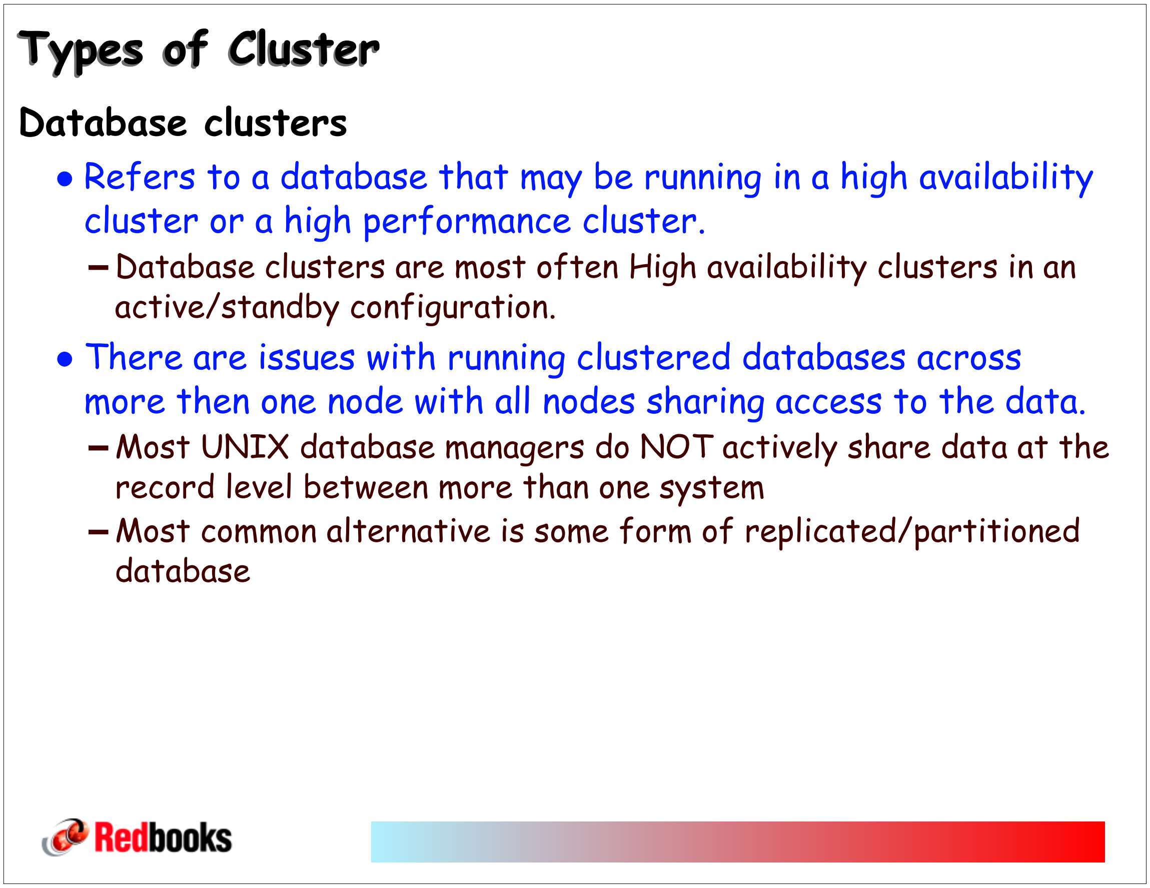
- High Performance clusters are used in numerically intensive environments, where the workloads tend to be very processor-, rather than data- intensive. These clusters are typically used in scientific areas like weather forecasting, earthquake analysis and in render and compilation farms, to mention a few.
- Used where processing can be split up into many discrete pieces and run in parallel. In this model each node is given its own piece of data and code to run and each node then returns its results to a central co-ordinating node. This is the basis of the popular "grid" computing
- Applications must be written specifically for this environment



Types of Cluster

High Availability clusters

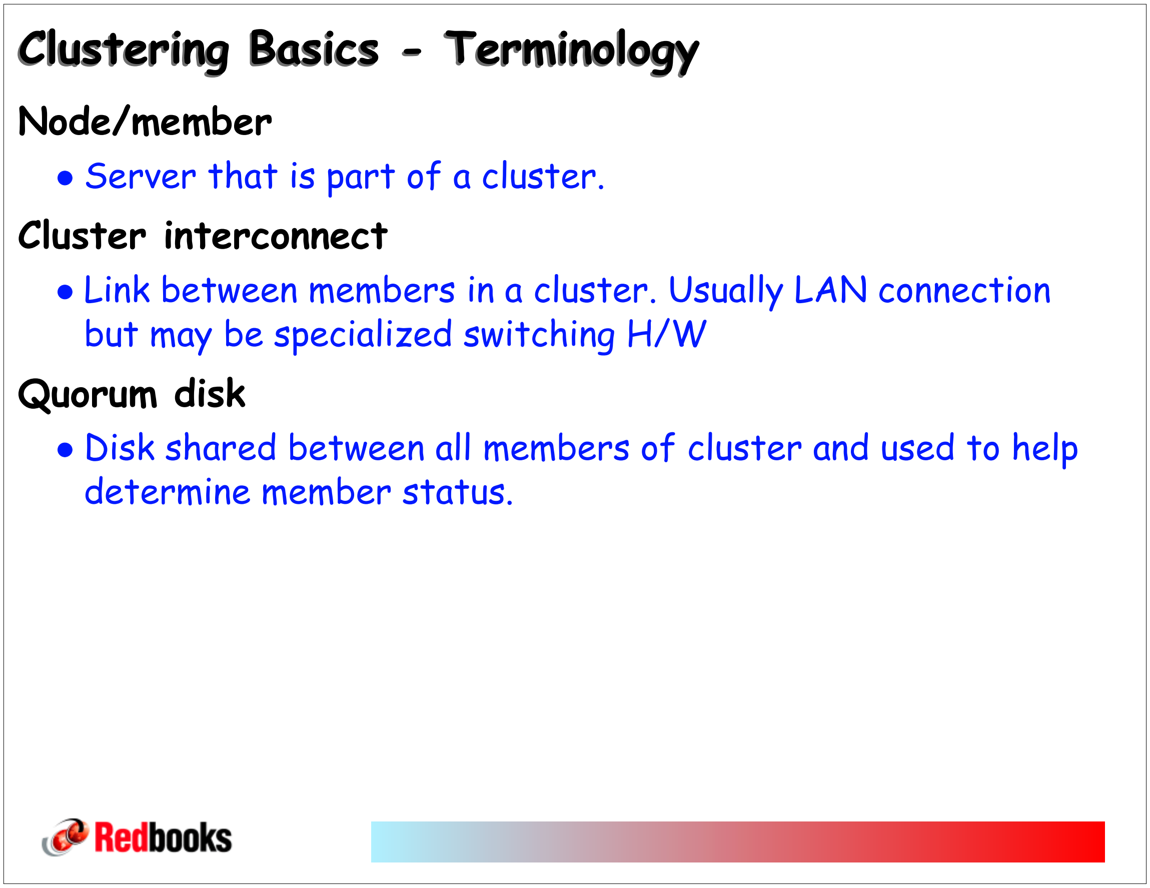
- High availability clusters are typically used to host commercial applications, typically running OLTP-type workloads.
- High availability clusters are used to provide fast system failover in the event of an outage.
- This is the traditional cluster type for Unix and Windows systems running in a commercial environment



Types of Cluster

Database clusters

- Refers to a database that may be running in a high availability cluster or a high performance cluster.
  - Database clusters are most often High availability clusters in an active/standby configuration.
- There are issues with running clustered databases across more then one node with all nodes sharing access to the data.
  - Most UNIX database managers do NOT actively share data at the record level between more than one system
  - Most common alternative is some form of replicated/partitioned database



Clustering Basics - Terminology

Node/member

- Server that is part of a cluster.

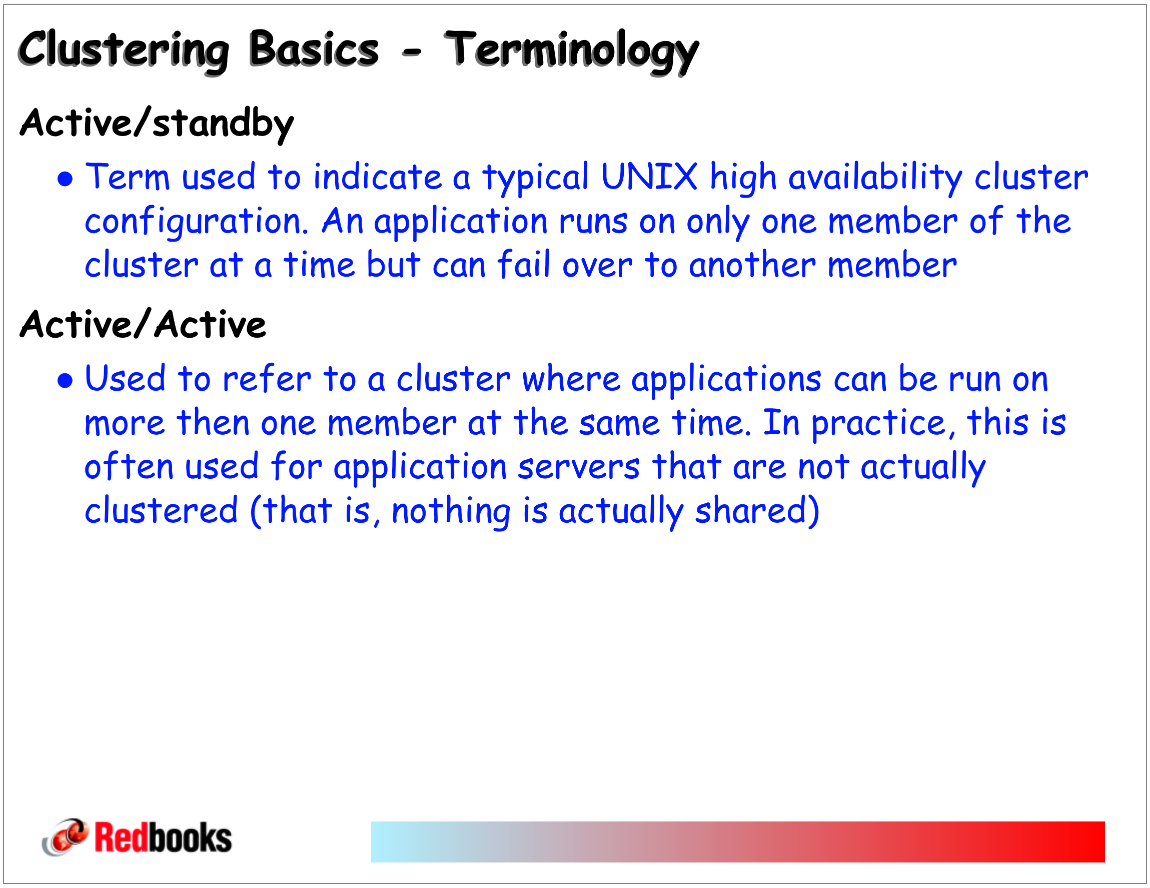
Cluster interconnect

- Link between members in a cluster. Usually LAN connection but may be specialized switching H/W

Quorum disk

- Disk shared between all members of cluster and used to help determine member status.

Redbooks



Clustering Basics - Terminology

Active/standby

- Term used to indicate a typical UNIX high availability cluster configuration. An application runs on only one member of the cluster at a time but can fail over to another member

Active/Active

- Used to refer to a cluster where applications can be run on more then one member at the same time. In practice, this is often used for application servers that are not actually clustered (that is, nothing is actually shared)

Redbooks

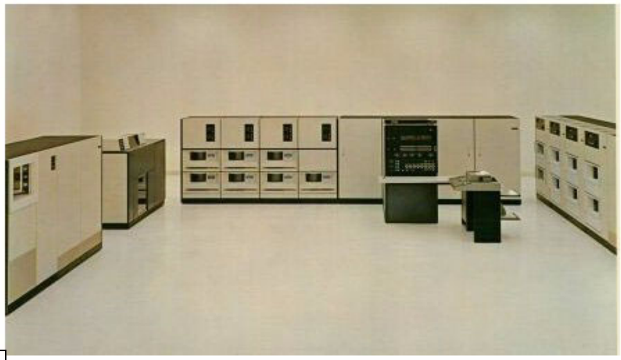
Cluster storage models

Shared nothing

- Shared nothing means that the nodes in a cluster do not actively share data. Data (whether it is a file or a database) is only owned by one node at a time.
- Disks may be defined as shared and accessible to all systems but they are only on-line to one system at a time.
- This is the simplest cluster storage model to implement because it needs the least amount of co-ordination between nodes.
- This is the model used in an active/standby high availablity cluster.
- Example - AIX with HACMP

Cluster models

"Shared nothing" in MVS terms



Active
MVS 1.3.2
1 CICS Region
VSAM
Reserve/Release
DASD ONLINE

Standby
MVS 1.3.2

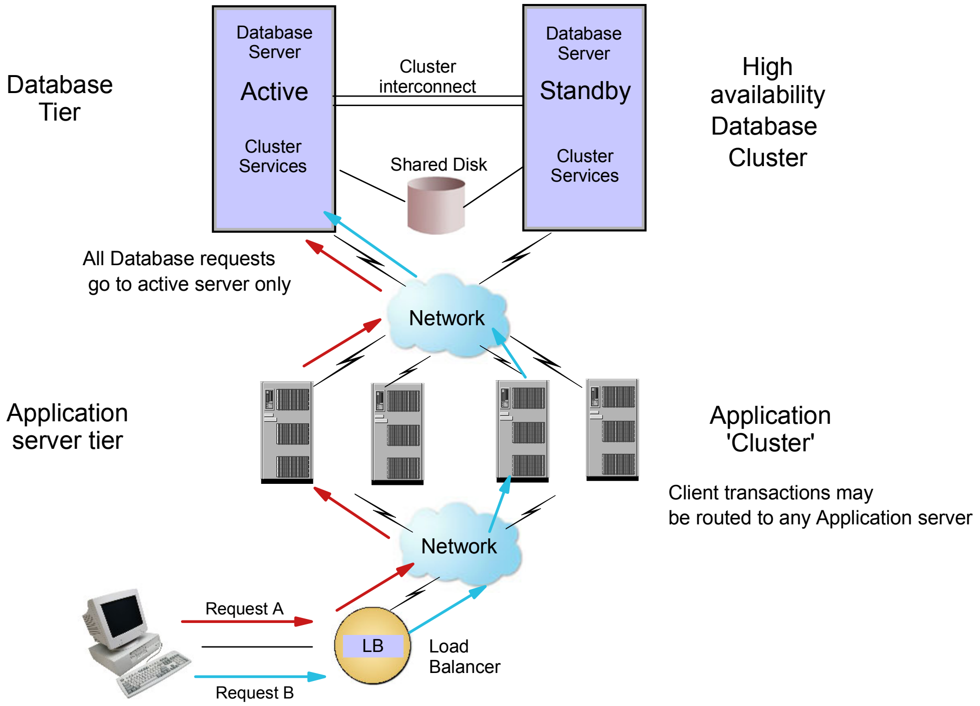
ARM
Reserve/Release
DASD OFFLINE

Cluster storage models - Shared disk

Shared disk

- The filesystem is logically shared among all the nodes, each node is able to access the same data on disk. The shared disk model is appropriate for fault tolerant systems. If a node fails for some reason, the other nodes will still be able to access the data.
- In the shared disk model a mechanism must be in place to be able to serialize the access to shared resources. Only one node may update a file at a time.
- In order to ensure all systems can see all updates,all update activity must be written directly to disk (no buffering or caches); this has performance implications especially for database servers which can only achieve good performance if data is buffered in local cache.
- Example: AIX with HACMP and GPFS

Typical Unix two tier architecture



Cluster models

"Shared disk, active/standby" in MVS terms

Active
MVS 1.3.2
1 CICS Region
VSAM
GRS
DASD ONLINE

Standby
MVS 1.3.2

GRS
ARM
Reserve/Release
DASD ONLINE

Cluster models

"Shared disk, active/active" in MVS terms

Active
MVS 1.3.2
1 CICS Region
VSAM
GRS
ARM
DASD ONLINE

Active
MVS 1.3.2
1 CICS Region
VSAM
GRS
ARM
DASD ONLINE

Cluster storage models - shared cache

Shared Cache

- Shared Disk models have a limitation in that all updates need to be written to disk so that all nodes can see updated data.
- For performance reasons, it is desirable to be able to cache writes in memory but this requires a mechanism to co-ordinate cache entries amongst the nodes in a cluster.
- A shared cache model is the most difficult to implement, it requires both lock management across the nodes and a cache coherency manager.
- We will look some more at how cache entries can be co-ordinated across nodes in a cluster

Performance considerations

To scale well, any shared cache cluster must be able to do lock management and cache management across the nodes in a cluster efficiently.

The test of a good shared cache architecture is how efficiently it scales when requests for data are spread randomly across all nodes in the cluster.

- Think back to IMS data sharing prior to Parallel Sysplex - did not scale very well beyond 2 systems

<u>Any</u> shared cache cluster can scale well if there is a high degree of data affinity for each node.

- Beware of benchmarks or performance papers claiming impressive scaleability where the numbers have been produced using an application environment tuned to ensure a very high degree of data affinity (TPC-C, for example).

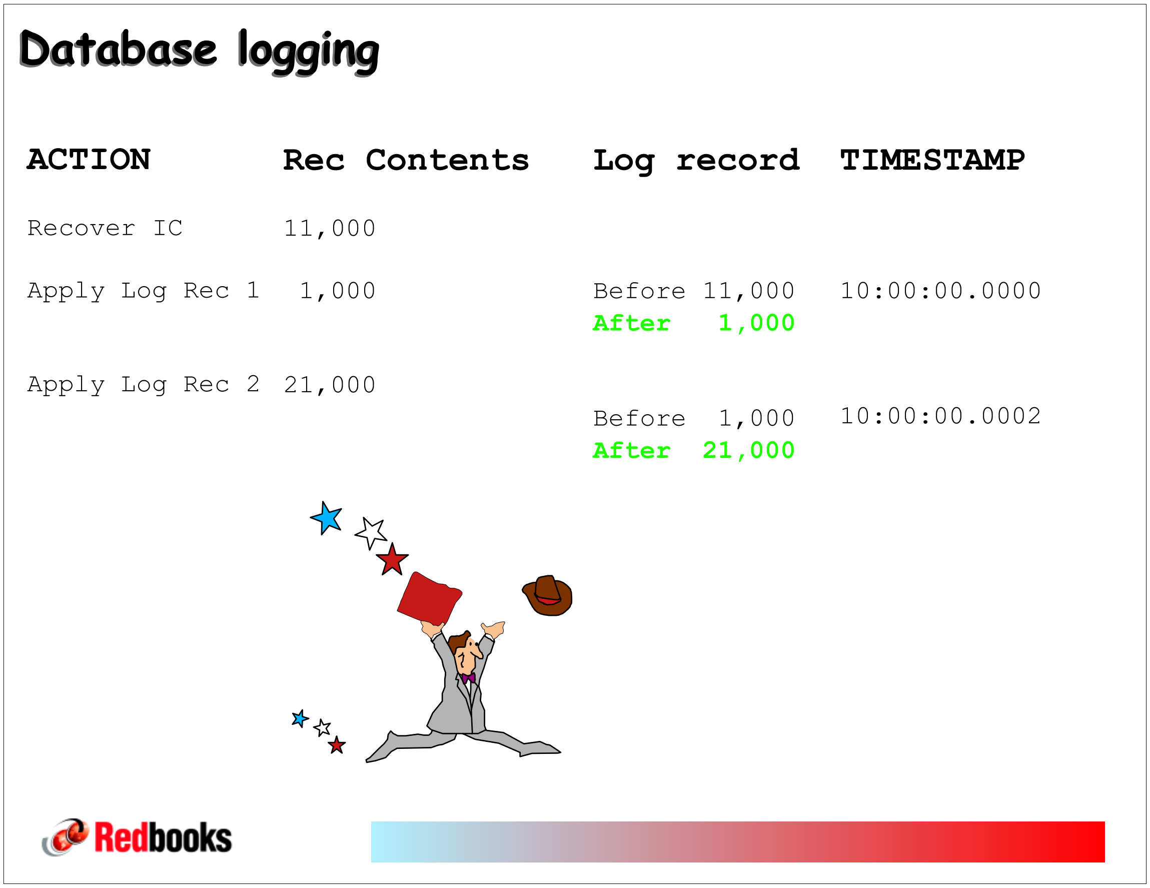
Performance considerations

How well a shared cache cluster scales depends on:

- Amount of inter-node communication required for lock and cache management:
  - How efficient this is relates to the lock and cache topologies
  - The degree of data affinity has a big impact on how much inter-node communication will occur. This is application specific.
- Mechanism for providing a common timestamp or unique identifier that can be used to order events.
  - All database systems need some way of identifying the order in which events occur. This can result in serialization on a global resource across the cluster.
- Speed and latency of the cluster interconnects.

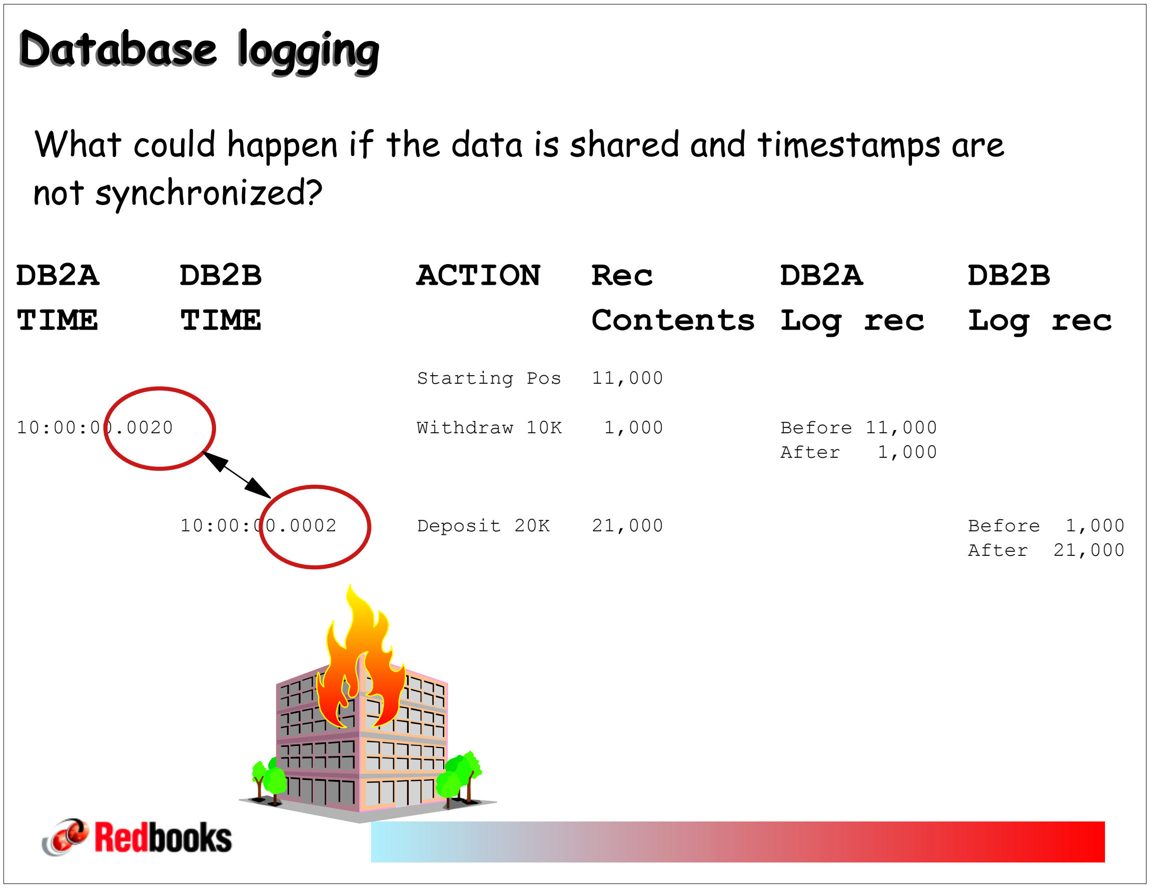
Database logging

TIME	ACTION	Rec Contents	Log record
	Starting Pos	11,000	
10:00:00.0000	Withdraw 10K	1,000	Before 11,000 After 1,000
10:00:00.0002	Deposit 20K	21,000	Before 1,000 After 21,000



Database logging

ACTION	Rec Contents	Log record	TIMESTAMP
Recover IC	11,000		
Apply Log Rec 1	1,000	Before 11,000 **After 1,000**	10:00:00.0000
Apply Log Rec 2	21,000	Before 1,000 **After 21,000**	10:00:00.0002



Database logging

What could happen if the data is shared and timestamps are not synchronized?

DB2A TIME	DB2B TIME	ACTION	Rec Contents	DB2A Log rec	DB2B Log rec
		Starting Pos	11,000		
10:00:00.0020		Withdraw 10K	1,000	Before 11,000 After 1,000	
	10:00:00.0002	Deposit 20K	21,000		Before 1,000 After 21,000

Database logging

ACTION	Rec Contents	Log record	TIMESTAMP
Recover IC	11,000		
Apply Log Rec	21,000	After 21,000	10:00:00.0002
Apply Log Rec	1,000	After 1,000	10:00:00.0020

Shared cache management

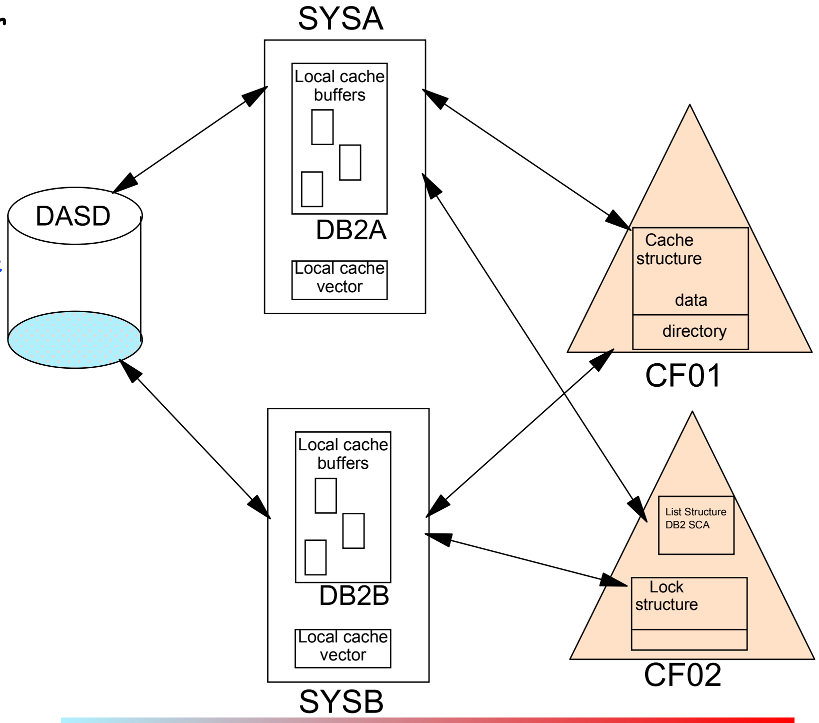
Shared Cache topologies

- Replicated cache
  - All cache entries are replicated across the caches in all nodes. This works fine for read only data but if there is a lot of cache updates this model does not scale well
    - Not used by any major DB manager
- Partitioned cache
  - Each cached data block is “owned” by one node. Each and every node is aware of which node the data blocks in the caches belongs to. Read and write operations usually involves only two nodes , so partitioned cache scales much better than the replicated cache.
    - Used by DB2 for AIX, Oracle RAC, others
- Centralised cache
  - All nodes store and retrieve updated data blocks in a centralised cache. The centralised cache tracks what data blocks are held in local cache in each node and invalidates the local cache copy in the event of the data block being updated by another node.
    - Used by DB2 for z/OS

DB2 for z/OS Data sharing

How does DB2 for z/OS deliver scalability:

- CF Lock structure
- GBP Structure
- CF Cross invalidate



Failure handling

A cluster must be able to identify and handle various failures. Failure types can be:

- Node failure
  - Complete failure of operating system or H/W
- Interconnect failure
  - Node may still be active but cannot communicate with other nodes in the cluster.
- Loss of access to shared disk
  - Node may still be active but cannot access shared disk.
- Loss of network access
  - Node may still be active but cannot communicate with network.

Failure handling

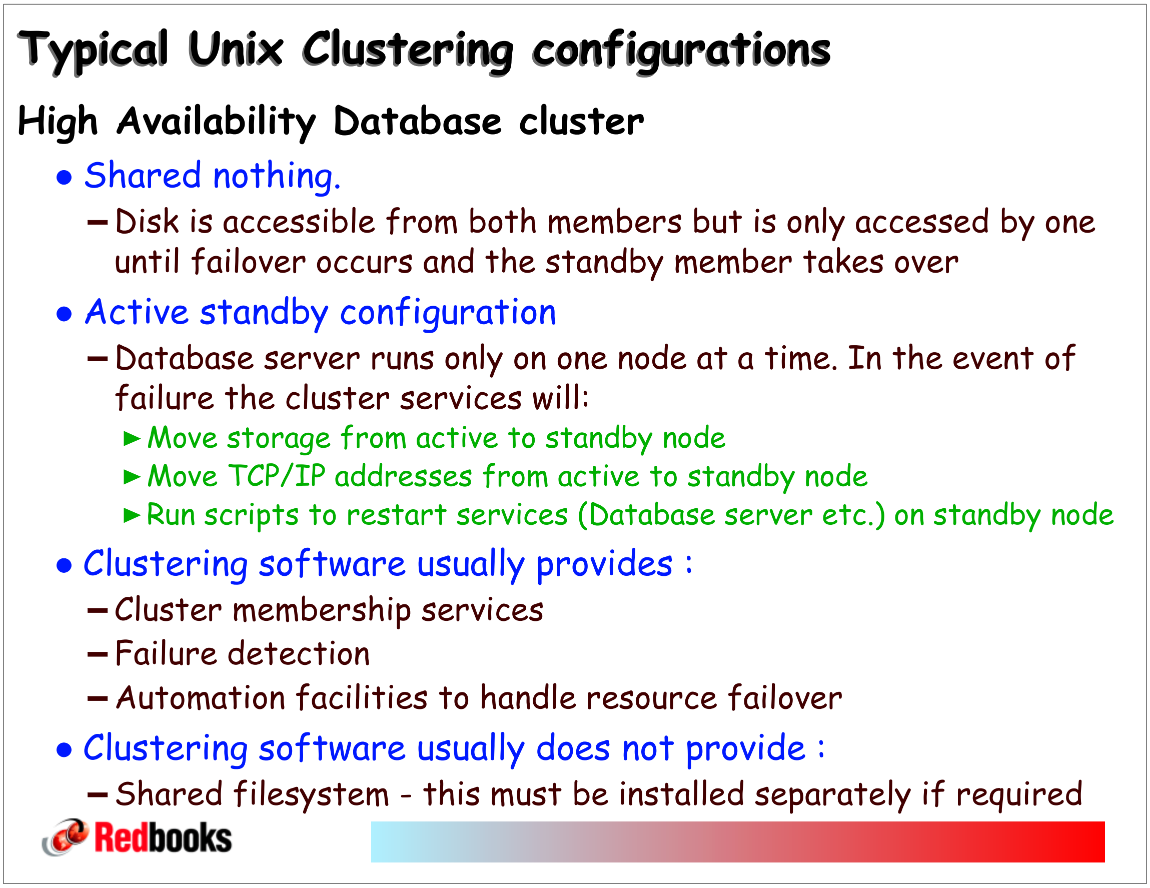
How well a shared cache cluster handles failures depends on:

- Speed of detection of failures.
- How quickly a failed node can be isolated from the cluster
- How the cluster determines which members to partition out of a cluster in the event of loss of communication between members.
- Cluster reconfiguration required to rebuild the shared lock and cache mechanisms to enable full access to data from the remaining nodes.
- The amount of data recovery actions that may be required in the case of multiple failures

Typical Unix Clustering configurations

Application Cluster

- Not really a cluster but often called a cluster !
- Each node is independent with no interconection
  - Runs copy of application
  - usually has common configuration information
  - does not contain any persistant application data - just session information
- Usually fronted by load balancer service
  - may be a web server that load balances across all available application servers.
  - service may be part of the client connection sofware (Eg. DB2 client or SAP client)
- All nodes usually connect to a common backend server
  - usually a database server

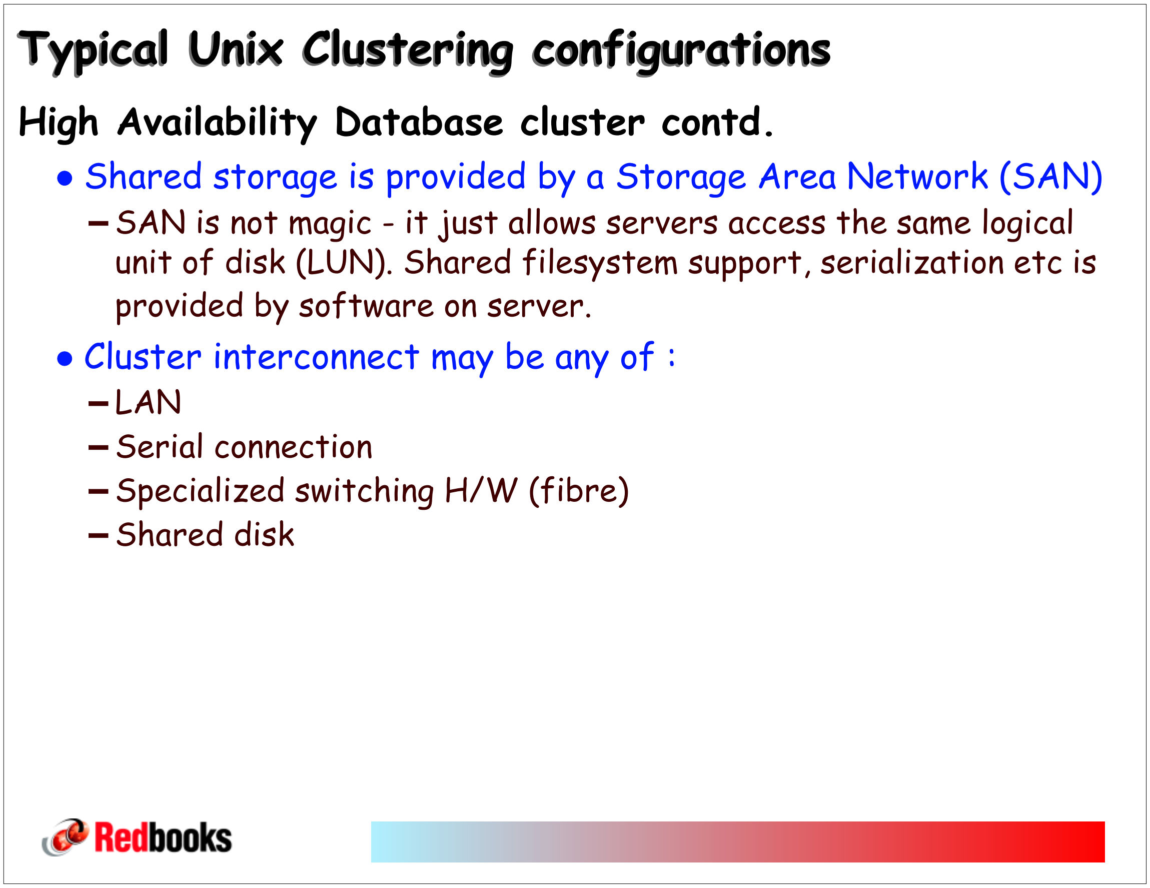


Typical Unix Clustering configurations

High Availability Database cluster

- Shared nothing.
  - Disk is accessible from both members but is only accessed by one until failover occurs and the standby member takes over
- Active standby configuration
  - Database server runs only on one node at a time. In the event of failure the cluster services will:
    - ► Move storage from active to standby node
    - ► Move TCP/IP addresses from active to standby node
    - ► Run scripts to restart services (Database server etc.) on standby node
- Clustering software usually provides :
  - Cluster membership services
  - Failure detection
  - Automation facilities to handle resource failover
- Clustering software usually does not provide :
  - Shared filesystem - this must be installed separately if required

Redbooks

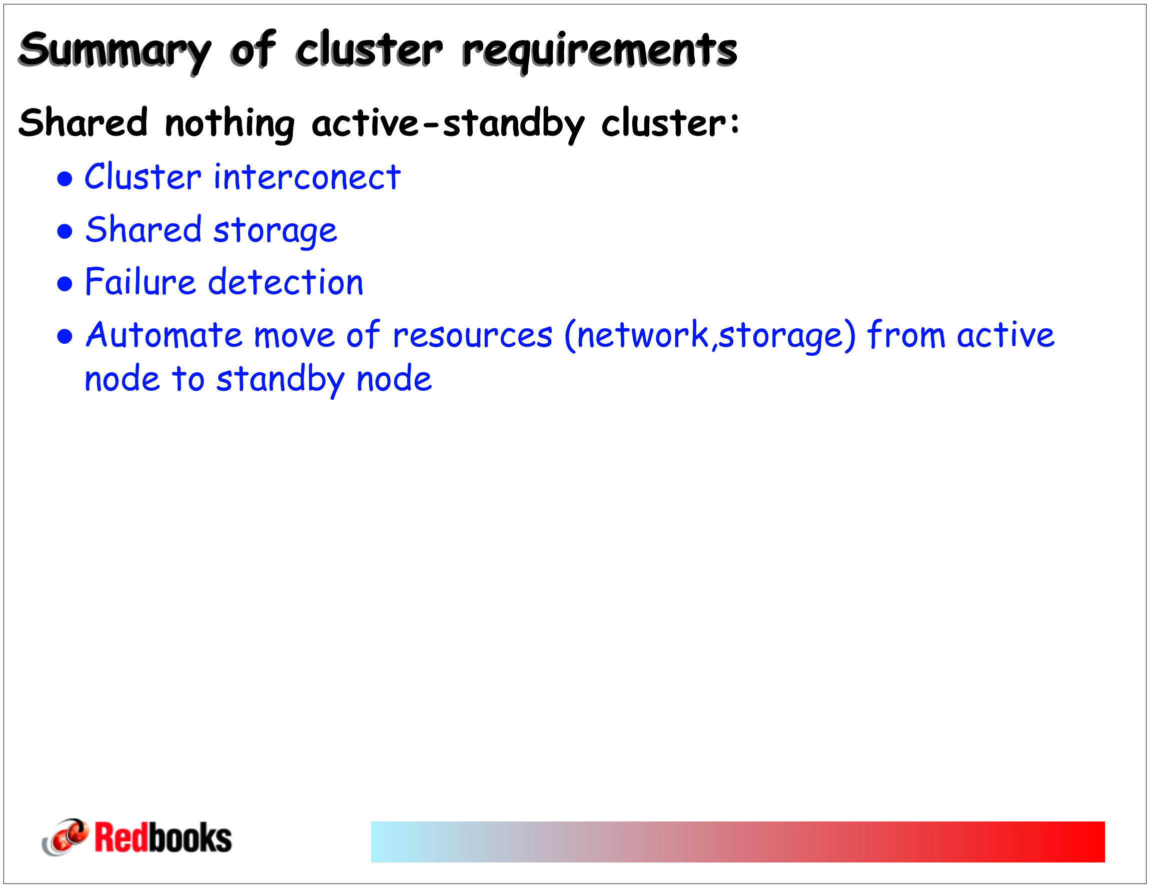


Typical Unix Clustering configurations

High Availability Database cluster contd.

- Shared storage is provided by a Storage Area Network (SAN)
  - SAN is not magic - it just allows servers access the same logical unit of disk (LUN). Shared filesystem support, serialization etc is provided by software on server.
- Cluster interconnect may be any of :
  - LAN
  - Serial connection
  - Specialized switching H/W (fibre)
  - Shared disk

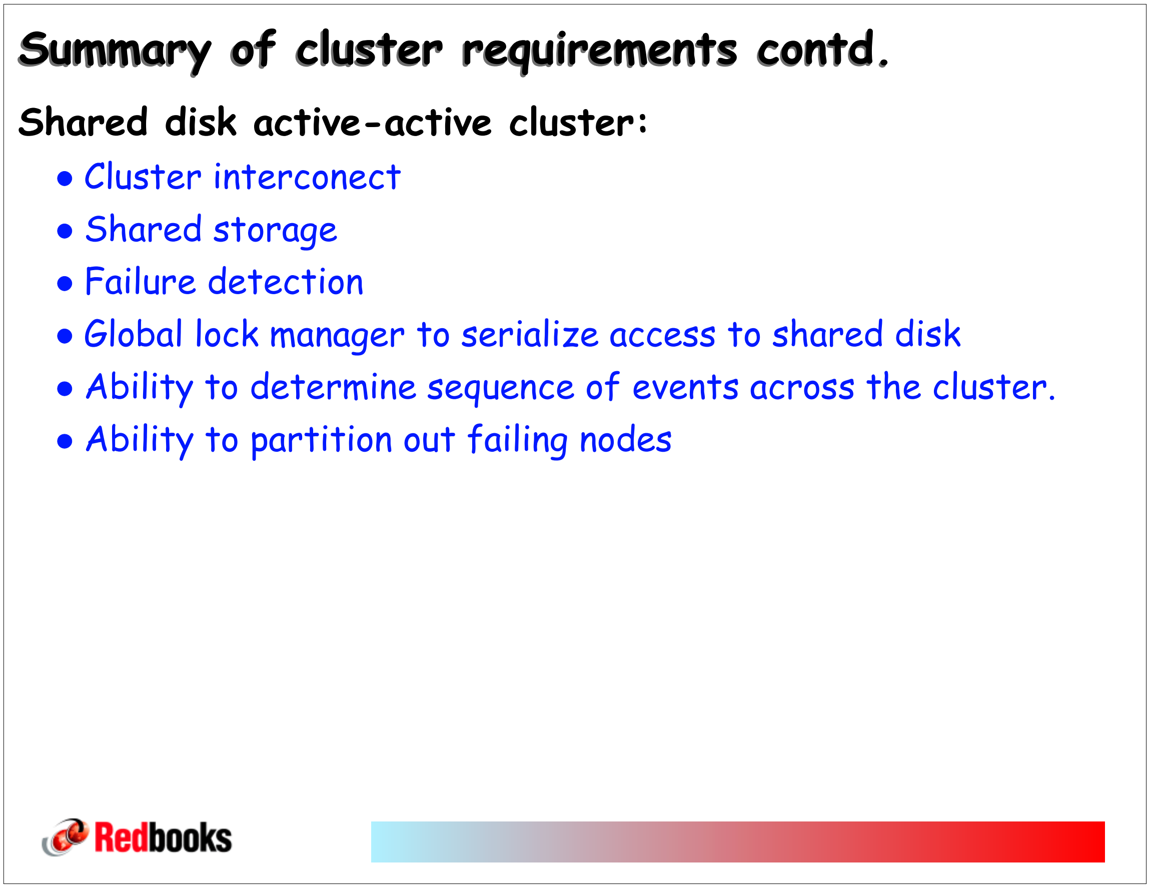
Redbooks



Summary of cluster requirements

Shared nothing active-standby cluster:

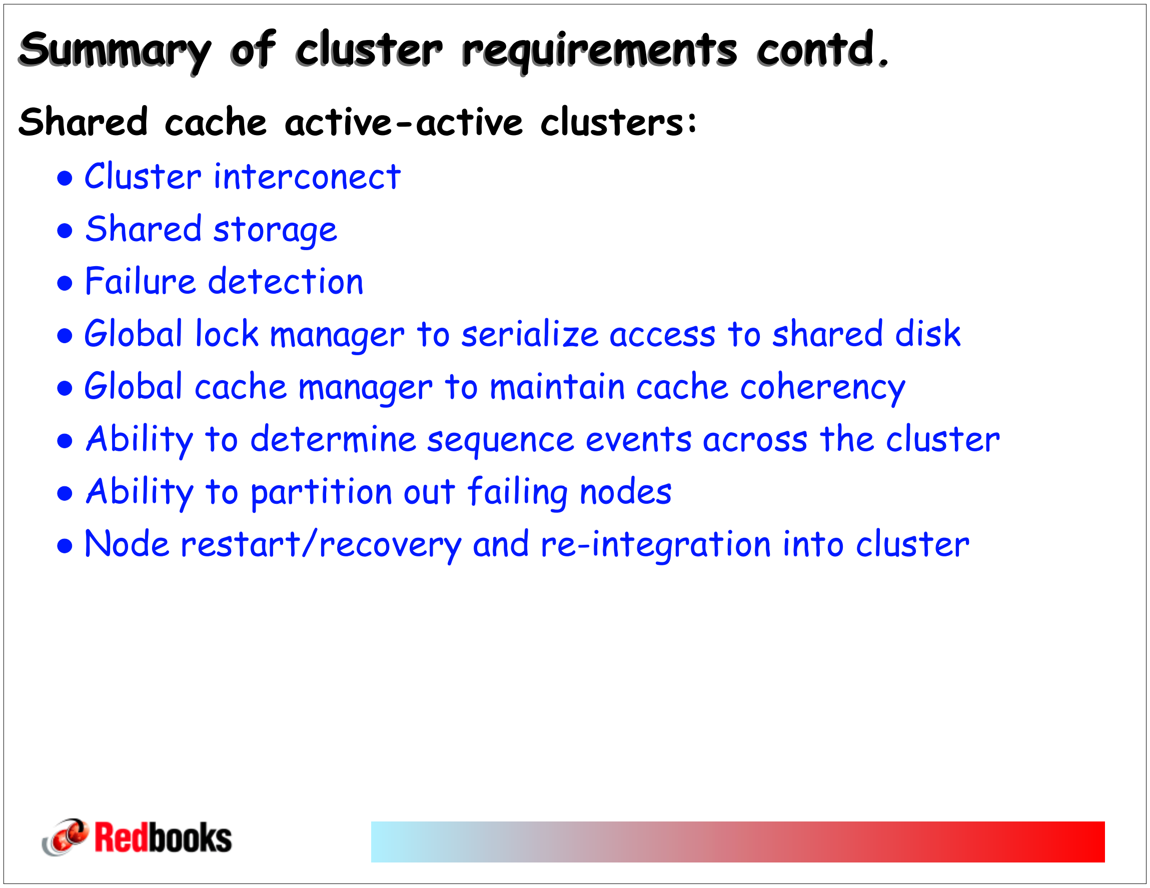
- Cluster interconect
- Shared storage
- Failure detection
- Automate move of resources (network,storage) from active node to standby node



Summary of cluster requirements contd.

Shared disk active-active cluster:

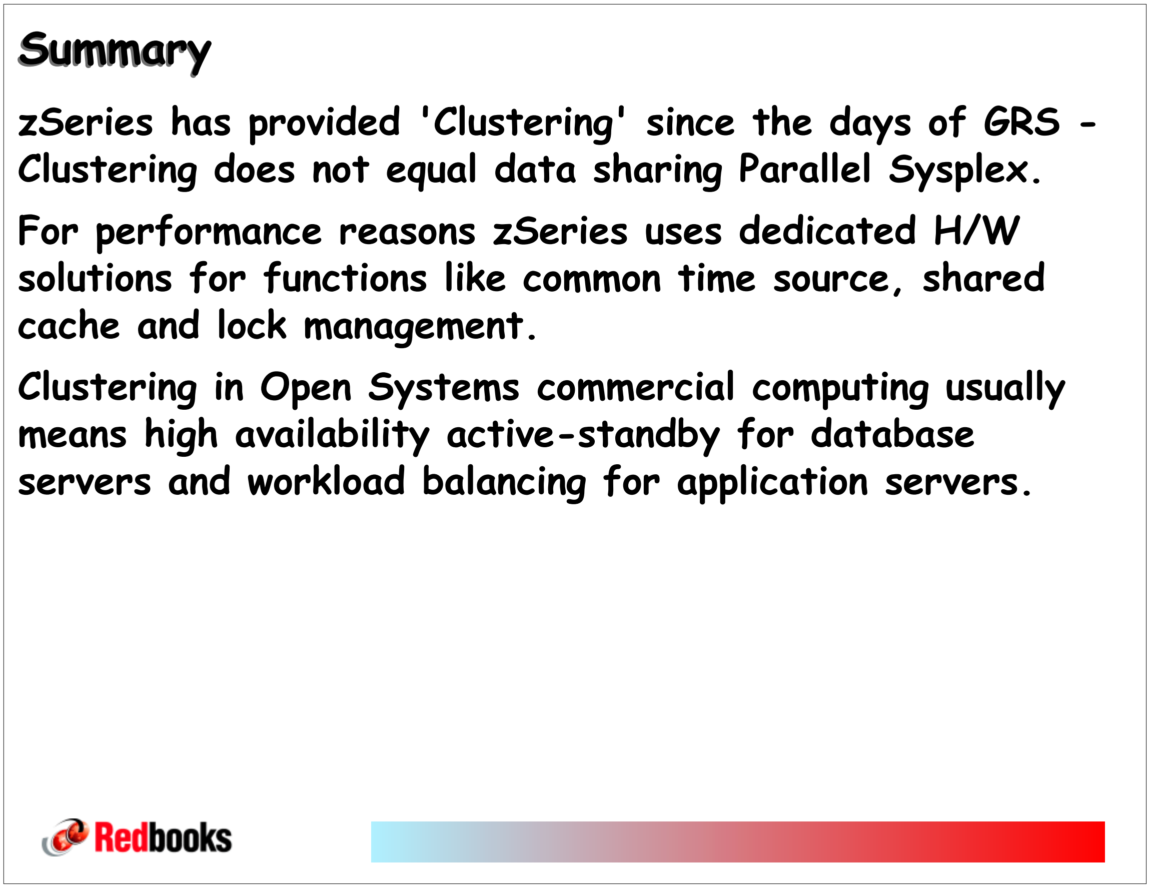
- Cluster interconect
- Shared storage
- Failure detection
- Global lock manager to serialize access to shared disk
- Ability to determine sequence of events across the cluster.
- Ability to partition out failing nodes



Summary of cluster requirements contd.

Shared cache active-active clusters:

- Cluster interconect
- Shared storage
- Failure detection
- Global lock manager to serialize access to shared disk
- Global cache manager to maintain cache coherency
- Ability to determine sequence events across the cluster
- Ability to partition out failing nodes
- Node restart/recovery and re-integration into cluster

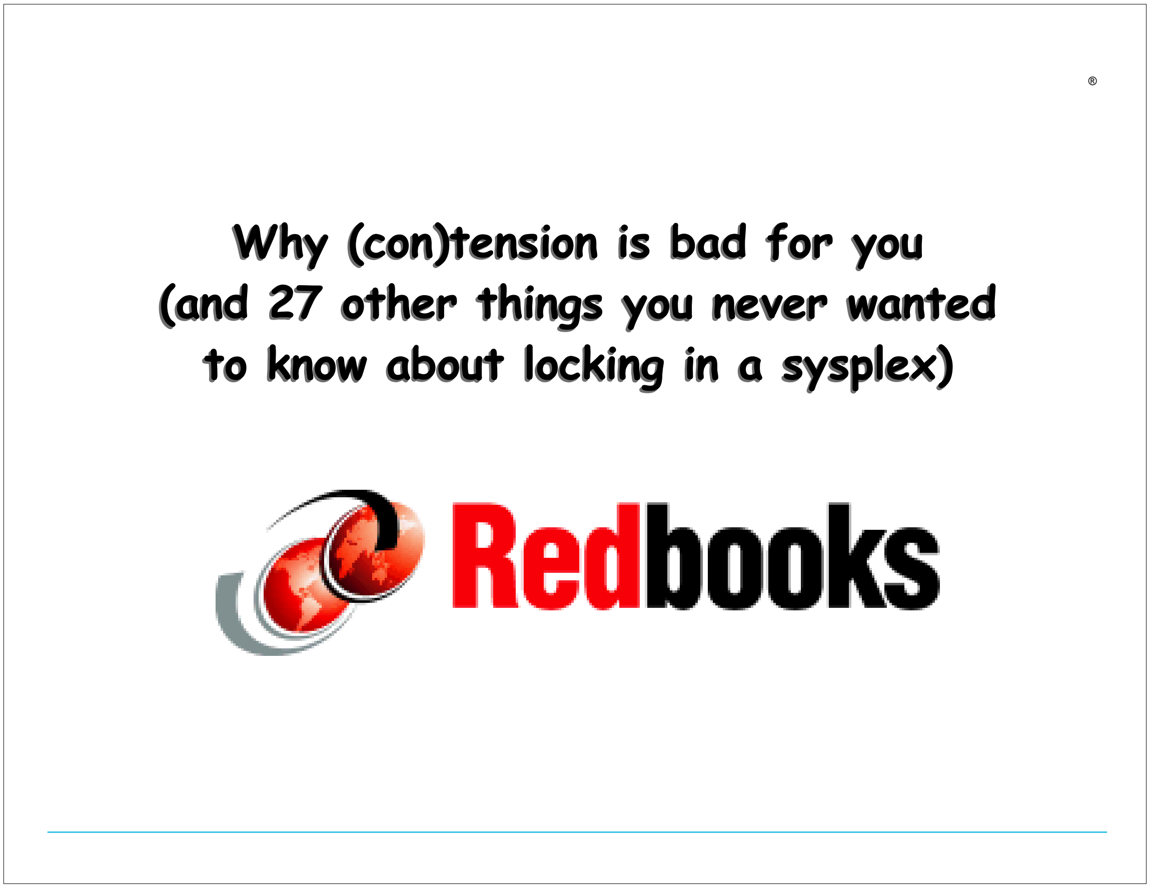


Summary

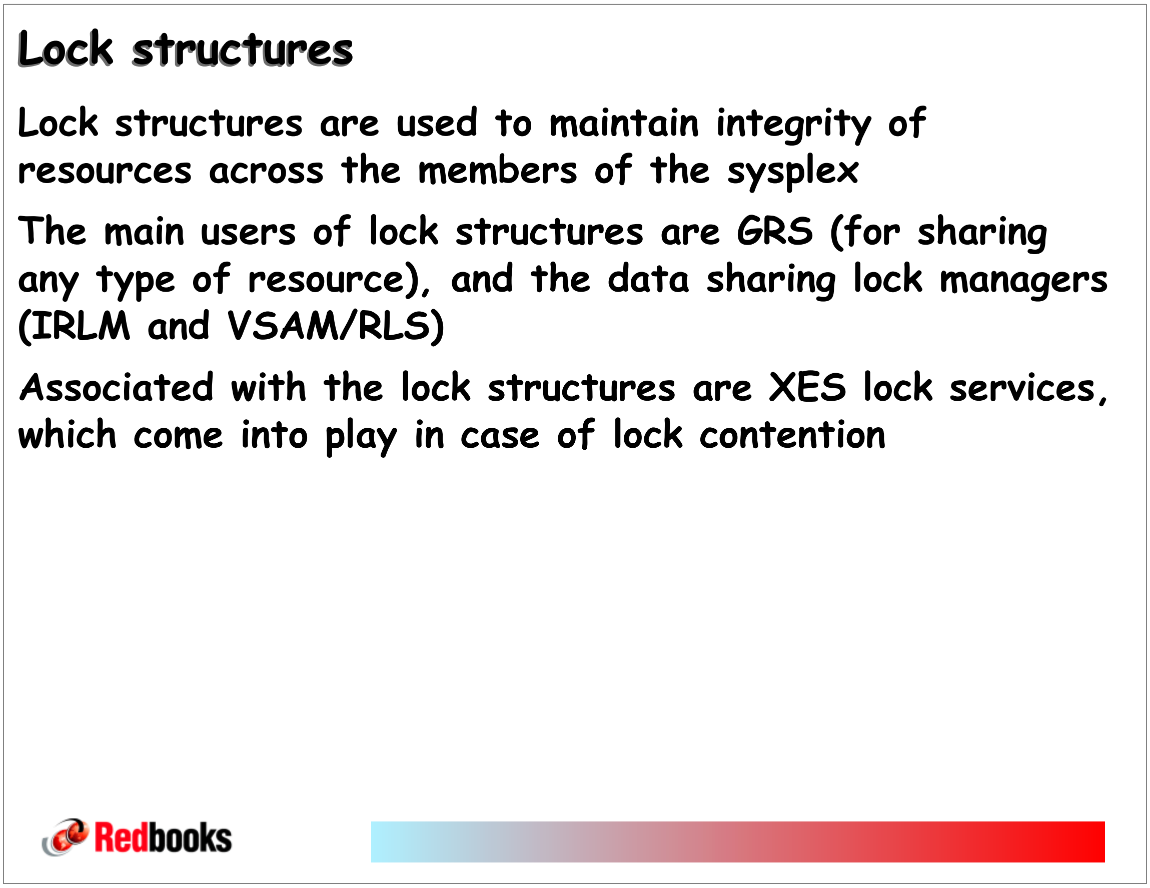
zSeries has provided 'Clustering' since the days of GRS - Clustering does not equal data sharing Parallel Sysplex.

For performance reasons zSeries uses dedicated H/W solutions for functions like common time source, shared cache and lock management.

Clustering in Open Systems commercial computing usually means high availability active-standby for database servers and workload balancing for application servers.



Why (con)tension is bad for you (and 27 other things you never wanted to know about locking in a sysplex)

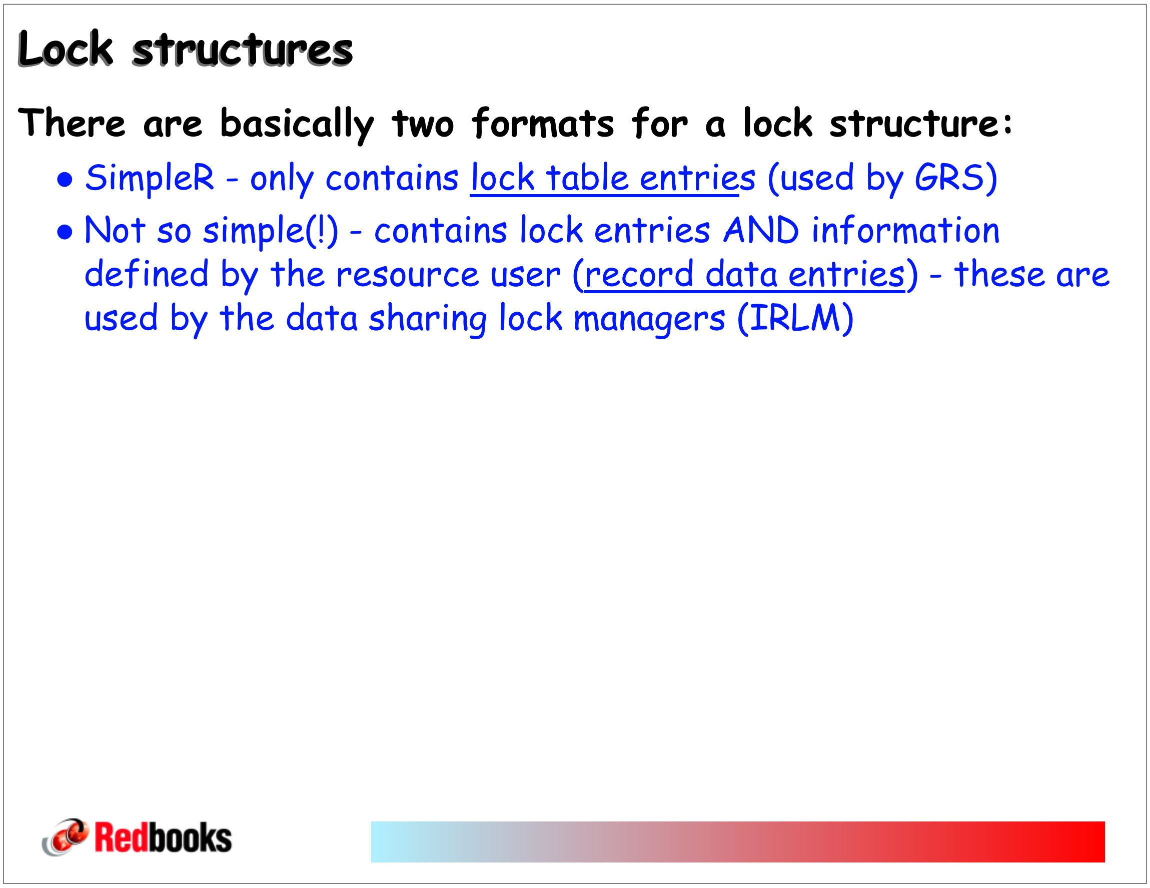


Lock structures

Lock structures are used to maintain integrity of resources across the members of the sysplex

The main users of lock structures are GRS (for sharing any type of resource), and the data sharing lock managers (IRLM and VSAM/RLS)

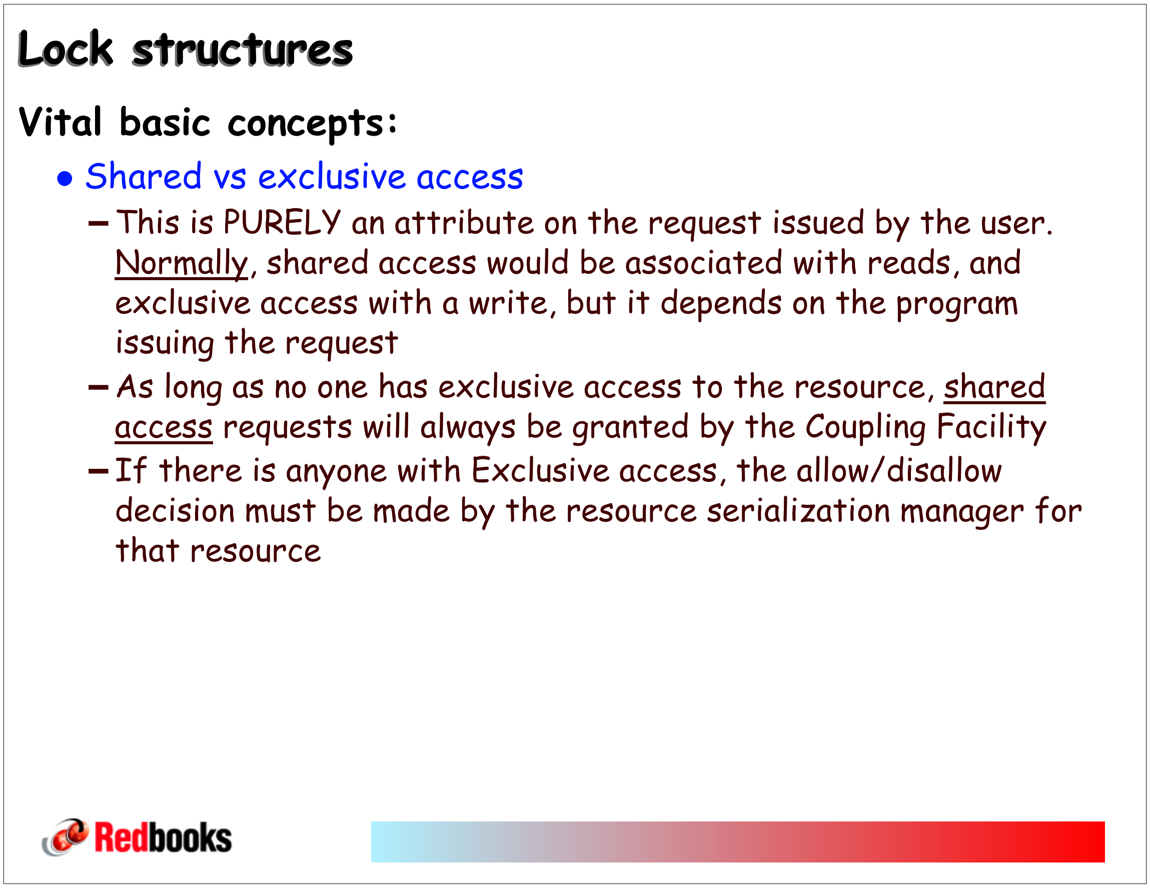
Associated with the lock structures are XES lock services, which come into play in case of lock contention



Lock structures

There are basically two formats for a lock structure:

- SimpleR - only contains lock table entries (used by GRS)
- Not so simple(!) - contains lock entries AND information defined by the resource user (record data entries) - these are used by the data sharing lock managers (IRLM)



Lock structures

Vital basic concepts:

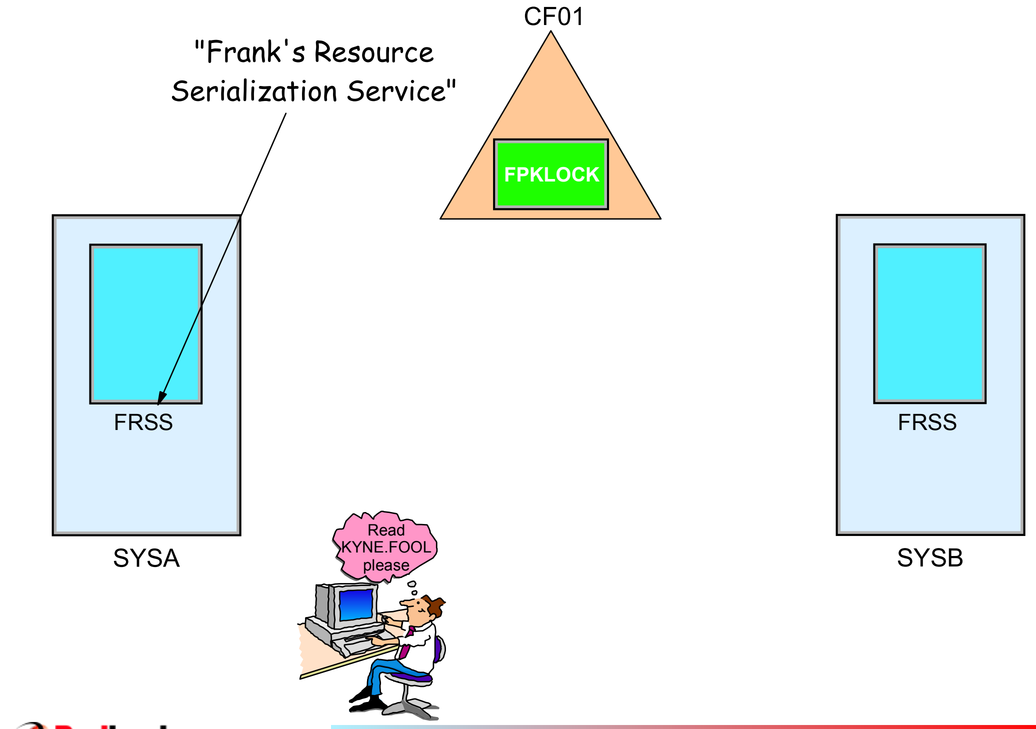
- Shared vs exclusive access
  - This is PURELY an attribute on the request issued by the user. Normally, shared access would be associated with reads, and exclusive access with a write, but it depends on the program issuing the request
  - As long as no one has exclusive access to the resource, shared access requests will always be granted by the Coupling Facility
  - If there is anyone with Exclusive access, the allow/disallow decision must be made by the resource serialization manager for that resource

Lock structures

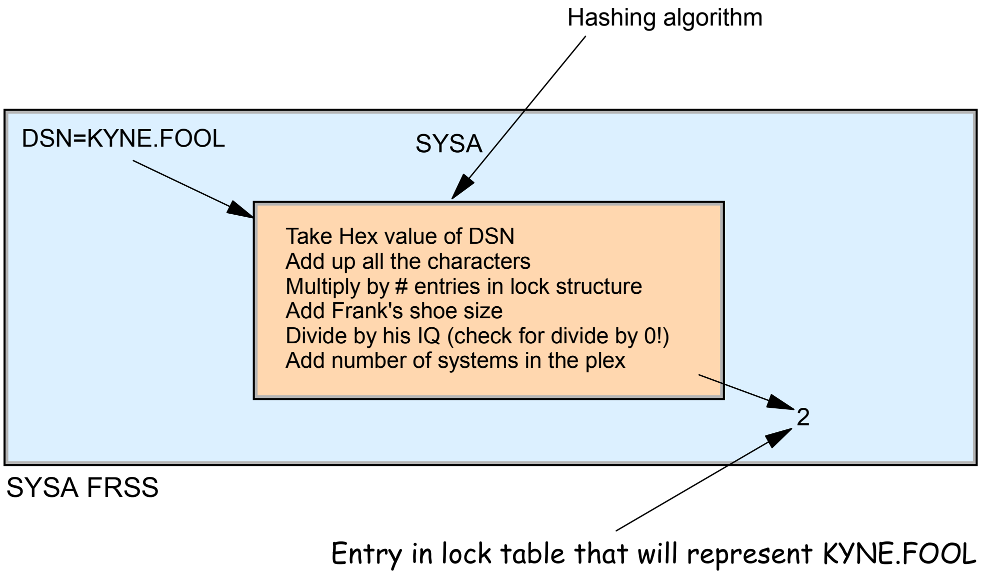
Vital basic concepts:

- Global lock manager
  - Every lock entry has two parts - an Exclusive access part, and a Shared access part
  - When the first requestor gets Exclusive access, XES on that system becomes the Global lock manager for that lock *entry* (NOT for the *resource* - for the lock *entry*)
  - From that point until there is no one with Exclusive access to that *lock entry*, all access requests for that any resource that hashes to that lock entry are passed to the Global lock manager to decide if access should be granted. The lock entry in the lock structure is NOT updated during this time - the Global lock manager keeps track of all lock requests and is responsible for updating the lock structure before the last exclusive access is released

Lock structures



Lock structures



Hashing algorithm

DSN=KYNE.FOOL

SYSA

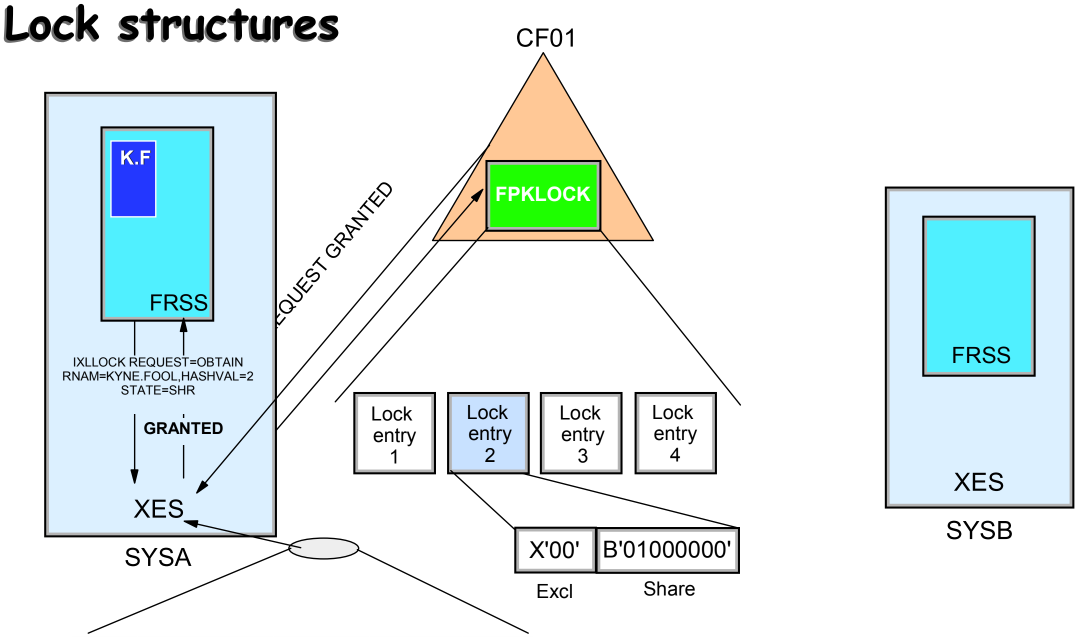
Take Hex value of DSN
Add up all the characters
Multiply by # entries in lock structure
Add Frank's shoe size
Divide by his IQ (check for divide by 0!)
Add number of systems in the plex

2

SYSA FRSS

Entry in lock table that will represent KYNE.FOOL

Lock structures



CF01

FPKLOCK

K.F

FRSS

IXLLOCK REQUEST=OBTAIN
RNAM=KYNE.FOOL,HASHVAL=2
STATE=SHR

GRANTED

XES

SYSA

EQUEST GRANTED

Lock entry 1

Lock entry 2

Lock entry 3

Lock entry 4

X'00'

B'01000000'

Excl

Share

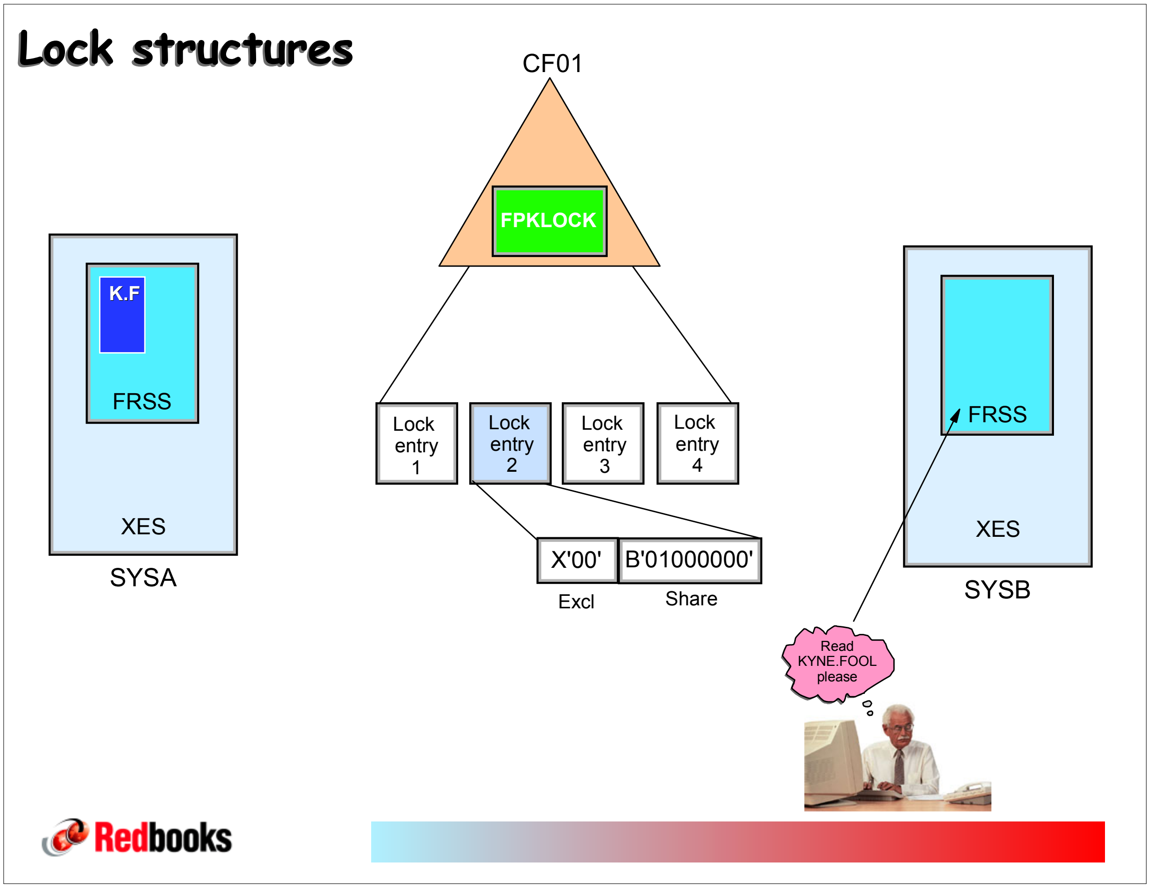
FRSS

XES

SYSB

Lock entry	Res Name	Access	Status	Conn ID
2	KYNE.FOOL	SHR	Granted	01

Resource Request Queue



Lock structures

CF01

FPKLOCK

Lock entry 1

Lock entry 2

Lock entry 3

Lock entry 4

X'00'

B'01000000'

Excl

Share

K.F

FRSS

XES

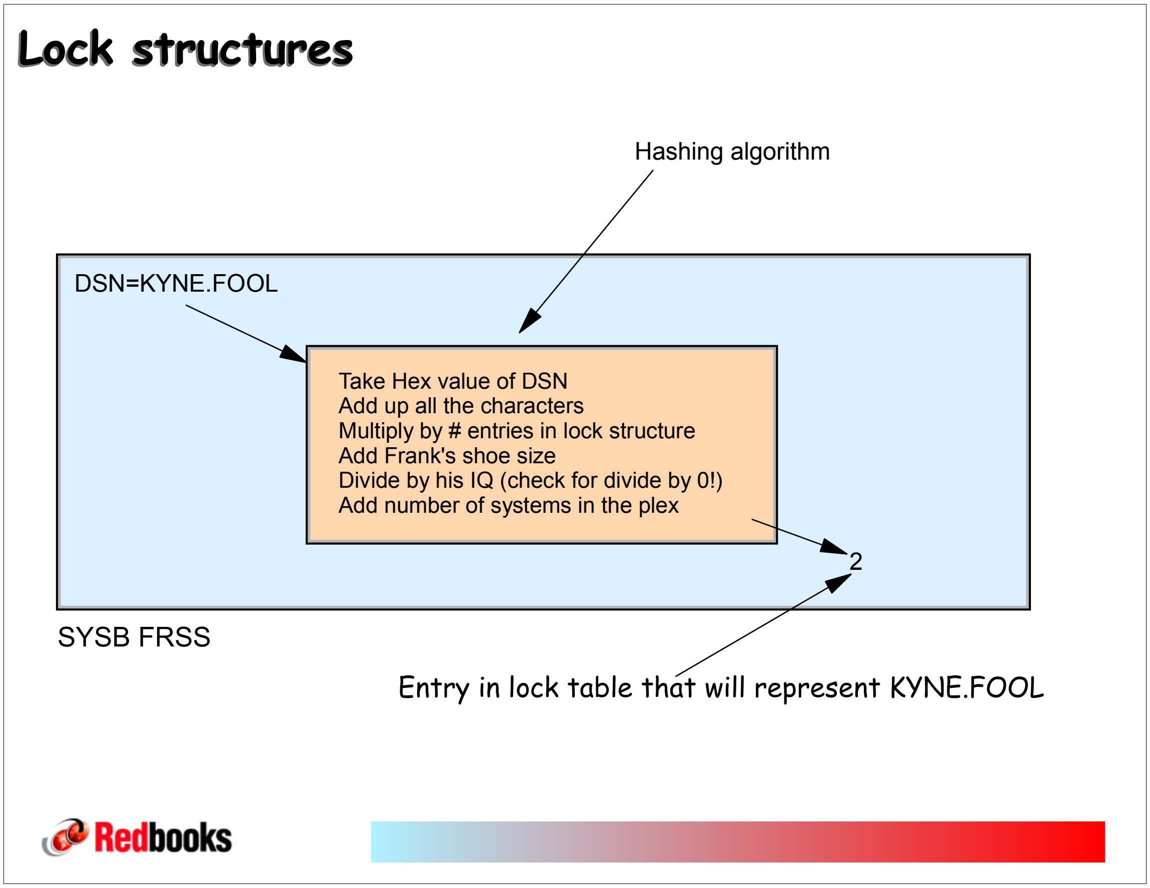
SYSA

FRSS

XES

SYSB

Read KYNE.FOOL please



Lock structures

Hashing algorithm

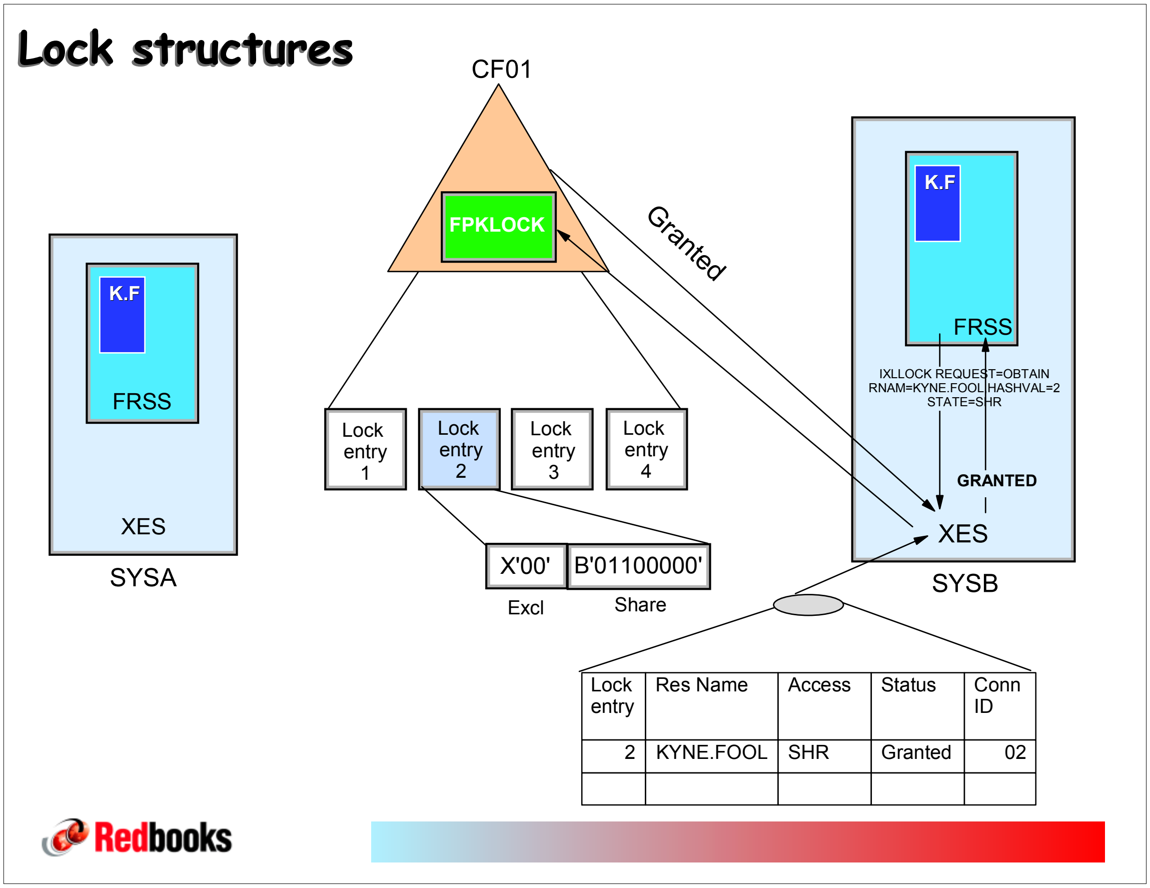
DSN=KYNE.FOOL

Take Hex value of DSN
Add up all the characters
Multiply by # entries in lock structure
Add Frank's shoe size
Divide by his IQ (check for divide by 0!)
Add number of systems in the plex

2

SYSB FRSS

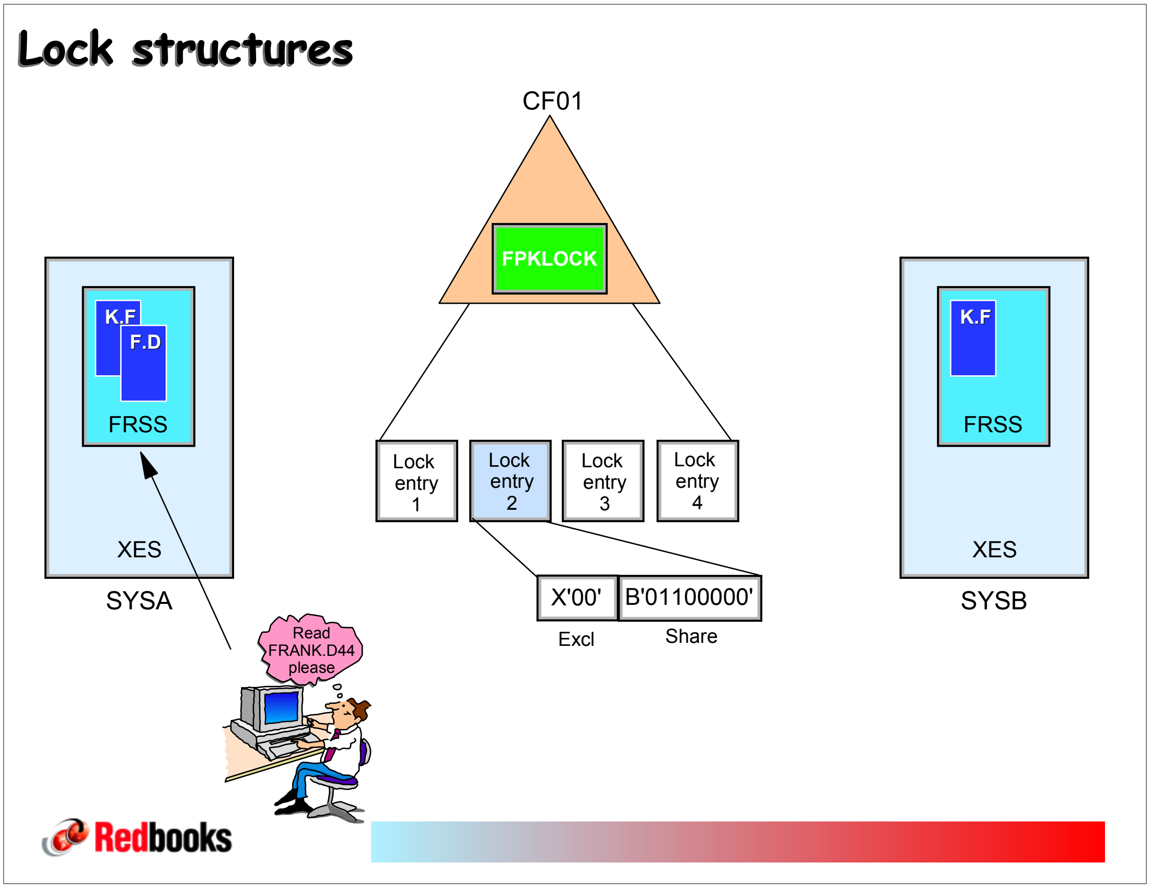
Entry in lock table that will represent KYNE.FOOL



Lock structures

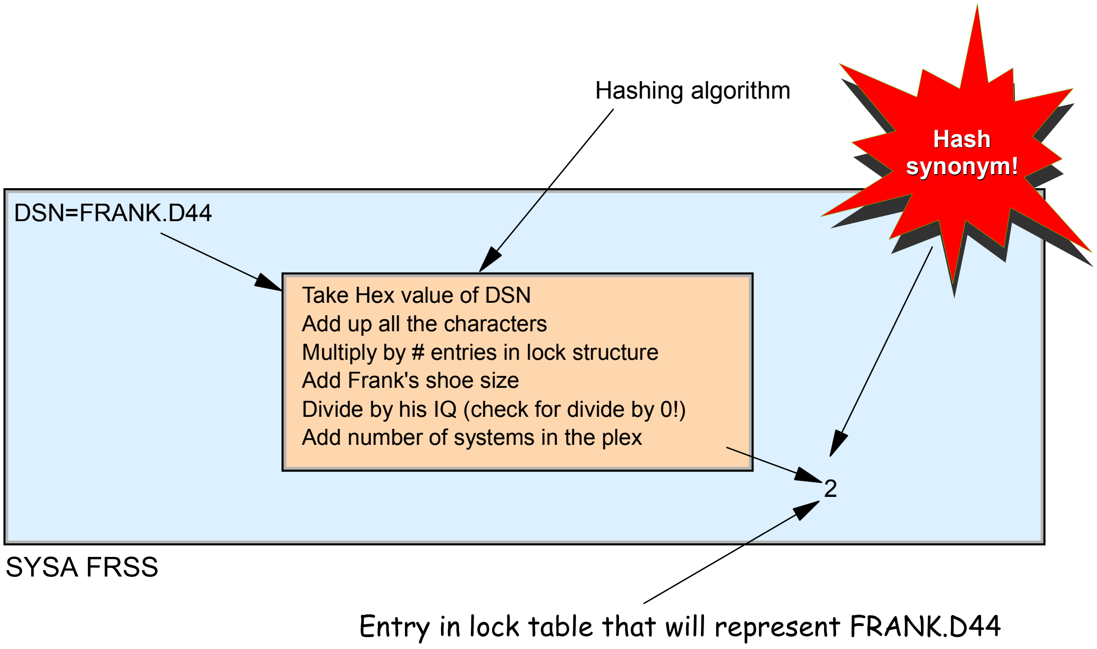
IXLLOCK REQUEST=OBTAIN
RNAM=KYNE.FOOL HASHVAL=2
STATE=SHR

Lock entry	Res Name	Access	Status	Conn ID
2	KYNE.FOOL	SHR	Granted	02



Lock structures

Lock structures



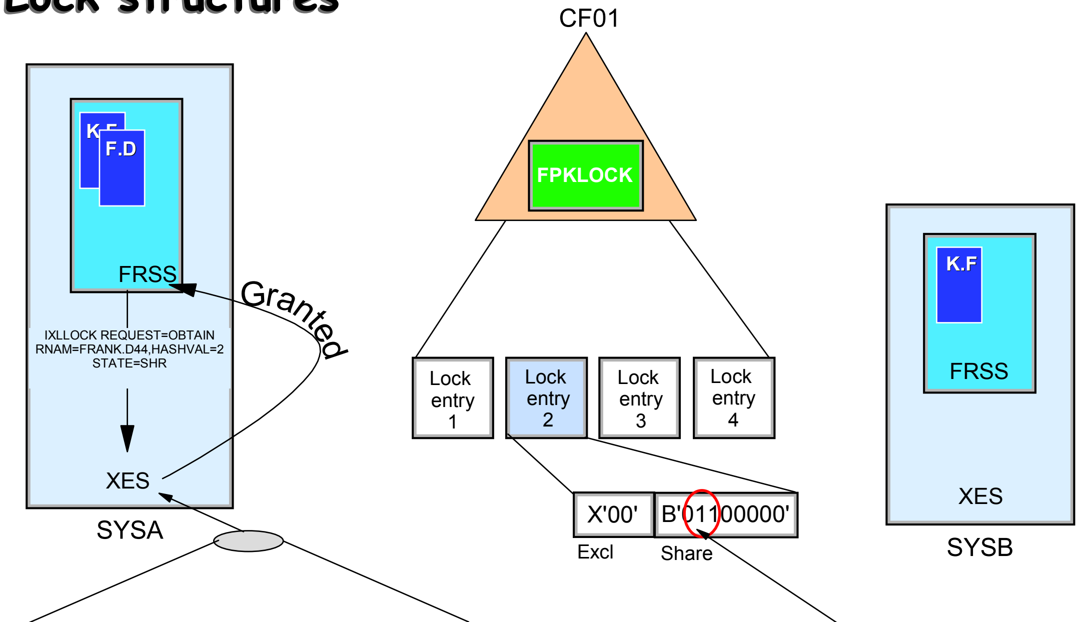
Hashing algorithm

DSN=FRANK.D44

Take Hex value of DSN
Add up all the characters
Multiply by # entries in lock structure
Add Frank's shoe size
Divide by his IQ (check for divide by 0!)
Add number of systems in the plex

Entry in lock table that will represent FRANK.D44

Lock structures

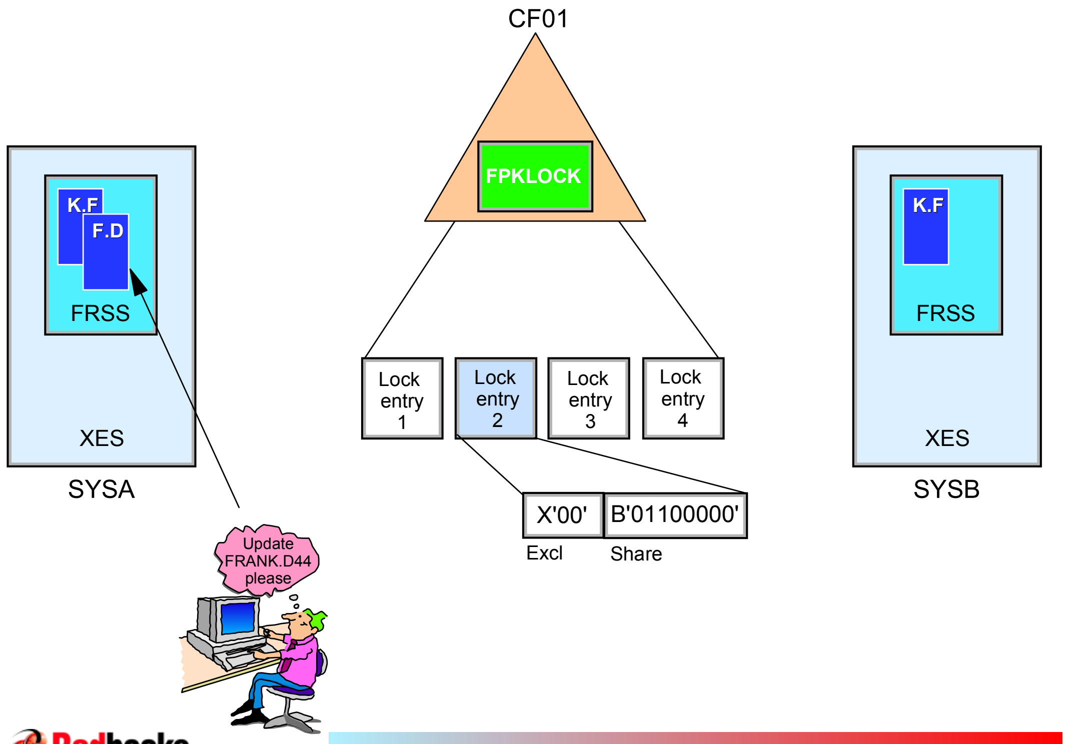


IXLLOCK REQUEST=OBTAIN
RNAM=FRANK.D44,HASHVAL=2
STATE=SHR

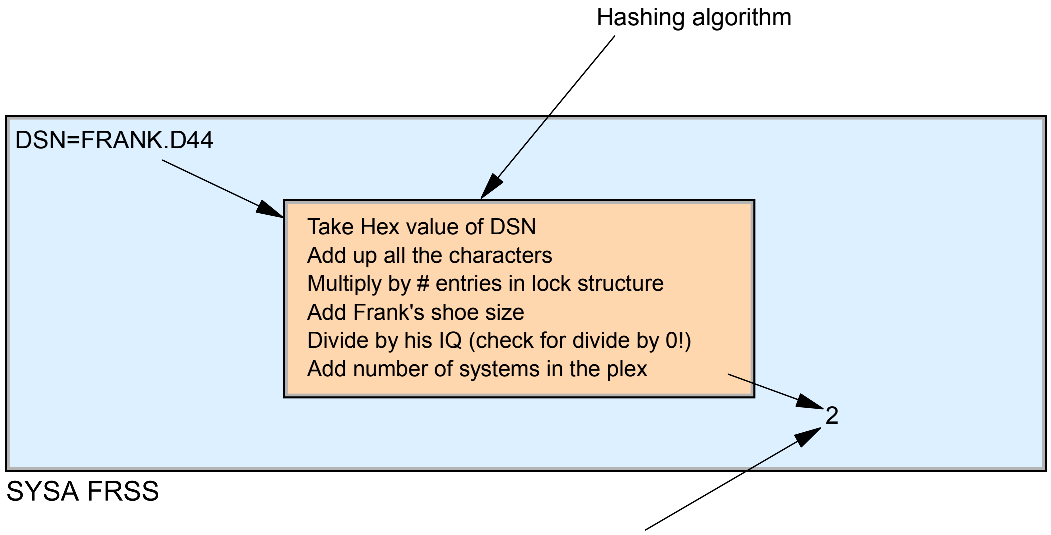
Lock entry	Res Name	Access	Status	Conn ID
2	KYNE.FOOL	SHR	Granted	01
2	FRANK.D44	SHR	Granted	01

Still the same..
No CF request reqd

Lock structures

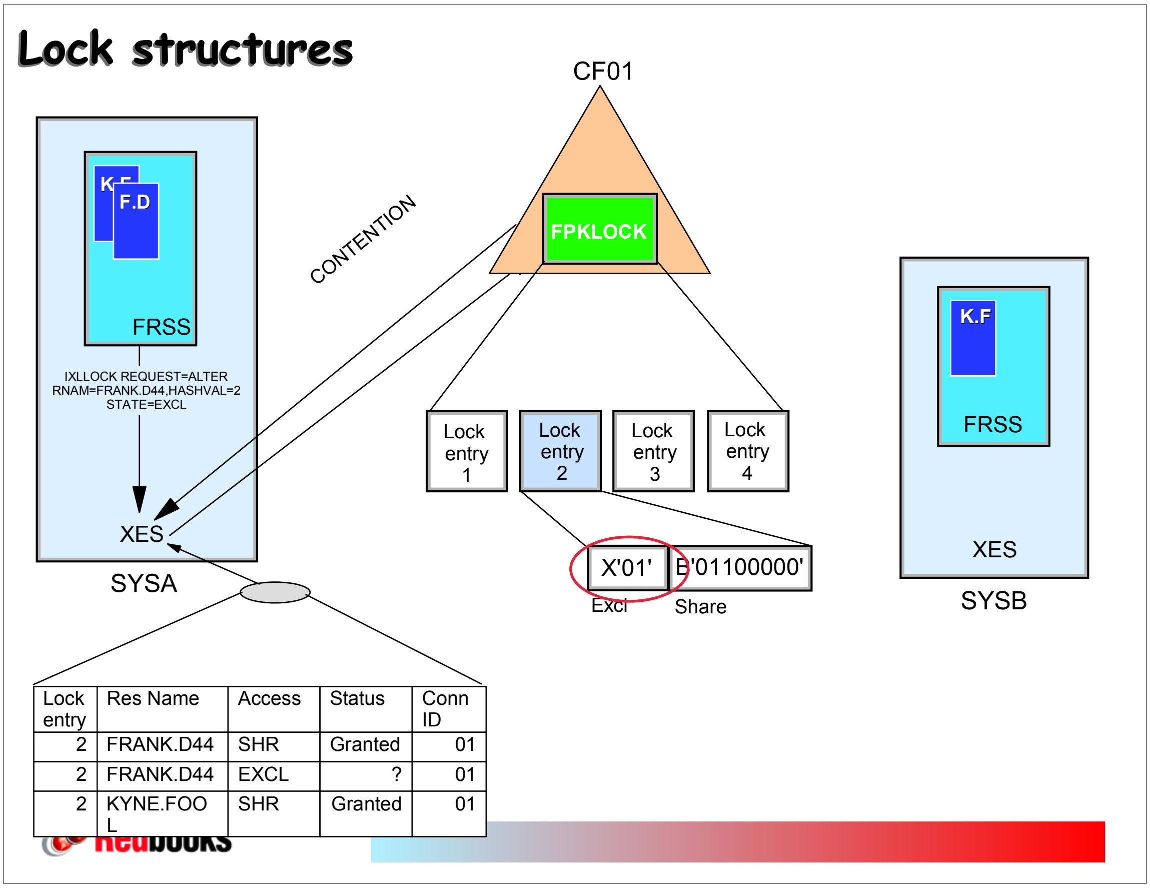


Lock structures



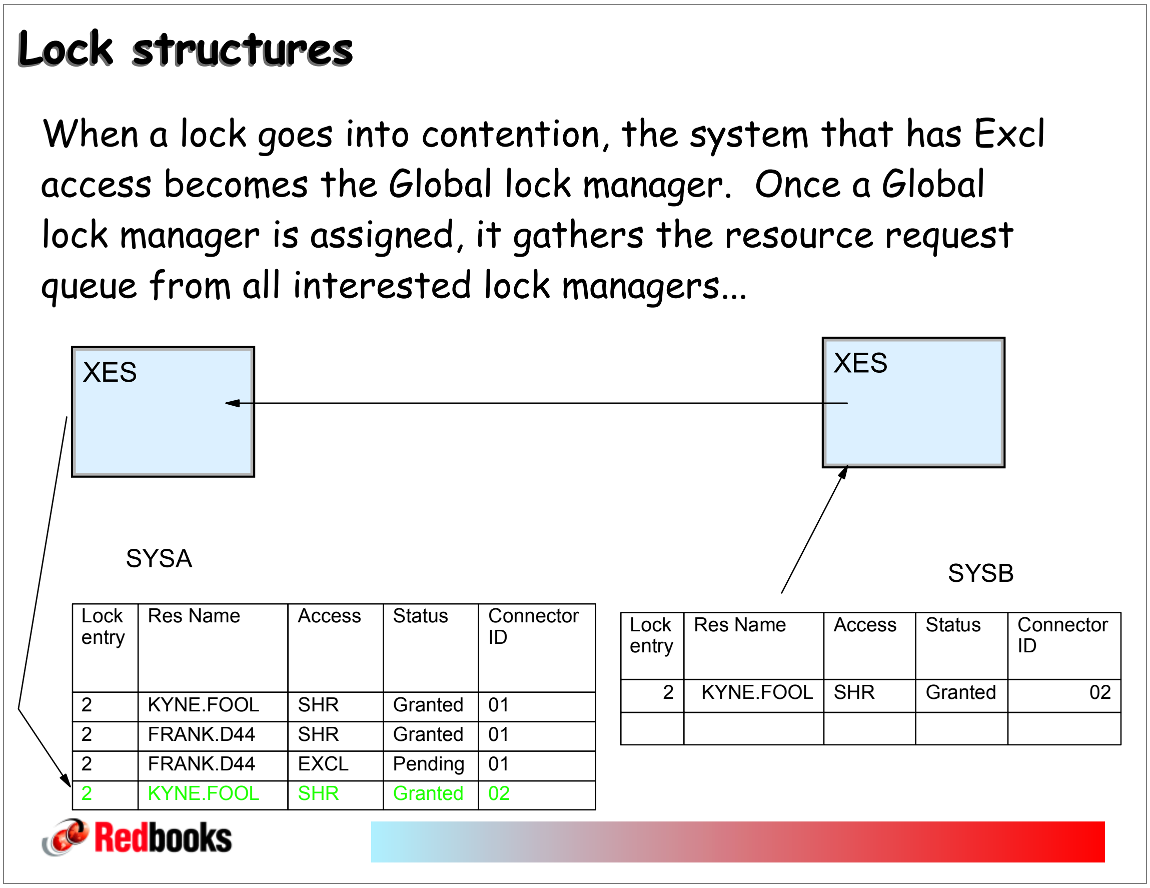
Take Hex value of DSN
Add up all the characters
Multiply by # entries in lock structure
Add Frank's shoe size
Divide by his IQ (check for divide by 0!)
Add number of systems in the plex

Entry in lock table that will represent FRANK.D44



Lock structures

Lock entry	Res Name	Access	Status	Conn ID
2	FRANK.D44	SHR	Granted	01
2	FRANK.D44	EXCL	?	01
2	KYNE.FOOL	SHR	Granted	01



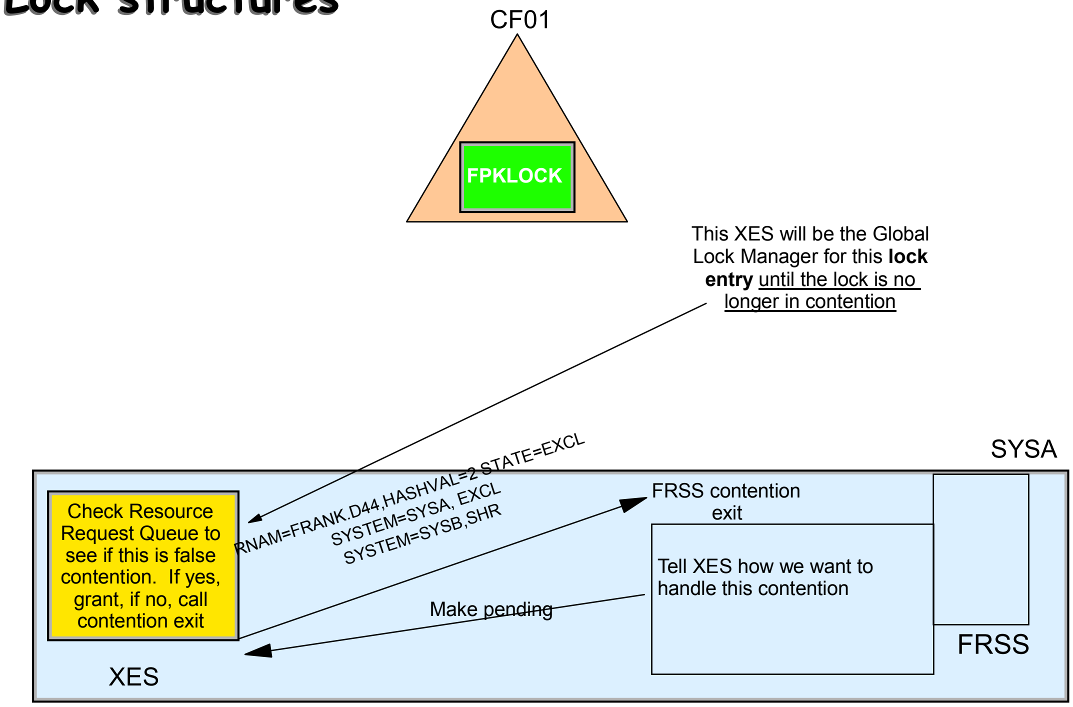
Lock structures

When a lock goes into contention, the system that has Excl access becomes the Global lock manager. Once a Global lock manager is assigned, it gathers the resource request queue from all interested lock managers...

Lock entry	Res Name	Access	Status	Connector ID
2	KYNE.FOOL	SHR	Granted	01
2	FRANK.D44	SHR	Granted	01
2	FRANK.D44	EXCL	Pending	01
2	KYNE.FOOL	SHR	Granted	02

Lock entry	Res Name	Access	Status	Connector ID
2	KYNE.FOOL	SHR	Granted	02

Lock structures

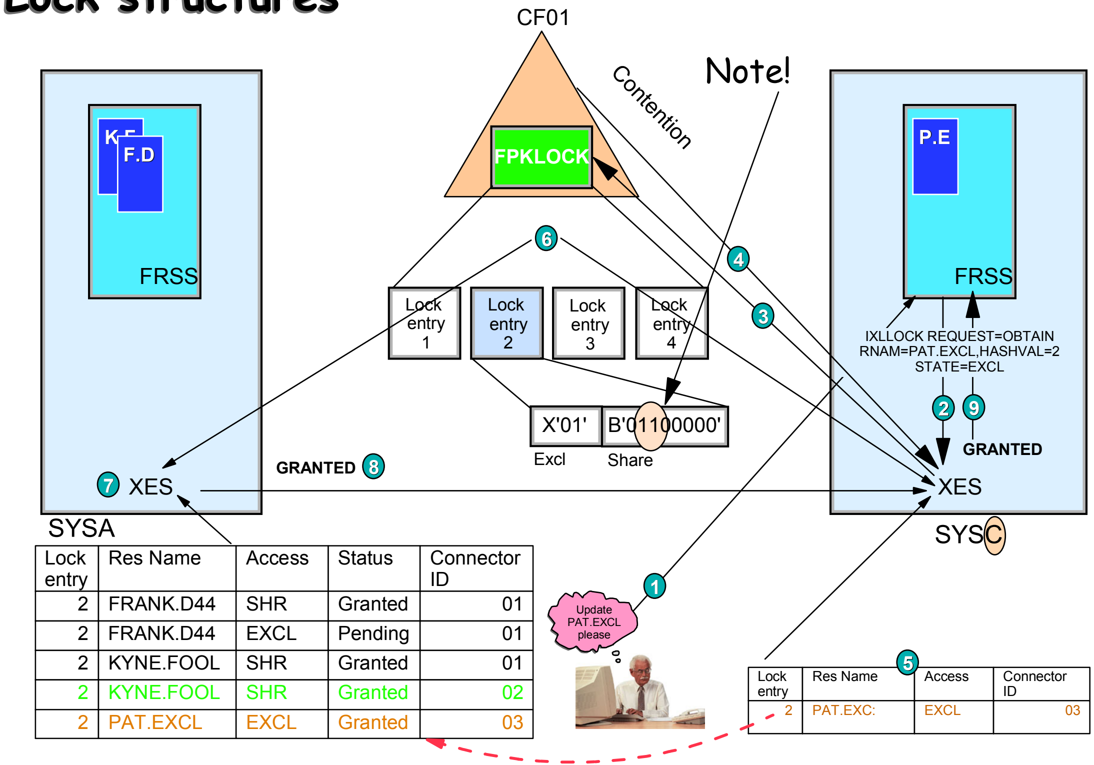


This XES will be the Global Lock Manager for this **lock entry** until the lock is no longer in contention

Check Resource Request Queue to see if this is false contention. If yes, grant, if no, call contention exit

Tell XES how we want to handle this contention

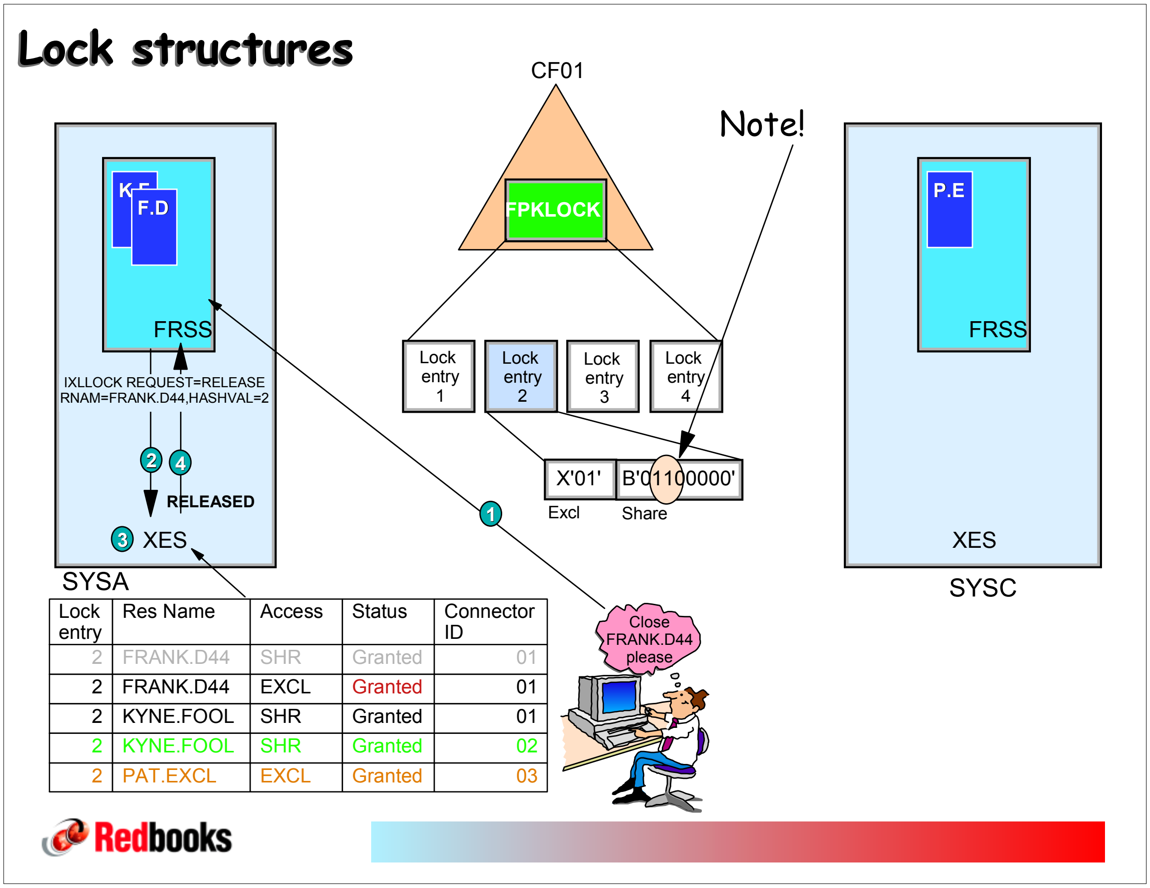
Lock structures



IXLLOCK REQUEST=OBTAIN
RNAM=PAT.EXCL,HASHVAL=2
STATE=EXCL

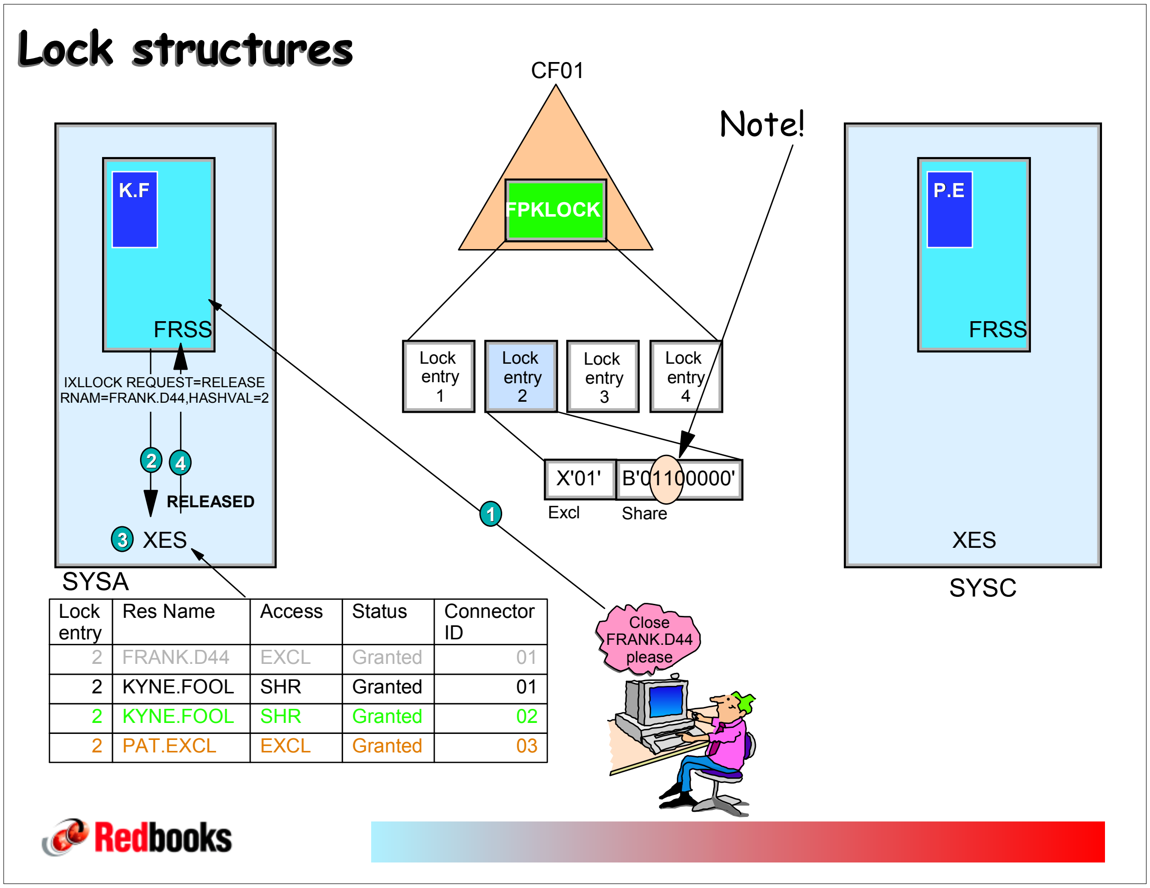
Lock entry	Res Name	Access	Status	Connector ID
2	FRANK.D44	SHR	Granted	01
2	FRANK.D44	EXCL	Pending	01
2	KYNE.FOOL	SHR	Granted	01
2	KYNE.FOOL	SHR	Granted	02
2	PAT.EXCL	EXCL	Granted	03

Lock entry	Res Name	Access	Connector ID
2	PAT.EXC:	EXCL	03



Lock structures

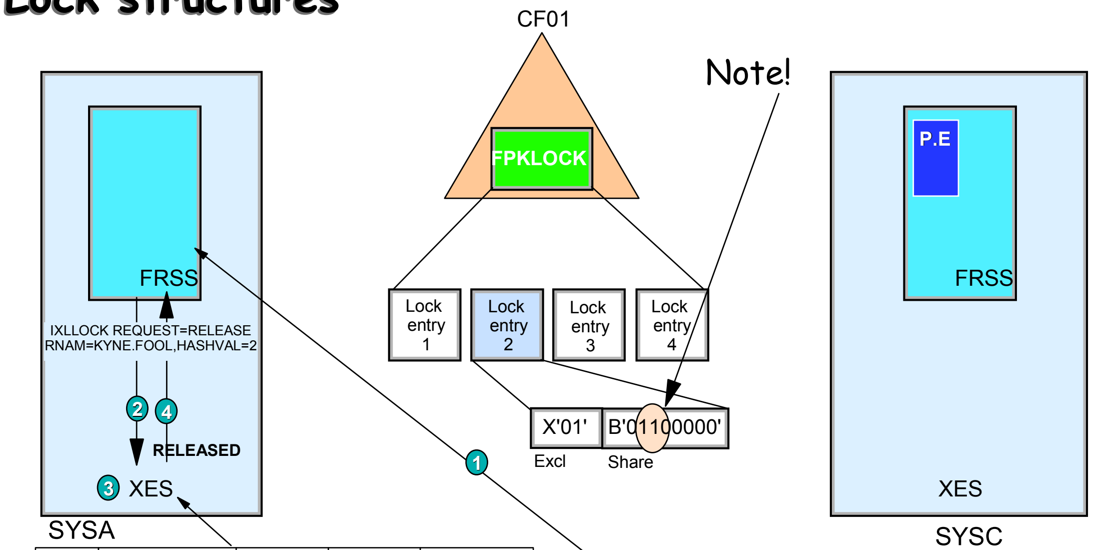
Lock entry	Res Name	Access	Status	Connector ID
2	FRANK.D44	SHR	Granted	01
2	FRANK.D44	EXCL	Granted	01
2	KYNE.FOOL	SHR	Granted	01
2	KYNE.FOOL	SHR	Granted	02
2	PAT.EXCL	EXCL	Granted	03



Lock structures

Lock entry	Res Name	Access	Status	Connector ID
2	FRANK.D44	EXCL	Granted	01
2	KYNE.FOOL	SHR	Granted	01
2	KYNE.FOOL	SHR	Granted	02
2	PAT.EXCL	EXCL	Granted	03

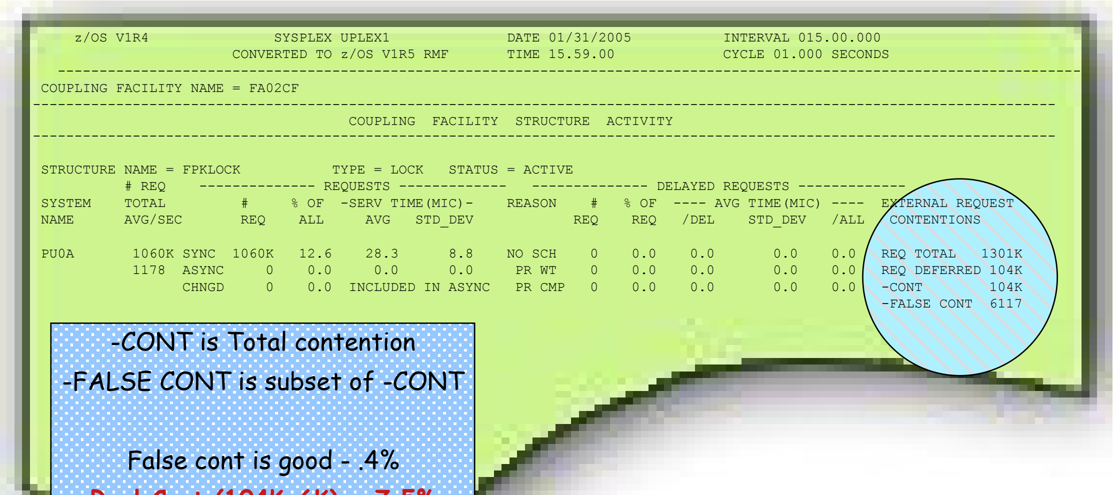
Lock structures



Lock entry	Res Name	Access	Status	Connector ID
2	KYNE.FOOL	SHR	Granted	01
2	KYNE.FOOL	SHR	Granted	02
2	PAT.EXCL	EXCL	Granted	03

Note that SYSA no longer has ANY interest in lock entry 2, but he is STILL the Global Lock Manager for this lock entry

RMF activity report for lock structures



```
    z/OS V1R4              SYSPLEX UPLEX1              DATE 01/31/2005          INTERVAL 015.00.000
                      CONVERTED TO z/OS V1R5 RMF       TIME 15.59.00            CYCLE 01.000 SECONDS
   ---------------------------------------------------------------------------------------------------------------
COUPLING FACILITY NAME = FA02CF
---------------------------------------------------------------------------------------------------------------
                                     COUPLING  FACILITY  STRUCTURE  ACTIVITY
---------------------------------------------------------------------------------------------------------------

STRUCTURE NAME = FPKLOCK             TYPE = LOCK   STATUS = ACTIVE
          # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
SYSTEM    TOTAL          #     % OF  -SERV TIME(MIC)-    REASON    #    % OF  ---- AVG TIME(MIC) ----   EXTERNAL REQUEST
NAME      AVG/SEC       REQ    ALL      AVG   STD_DEV              REQ   REQ   /DEL    STD_DEV   /ALL    CONTENTIONS

PU0A       1060K SYNC  1060K   12.6     28.3       8.8    NO SCH    0    0.0    0.0       0.0    0.0    REQ TOTAL   1301K
           1178  ASYNC     0    0.0      0.0       0.0     PR WT    0    0.0    0.0       0.0    0.0    REQ DEFERRED 104K
                 CHNGD     0    0.0  INCLUDED IN ASYNC     PR CMP   0    0.0    0.0       0.0    0.0    -CONT        104K
                                                                                                       -FALSE CONT  6117
```

-CONT is Total contention

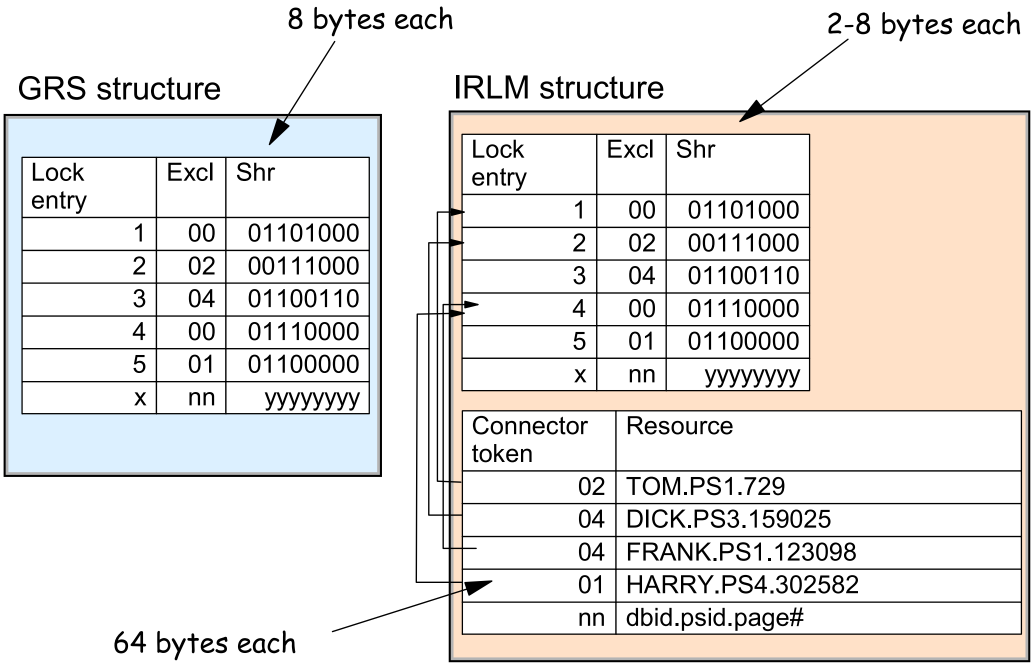
-FALSE CONT is subset of -CONT

False cont is good - .4%

Real Cont (104K-6K) - 7.5%

Any thoughts on why REQ TOTAL is > #REQ?

Lock structures with record data areas



GRS structure

Lock entry	Excl	Shr
1	00	01101000
2	02	00111000
3	04	01100110
4	00	01110000
5	01	01100000
x	nn	yyyyyyyy

IRLM structure

Lock entry	Excl	Shr
1	00	01101000
2	02	00111000
3	04	01100110
4	00	01110000
5	01	01100000
x	nn	yyyyyyyy

Connector token	Resource
02	TOM.PS1.729
04	DICK.PS3.159025
04	FRANK.PS1.123098
01	HARRY.PS4.302582
nn	dbid.psid.page#

Record data entries

Information in the record data entries (IRLM calls them Record List Entries) is kept purely for recovery.

When an IRLM is stopping, it asks the CF for a list of all RLEs with his connection token. Each of these is then changed to a Retained Lock Entry.

If an IRLM disappears, the other members of the data sharing group do the same processing on his behalf - this ensures that no one else can update any items that were locked by the failing IRLM at the time he died.

- This is one of the reasons failure isolation is required for the lock structure

®

Bits and pieces

®

System z9

System z9 considerations

System z9 is first processor not to support compatability mode CF links:

- No ICB2 support on z9. No ISC3 in compatability mode support
- This means that any sysplex containing a z9 CANNOT have any pre-zSeries (9672 or 9674) CFs
- If the CF is in the z9, the 9672s cannot connect to it
- You CAN have 9672s and z9s in the same sysplex, as long as neither are the CF.
- There are no coesixtence PTFs above those already provided for z990.

System z9 enhancements

System z9 is first processor to provide the ability to have CBU for ICFs, IFLs, and zAAPs.

- You must decide in advance whether a given PU will act as CBU for ICF, IFL, zAAP, or CP
  - Each PU type has a different CBU feature code (#7822 for ICF CBU)
  - No support in GDPS yet

On/Off Capacity on Demand (OOCoD) also supports ICF, IFL, zAAP, and CP. Both OOCoD and CBU can be installed, but only one can be active at a given time

System z9 enhancements

Previously you could only define one weight for an LPAR - this dictated the share of both CPs and zAAPs that the LPAR would be granted

System z9 adds the ability to specify separate weights for CPs and zAAPs, so the share of the zAAPs that is given to an LPAR might be different to the share of CPs it is entitled to

You can also specify that you wish to use hard capping with the zAAPs

- This could be important from a sysplex aggregation perspective

System z9 enhancements

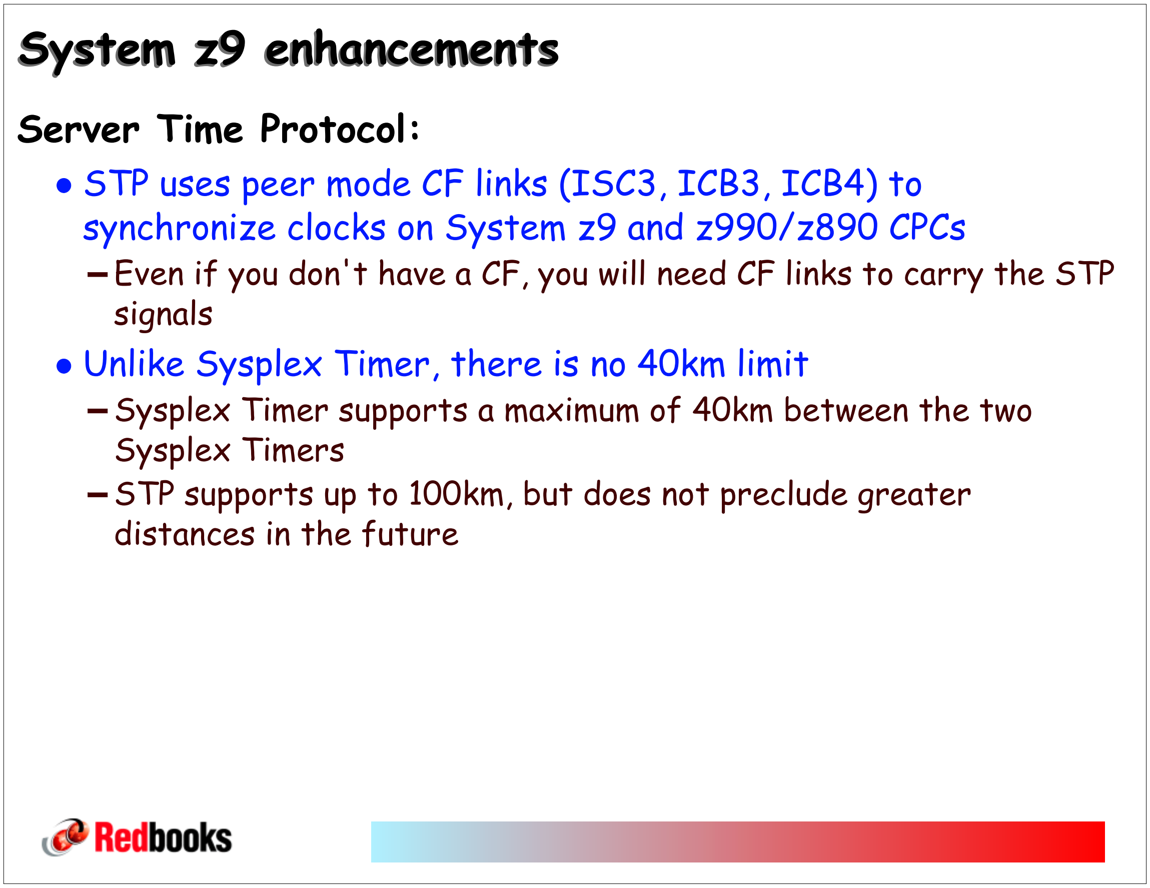
On zSeries CPCs prior to z9, there were two pools of PUs:

- CPs (used by z/OS, z/VSE, and z/VM)
- Everything else - IFLs, zAAPs, ICFs

As a result, it was possible that capacity on shared IFLs, ICFs, or zAAPs could be used for a workload other than you intended

On z9, there is a separate pool for each PU type

- Should make capacity and performance management simpler when using multiple (shared) PU types

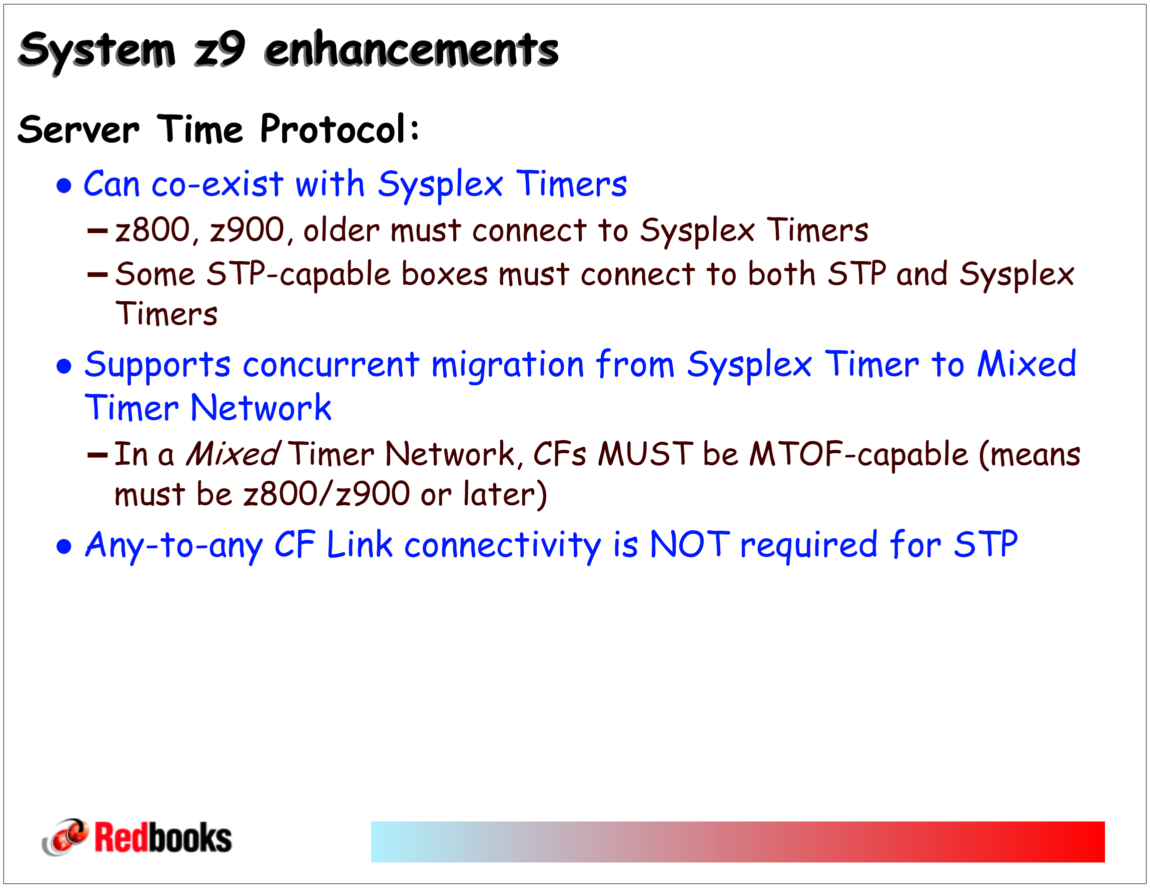


System z9 enhancements

Server Time Protocol:

- STP uses peer mode CF links (ISC3, ICB3, ICB4) to synchronize clocks on System z9 and z990/z890 CPCs
  - Even if you don't have a CF, you will need CF links to carry the STP signals
- Unlike Sysplex Timer, there is no 40km limit
  - Sysplex Timer supports a maximum of 40km between the two Sysplex Timers
  - STP supports up to 100km, but does not preclude greater distances in the future

Redbooks



System z9 enhancements

Server Time Protocol:

- Can co-exist with Sysplex Timers
  - z800, z900, older must connect to Sysplex Timers
  - Some STP-capable boxes must connect to both STP and Sysplex Timers
- Supports concurrent migration from Sysplex Timer to Mixed Timer Network
  - In a *Mixed* Timer Network, CFs MUST be MTOF-capable (means must be z800/z900 or later)
- Any-to-any CF Link connectivity is NOT required for STP

Redbooks

System z9 enhancements

Server Time Protocol:

- Software pre-reqs:
  - If z/OS is on STP CPC and is member of multi-CPC plex, must be running z/OS 1.7
  - If z/OS is on STP CPC and is a monoplex, it can be running any z/OS release
  - If z/OS is on a CPC that is in a Mixed Timer Network and is connected to 9037, it must be 1.4 or later plus toleration service
- Can use External Time Sources
- Can be used to adjust time zones and summer/winter time
- There is an Implementation Assistance Program available starting Oct 2005 for those that want early access

System z9 announcement

IBM statement of direction:

- IBM plans to withdraw 9037-002 (Sysplex Timer) from marketing in 2006
- "IBM intends to phase out Integrated Cluster Bus-3 links (ICB-3 links) over time. IBM plans to support ICB-3 links on z9-109 through the lifecycle of z9-109."

®

z/OS 1.7

z/OS 1.7 enhancements (for sysplex)

Maximum number of locks per lock structure connector increased from roughly 2.5 to 3.5 to about 16 times that amount.

No externals or toleration service required - support is automatically available on the 1.7 systems

z/OS 1.7 enhancements (for sysplex)

Number of *connected* DASDONLY log streams increased from 1024 to 16,384.

Prior to z/OS 1.7 there was one Logger task to manage all DASDONLY log streams

- With 1.7, there are up to 256 tasks

No migration or coexistence considerations

z/OS 1.7 enhancements (for sysplex)

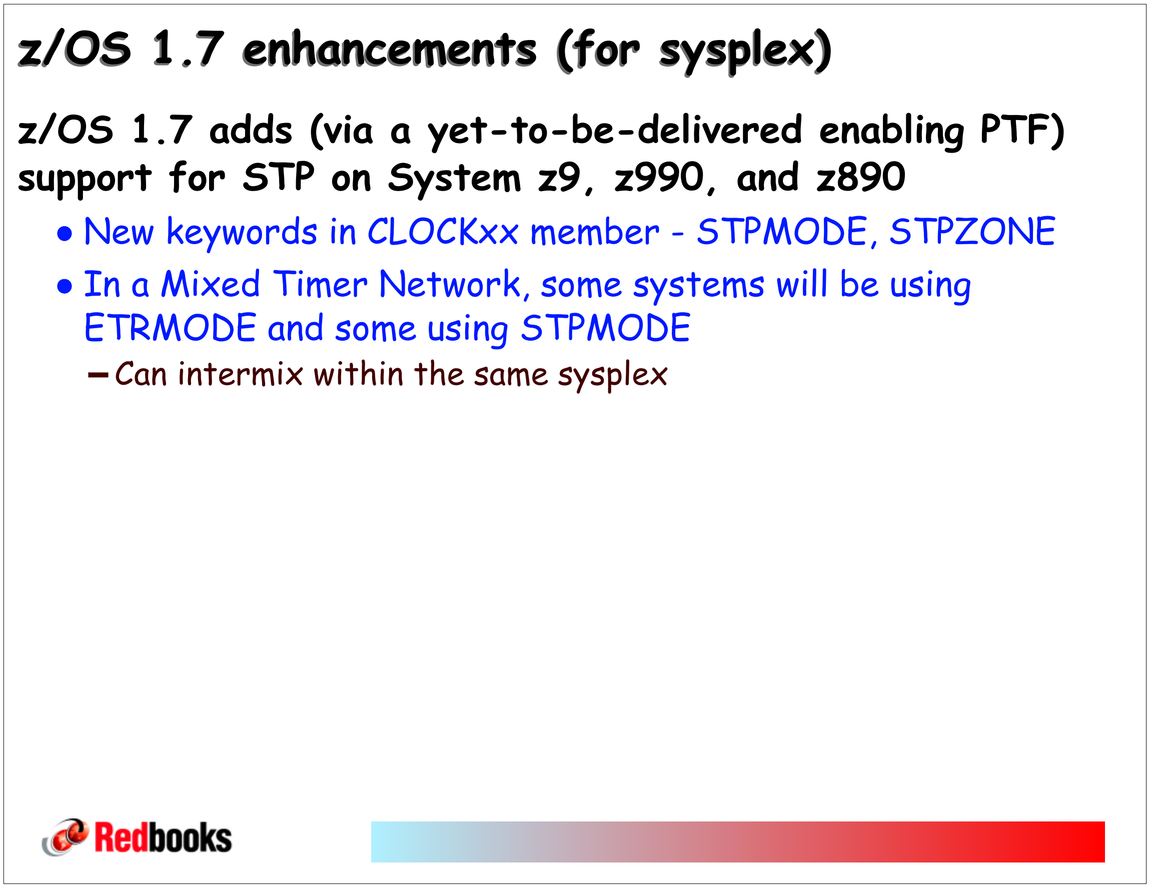
DEFAULT MAXMSG value increased from 750 to 2000 in z/OS 1.7

- This is the MAXIMUM amount of storage that can be fixed for XCF message buffers

Recommend removing MAXMSG values from 1.7 COUPLExx members unless you use larger values than this

Recommend increasing MAXMSG to 2000 on older systems

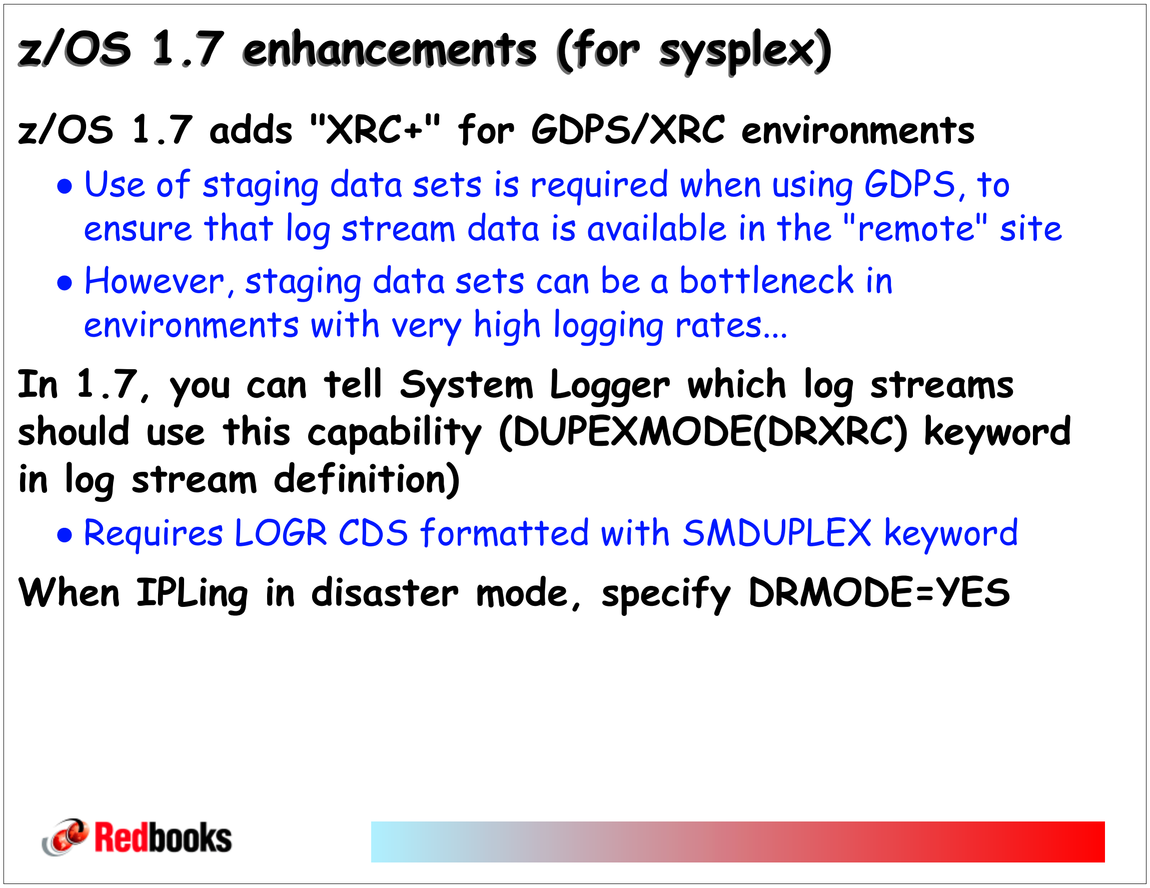
- If systems don't need this much, there is no cost
- If they do, you should see improved performance/fewer disruptions



z/OS 1.7 enhancements (for sysplex)

z/OS 1.7 adds (via a yet-to-be-delivered enabling PTF) support for STP on System z9, z990, and z890

- New keywords in CLOCKxx member - STPMODE, STPZONE
- In a Mixed Timer Network, some systems will be using ETRMODE and some using STPMODE
  - Can intermix within the same sysplex



z/OS 1.7 enhancements (for sysplex)

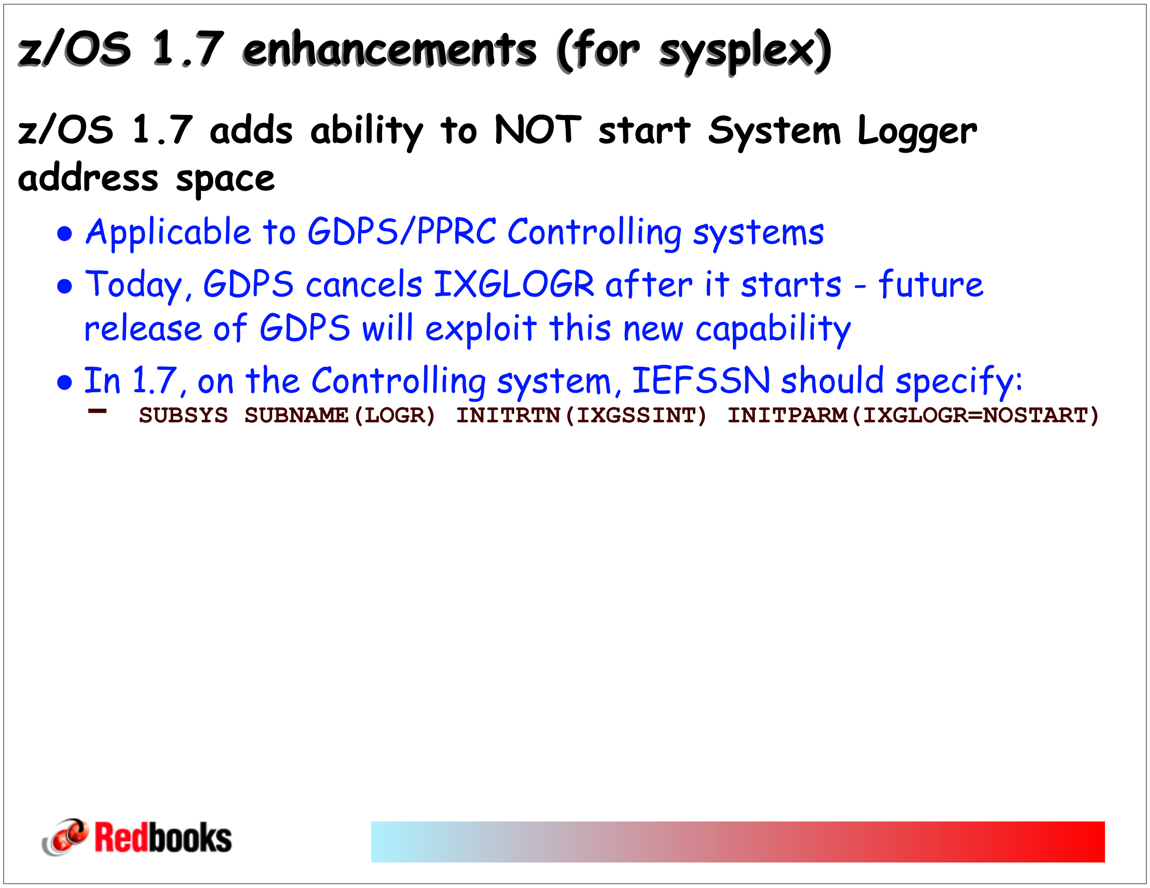
z/OS 1.7 adds "XRC+" for GDPS/XRC environments

- Use of staging data sets is required when using GDPS, to ensure that log stream data is available in the "remote" site
- However, staging data sets can be a bottleneck in environments with very high logging rates...

In 1.7, you can tell System Logger which log streams should use this capability (DUPEXMODE(DRXRC) keyword in log stream definition)

- Requires LOGR CDS formatted with SMDUPLEX keyword

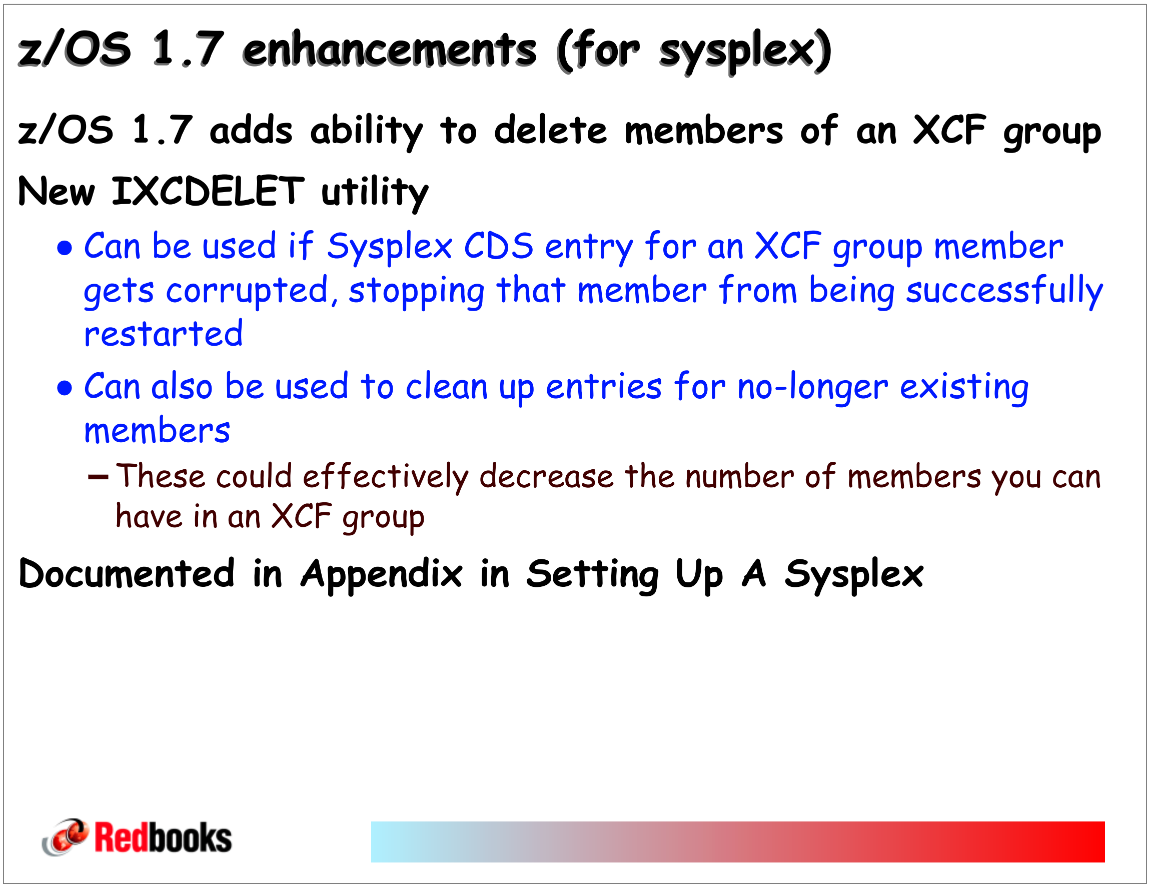
When IPLing in disaster mode, specify DRMODE=YES



z/OS 1.7 enhancements (for sysplex)

z/OS 1.7 adds ability to NOT start System Logger address space

- Applicable to GDPS/PPRC Controlling systems
- Today, GDPS cancels IXGLOGR after it starts - future release of GDPS will exploit this new capability
- In 1.7, on the Controlling system, IEFSSN should specify:
  - `SUBSYS SUBNAME(LOGR) INITRTN(IXGSSINT) INITPARM(IXGLOGR=NOSTART)`



z/OS 1.7 enhancements (for sysplex)

z/OS 1.7 adds ability to delete members of an XCF group

New IXCDELET utility

- Can be used if Sysplex CDS entry for an XCF group member gets corrupted, stopping that member from being successfully restarted
- Can also be used to clean up entries for no-longer existing members
  - These could effectively decrease the number of members you can have in an XCF group

Documented in Appendix in Setting Up A Sysplex

z/OS 1.7 announcements

z/OS 1.7 also adds:

- Support for up to 32 CPs in a single z/OS image
- Support for multiple subchannel sets on System z9, freeing up subchannels by moving PAV aliases to a seperate subchannel set
- Increased number of subchannels in System z9 - increased by 768 up to 65,280
- Support for non-extended format sequential data sets larger than 64K tracks
- Increased number of extents for VSAM data sets - up from 255 extents (for non-extended format) to 123 extents per volume x 59 volumes (must be SMS-managed)
- Enhanced parallelism in Vary processing
  - Queueing against Q4 SYSIEFSD reduced
  - Vary OFF processing parallelized

z/OS 1.7 denouncements

z/OS 1.7 will be the last release to support the z/OS-related plugins for msys for Setup

- DB2 V8 plugin is unaffected
- TCP/IP plugin will continue to be available but will no longer use msys for Setup

z/OS 1.7 will be the last release to support msys for Operations

- "IBM plans to transition many of the current msys for Operations functions to a new user interface and infrastructure in a future release of z/OS."

Communications Server enhancements

In z/OS 1.7, there are a number of Comms Server enhancements specifically relating to sysplex..

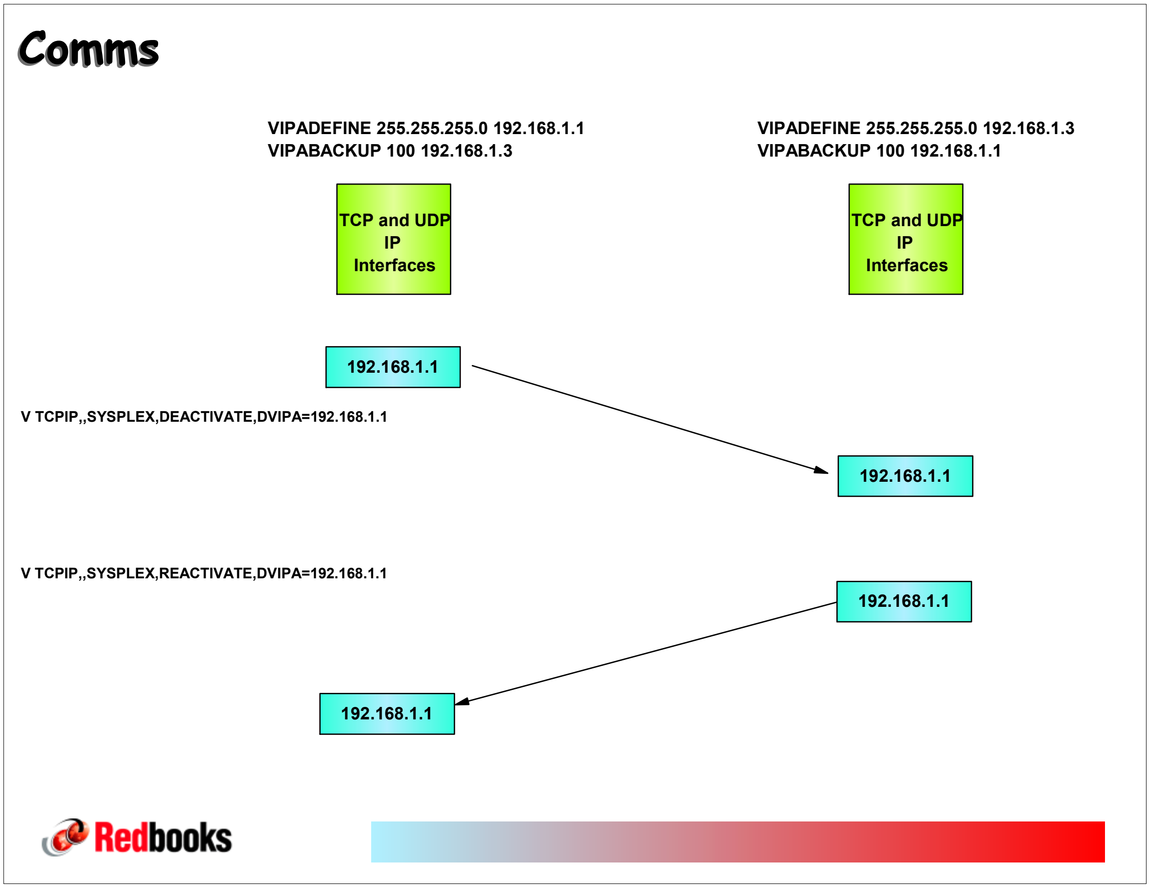
The first is Optimized Routing:

- Prior to 1.7 you could use Sysplex routing (over XCF) or traditional TCP routing (over OSA, for example), but not both
- In 1.7, you can tell IP to use the faster of XCF or an alternate route for DVIPA packets, using new VIPAROUTE statement
- Recommend that you wait until all systems are running z/OS 1.7 or later before enabling this feature
- For more information, refer to:
  - z/OS Communications Server: IP New function summary, GC31-8771

Communications Server enhancements

Next enhancement is server-level WLM recommendations:

- There are now 4 options for workload balancing:
  - BASEWLM - requests distributed based on available capacity in each target system (enhanced in 1.7 to be more granular)
  - ROUNDROBIN - balance requests around all system equally
  - SERVERWLM (new) - use goal achievement information to determine target server
    - If there are multiple servers per stack, all listening on the same port, you can use WLM to determine which instance requests are routed to
  - Quality of Service - use Policy Agent with BASEWLM or SERVERWLM to adjust WLM recommendations based on network performance
- This controlled by DISTMETHOD keyword on VIPADISTRIBUTE
- You can use different methods for different servers on same system



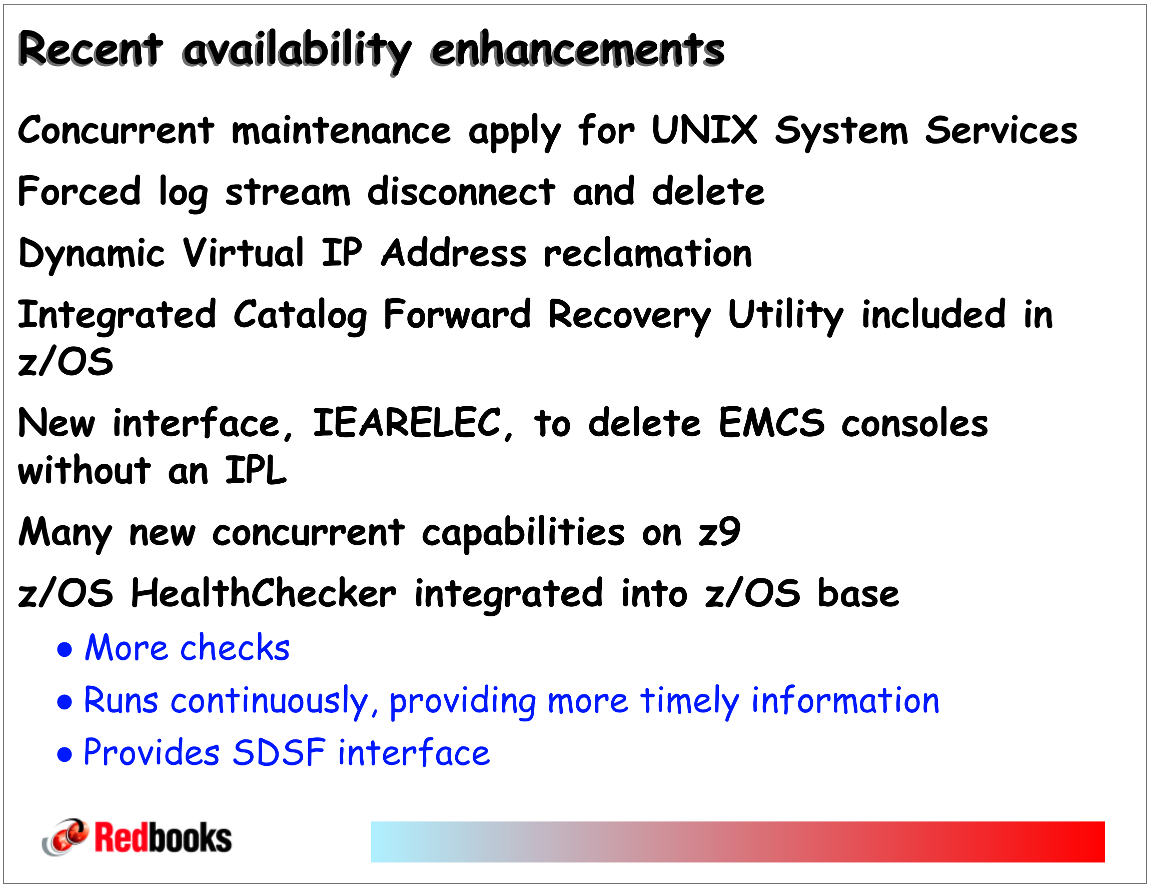
Comms

```
VIPADEFINE 255.255.255.0 192.168.1.1
VIPABACKUP 100 192.168.1.3
```

```
VIPADEFINE 255.255.255.0 192.168.1.3
VIPABACKUP 100 192.168.1.1
```

```
V TCPIP,,SYSPLEX,DEACTIVATE,DVIPA=192.168.1.1
```

```
V TCPIP,,SYSPLEX,REACTIVATE,DVIPA=192.168.1.1
```



Recent availability enhancements

Concurrent maintenance apply for UNIX System Services

Forced log stream disconnect and delete

Dynamic Virtual IP Address reclamation

Integrated Catalog Forward Recovery Utility included in z/OS

New interface, IEARELEC, to delete EMCS consoles without an IPL

Many new concurrent capabilities on z9

z/OS HealthChecker integrated into z/OS base

- More checks
- Runs continuously, providing more timely information
- Provides SDSF interface

®

CF Monitoring/Planning Tools

Tools

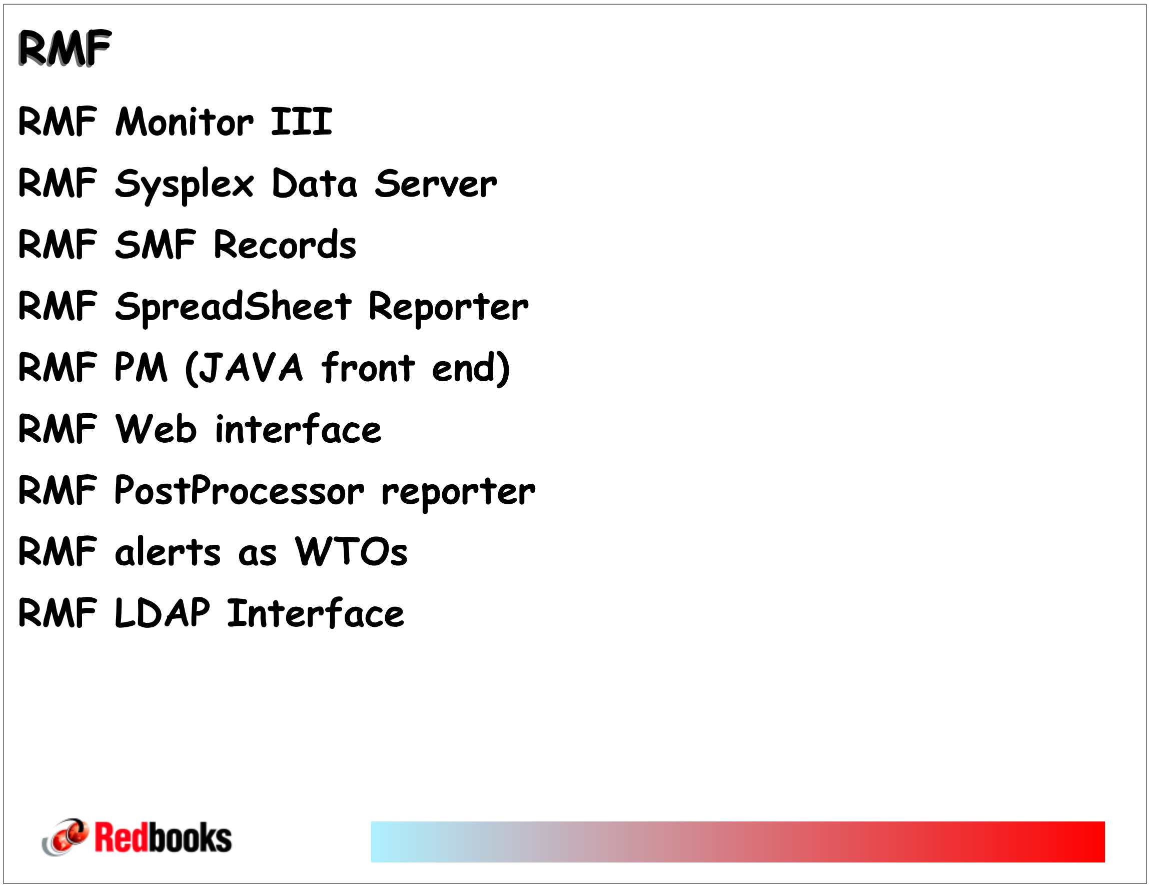
RMF

IBM Tivoli Omegamon

Tivoli Decision Support (great-grandson of SLR)

Mario Bezzi tools

Eric Frederiksen tools



RMF

RMF Monitor III

RMF Sysplex Data Server

RMF SMF Records

RMF SpreadSheet Reporter

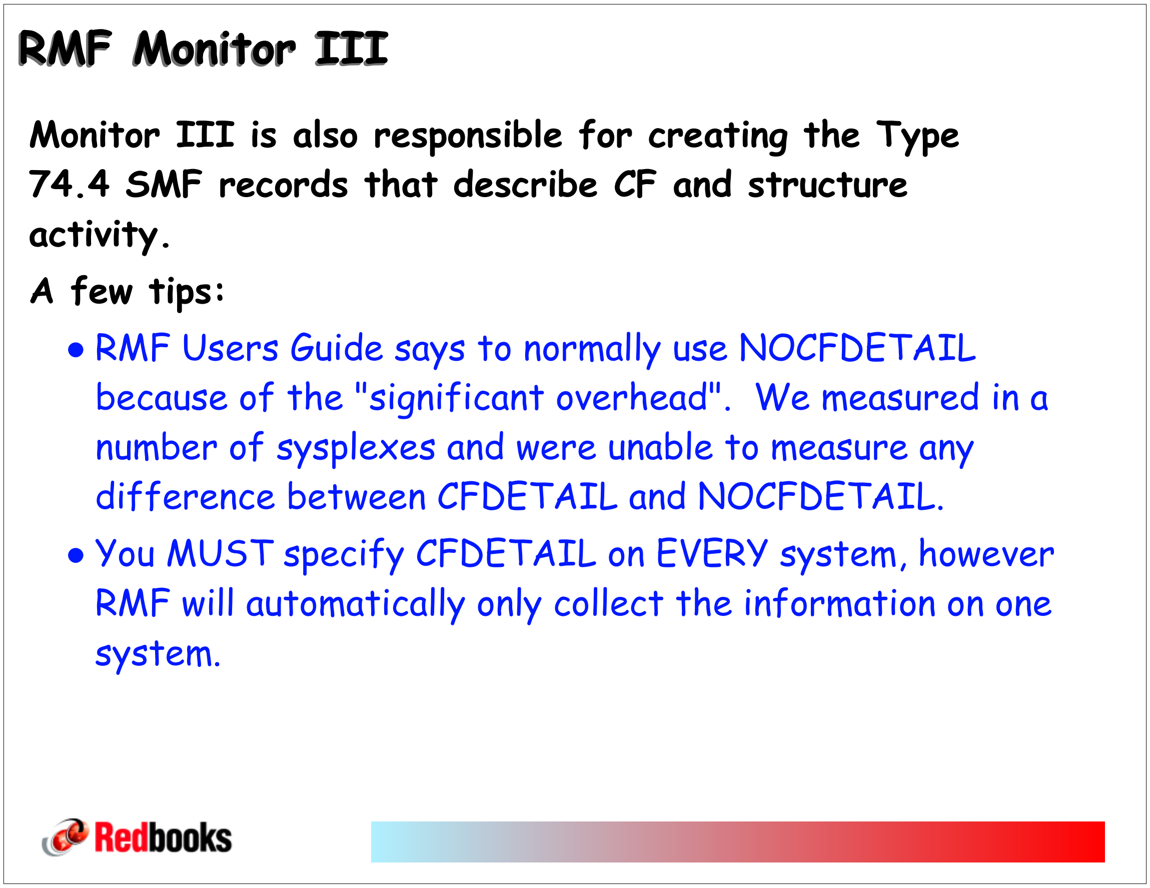
RMF PM (JAVA front end)

RMF Web interface

RMF PostProcessor reporter

RMF alerts as WTOs

RMF LDAP Interface



RMF Monitor III

Monitor III is also responsible for creating the Type 74.4 SMF records that describe CF and structure activity.

A few tips:

- RMF Users Guide says to normally use NOCFDETAIL because of the "significant overhead". We measured in a number of sysplexes and were unable to measure any difference between CFDETAIL and NOCFDETAIL.
- You MUST specify CFDETAIL on EVERY system, however RMF will automatically only collect the information on one system.

RMF Monitor III

Monitor III supports the use of VSAM data sets as a wrap around buffer to extend the amount of data(time) that is viewable in Mon III.

We strongly recommend the use of these data sets to keep up to 3 days of data if possible.

- Allocate the data sets (recommend 8-10 per system at about 50 cyls each) using clist:
- ERBVSDEF vsam_dsn VSAMVOL(volume) [TRACKS(num_tracks)]

Specify the data set names in the ERBRMF04 member of parmlib using:

- DATASET ADD(dsn,dsn...)

RMF SMF record

Just about everything you can display in any of the RMF interfaces is saved in the RMF SMF records. In fact, there is some information that RMF writes in his SMF records that he doesn't report on himself!

From a sysplex perspective, the interesting records are:

- Records about XCF group, transport class, and path usage - Type 74, Subtype 2
- Records about CFs and structures, Type 74 Subtype 4

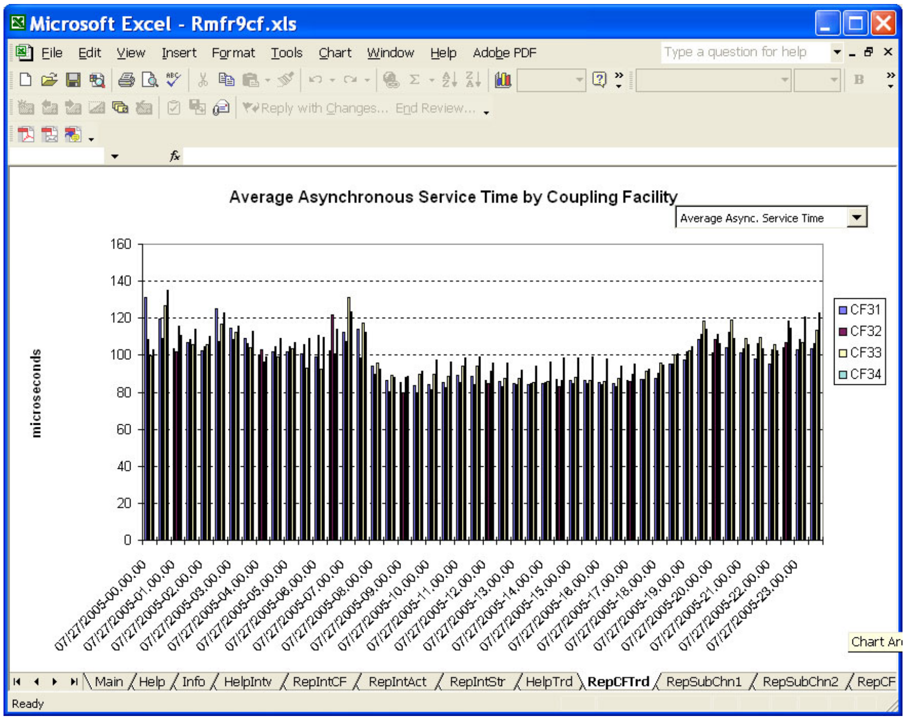
There are a variety of tools to postprocess these records, both from IBM and other vendors

RMF SpreadSheet Reporter

What is it?

- Tool to convert select RMF PostProcessor reports into a format that can be processed in a spreadsheet program
- Set of spreadsheet macros to produce common reports
- Easy way to spot out-of-line situations, detect trends
- Available in SERBPWSV library member ERB9R2SW
  - Also downloadable from RMF Home page

RMF Spreadsheet Reporter



RMF Spreadsheet Reporter

For CF-related reports, SR provides:

- A set of reports showing data for one interval for:
- CF reports, showing:
  - CF Utilization
  - Storage allocation in MB and %
  - Average number of requests per second
  - Average Sync and Async service times
- Each structure broken out by connected system:
  - Number of Total, Synch, Asynch, and Changed requests
  - Synch and Asynch service times
  - Number of no subch avail, and delay due to no subch avail
  - Number of external requests and external requests deferred
  - Real and False contention

RMF Spreadsheet Reporter

For CF-related reports, SR provides:

- The following info for all/list/cache/lock/top 5/top 10 strs:
  - Allocated size, % of CF storage
  - Number of requests. % of all CF requests, avg req/sec
  - Current and Total dir/entries
  - Current and Total elements and lock entries
  - Number of directory reclaims

RMF Spreadsheet Reporter

For CF-related reports, SR provides:

- A number of trend reports, showing trends over the period of data in the SMF data set(s) and the PostProcessor report:
- The following every CF over the duration:
  - CF CPU Utilization
  - Average number of requests (combined Synch and Asynch)
  - Average Synch response time
  - Average Asynch response time
- The following for a combination of CF/System for the duration
  - Path and Subchannel Busy
  - Number of Synch and Asynch requests
  - Number of delayed requests
  - Percent of requests that are delayed

RMF Spreadsheet Reporter

For CF-related reports, SR provides:

- The following for a combination of CF/System for the duration
  - Synch and Asynch response times plotted against number of requests
- The following information, for 1 CF, broken out by system, for every interval in the duration:
  - Avg number of requests/sec
  - Number of Synch requests
  - Number of Asynch requests
  - Number of Path Busy
  - Number of Changed Requests
  - Average Synch Service Time
  - Average Asynch Service Time

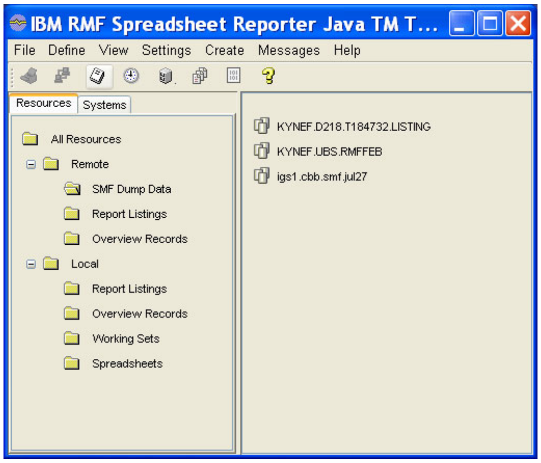
RMF Spreadsheet Reporter

For CF-related reports, SR provides:

- The following information for each structure, for all systems or a selected system for the duration:
  - Total requests
  - Directory Entry to Element ratio
  - Number of Synch and Asynch requests (plotted together)
  - Number of Synch and Changed requests (plotted together)
  - Number of Asynch, Changed, and no subchan requests
  - Synch requests and Synch Service Time
  - Asynch requests and Asynch Service Time
  - True and False contention

RMF Spreadsheet Reporter

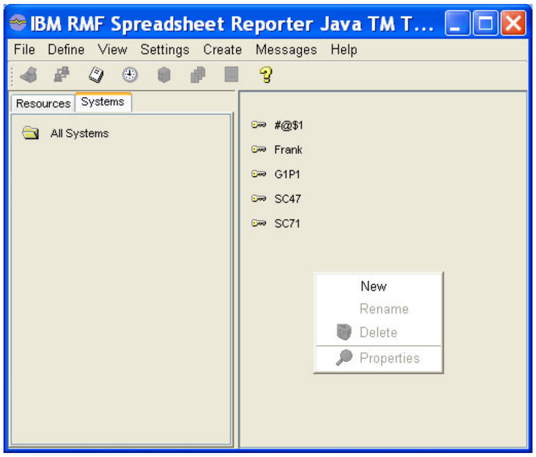
Spreadsheet Reporter initial screen....

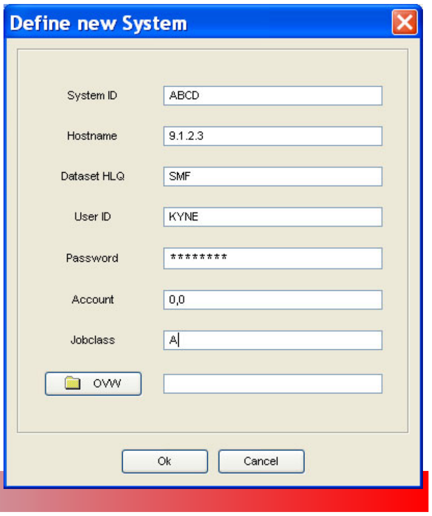


RMF Spreadsheet Reporter

Using the SR:

- Define your systems and userids:
  - Click on Systems tab
  - Right mouse in work area (right panel)
  - Select "New"


IBM RMF Spreadsheet Reporter Java TM T...
File Define View Settings Create Messages Help
Resources
Systems
All Systems
#@$1
Frank
G1P1
SC47
SC71
New
Rename
Delete
Properties

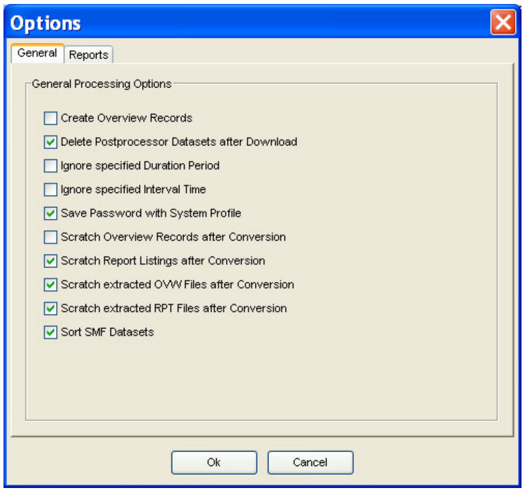

Define new System

System ID: ABCD
Hostname: 9.1.2.3
Dataset HLQ: SMF
User ID: KYNE
Password: ********
Account: 0,0
Jobclass: A
OVW:

RMF Spreadsheet Reporter

Using the SR:

- Customize jobs that will be run:
  - Click on Settings, then Options
  - Customize job as required (suggest using these settings)
  - Changes apply to all systems



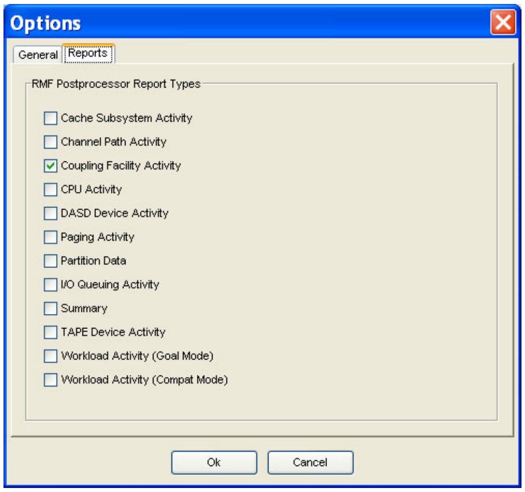
General Processing Options

- □ Create Overview Records
- ☑ Delete Postprocessor Datasets after Download
- □ Ignore specified Duration Period
- □ Ignore specified Interval Time
- ☑ Save Password with System Profile
- □ Scratch Overview Records after Conversion
- ☑ Scratch Report Listings after Conversion
- ☑ Scratch extracted OVW Files after Conversion
- ☑ Scratch extracted RPT Files after Conversion
- ☑ Sort SMF Datasets

RMF Spreadsheet Reporter

Using the SR:

- Customize jobs that will be run:
  - Decide which PP reports you want



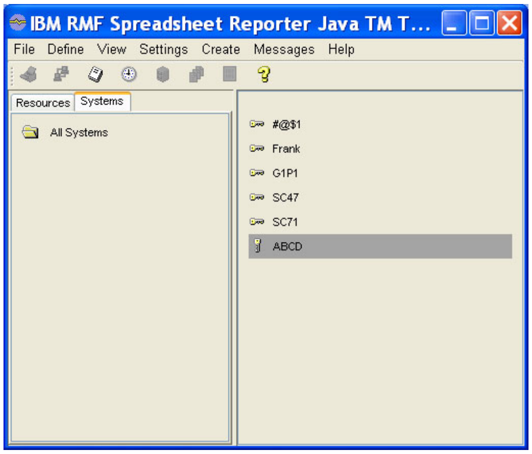
RMF Postprocessor Report Types

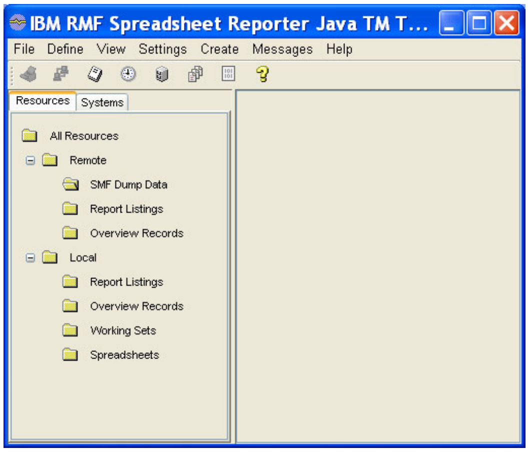
☐ Cache Subsystem Activity
☐ Channel Path Activity
☑ Coupling Facility Activity
☐ CPU Activity
☐ DASD Device Activity
☐ Paging Activity
☐ Partition Data
☐ I/O Queuing Activity
☐ Summary
☐ TAPE Device Activity
☐ Workload Activity (Goal Mode)
☐ Workload Activity (Compat Mode)

RMF Spreadsheet Reporter

Using the SR:

- Define your SMF data sets:
  - Select system you are defining for
  - Click Resources tab, then click on "SMF Dump Data"

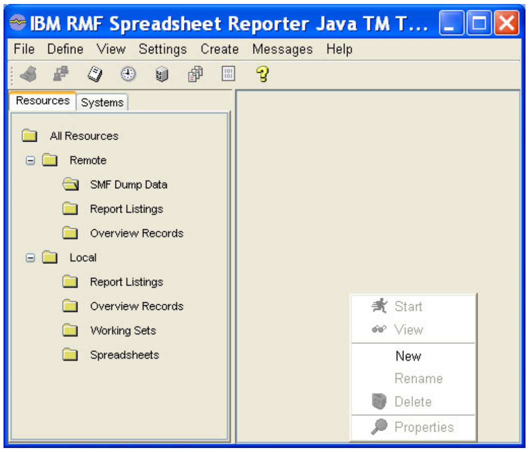


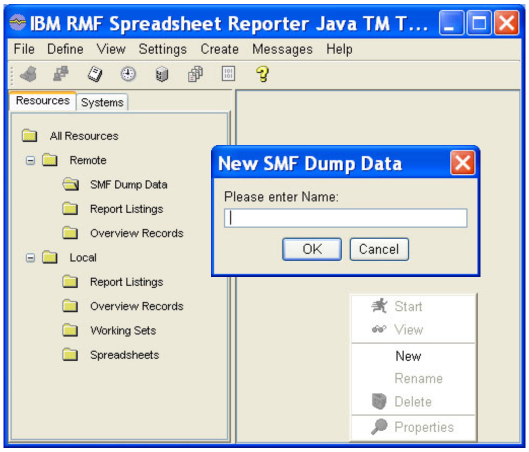


RMF Spreadsheet Reporter

Using the SR:

- Define your SMF data sets (cont):
  - Right mouse in work area, select "New"
  - Enter complete MVS data set name, without quotes





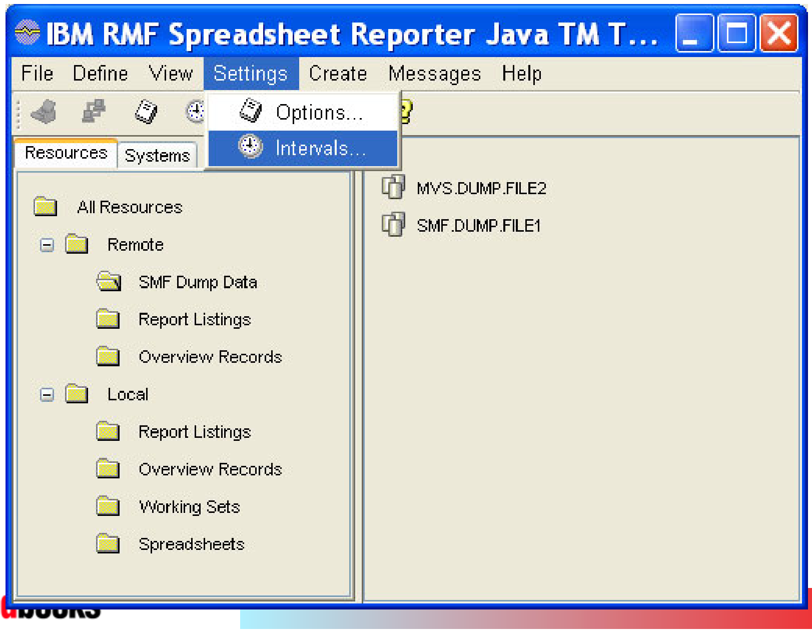
New SMF Dump Data

Please enter Name:

RMF Spreadsheet Reporter

Using the SR:

- Now you want to specify the PP time and date parameters..



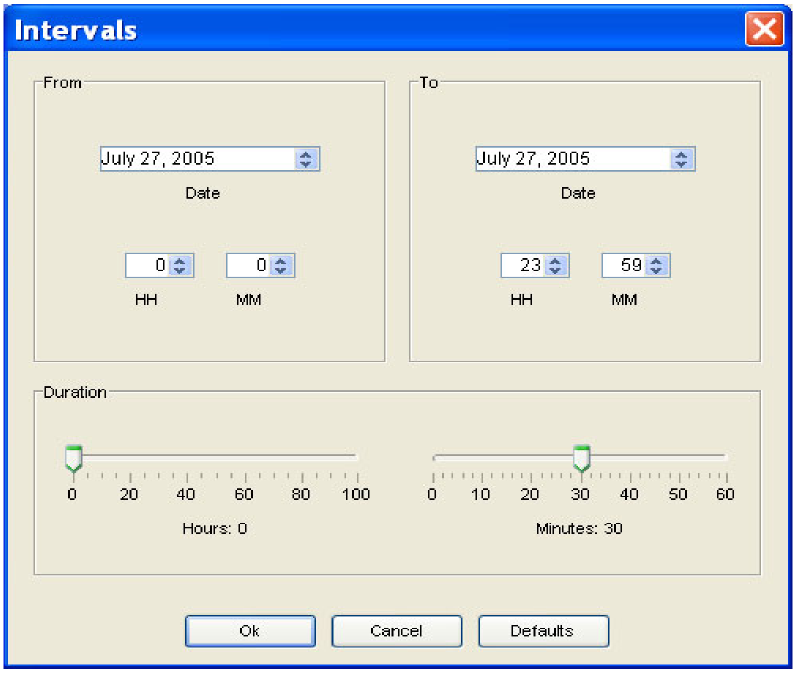
MVS.DUMP.FILE2

SMF.DUMP.FILE1

RMF Spreadsheet Reporter

Using the SR:

- Intervals and the durations



From: Date July 27, 2005; HH 0; MM 0

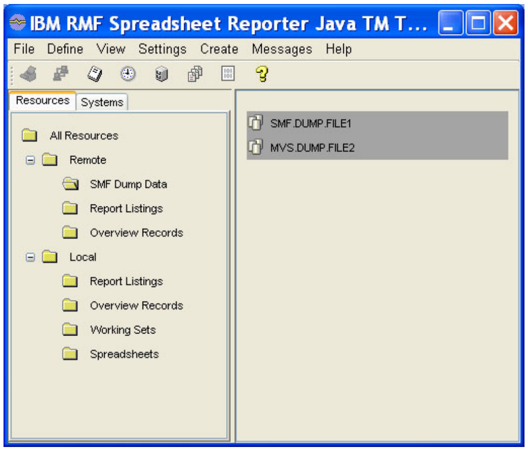
To: Date July 27, 2005; HH 23; MM 59

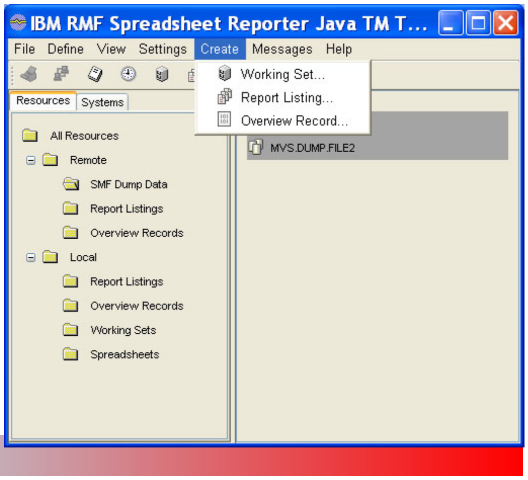
Duration: Hours: 0; Minutes: 30

RMF Spreadsheet Reporter

Using the SR:

- Getting the data to your PC:
  - Select all the SMF file(s) you want to process (SR will sort)
  - Click on Create, then Working Set to create RMF PP reports, download to PC, and format for use by spreadsheet program

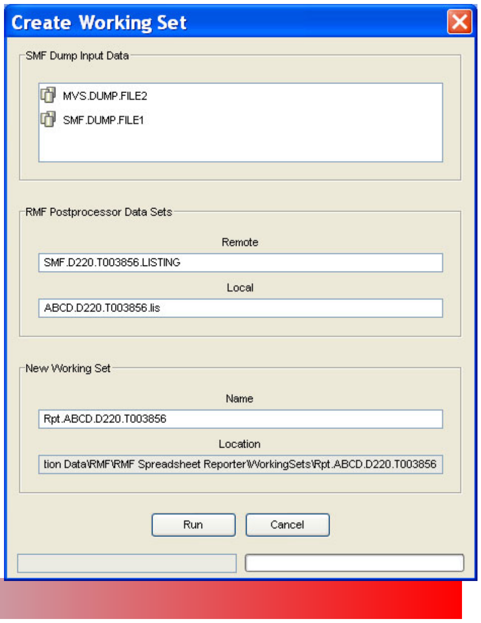




RMF Spreadsheet Reporter

Using the SR:

- Getting the data to your PC (cont):
  - Verify files names (should be OK)
  - Click on Run. This:
    - ► Submits job using FTP
    - ► Downloads resulting PP report
    - ► Processes into S/Sheet format
    - ► Status is shown under "Run" button



SMF Dump Input Data

MVS.DUMP.FILE2
SMF.DUMP.FILE1

RMF Postprocessor Data Sets

Remote: SMF.D220.T003856.LISTING

Local: ABCD.D220.T003856.lis

New Working Set

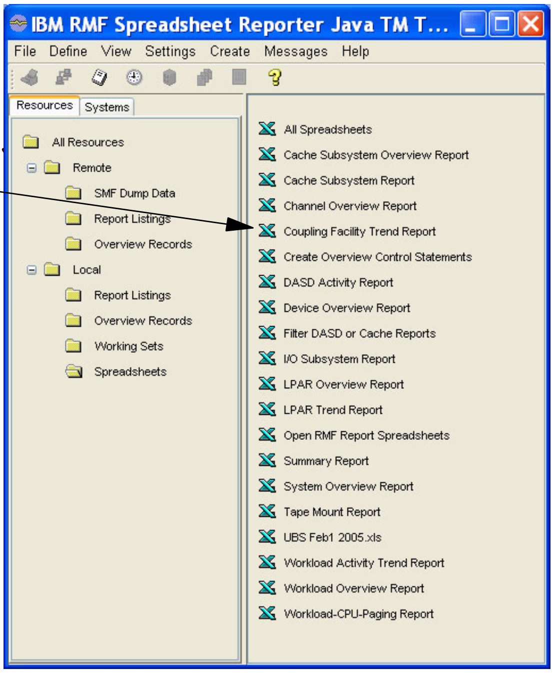
Name: Rpt.ABCD.D220.T003856

Location: tion Data\RMF\RMF Spreadsheet Reporter\WorkingSets\Rpt.ABCD.D220.T003856

RMF Spreadsheet Reporter

Using the SR:

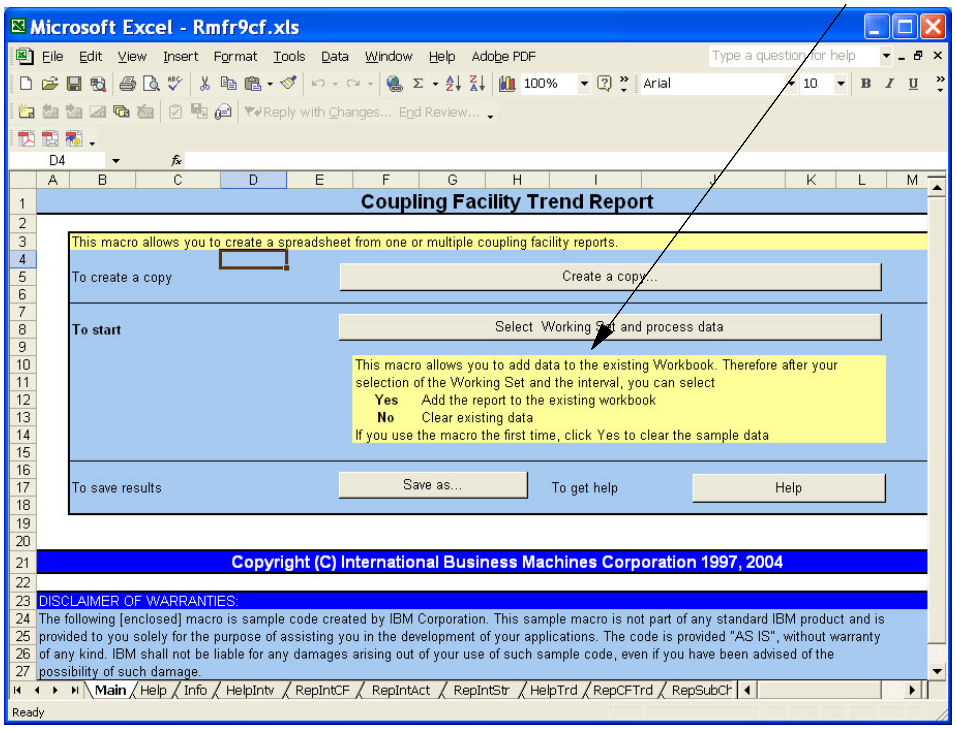
- When working set created:
  - Click "Spreadsheets"
  - Then double-click the report



All Resources
Remote
SMF Dump Data
Report Listings
Overview Records
Local
Report Listings
Overview Records
Working Sets
Spreadsheets

All Spreadsheets
Cache Subsystem Overview Report
Cache Subsystem Report
Channel Overview Report
Coupling Facility Trend Report
Create Overview Control Statements
DASD Activity Report
Device Overview Report
Filter DASD or Cache Reports
I/O Subsystem Report
LPAR Overview Report
LPAR Trend Report
Open RMF Report Spreadsheets
Summary Report
System Overview Report
Tape Mount Report
UBS Feb1 2005.xls
Workload Activity Trend Report
Workload Overview Report
Workload-CPU-Paging Report

RMF Spreadsheet Reporter


Microsoft Excel - Rmfr9cf.xls
File Edit View Insert Format Tools Data Window Help Adobe PDF
Type a question for help
100%
Arial
10
Reply with Changes... End Review...
D4
Create a copy...
Select Working Set and process data
Save as...
Help
Main Help Info HelpIntv RepIntCF RepIntAct RepIntStr HelpTrd RepCFTrd RepSubCh
Ready

Coupling Facility Trend Report

This macro allows you to create a spreadsheet from one or multiple coupling facility reports.

To create a copy

To start

This macro allows you to add data to the existing Workbook. Therefore after your selection of the Working Set and the interval, you can select

Yes Add the report to the existing workbook

No Clear existing data

If you use the macro the first time, click Yes to clear the sample data

To save results

To get help

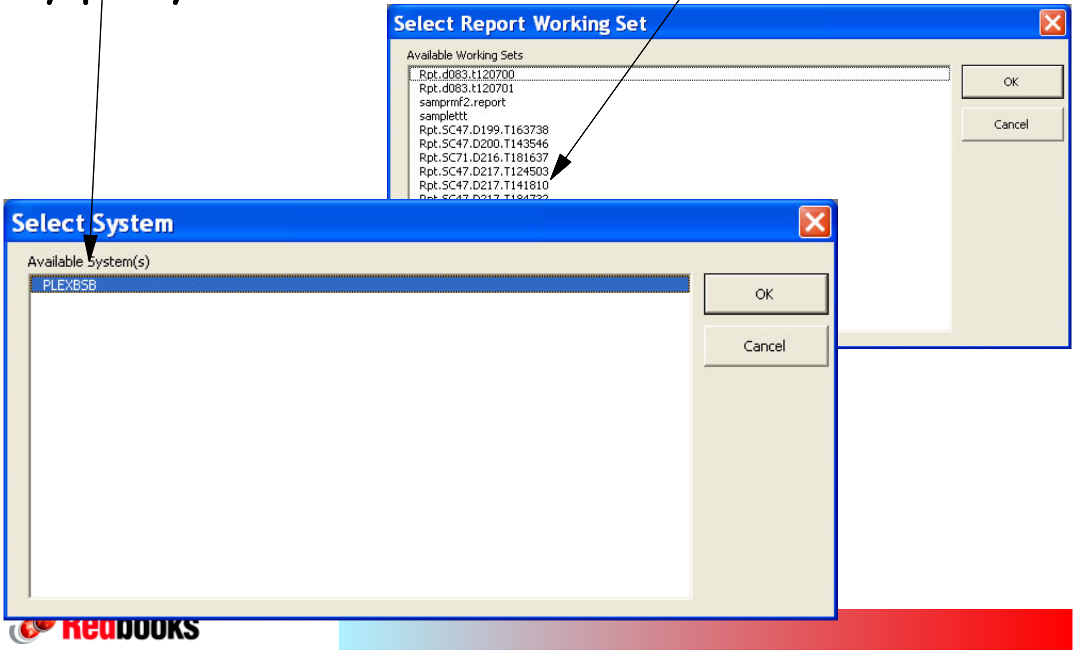
Copyright (C) International Business Machines Corporation 1997, 2004

DISCLAIMER OF WARRANTIES:

The following [enclosed] macro is sample code created by IBM Corporation. This sample macro is not part of any standard IBM product and is provided to you solely for the purpose of assisting you in the development of your applications. The code is provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of such sample code, even if you have been advised of the possibility of such damage.

RMF Spreadsheet Reporter

Select the working set you just created, then the sysplex you are interested in...



Available Working Sets

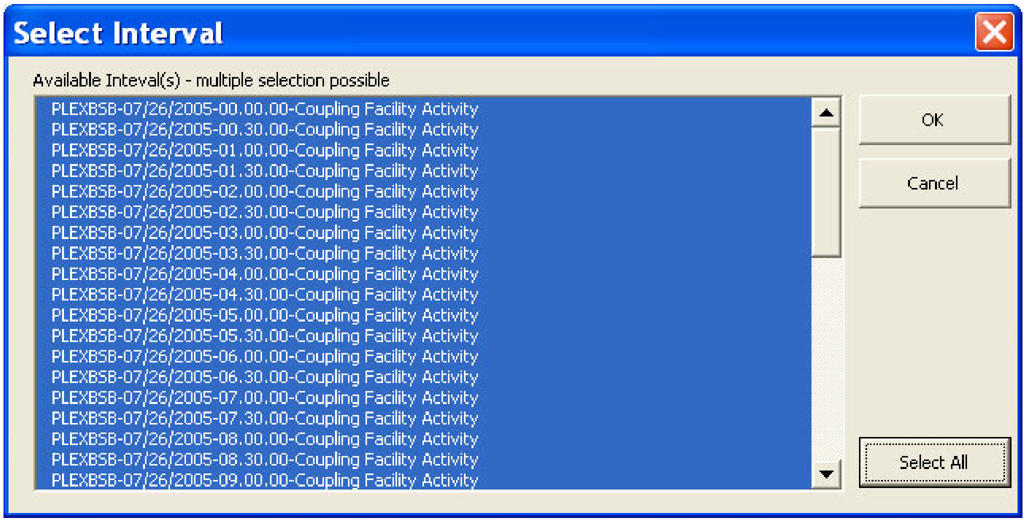
Rpt.d083.t120700
Rpt.d083.t120701
samprmf2.report
samplettt
Rpt.SC47.D199.T163738
Rpt.SC47.D200.T143546
Rpt.SC71.D216.T181637
Rpt.SC47.D217.T124503
Rpt.SC47.D217.T141810
Rpt.SC47.D217.T184732

Available System(s)

PLEXBSB

RMF Spreadsheet Reporter

And finally the intervals you are interested in....



Available Inteval(s) - multiple selection possible

```
PLEXBSB-07/26/2005-00.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-00.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-01.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-01.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-02.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-02.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-03.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-03.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-04.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-04.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-05.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-05.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-06.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-06.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-07.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-07.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-08.00.00-Coupling Facility Activity
PLEXBSB-07/26/2005-08.30.00-Coupling Facility Activity
PLEXBSB-07/26/2005-09.00.00-Coupling Facility Activity
```

Redbooks

RMF Spreadsheet Reporter

A few hints and tips:

- DON'T save the spreadsheet you just worked with - do a Save As to save to a different name
- Download the latest version (currently 5.1.5) of the SR (APAR OA10346) from SERBPWS data set or from RMF Web site:
  http://www.ibm.com/servers/eserver/zseries/rmf/
- Update the provided skeleton JCL to have more realistic data set sizes and blksizes (program file/RMF/RMF Spreadsheet Reporter/Connect/RMFPP1.JCL)
- If you find that it is taking a looooong time to download the reports, try disabling local firewall and set priority of javaw to "belownormal".

Redbooks

Mario Bezzi tools

Mario Bezzi works for IGS in Italy and has developed a number of tools to process SMF records into DB2, run SQL commands through FTP to download the data to PC, and from there into a spreadsheet.

He currently has tools for:

- SMF Type 33 APPC records
- SMF Type 88 (System Logger)
- SMF Type 74 subtype 2 (XCF)
- SMF Type 74 subtype 4 (CF) (still under development)

Mario Bezzi tools

SMF Type 33 (APPC) reports

Sheet name	Description
WQATB09R	selected measurement intervals
WQATB10R	all conversations by synclevel
WQATB11R	all conversations by smf33sid
WQATB12R	all conversations by smf33jid
WQATB13R	all conversations by smf33cll
WQATB14R	all conversations by smf33cpl
WQATB15R	all conversations by smf33tpn L
WQATB16R	all conversations by relationships
WQATB21R	sl-none conversations by smf33sid
WQATB22R	sl-none conversations by smf33jid
WQATB23R	sl-none conversations by smf33cll
WQATB24R	sl-none conversations by smf33cpl
WQATB25R	sl-none conversations by smf33tpn L
WQATB26R	sl-none conversations by relationships
WQATB31R	sl-confirm conversations by smf33sid
WQATB32R	sl-confirm conversations by smf33jid
WQATB33R	sl-confirm conversations by smf33cll
WQATB34R	sl-confirm conversations by smf33cpl
WQATB35R	sl-confirm conversations by smf33tpn L
WQATB36R	sl-confirm conversations by relationships
WQATB41R	sl-syncpt conversations by smf33sid
WQATB42R	sl-syncpt conversations by smf33jid
WQATB43R	sl-syncpt conversations by smf33cll
WQATB44R	sl-syncpt conversations by smf33cpl
WQATB45R	sl-syncpt conversations by smf33tpn L
WQATB46R	sl-syncpt conversations by relationships
WQATB51R	local conversations by smf33sid
WQATB52R	local conversations by smf33jid
WQATB53R	local conversations by smf33cll
WQATB54R	local conversations by smf33cpl
WQATB55R	local conversations by smf33tpn L
WQATB56R	local conversations by relationships

Mario Bezzi tools

SMF Type 88 (System Logger) reports

Sheet Name	Description
WQIXG10O	Overview - Selected Interval
WQIXG11O	Overview - LogStream Counters
WQIXG12O	Overview - LogStream Counters
WQIXG13O	Overview - LogStream Counters
WQIXG14O	Overview - System Activity Summary
WQIXG15O	Overview - System Activity Details by Interval
WQIXG16O	Overview - System Activity Summary by Logstream
WQIXG20O	Overview - Logstream Activity Summary
WQIXG21O	Overview - Logstream Activity Details - Selected Logstream
WQIXG22O	Overview - Logstream Activity Details - All the Logstreams
WQIXG23O	Overview - Logstream Connectors - All the Logstreams by Interval
WQIXG24O	Overview - LogStream Blocksize and Hot Air factors
WQIXG50E	Exceptions - SMF88ESF GT 0 Summary
WQIXG51E	Exceptions - SMF88ESF GT 0 Details
WQIXG52E	Exceptions - SMF88ETF GT 0 Summary
WQIXG53E	Exceptions - SMF88ETF GT 0 Details
WQIXG54E	Exceptions - SMF88EFS GT 0 Summary
WQIXG55E	Exceptions - SMF88EFS GT 0 Details
WQIXG56E	Exceptions - SMF88SC3 GT 0 Summary
WQIXG57E	Exceptions - SMF88SC3 GT 0 Details
WQIXG58E	Exceptions - SMF88SAB GT 0 Summary
WQIXG59E	Exceptions - SMF88SAB GT 0 Details
WQIXG60E	Exceptions - SMF88SIB EQ 0 Summary
WQIXG61E	Exceptions - SMF88SIB EQ 0 Details
WQIXG62E	Exceptions - SMF88EO GT 0 Summary
WQIXG63E	Exceptions - SMF88EO GT 0 Details
WQIXG64E	Exceptions - SMF88EDS GT 0 Summary
WQIXG65E	Exceptions - SMF88EDS GT 0 Details
WQIXG66E	Exceptions - SMF88ERI GT 0 Summary
WQIXG67E	Exceptions - SMF88ERI GT 0 Details

Redbooks

Mario Bezzi tools

SMF Type 74.4 (CF) reports

Sheet Name	Description
WQIXL10M	**XES Activity - Selected measurement intervals**
WQIXL11S	**XES Activity - Coupling Facility Connectivity by system**
WQIXL20S	**XES Activity - Sysplex Summary**
WQIXL21S	**XES Activity - Coupling Facility Summary**
WQIXL22S	**XES Activity - System Summary**
WQIXL23S	**XES Activity - System by CF Summary**
WQIXL24D	**XES Activity - Subchannel data by System**
WQIXL25D	**XES Activity - Coupling Facility Resources Utilization**
WQIXL30S	**XES Activity - Structure Activity, Sysplex Summary**
WQIXL31S	**XES Activity - Structure Activity, System Summary**
WQIXL32S	**XES Activity - CF Activity, Sysplex Summary**
WQIXL33S	**XES Activity - CF Activity, System Summary**
WQIXL34D	**XES Activity - Structure Activity, System Details**
WQIXL35D	**XES Activity - Structure Resource Utilization, System Details**
WQIXL50E	**XES Activity - Lock Contention Analysis**

Redbooks

Mario Bezzi tools

SMF Type 74.2
(XCF) reports

Sheet name	Description
WQIXC20O	System Activity - OutBound Traffic - Sysplex Overview
WQIXC21O	System Activity - OutBound Traffic - System Overview
WQIXC22O	System Activity - InBound Traffic - System Overview
WQIXC23O	System Activity - Local Traffic - System Overview
WQIXC24O	System Activity - OutBound Transport Class Overview
WQIXC25O	System Activity - Local Transport Class Overview
WQIXC26D	System Activity - OutBound Transport Class Details
WQIXC27D	System Activity - Local Transport Class Details
WQIXC28D	System Activity - InBound System Details
WQIXC30E	System Exceptions - Abnormal System Status
WQIXC31E	System Exceptions - No OutBound Paths
WQIXC32E	System Exceptions - No OutBound Buffers
WQIXC33E	System Exceptions - No InBound Buffers
WQIXC34E	System Exceptions - No Local Buffers
WQIXC40D	Path Activity - Outbound Path Details
WQIXC41D	Path Activity - Inbound Path Details
WQIXC42O	Path Activity - Inbound Path Structure Performances
WQIXC43O	Path Activity - Inbound Path CTC Performances
WQIXC51E	Path Exceptions - OutBound Paths with retries
WQIXC52E	Path Exceptions - InBound Paths with rejects
WQIXC53E	Path Exceptions - Abnormal Path Status
WQIXC60O	Member Activity - Group Summary
WQIXC61O	Member Activity - Group Overview
WQIXC62D	Member Activity - Group Details
WQIXC63O	Member Activity - System Overview
WQIXC64D	Member Activity - System Details
WQIXC65D	Member Activity - Member Details

Redbooks

Erik Frederiksen tools

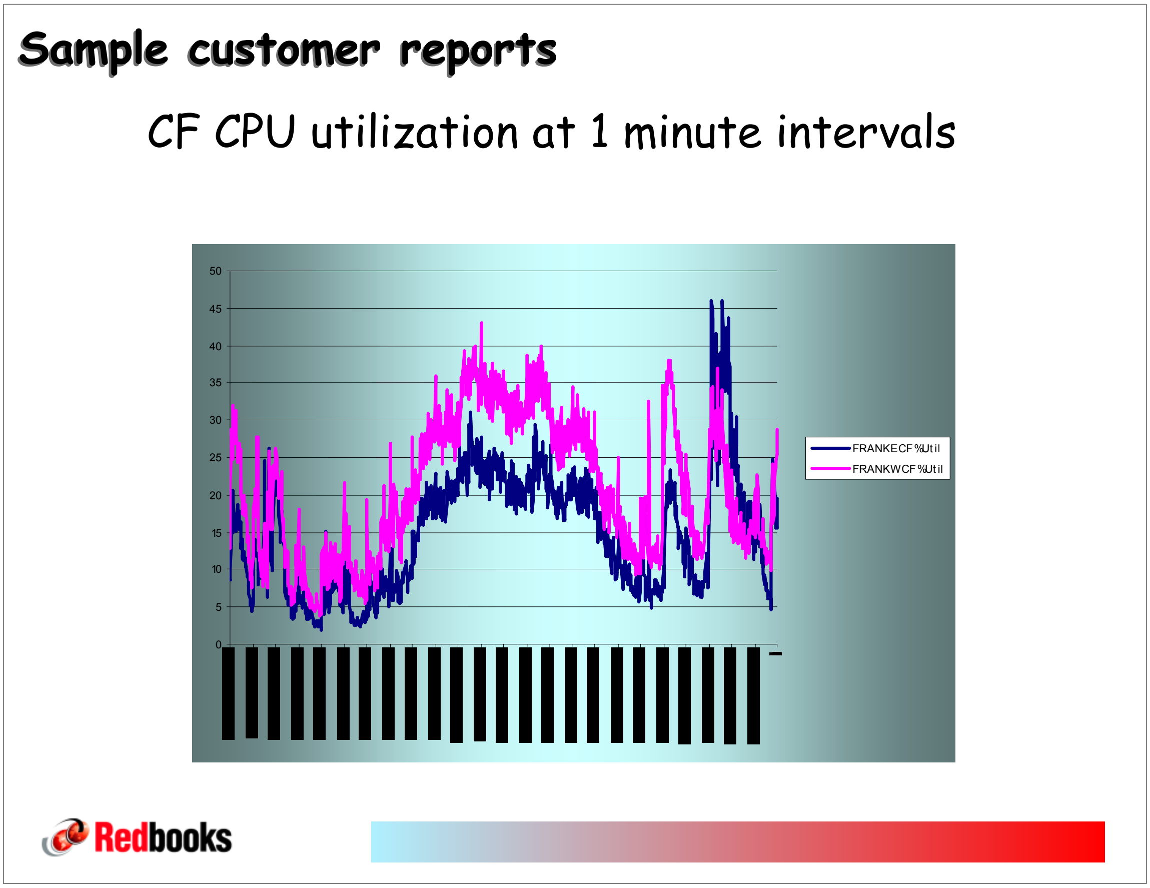
Erik works for IBM SO in Denmark, and has developed a number of tools for short term trending and reporting on CF activity

He has one set of tools to extract information from RMF Mon III, at one minute intervals, and build in a form that can be loaded into a spreadsheet for graphing

He has another set of tools to create and e-mail exception reports on a daily basis

And he has a tool that runs as an STC, putting out WTOs whenever RMF detects that any of a set of thresholds have been exceeded. The WTO can then be acted on by automation.

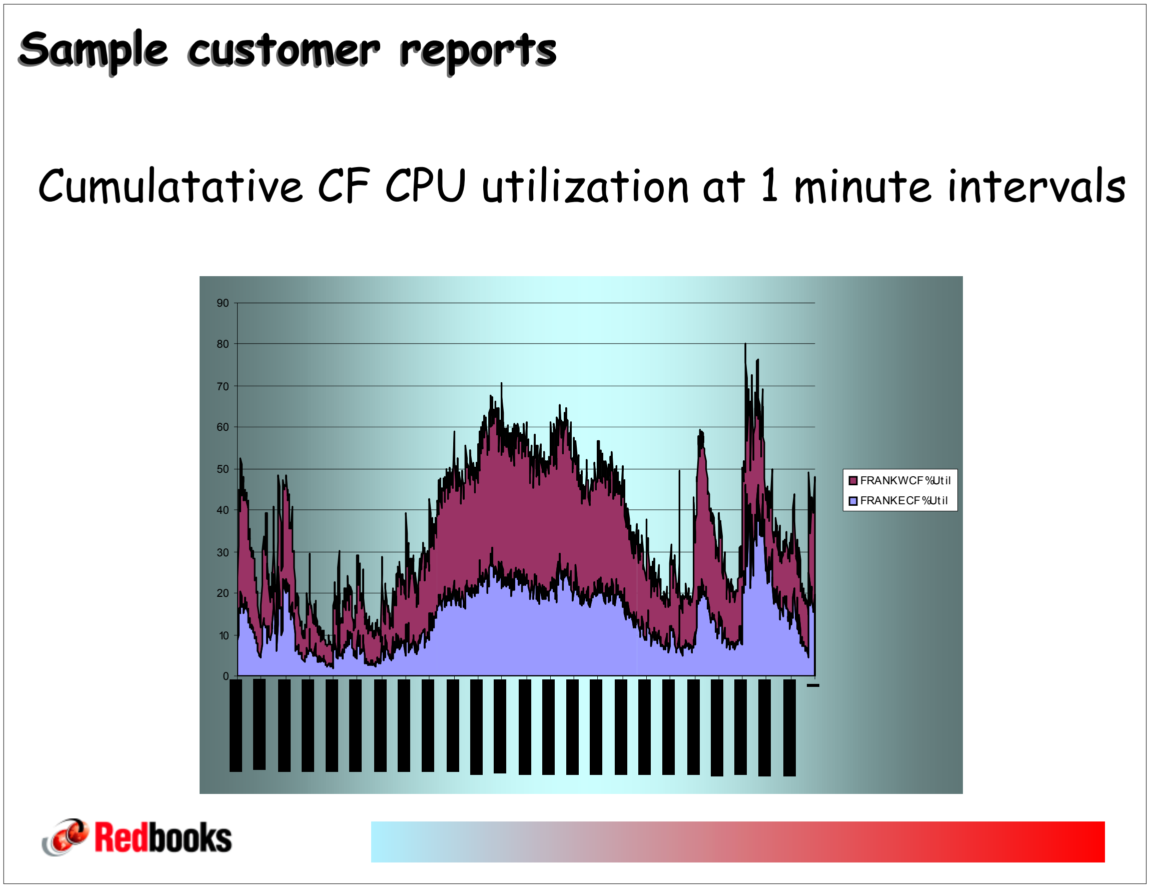
Redbooks



Sample customer reports

CF CPU utilization at 1 minute intervals

[redacted]



Sample customer reports

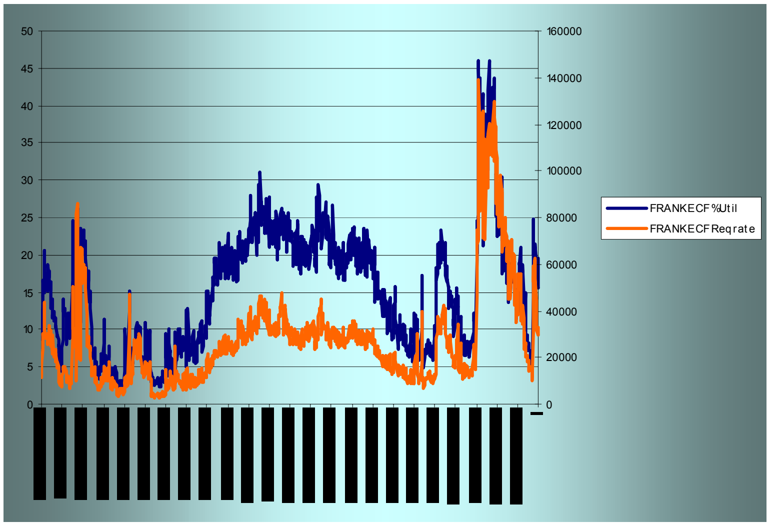
Cumulatative CF CPU utilization at 1 minute intervals

[redacted]

Sample customer reports

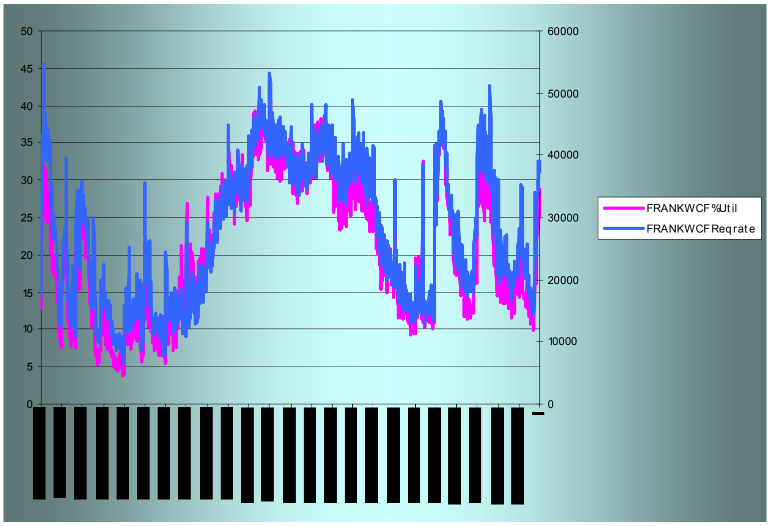
CF request rate plotted against CPU util

- Provides indication of CF CPU per request



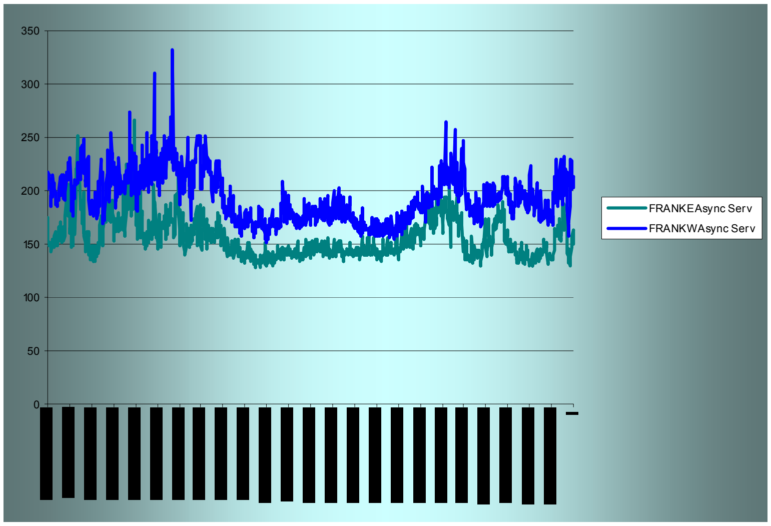
Sample customer reports

Same for other CF...



Sample customer reports

Tracking asynch response times



Shameless advertising

Recent Redbooks:

- JES2 Performance and Availability Considerations, REDP-3940
- Systems Programmers Guide to Workload Manager, SG24-6472
- Effective zSeries Performance Monitoring using RMF, SG24-6645
- GDPS - An Introduction to Concepts and Capabilities, SG24-6374
- z/OS Systems Programmers Guide to Sysplex Aggregation, REDP-3967 (in progress)
- Planned Outage Avoidance checklist, REDP-4069 (in progress)
- Introduction to Clustering Techniques, REDP-4072 (in progress)
- Parallel Sysplex Implementation - a customer's experience, due in 1Q2006

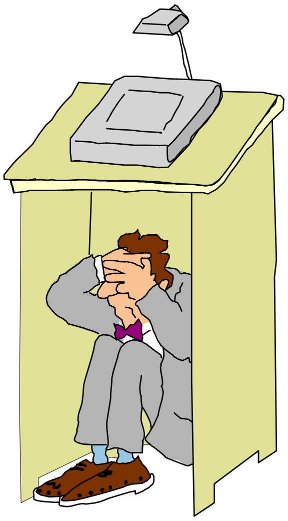
Shameless advertising

ITSO Residencies:

- Open to IBMers, BPs, customers
- ITSO covers all expenses (but not salary and not "utilization" (for IGS people!)
- Duration is typically 4-6 weeks
  - Long days, short weekends!
- Objective is to learn as much as possible about your assigned topic and then pass on your experiences to other users
- Unsurpassed education opportunity - what would you pay to attend a class where you have access to software, hardware, and the product developers?
- Residencies are announced on www.redbooks.ibm.com and you can subscribe based on areas of interest

Redbooks

Questions?



Redbooks

Thanks!!

Redbooks

®

MVS Automatic Restart Manager (ARM)

Redbooks

Automatic Restart Manager (ARM)

Agenda

- What is ARM?
- Who is using ARM?
- How do applications exploit ARM?
- How do you control ARM?
- ARM support for any application with JCL
- Appendix

ARM

What is Automatic Restart Manager?

- Sysplex wide recovery restart function for batch jobs and started tasks
- Part of z/OS
- Performs same system restarts
- Performs cross system restarts
- Policy based

What does ARM do?

ARM Manages

- When restarts should occur
- Where restarts should occur
- Grouping work that belongs together on the same system
- Dependencies that elements in a group have on each other in order to complete their initialization
- Restarting thresholds and intervals

Redbooks

What are the benefits of ARM?

- Integrated parallel sysplex solution
- Minimized outage time
  - swift detection
  - not message driven
  - no operator intervention required
  - awareness of the state of the sysplex
- Distributes work across systems in the sysplex
- Work that belongs together stays together
- Propogates symbolic substitution tables
- Restarts work in parallel

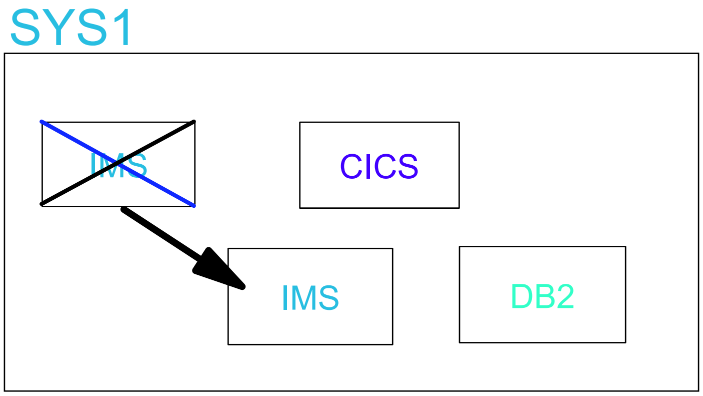
Redbooks

Same system restart

If a job or started tasks abends, ARM restarts the job or

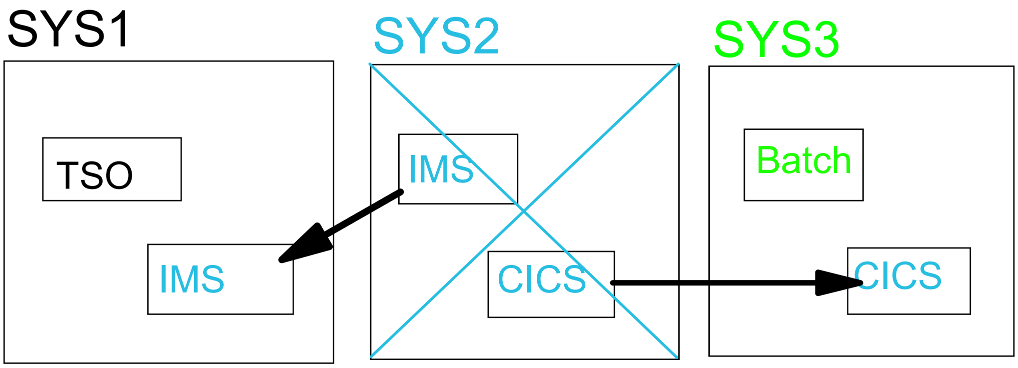
started task on the same system

- restart is based on
  - policy
  - installation and/or application exits



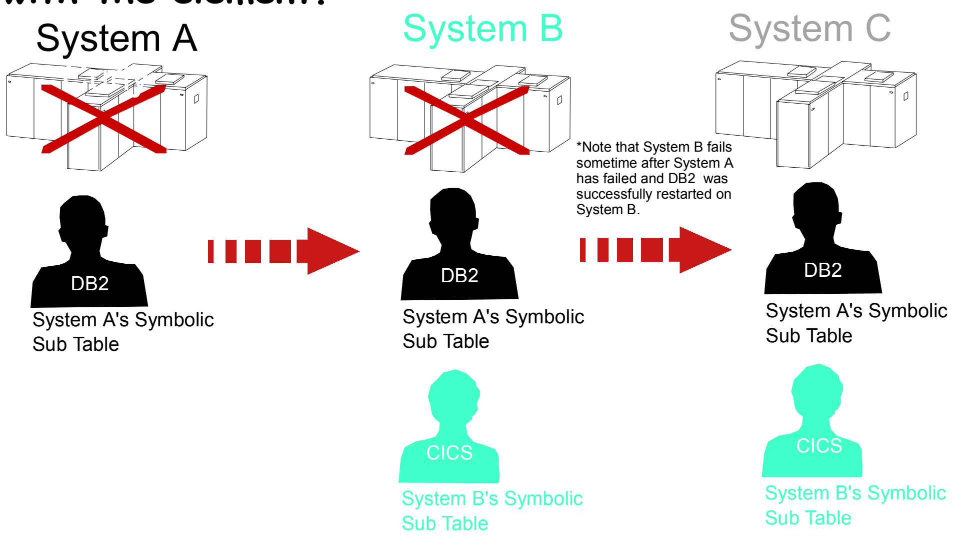
Cross system restart for system failures

If z/OS fails, ARM restarts jobs and started tasks on other systems in the sysplex



System symbols move with restarts

When ARM moves an element from one system to another, it moves the initial system's symbolic substitution table along with the element.



System A

System B

System C

*Note that System B fails sometime after System A has failed and DB2 was successfully restarted on System B.

DB2

System A's Symbolic Sub Table

DB2

System A's Symbolic Sub Table

DB2

System A's Symbolic Sub Table

CICS

System B's Symbolic Sub Table

CICS

System B's Symbolic Sub Table

Restart Groups

Restart groups contains elements that are:

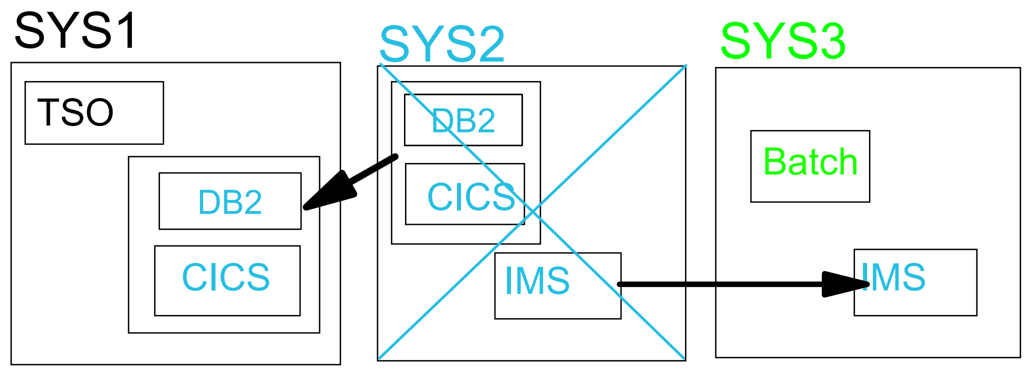
- related to one another
- need to be restarted on the same system
- dependent on each other for initialization and services

Example:

- CICS and the DB2 to which it attaches

System Failure with Restart Groups

Jobs and/or Started Tasks can be group to restart together on the same system



What are the configuration requirements?

- ARM couple data set needs to be formatted and made known to the system
- ARM policy must be started
  - z/OS provides a default policy
- Jobs and Started Tasks can NOT use Checkpoint/Restart or Step Restart
- Systems must be in the same sysplex
- Systems must be in the same JES2 MAS or JES3 Complex

Automatic Restart Manager (ARM)

Agenda

- What is ARM?
- Who is using ARM?
- How do applications exploit ARM?
- How do you control ARM?
- ARM support for any application with JCL
- Appendix

Who is using ARM?

- CICS Transaction Server
- DB2
- IMS
- IMS/CQS
- IMS Fast Database Recovery (FDBR)
- IMS XRF
- IRLM
- VTAM
- IBM Tivoli Netview for z/OS
- RRMS
- MQSERIES
- TCP/IP

Who is using ARM?

Automation Products integrated with ARM

- IBM Tivoli Systems Automation for z/OS
- Computer Associates - OPS/MVS

Automatic Restart Manager (ARM)

Agenda

- What is ARM?
- Who is using ARM?
- **How do applications exploit ARM?**
- How do you control ARM?
- ARM support for any application with JCL
- Appendix

How do applications exploit ARM?

Batch jobs and started tasks must:

- Register as an element to be restarted
- May indicate dependency on other elements in the restart group to complete initialization
- Indicate readiness to accept work
- Listen for ENFs
- Deregister on completion

How do applications exploit ARM?

Marcos used to provide those services

- Changing states

```
IXCARM REQUEST=REGISTER
              =READY
              =DEREGISTER
```

- Waiting for other elements

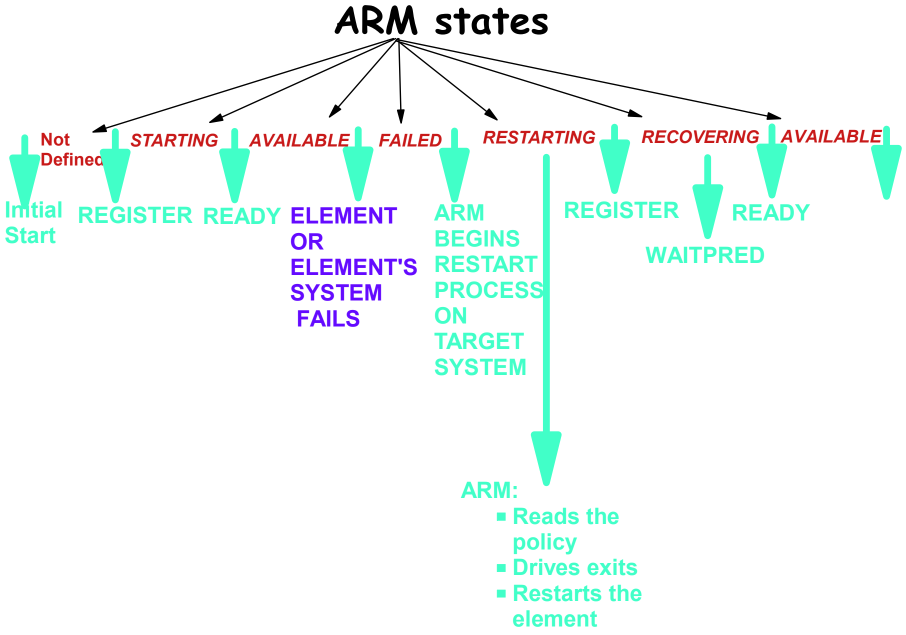
```
IXCARM REQUEST= WAITPRED
              = READY (implied WAITPRED)
```

How do applications exploit ARM?

Application Exits

- Application can provide Event Exit
  - specified when application registers with ARM
  - called on system where failed element is to be restarted
  - gets control AFTER installation exits have completed
  - performs whatever preparation is needed to restart the element
  - may prohibit the restart

Life of an ARM element



ARM states

Not Defined

STARTING

AVAILABLE

FAILED

RESTARTING

RECOVERING

AVAILABLE

Initial Start

REGISTER

READY

ELEMENT OR ELEMENT'S SYSTEM FAILS

ARM BEGINS RESTART PROCESS ON TARGET SYSTEM

REGISTER

WAITPRED

READY

ARM:

- Reads the policy
- Drives exits
- Restarts the element

Automatic Restart Manager (ARM)

Agenda

- What is ARM?
- Who is using ARM?
- How do applications exploit ARM?
- How do you control ARM?
- ARM support for any application with JCL
- Appendix

What can the installation control?

Using the policy and/or exits, the installation can control:

- Whether restarts will occur for none, some, or all elements
- Restart methods
- Groupings/dependencies
- Candidate systems for restarting elements
- Restarting intervals and thresholds

Element related statements in the policy

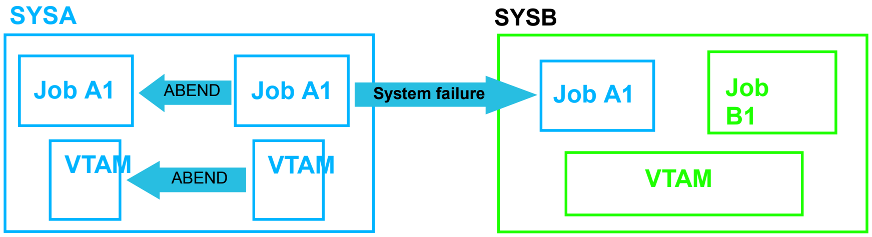
For elements, you define in the policy:

- When to restart
- How to restart
- Number of restarts in an interval
- Maximum wait times on restart
- Time to Re-register
- Time to become ready

TERMTYPE statement

Defines when the element should be restarted:

- ELEMTERM - restart when just the element fails and the system is still active
- SYSTERM - restart when the entire system fails



RESTART_METHOD policy statement

Three choices:

- PERSISTENT
  - same JCL or command as when originally started
- JOB
  - specifies data set that contains JCL to restart the element
- STC
  - specifies command text to restart the element

Choices may be different for ELEMTERM (single element failure) and ALLTERM (complete system failure)

RESTART_ATTEMPTS

Maximum number of restart attempts within an interval

Example: If there are more than 2 restart attempts in 3 minutes, do not restart the element.

Default is 3 restart attempts within 5 minuts

- Zero indicates the element should not be restarted at all

RESTART_TIMEOUT

Greater than the maximum amount of time that an element is expected to take to reregister after it fails.

Example: If the element does not reregister with ARM within 5 minutes, ARM will deregister the element.

READY_TIMEOUT

Greater than the maximum amount of time that an element would take to complete it's initialization, which is when it would issue the IXCARM macro indicating ready to accept work.

Example: If the element does not READY with ARM within 2 minutes, ARM will assume that the element is ready.

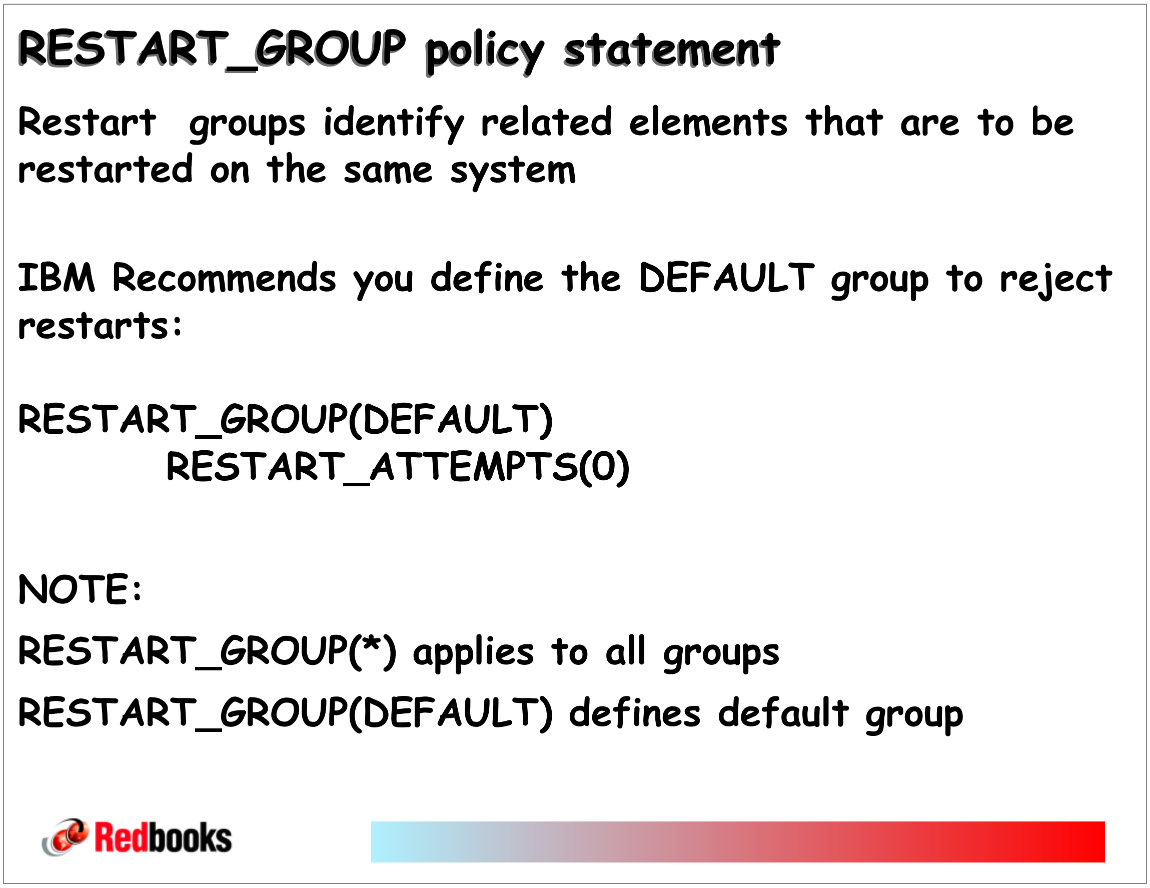
Default values for element keywords

- TERMTYPE
  - ALLTERM
- RESTART_METHOD
  - persistent JCL/Start text
- RESTART_ATTEMPTS
  - 3 times within 5 minutes
- RESTART_TIMEOUT
  - 5 minutes
- READY_TIMEOUT
  - 5 minutes

Restart Group keywords

For restart groups, the policy defines:

- A name to identify the group
- Where restarts should occur
- How elements relate to each other
- Minimum CSA/ECSA required on systems for restarts
- Pacing intervals between restarts of elements



RESTART_GROUP policy statement

Restart groups identify related elements that are to be restarted on the same system

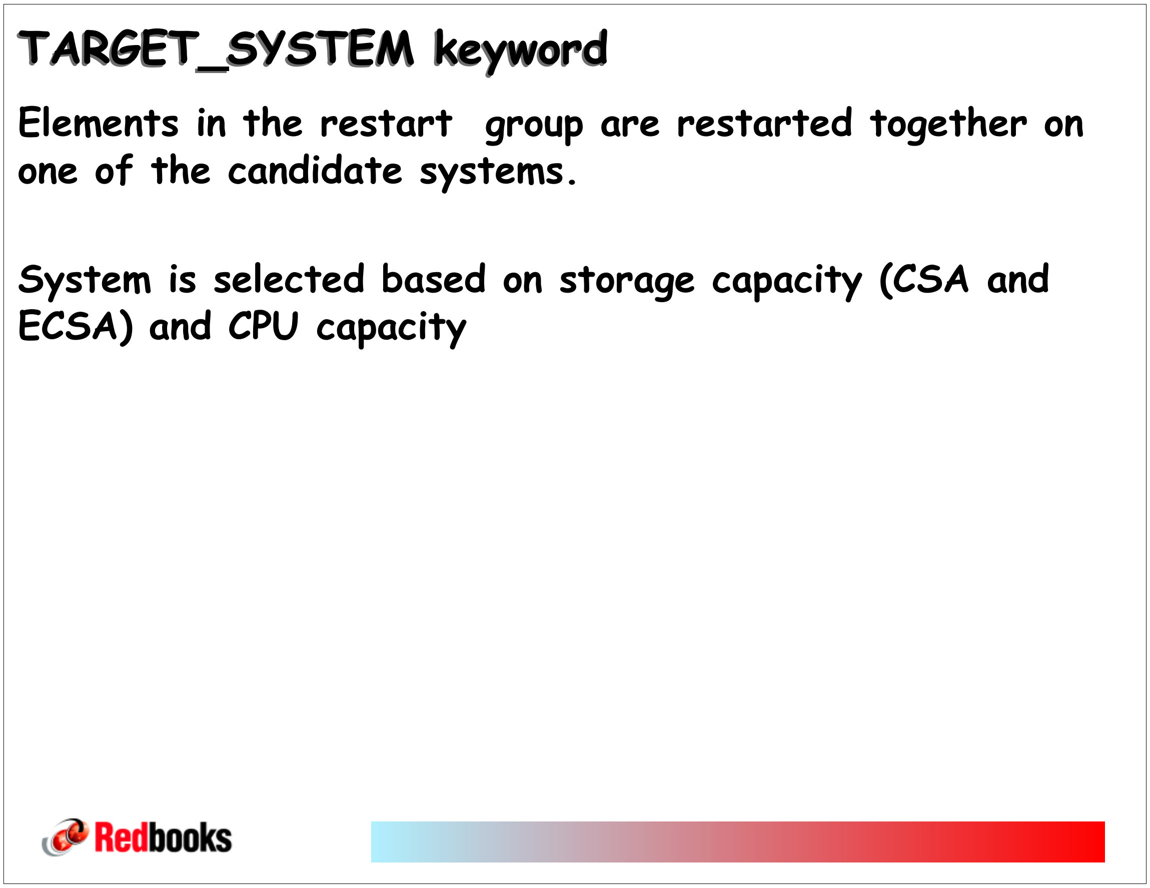
IBM Recommends you define the DEFAULT group to reject restarts:

```
RESTART_GROUP(DEFAULT)
        RESTART_ATTEMPTS(0)
```

NOTE:

RESTART_GROUP(*) applies to all groups

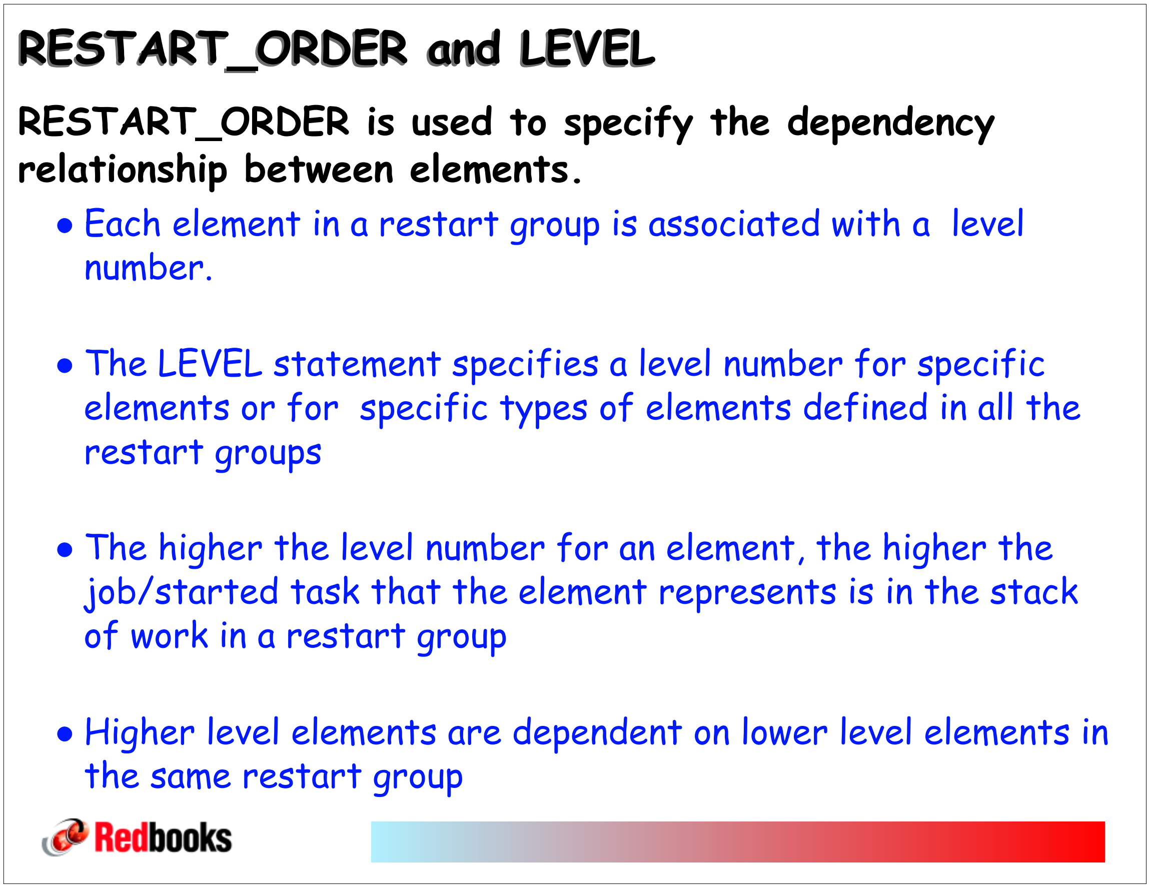
RESTART_GROUP(DEFAULT) defines default group



TARGET_SYSTEM keyword

Elements in the restart group are restarted together on one of the candidate systems.

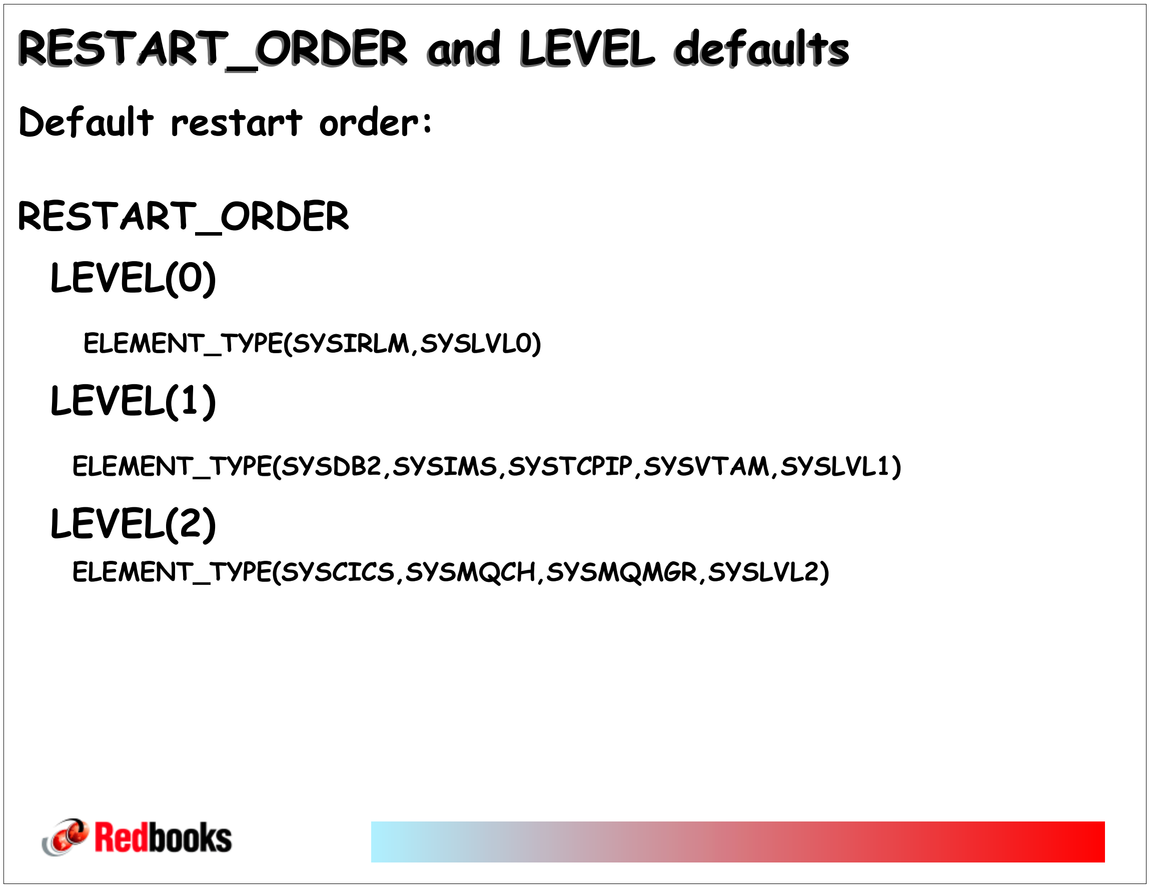
System is selected based on storage capacity (CSA and ECSA) and CPU capacity



RESTART_ORDER and LEVEL

RESTART_ORDER is used to specify the dependency relationship between elements.

- Each element in a restart group is associated with a level number.
- The LEVEL statement specifies a level number for specific elements or for specific types of elements defined in all the restart groups
- The higher the level number for an element, the higher the job/started task that the element represents is in the stack of work in a restart group
- Higher level elements are dependent on lower level elements in the same restart group

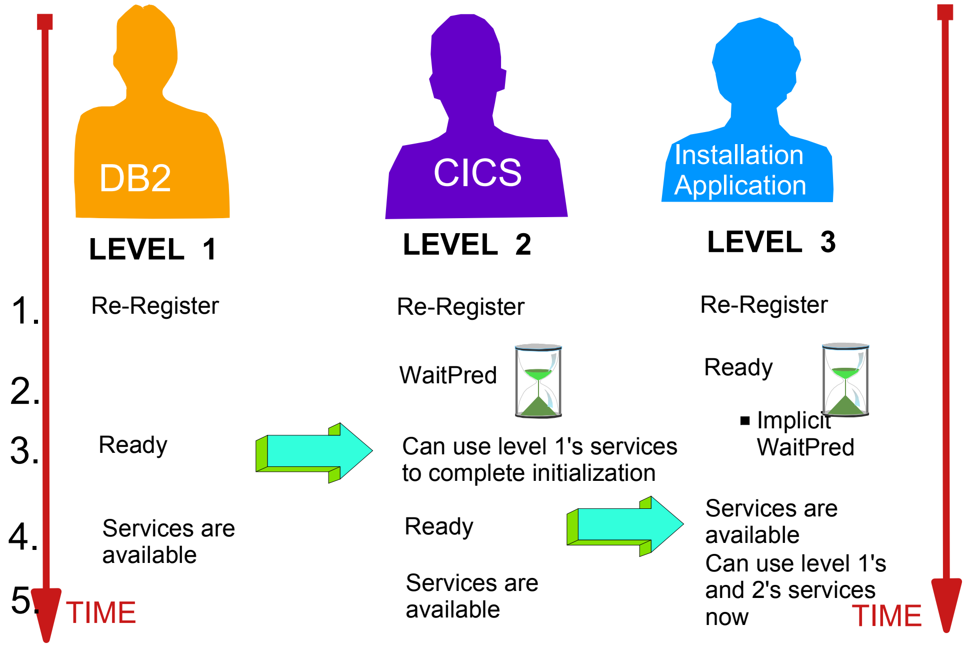


RESTART_ORDER and LEVEL defaults

Default restart order:

```
RESTART_ORDER
  LEVEL(0)
    ELEMENT_TYPE(SYSIRLM,SYSLVL0)
  LEVEL(1)
   ELEMENT_TYPE(SYSDB2,SYSIMS,SYSTCPIP,SYSVTAM,SYSLVL1)
  LEVEL(2)
   ELEMENT_TYPE(SYSCICS,SYSMQCH,SYSMQMGR,SYSLVL2)
```

Handling dependencies within restart group



LEVEL 1

1. Re-Register

2.

3. Ready

4. Services are available

5.

TIME

LEVEL 2

Re-Register

WaitPred

Can use level 1's services to complete initialization

Ready

Services are available

LEVEL 3

Re-Register

Ready

- Implicit WaitPred

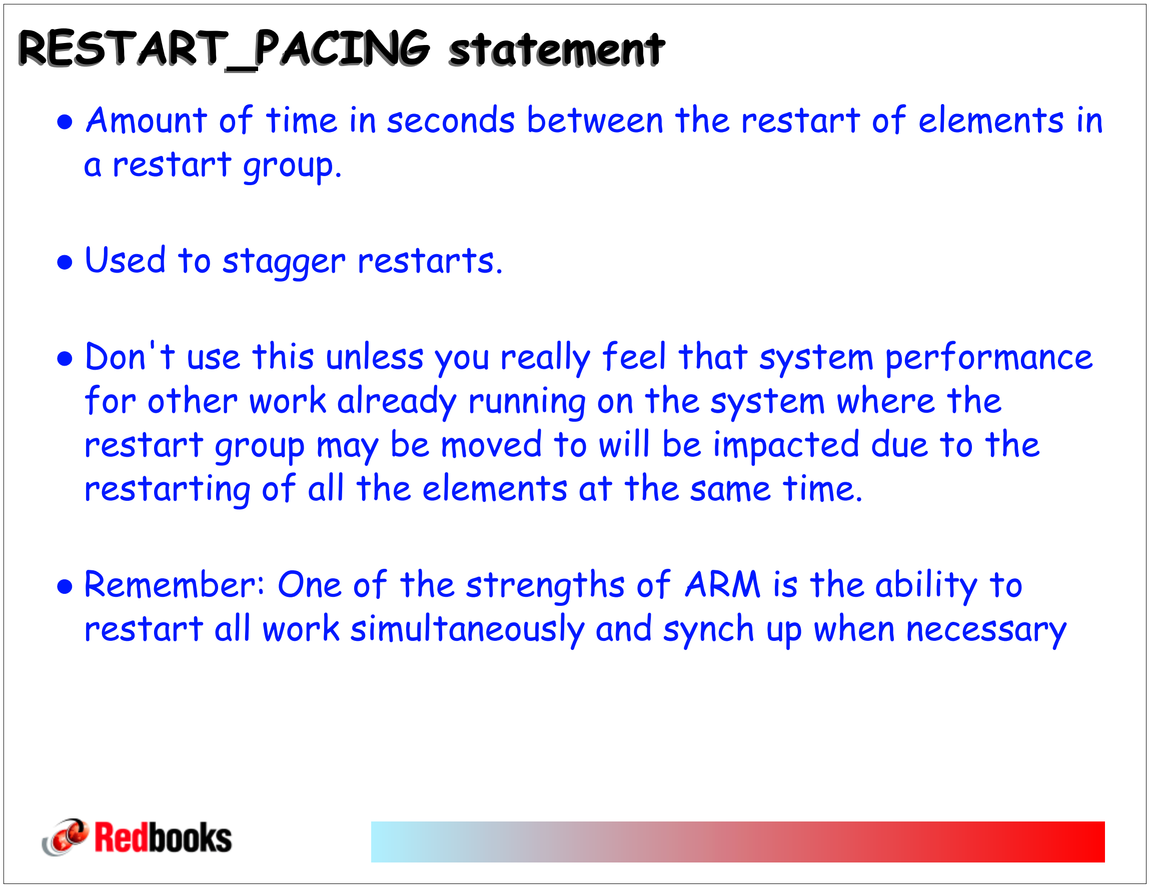
Services are available

Can use level 1's and 2's services now

TIME

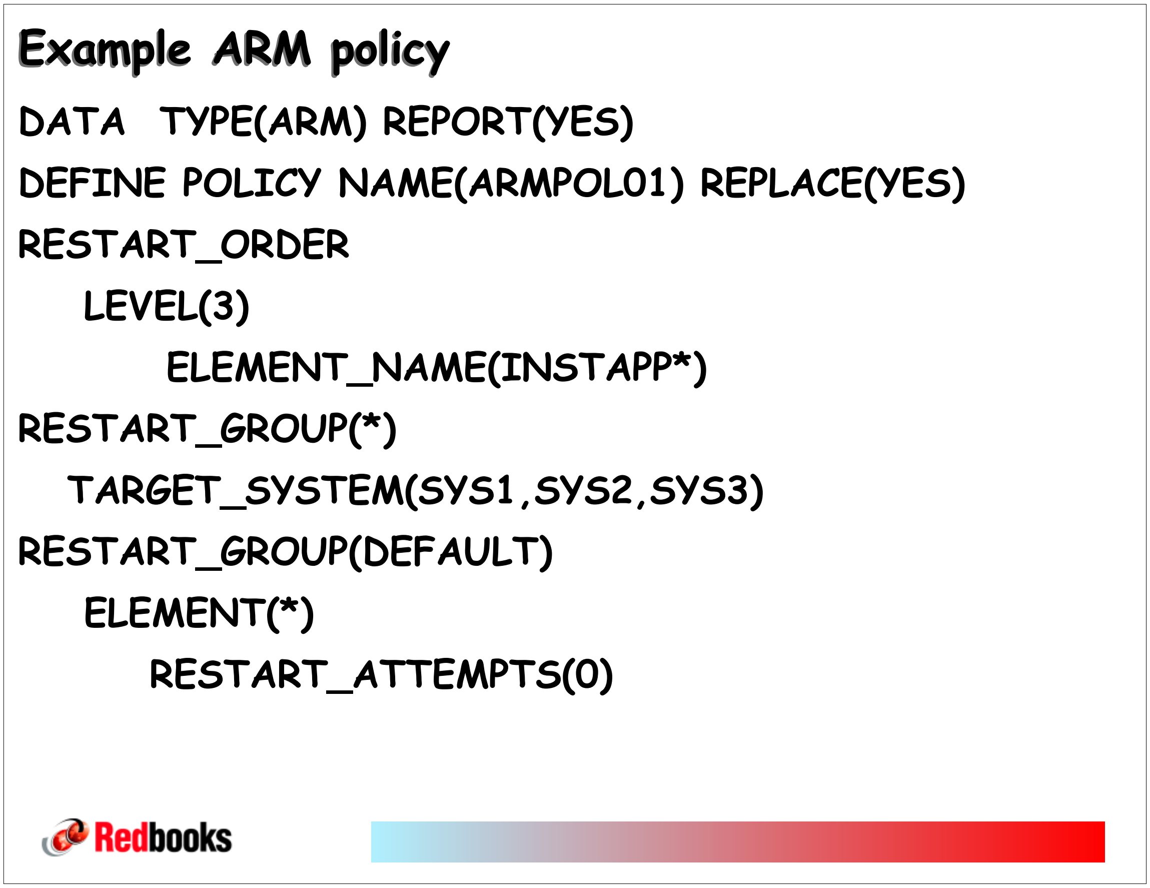
FREE_CSA statement

Minimum amount of CSA and ECSA in KB that must be available on a system for the restart group to be moved to that system



RESTART_PACING statement

- Amount of time in seconds between the restart of elements in a restart group.
- Used to stagger restarts.
- Don't use this unless you really feel that system performance for other work already running on the system where the restart group may be moved to will be impacted due to the restarting of all the elements at the same time.
- Remember: One of the strengths of ARM is the ability to restart all work simultaneously and synch up when necessary



Example ARM policy

```
DATA  TYPE(ARM) REPORT(YES)
DEFINE POLICY NAME(ARMPOL01) REPLACE(YES)
RESTART_ORDER
    LEVEL(3)
         ELEMENT_NAME(INSTAPP*)
RESTART_GROUP(*)
   TARGET_SYSTEM(SYS1,SYS2,SYS3)
RESTART_GROUP(DEFAULT)
    ELEMENT(*)
        RESTART_ATTEMPTS(0)
```

Example ARM policy ...

```
RESTART_GROUP(GROUP1)
   TARGET_SYSTEM(SYS1,SYS2)
    FREE_CSA(100,500)
    ELEMENT(DSNDB1GDB81)
    ELEMENT(SYSCICS_CIC3A8)
    ELEMENT(INSTAPP1)
```

Example ARM policy ...

```
RESTART_GROUP(GROUP2)
   TARGET_SYSTEM(SYS1,SYS3)
    ELEMENT(IRLMGRP2IRL8002)
    ELEMENT(IMS8)
     RESTART_ATTEMPTS(3,1200)
     RESTART_TIMEOUT(600)
     READY_TIMEOUT(600)
     RESTART_METHOD(ELEMTERM,PERSIST)
     RESTART_METHOD(SYSTERM,STC,
                              'S IMSR6.IMS8,IMAGE=8,'

                              'APPL=IMSPETJ8,IRLMNM=IRL8')
    ELEMENT(SYSCICS*)
```

Installation Exits

- Workload Restart Exit(s)
  - Called once on each system where failing workload is to be restarted
  - Prepares to receive additional workload from failing system
  - IXC_WORK_RESTART

- Element Restart Exit(s)
  - Called once for each failed element
  - Called on the system where element is to be restarted
  - Can modify or cancel ARM initiated restart of failed element
  - Helps coordinate ARM with other automation routines
  - IXC_ELEM_RESTART

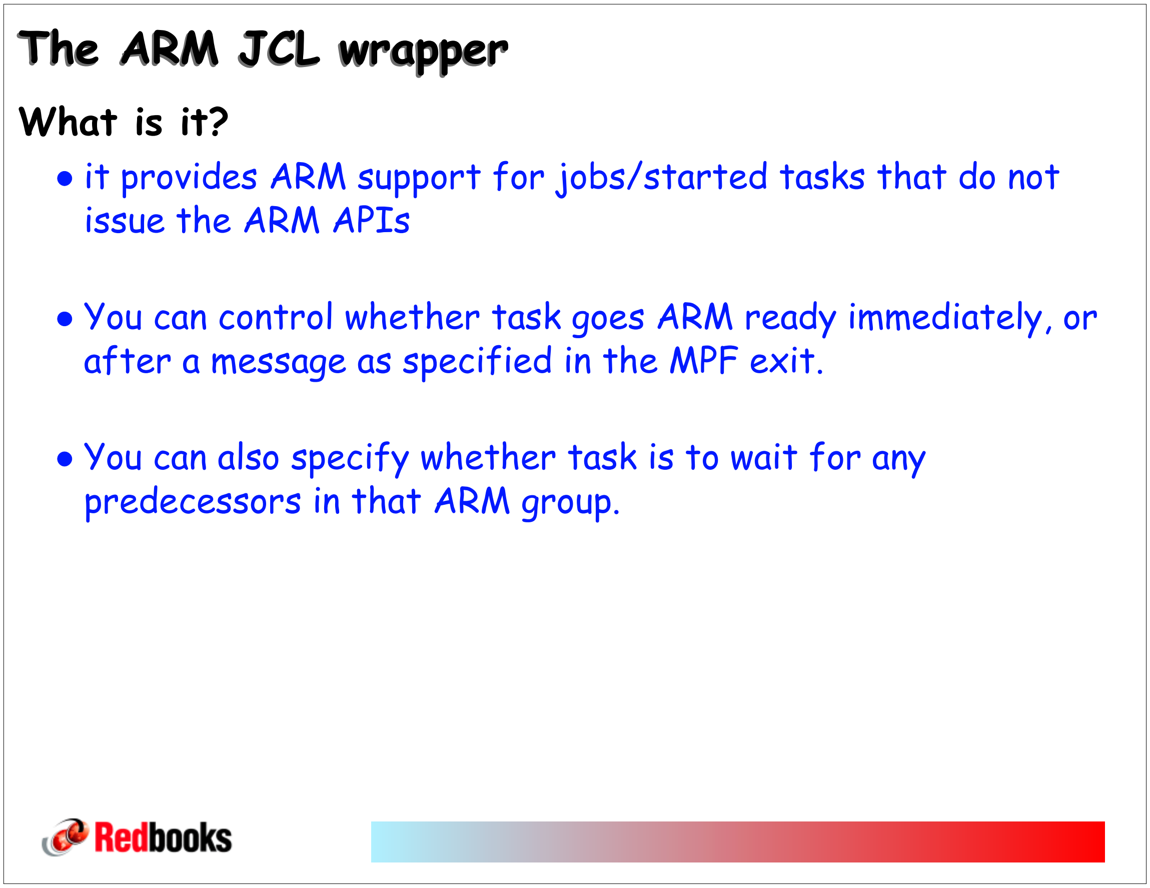
Redbooks

Automatic Restart Manager (ARM)

Agenda

- What is ARM?
- Who is using ARM?
- How do applications exploit ARM?
- How do you control ARM?
- **ARM support for any application with JCL**
- Appendix

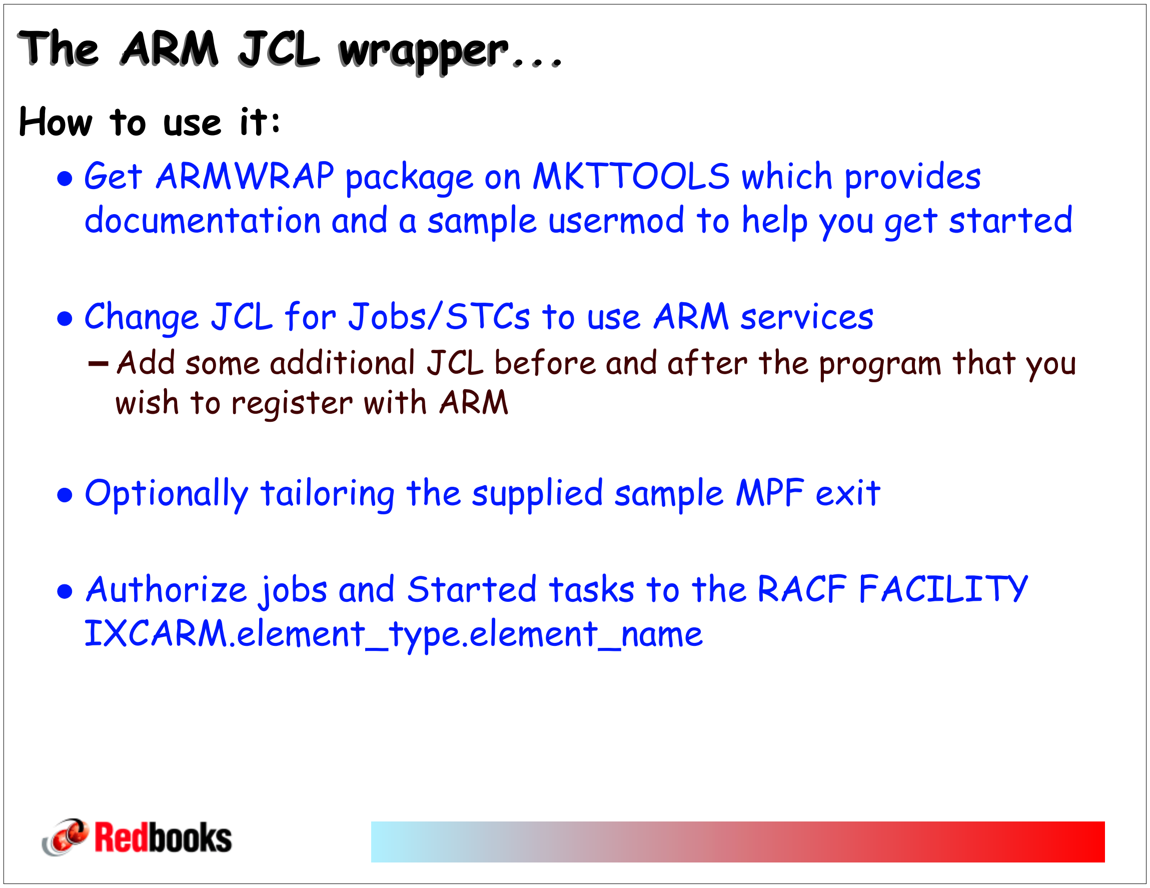
Redbooks



The ARM JCL wrapper

What is it?

- it provides ARM support for jobs/started tasks that do not issue the ARM APIs
- You can control whether task goes ARM ready immediately, or after a message as specified in the MPF exit.
- You can also specify whether task is to wait for any predecessors in that ARM group.



The ARM JCL wrapper...

How to use it:

- Get ARMWRAP package on MKTTOOLS which provides documentation and a sample usermod to help you get started
- Change JCL for Jobs/STCs to use ARM services
  - Add some additional JCL before and after the program that you wish to register with ARM
- Optionally tailoring the supplied sample MPF exit
- Authorize jobs and Started tasks to the RACF FACILITY IXCARM.element_type.element_name

The ARM JCL wrapper - JCL

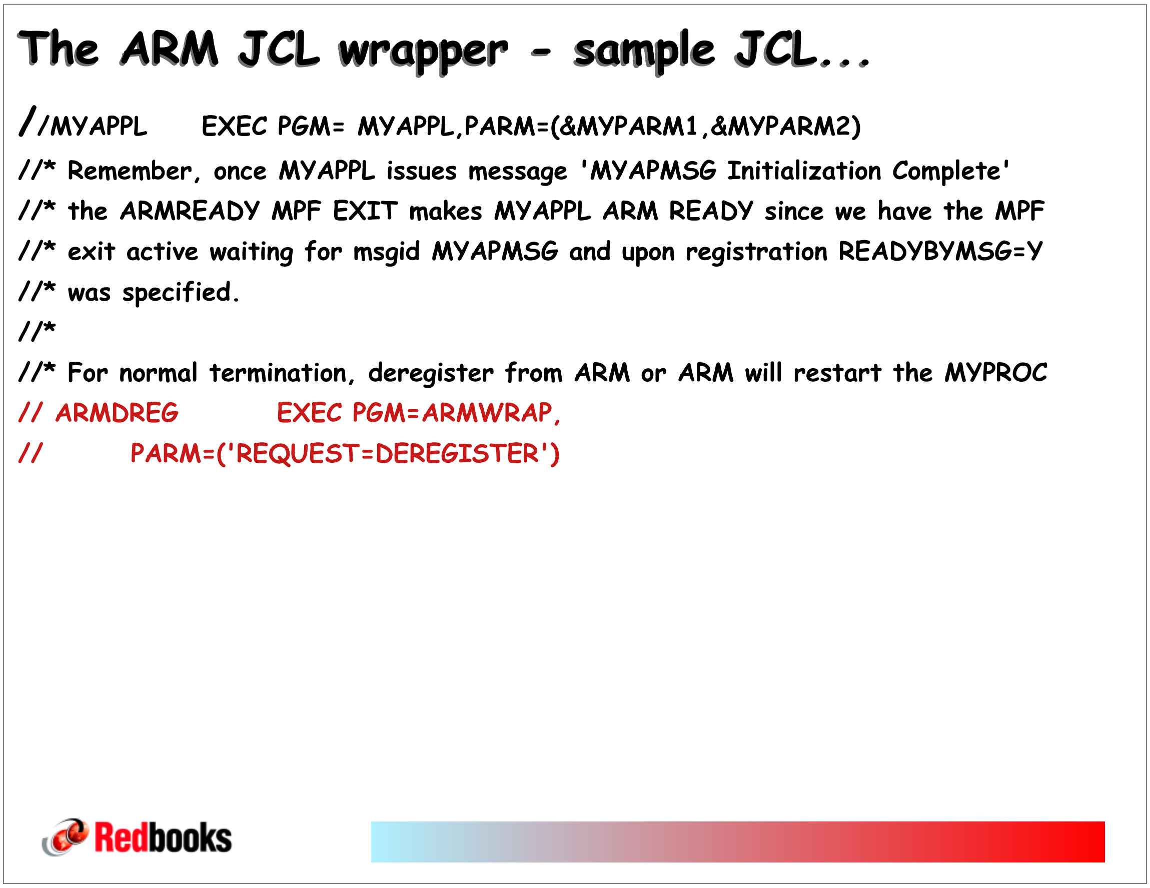
```
//nnnnnnnn   EXEC PGM=ARMWRAP,PARM=(

        REQUEST = REGISTER
                TERMTYPE =  ALLTERM | ELEMTERM
                ELEMENT = n16
                ELEMENTTYPE = n8
                READYBYMSG =   Y|N
        REQUEST =  WAITPRED
                     DEREGISTER)
```

Redbooks

The ARM JCL wrapper - sample JCL

```
//MYPROC PROC ...
//* Register element 'MYAPPLELEMENT'  element type 'APPLTYPE' with ARM
//ARMREG   EXEC PGM=ARMWRAP,
//         PARM=('REQUEST=REGISTER,READYBYMSG=Y,',
//                'TERMTYPE=ALLTERM,ELEMENT=MYAPPLELEMENT,',
//                'ELEMTYPE=APPLTYPE')
//*  Requires access to FACILITY IXCARM.APPLTYPE.MYAPPLELEMENT
//*  On a restart, wait for any predecessors in my ARM group
//ARMPRED  EXEC PGM=ARMWRAP,
//        PARM=('REQUEST=WAITPRED')
```

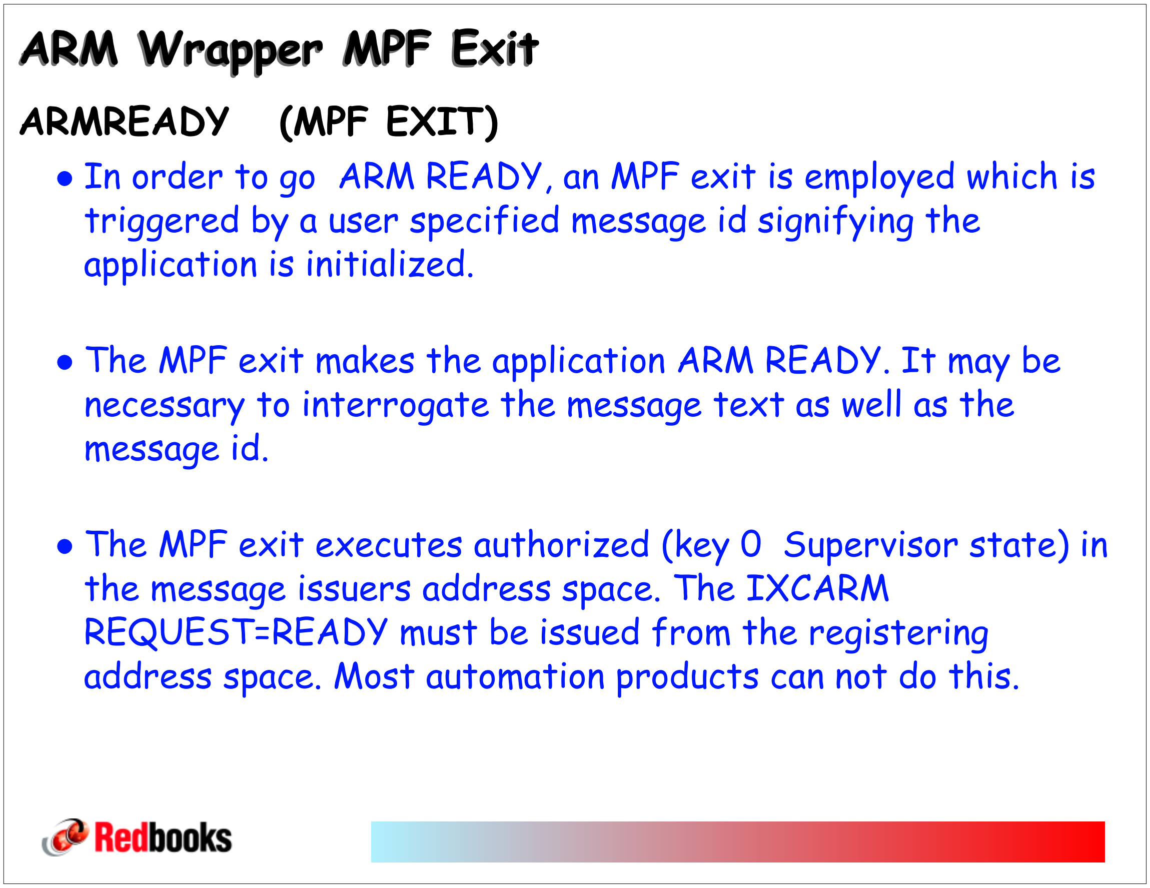
Redbooks



The ARM JCL wrapper - sample JCL...

```
//MYAPPL     EXEC PGM= MYAPPL,PARM=(&MYPARM1,&MYPARM2)
//* Remember, once MYAPPL issues message 'MYAPMSG Initialization Complete'
//* the ARMREADY MPF EXIT makes MYAPPL ARM READY since we have the MPF
//* exit active waiting for msgid MYAPMSG and upon registration READYBYMSG=Y
//* was specified.
//*
//* For normal termination, deregister from ARM or ARM will restart the MYPROC
// ARMDREG        EXEC PGM=ARMWRAP,
//        PARM=('REQUEST=DEREGISTER')
```

Redbooks

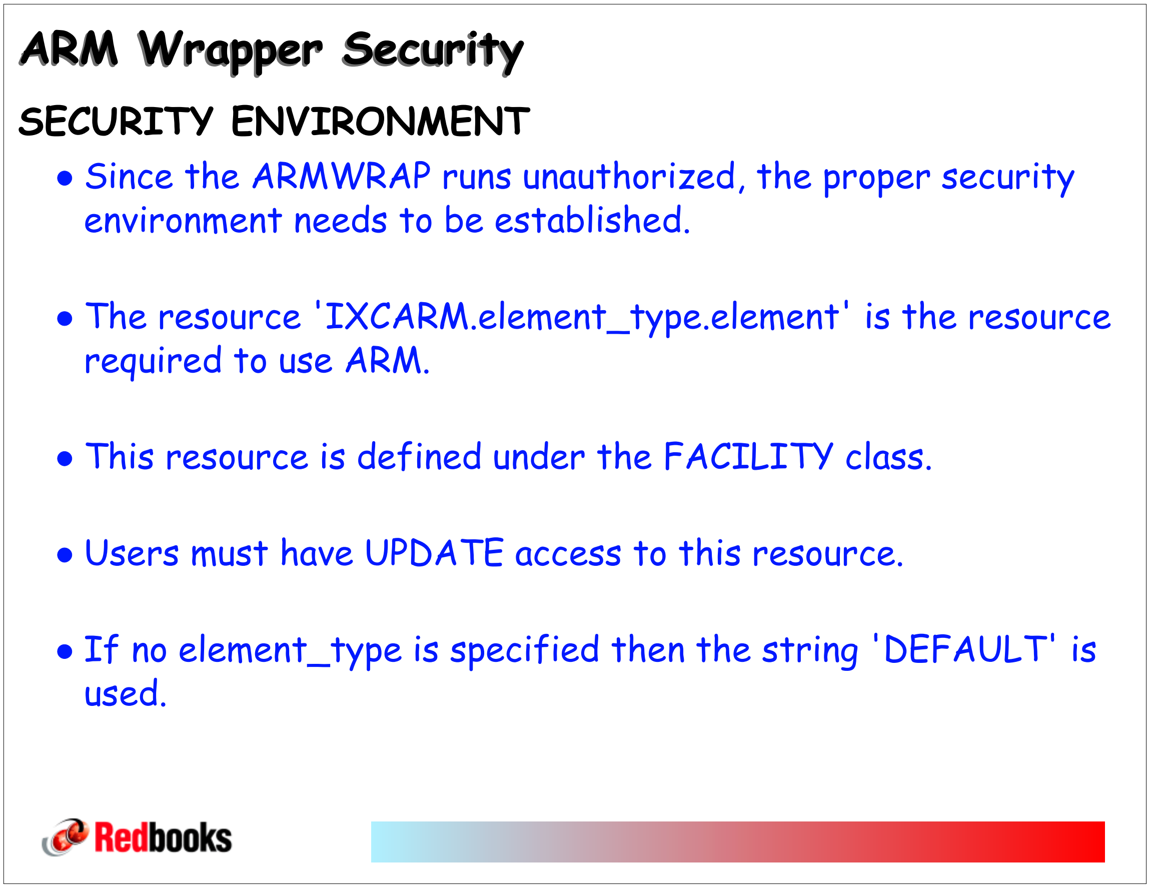


ARM Wrapper MPF Exit

ARMREADY   (MPF EXIT)

- In order to go ARM READY, an MPF exit is employed which is triggered by a user specified message id signifying the application is initialized.
- The MPF exit makes the application ARM READY. It may be necessary to interrogate the message text as well as the message id.
- The MPF exit executes authorized (key 0 Supervisor state) in the message issuers address space. The IXCARM REQUEST=READY must be issued from the registering address space. Most automation products can not do this.

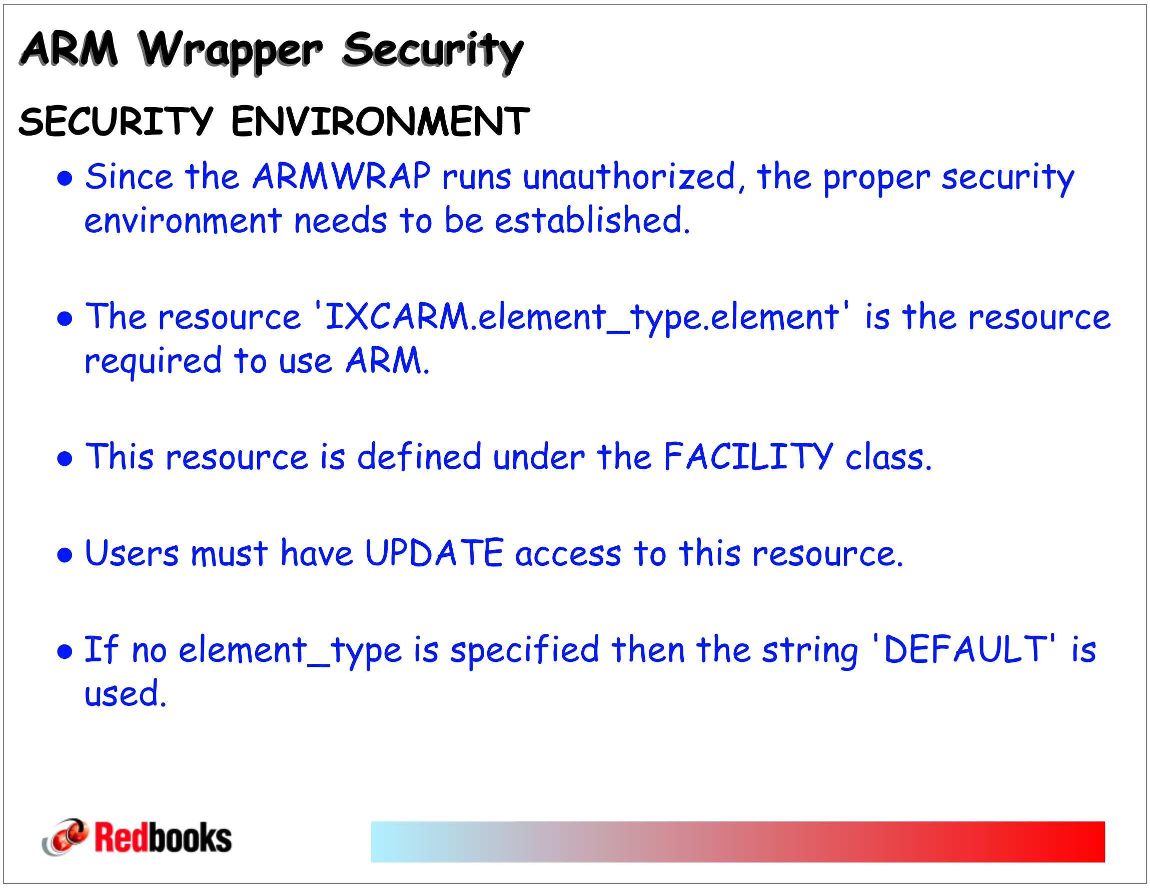
Redbooks



ARM Wrapper Security

SECURITY ENVIRONMENT

- Since the ARMWRAP runs unauthorized, the proper security environment needs to be established.
- The resource 'IXCARM.element_type.element' is the resource required to use ARM.
- This resource is defined under the FACILITY class.
- Users must have UPDATE access to this resource.
- If no element_type is specified then the string 'DEFAULT' is used.



ARM Wrapper Security

SECURITY ENVIRONMENT

- Since the ARMWRAP runs unauthorized, the proper security environment needs to be established.
- The resource 'IXCARM.element_type.element' is the resource required to use ARM.
- This resource is defined under the FACILITY class.
- Users must have UPDATE access to this resource.
- If no element_type is specified then the string 'DEFAULT' is used.

Automatic Restart Manager (ARM)

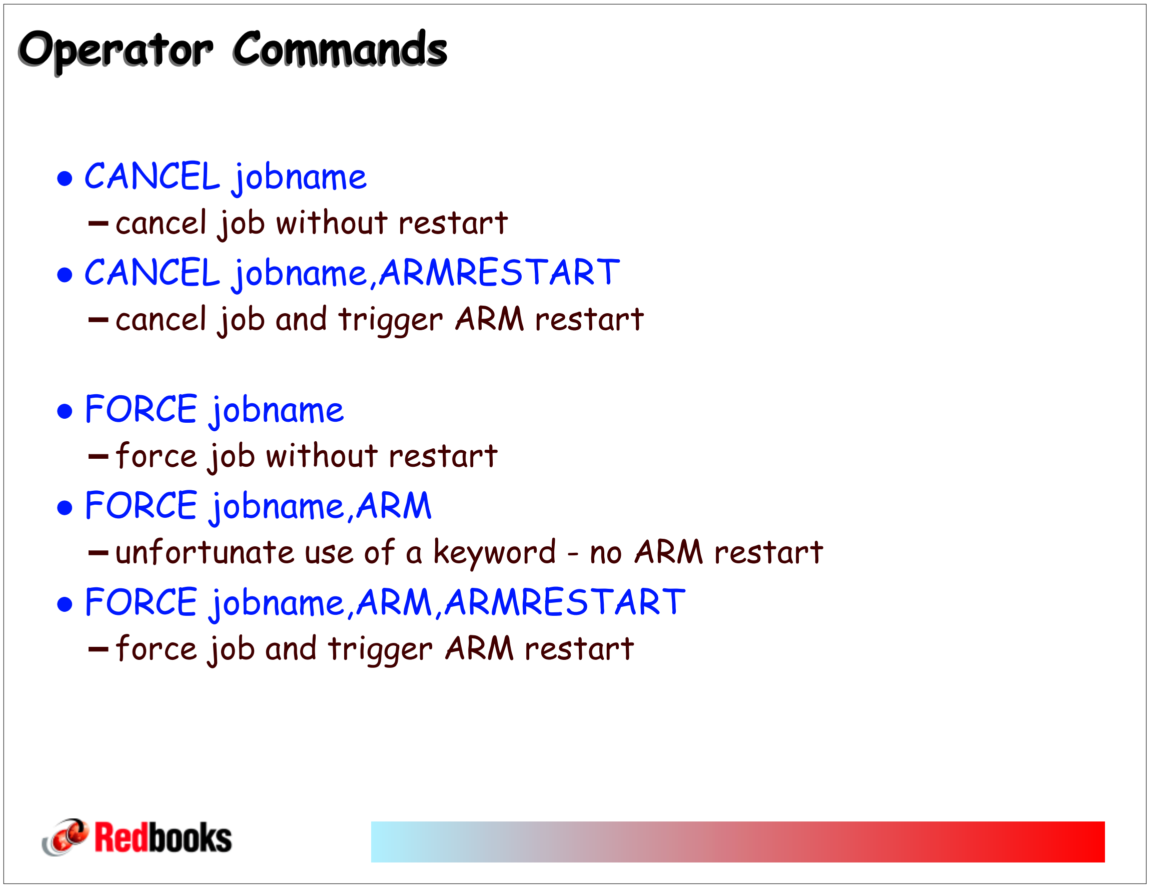
Agenda

- What is ARM?
- Who is using ARM?
- How do applications exploit ARM?
- How do you control ARM?
- ARM support for any application with JCL
- Appendix

Automatic Restart Manager (ARM)

Appendix

- Operator Commands
- Security
- Formatting an ARM couple data set
- Creating an ARM policy
- Publications

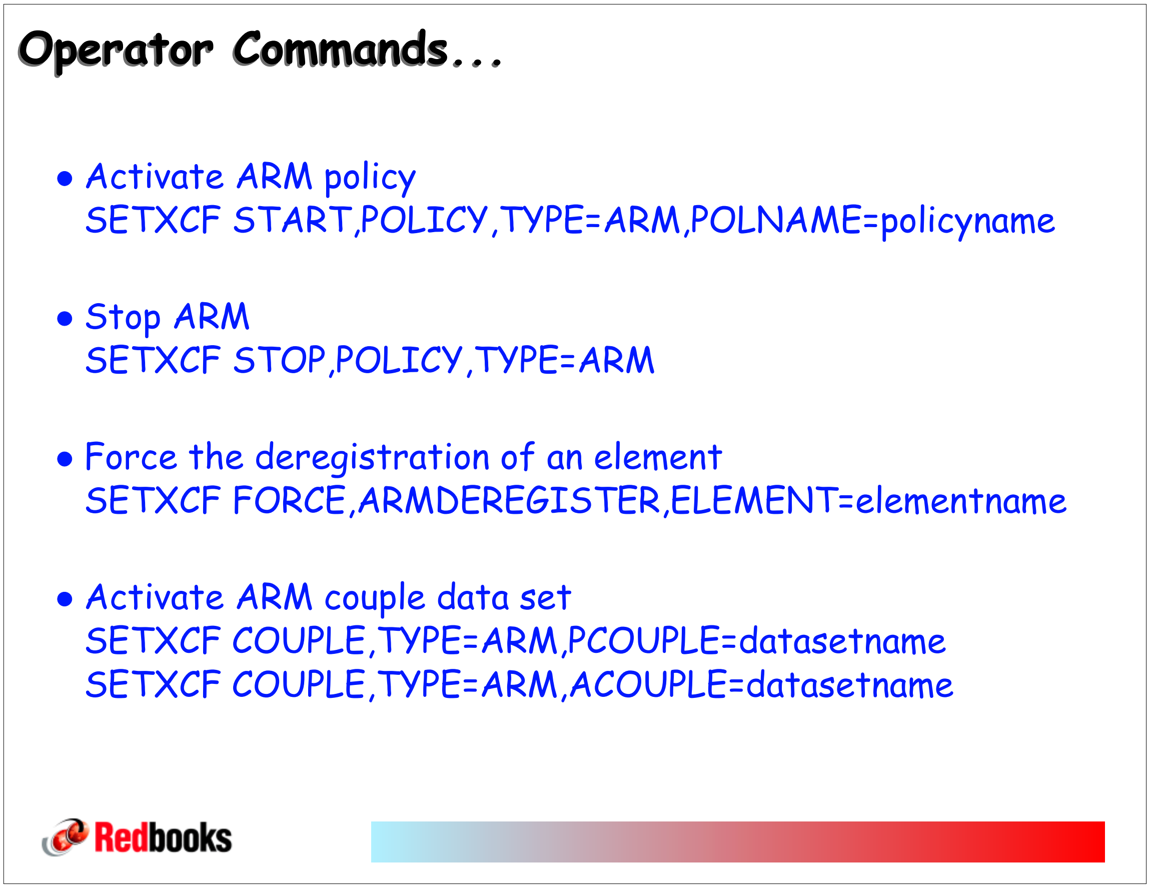


Operator Commands

- CANCEL jobname
  - cancel job without restart
- CANCEL jobname,ARMRESTART
  - cancel job and trigger ARM restart

- FORCE jobname
  - force job without restart
- FORCE jobname,ARM
  - unfortunate use of a keyword - no ARM restart
- FORCE jobname,ARM,ARMRESTART
  - force job and trigger ARM restart

Redbooks



Operator Commands...

- Activate ARM policy
  SETXCF START,POLICY,TYPE=ARM,POLNAME=policyname

- Stop ARM
  SETXCF STOP,POLICY,TYPE=ARM

- Force the deregistration of an element
  SETXCF FORCE,ARMDEREGISTER,ELEMENT=elementname

- Activate ARM couple data set
  SETXCF COUPLE,TYPE=ARM,PCOUPLE=datasetname
  SETXCF COUPLE,TYPE=ARM,ACOUPLE=datasetname

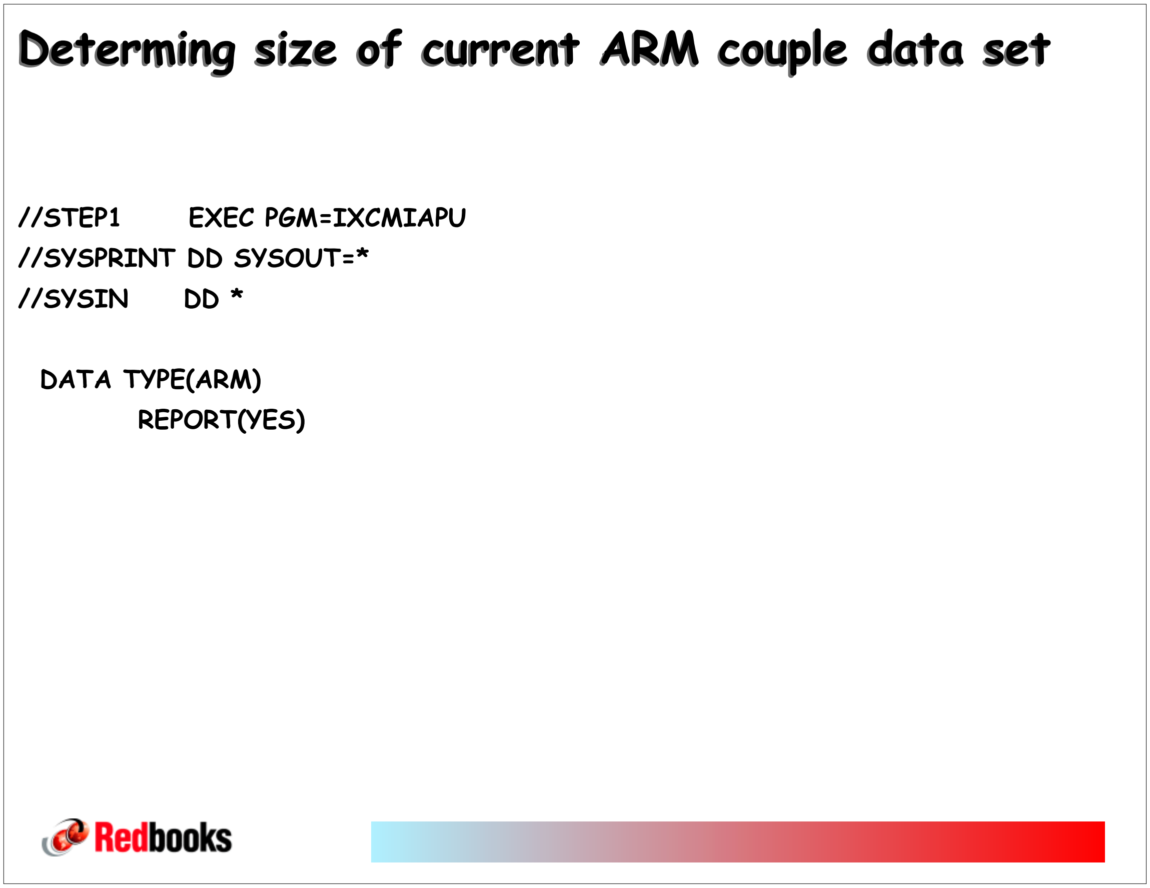
Redbooks

Operator Commands...

- Display ARM status
  ```
  D XCF,ARMSTATUS
  ```
- Display ARM restart group
  ```
  D XCF,ARMSTATUS,RESTARTGRP=groupname
  ```
- Display ARM element
  ```
  D XCF,ARMSTATUS,ELEMENT=elementname
  ```
- Display everything
  ```
  D XCF,ARMSTATUS,DETAIL
  ```

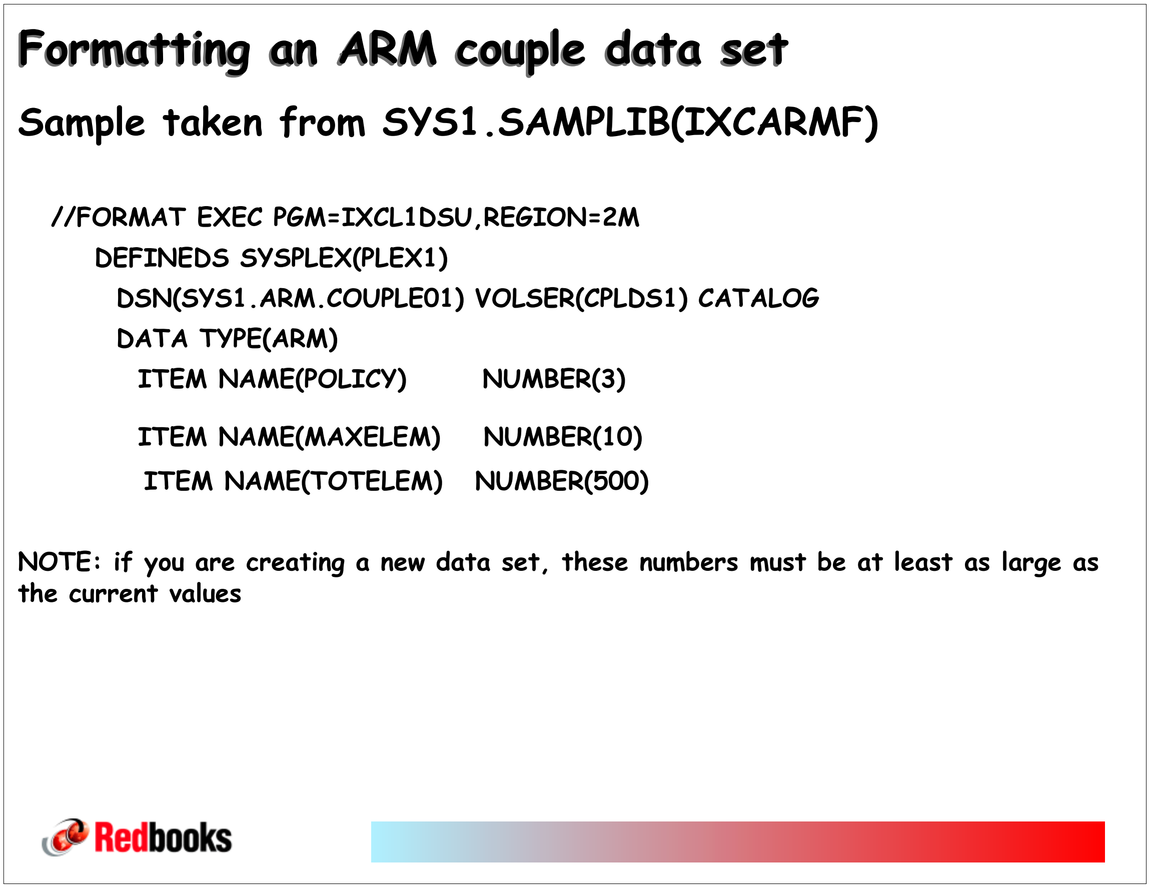
ARM Exploiters

Exploiter	Element Name	Element Type
CICS	SYSCICSapplid	SYSCICS
CPSM	SYSCICSapplid	SYSCICS
DB2 (datasharing)	datasharinggrpname memname	SYSDB2
DB2 (non-datasharing)	DB2$subsystemname	SYSDB2
IMS	IMSID	SYSIMS
IMS CQS	IMSID	SYSIMS
IRLM (local mode)	subsystemname IRLMID	SYSIRLM
IRLM (global mode)	datasharinggrpname subsystemname IRLMID	SYSIRLM
RRMS	SYS_RRS_sysname	NA
Netview	NETVIEW@domainid	NA
VTAM	NET@cpname	SYSVTAM
Websphere		



Determing size of current ARM couple data set

```
//STEP1       EXEC PGM=IXCMIAPU
//SYSPRINT DD SYSOUT=*
//SYSIN      DD *

  DATA TYPE(ARM)
          REPORT(YES)
```



Formatting an ARM couple data set

Sample taken from SYS1.SAMPLIB(IXCARMF)

```
   //FORMAT EXEC PGM=IXCL1DSU,REGION=2M
       DEFINEDS SYSPLEX(PLEX1)
         DSN(SYS1.ARM.COUPLE01) VOLSER(CPLDS1) CATALOG
         DATA TYPE(ARM)
           ITEM NAME(POLICY)       NUMBER(3)

           ITEM NAME(MAXELEM)     NUMBER(10)
            ITEM NAME(TOTELEM)   NUMBER(500)
```

NOTE: if you are creating a new data set, these numbers must be at least as large as the current values

Defining an ARM policy

The administrative policy utility (IXCMIAPU) program maintains policies:

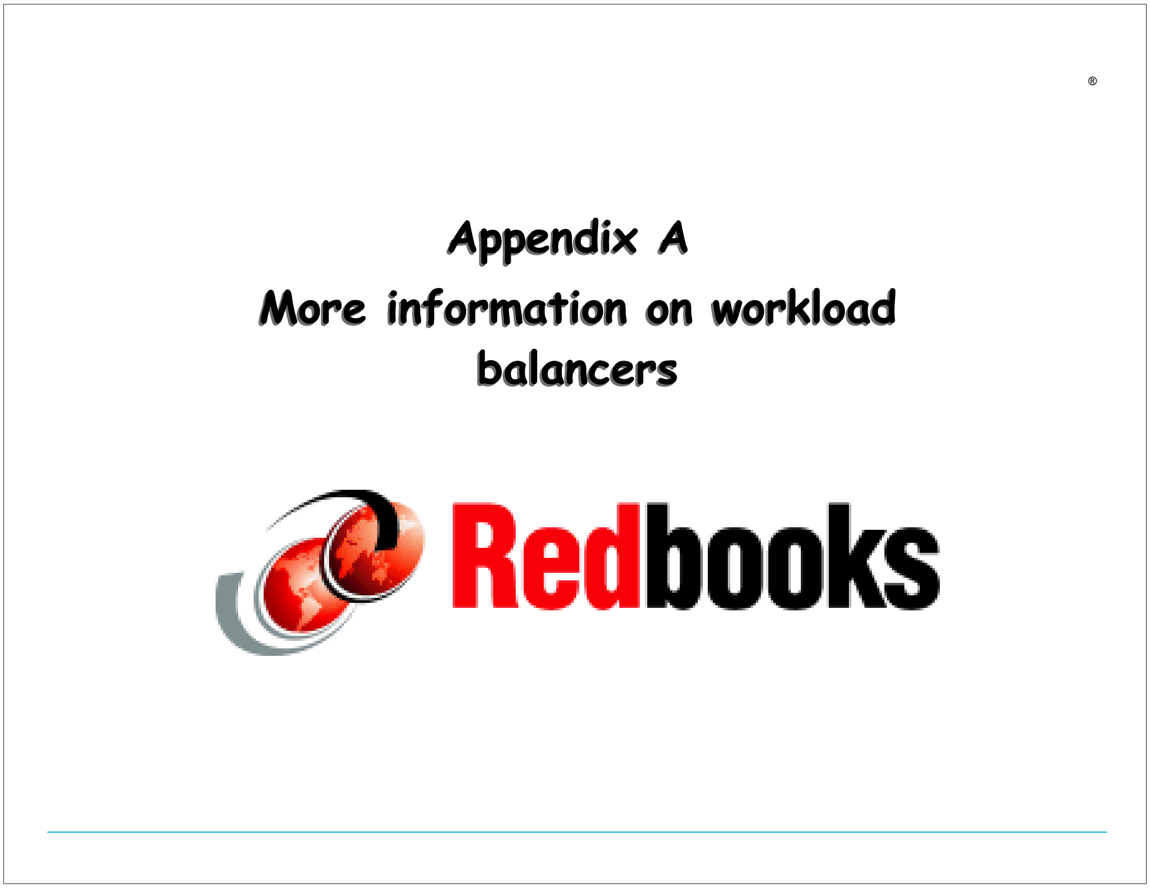
- Adds new policies
- Replaces existing policies
- Deletes policies
- Produces reports of the policy contents

IXCMIAPU is also used for LOGR,CFRM, and SFM.

See z/OS MVS Setting Up a Sysplex

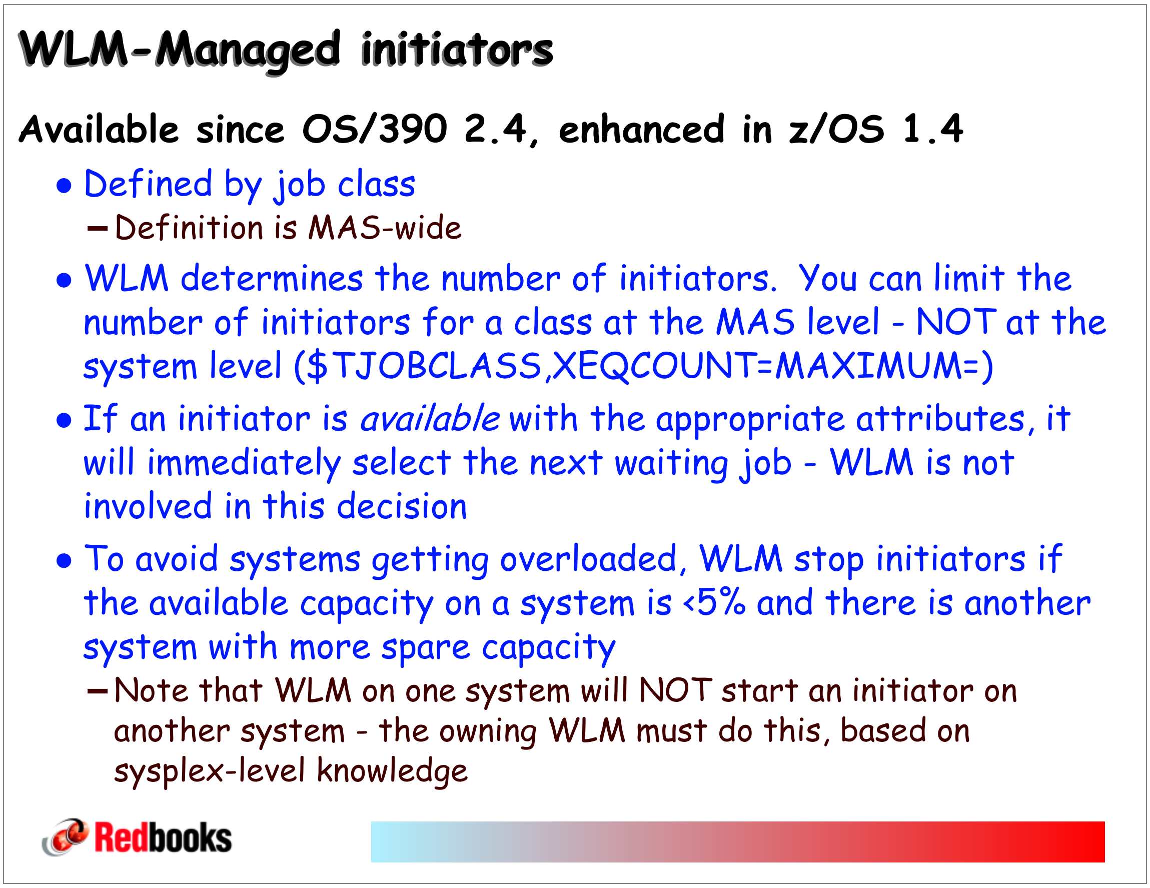
Publications

- SA22-7661 Parallel Sysplex Overview - An Introduction to Data Sharing and Parallelism
- SA22-7630 z/OS MVS System Management Facilities (SMF)
- SA22-7606 z/OS MVS Authorized Assembler Services Guilde (for ENFREQ)
- SA22-7617 z/OS MVS Sysplex Services Guide
- SA22-7618 z/OS MVS Sysplex Services Reference
- SA22-7622 z/OS MVS Setting up a Sysplex
- SA22-7593 z/OS MVS Installation Exits



®

Appendix A
More information on workload balancers



WLM-Managed initiators

Available since OS/390 2.4, enhanced in z/OS 1.4

- Defined by job class
  - Definition is MAS-wide
- WLM determines the number of initiators. You can limit the number of initiators for a class at the MAS level - NOT at the system level ($TJOBCLASS,XEQCOUNT=MAXIMUM=)
- If an initiator is *available* with the appropriate attributes, it will immediately select the next waiting job - WLM is not involved in this decision
- To avoid systems getting overloaded, WLM stop initiators if the available capacity on a system is <5% and there is another system with more spare capacity
  - Note that WLM on one system will NOT start an initiator on another system - the owning WLM must do this, based on sysplex-level knowledge

WLM Scheduling Environments

Scheduling Environments - WLM facility to control which systems a batch job can potentially run on

- Defined in WLM policy, made available (or not) by operator command
- Names are up to 16 chars long, with national characters (#,@, $), and _
- Specify on JOB CARD with SCHENV=whatever
  - Each job can only specify one SCHENV
- If job specifies SCHENV, it will only run on whichever system has that SE active
  - Can make a SCH ENV active on multiple systems
  - Could link to automation to make them unavailable/available when given thresholds are exceeded or missed

WLM Resource Groups

WLM Resource Groups are used to limit the capacity given to the associated workload or to guarantee it a minimum amount of capacity

- However, the specified capacity value is at the sysplex level - there is no way to limit the capacity for a workload *on a particular system*
- Performance of workloads that are in Resource Group can be erratic

Probably not a good candidate for your sysplex aggregation toolkit

VTAM GR

CICS, IMS, DB2, MQ, and TSO all support VTAM GR

Default mode of operation is to select a local appl instance

- If multiple local instances, WLM picks the "best" one

If overridden or none available, WLM will return to VTAM the one instance in the sysplex that is getting the best goal achievement.

- WLM updates his recommendation once every policy adjustment interval (10 seconds)

You can influence or completely override the WLM decision using the ISTEXCGR VTAM exit

- There is a sample ISTEXCUV exit in SG24-5451 that enables override of the WLM recommendations - code is conceptually similar

Owning VTAM *pushes* work to the selected instance

Sysplex Distributor

Works with TCP Dynamic VIPA support to have multiple servers with same IP address

Depending on the options you specify, WLM can return

- Weighted results based on the number of sessions to each server (DISTMETHOD=ROUNDROBIN)
- Weighted results based on the available capacity for each system (DISTMETHOD=BASEWLM)
- Weighted results based on the goal achievement of each server instance (DISTMETHOD=SERVERWLM)

There is no exit similar to ISTEXCGR to let you override WLM recommendations

- However, you may be able to use the Sysplex Distributor routing policy to affect decisions

Sysplex Distributor routes request to the chosen server

Load balancing advisor

Intended to work with outboard load balancers (using SASP protocol) to intelligently route IP requests in a sysplex

- SASP protocol is also used by EWLM

Delivers recommendations to external load balancers using similar rules to Sysplex Distributor (BASEWLM, SERVERWLM, ROUNDROBIN)

Works back to z/OS 1.4

Supported by Cisco Content Switching Module

Redbooks

CICSPlex System Manager

CICSPlex System Manager (CP/SM) is standard part of CICS TS - no longer separately chargeable

- One main function of CP/SM is Dynamic Transaction Routing of CICS transactions
- CP/SM talks to WLM to get the goal objective for each Txn
- It talks to AORs to get actual response time info
- It has two dist methods:
  - Goal - tries to honour WLM average response time goals
  - Queue - send it to the AOR that has the shortest queue
  - Neither method uses WLM to decide where to route the transaction
- You can override the CP/SM decision by writing your own version of EYU9WRAM

Redbooks

DB2 Group Attach Facility

NOT a workload balancer

- But supports ability to run DB2 batch anywhere in the plex

Rather than specifying a DB2 instance name, you should specify the DB2 Group Attach Name. This allows jobs to connect to any DB2 in that data sharing group.

- Also works with CICS as of TS 2.3

If you have multiple DB2 subsystems in the same z/OS image all in the same DB2 data sharing group, default is for all local work to connect to the first DB2 on the subsystem interface

- There is a usermod available to override this behaviour
- Usermod distributes local connections in random manner across all instances

IMS data sharing group connect

IMS equivalent to DB2 Group Attach Facility - IMSGROUP

- Also *does not provide any workload balancing function*, but does support workload balancing by providing the ability to have an IMS job connect to any IMS in the data sharing group
- Unlike DB2 Group Attach, it is not possible to specify the IMS Group name in CICS - must point at a specific IMS instance

DB2 Connect

DB2 Connect Enterprise Edition:

- Used to route distributed DB2 requests to DB2 subsystems in a sysplex
- Communicates with WLM to determine which of the systems where a member of the target DB2 data sharing group resides has the most available capacity
- Must specify the "sysplex" parameter in DB2 Connect definition
- Can control whether WLM information is used for each transaction or just at initial CONNECT time

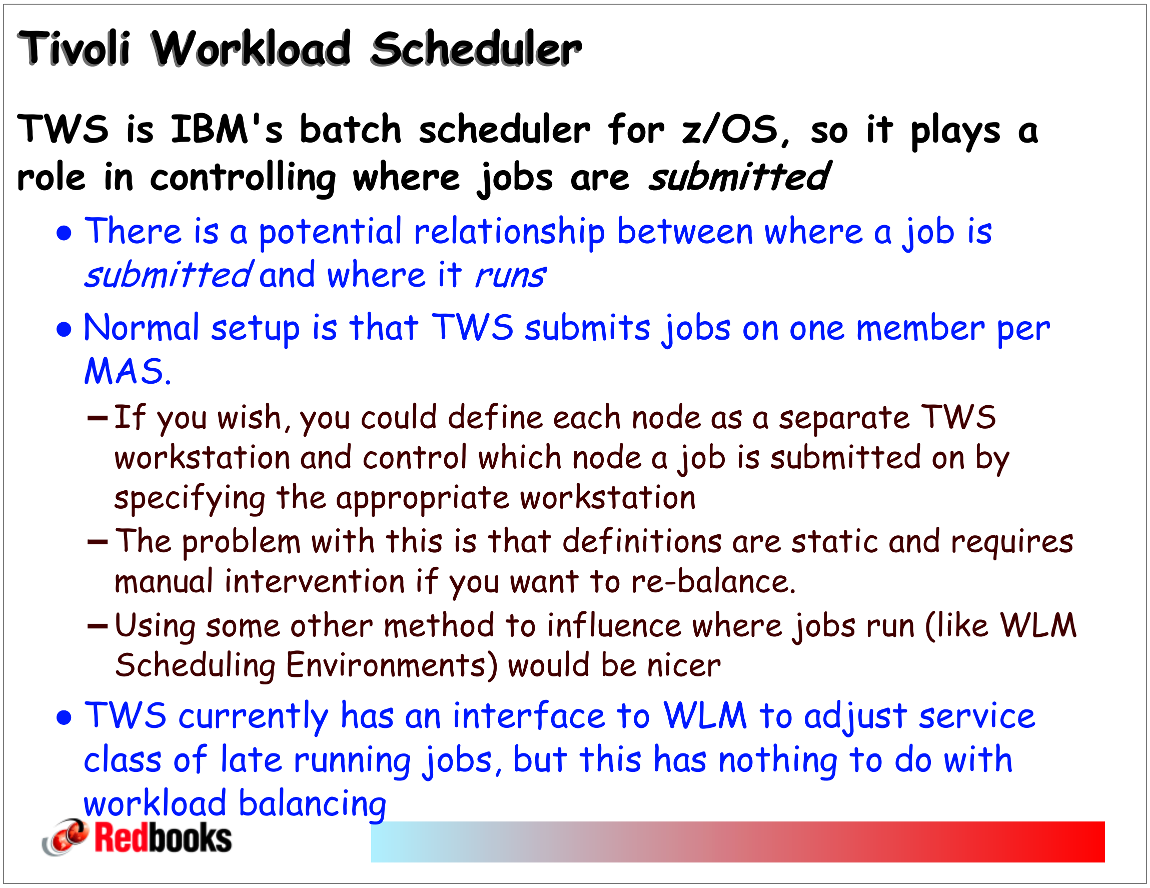
IMS Connect

IMS Connect (part of IMS as of IMS V9) acts as a "middle man" between TCP-based IMS clients, and IMS subsystems

Communication between clients and IMS Connect can exploit Sysplex Distributor to balance requests across multiple IMS Connect address spaces

IMS Connect communicates with IMS via XCF. You can write your own IMS Connect User Exit to round-robin requests across multiple IMS subsystems

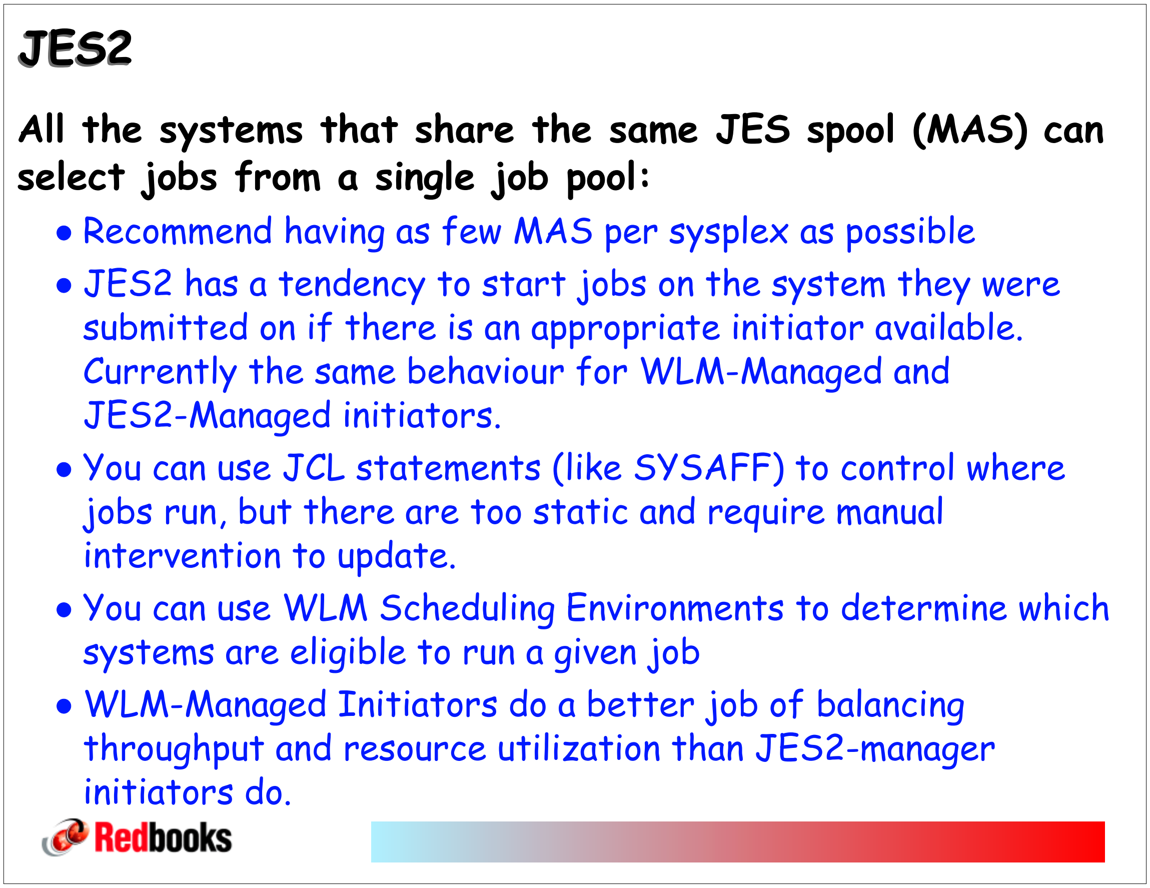
- Alternatively, use IMS Connect Extensions to route requests based on hard-coded relative weights for the different IMS subsystems



Tivoli Workload Scheduler

TWS is IBM's batch scheduler for z/OS, so it plays a role in controlling where jobs are *submitted*

- There is a potential relationship between where a job is *submitted* and where it *runs*
- Normal setup is that TWS submits jobs on one member per MAS.
  - If you wish, you could define each node as a separate TWS workstation and control which node a job is submitted on by specifying the appropriate workstation
  - The problem with this is that definitions are static and requires manual intervention if you want to re-balance.
  - Using some other method to influence where jobs run (like WLM Scheduling Environments) would be nicer
- TWS currently has an interface to WLM to adjust service class of late running jobs, but this has nothing to do with workload balancing



JES2

All the systems that share the same JES spool (MAS) can select jobs from a single job pool:

- Recommend having as few MAS per sysplex as possible
- JES2 has a tendency to start jobs on the system they were submitted on if there is an appropriate initiator available. Currently the same behaviour for WLM-Managed and JES2-Managed initiators.
- You can use JCL statements (like SYSAFF) to control where jobs run, but there are too static and require manual intervention to update.
- You can use WLM Scheduling Environments to determine which systems are eligible to run a given job
- WLM-Managed Initiators do a better job of balancing throughput and resource utilization than JES2-manager initiators do.

CPC Weights

What do weights have to do with workload balancing??

- When WLM is calculating the "available capacity" for an LPAR, it determines the delta between the current capacity of the LPAR and the "guaranteed" capacity, based on the lesser of the number of logical CPs or the weight. It then apportions any unused capacity based on the relative weight of this LPAR.
- So.... the relative weight of an LPAR (or group of LPARs if there is more than one sysplex member on the CPC) influences *some* decisions about where work may be directed
- The number of logical CPs and hard capping also play in this equation

Intelligent Resource Director

WLM Weight Management

- WLM can move weights between LPARs in order to help high important workloads achieve their objective
- However.... the total weight of the LPAR Cluster remains unchanged

WLM Vary CPU Management

- WLM will take logical CPs on and offline based on the capacity requirements of the work in the LPAR
  - Decision does *not* take into account PIs or work in other LPARs or available capacity on the CPC
- The number of currently online logical CPs is a factor in the available capacity calculation
  - Enabling WLM Vary CPU Management *could* have an impact on the volume of work directed to that image

DB2 query parallelism

Three levels of function in DB2 query parallelism:

- I/O parallelism (DB2 V3)
- CP parallelism (DB2 V4)
- Sysplex parallelism (DB2 V5)

While WLM is used to assign a service class to a query, WLM is not used in deciding how/where to split the query

Redbooks

IMS Workload Router

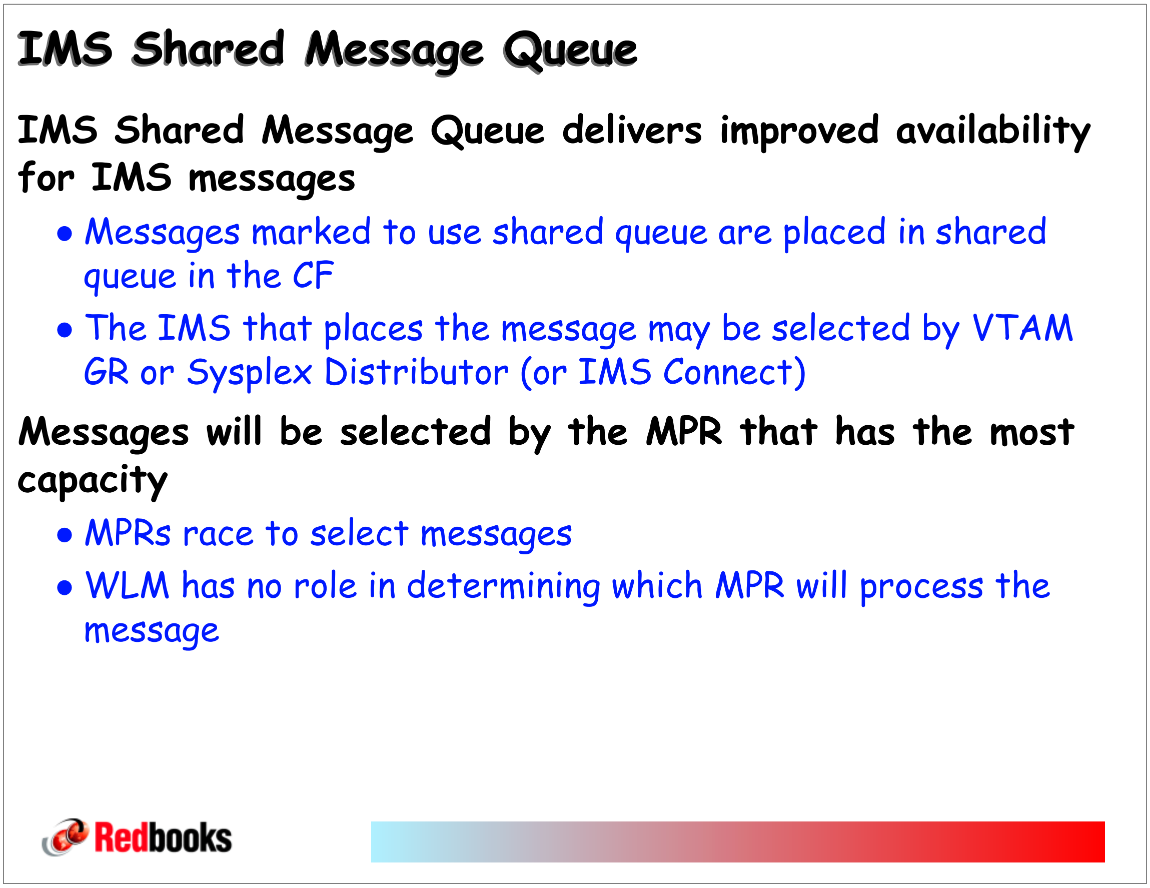
IMS WLR uses MSC to spread incoming transactions across multiple IMS subsystems

WLR does NOT use WLM to decide where to route transactions to

However you *can* assign weights to target IMS subsytems, as relative weights or as percentage of incoming IMS transactions

- If you had a 1000 MSU and a 500 MSU CPC, you could assign the 1000 MSU one a relative weight of 66 and a weight of 33 to the other CPC in an effort to get similar utilizations

Redbooks



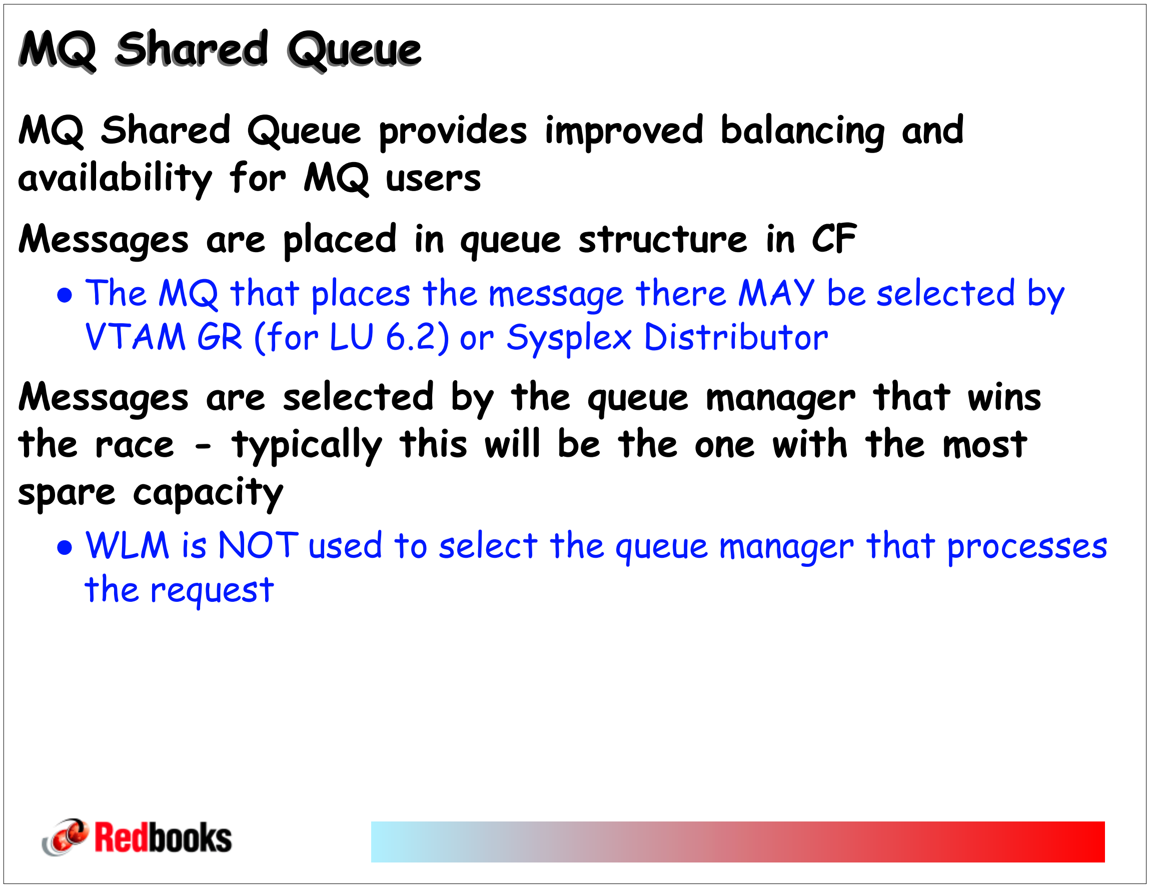
IMS Shared Message Queue

IMS Shared Message Queue delivers improved availability for IMS messages

- Messages marked to use shared queue are placed in shared queue in the CF
- The IMS that places the message may be selected by VTAM GR or Sysplex Distributor (or IMS Connect)

Messages will be selected by the MPR that has the most capacity

- MPRs race to select messages
- WLM has no role in determining which MPR will process the message



MQ Shared Queue

MQ Shared Queue provides improved balancing and availability for MQ users

Messages are placed in queue structure in CF

- The MQ that places the message there MAY be selected by VTAM GR (for LU 6.2) or Sysplex Distributor

Messages are selected by the queue manager that wins the race - typically this will be the one with the most spare capacity

- WLM is NOT used to select the queue manager that processes the request

CICS sysplex facilities

CICS MultiRegion Option (MRO) supports ability to distribute CICS transactions

- However, the actual routing must be handled by CP/SM or some other Dynamic Transaction Routing function

CICS Temp Storage in CF, CF Data Tables, Named Counter Server, and Global ENQ/DEQ all support the ability to route transactions anywhere in the system

- These are enabling technologies, not routing mechanisms

WebSphere Application Server

Sysplex Distributor should be used to select an appropriate WAS instance

Within a given WAS instance, WAS may work with WLM to determine the apprpriate number of servant address spaces to be started

- This only impacts WAS bandwidth on one system

zAAPs

Work that runs on zAAPs is not included in Sysplex Aggregation calculations

- Moving too much work to zAAPs could result in a sysplex dropping below 50% is used MSUs on that CPC
- You can specify if a system should use zAAPs or not, but you cannot limit how much work is routed to the zAAPs
  - On System z9, you CAN specify a hard (PR/SM) cap on the amount of zAAP capacity that can be used by the associated LPAR
  - The cap is specified on the HMC and is (relatively) static

Defined capacity (Softcap)

Defined capacities (Soft capping) are intended as a way to control software bills

Capping is turned on or off by WLM, for a single system, based on rolling 4-hour average utilization compared to the limit you specify on the HMC

- There is currently no way to specify a defined capacity for a group of LPARs

Ensure that the defined capacity could not result in the PrimaryPlex getting less than 50% of the CPC

- When soft capping is invoked, LPAR will be limited to the soft cap value until the rolling 4-hour average drops below the target.

PR/SM LPAR Capping

In general, PR/SM capping is NOT recommended

- Can result in workloads missing target while available capacity is unused

However, in a dynamic workload balancing configuration, judicious use of PR/SM capping MAY result in work being pushed to other CPCs, potentially helping the PrimaryPlex on those systems remaining above 50% of used capacity

Redbooks

ibm.com

e-business

Presentation end

Redbooks

Planned outage avoidance

JES2 changes to avoid IPLs

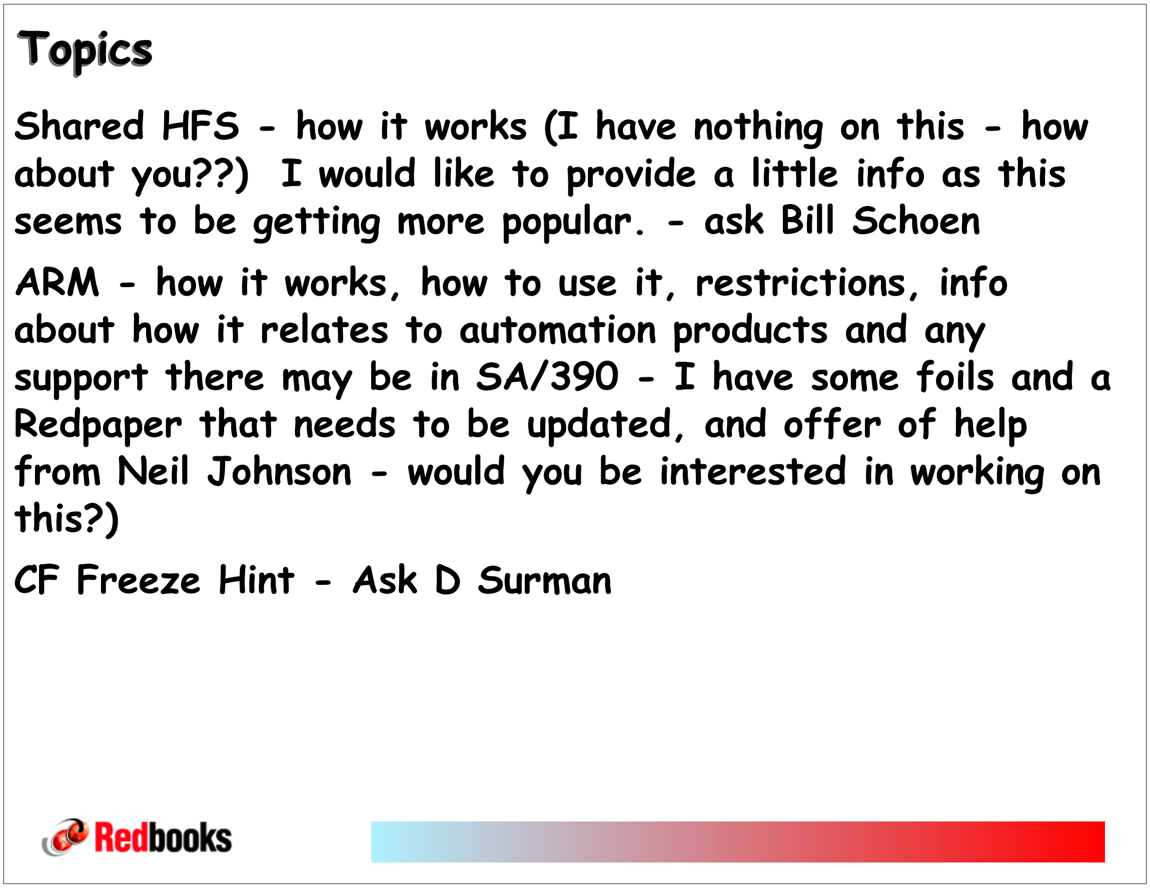
- What about testing JES2 exits?? Preferred option is to use a secondary JES2, together with a sample Exit 5 from CBT that adds new $REPEXIT and $ADDEXIT commands.
  - JES2 must be restarted once to enable $REPEXIT
  - Be careful if you try to use a JES2 Hot start to reload exits - see section entitled "Hot Start considerations" in JES2 Init & Tuning Guide
  - Make sure OA12725 is installed if you want to work with Exit 54 or higher

Redbooks

Topics

Planned outage avoidance (I will have some material for this)
Clustering comparison (from Alan)
SCRT update (couple of foils just - I will provide)
Sysplex aggregation (I have some stuff for this)
CF Capacity planning (I can scavange from another pres.)
Mario's SMF88 and 74.2 tools (I can provide)
Casual use z/OS licence (??? need latest status)
GDPS update (Can you provide?)
JES2 scalability topics (based on Redpaper - would you like to tackle this?)
Recent Redbooks
Advertise residencies for non-IBMers
Ask Bob Abrams if I can speak about Steve Heisig work

Redbooks

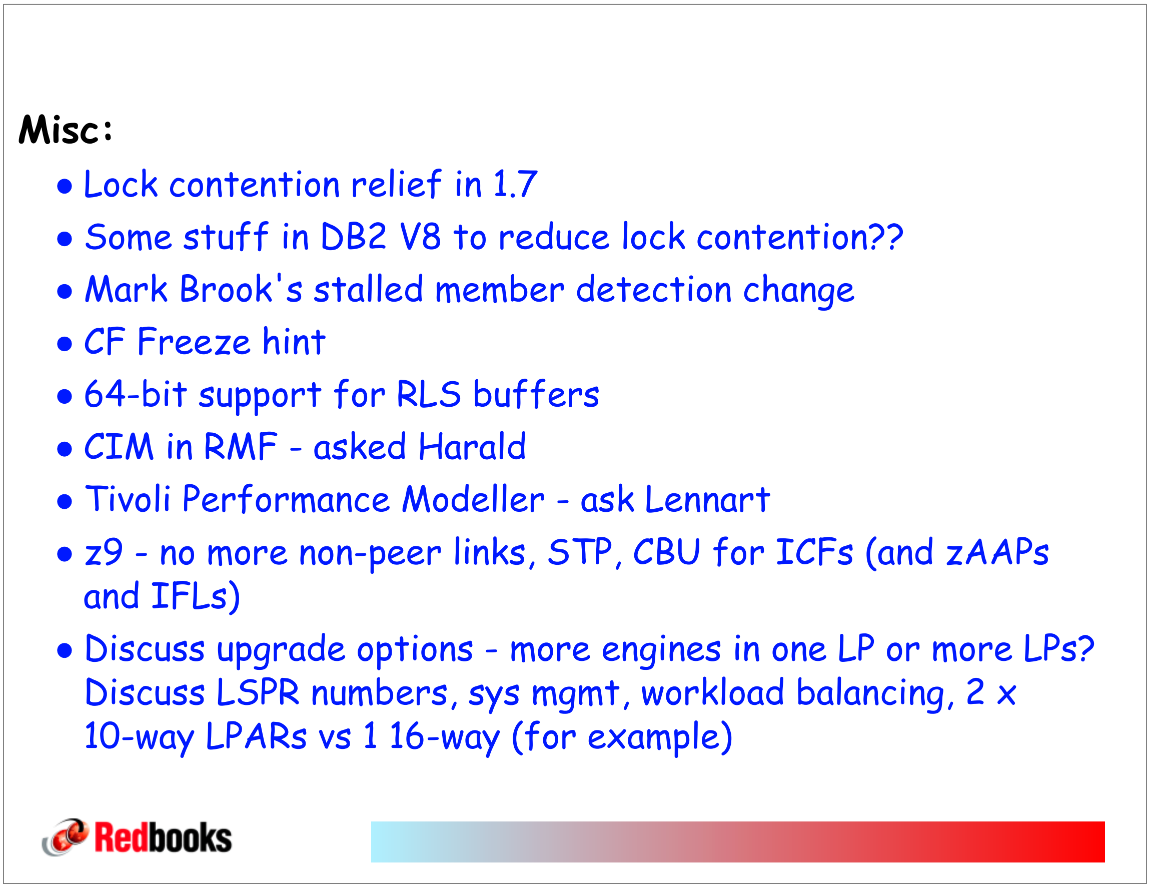


Topics

Shared HFS - how it works (I have nothing on this - how about you??) I would like to provide a little info as this seems to be getting more popular. - ask Bill Schoen

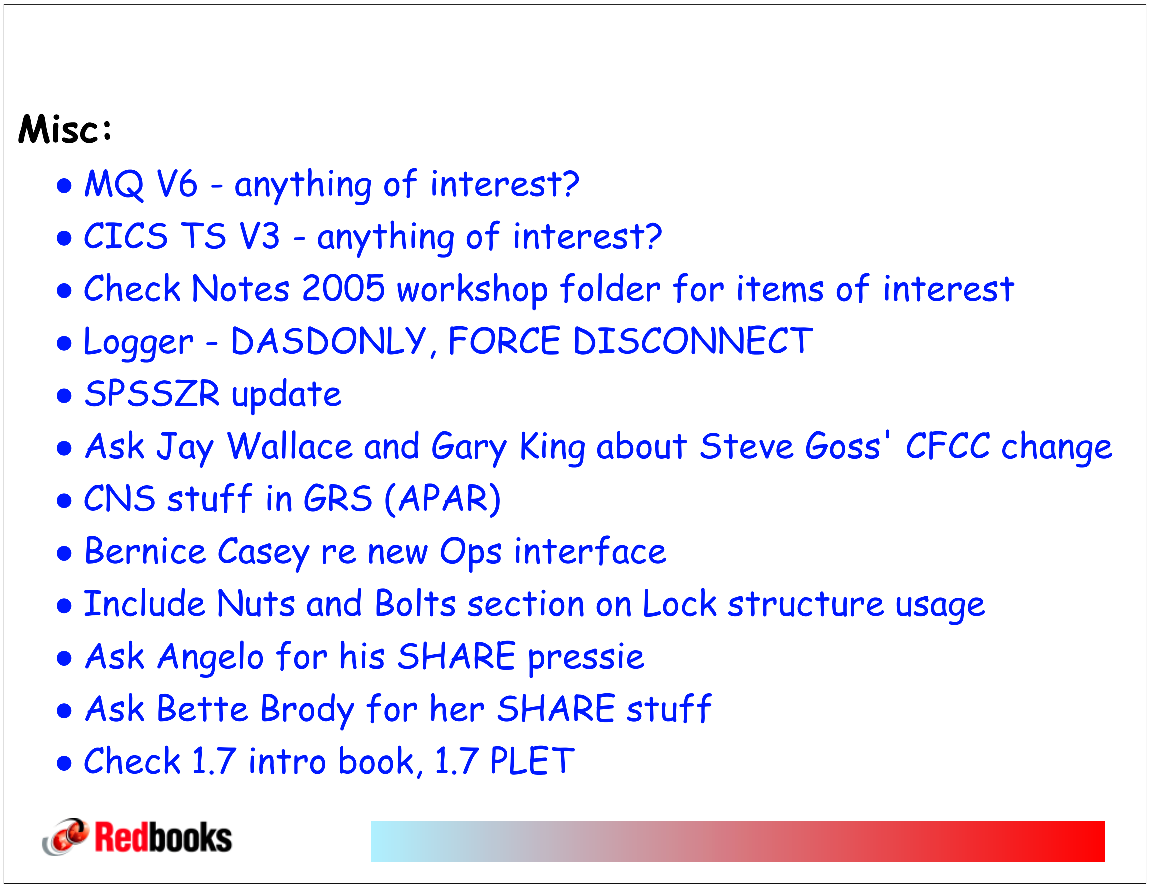
ARM - how it works, how to use it, restrictions, info about how it relates to automation products and any support there may be in SA/390 - I have some foils and a Redpaper that needs to be updated, and offer of help from Neil Johnson - would you be interested in working on this?)

CF Freeze Hint - Ask D Surman



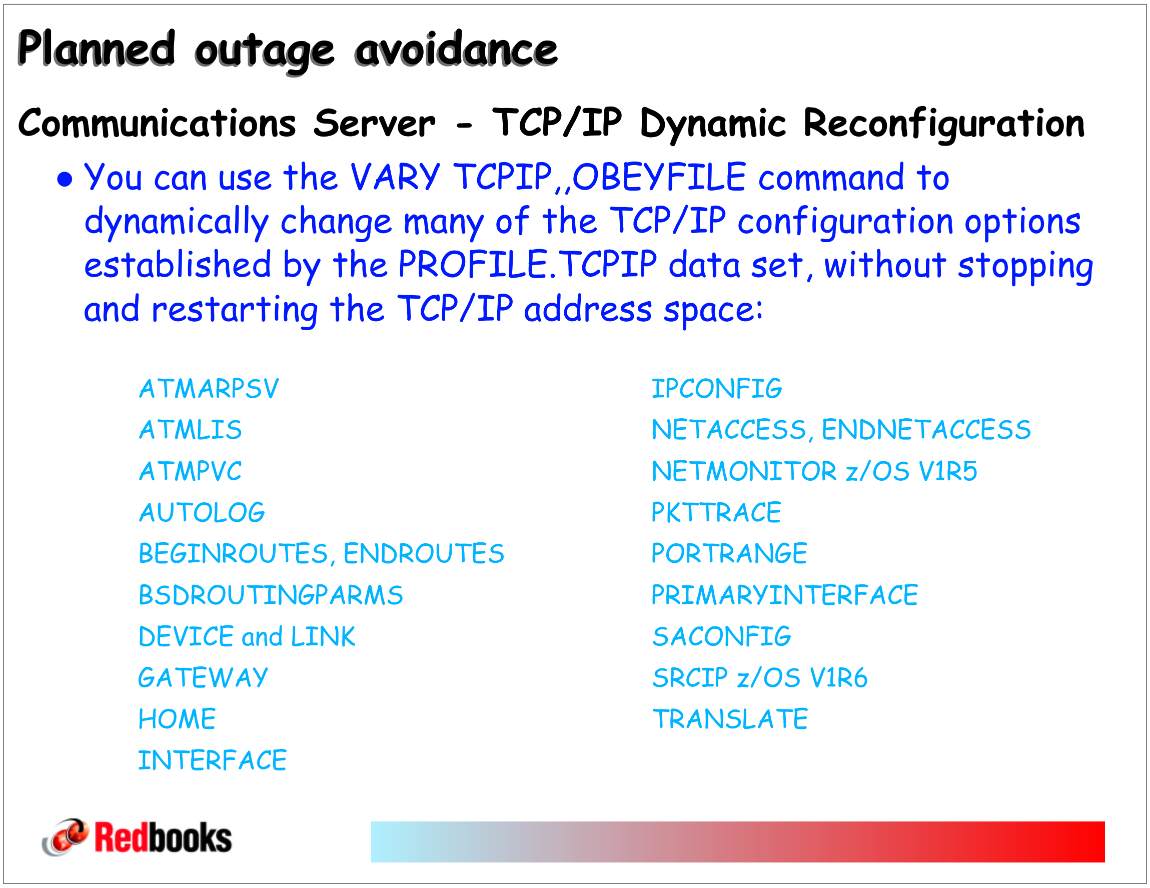
Misc:

- Lock contention relief in 1.7
- Some stuff in DB2 V8 to reduce lock contention??
- Mark Brook's stalled member detection change
- CF Freeze hint
- 64-bit support for RLS buffers
- CIM in RMF - asked Harald
- Tivoli Performance Modeller - ask Lennart
- z9 - no more non-peer links, STP, CBU for ICFs (and zAAPs and IFLs)
- Discuss upgrade options - more engines in one LP or more LPs? Discuss LSPR numbers, sys mgmt, workload balancing, 2 x 10-way LPARs vs 1 16-way (for example)



Misc:

- MQ V6 - anything of interest?
- CICS TS V3 - anything of interest?
- Check Notes 2005 workshop folder for items of interest
- Logger - DASDONLY, FORCE DISCONNECT
- SPSSZR update
- Ask Jay Wallace and Gary King about Steve Goss' CFCC change
- CNS stuff in GRS (APAR)
- Bernice Casey re new Ops interface
- Include Nuts and Bolts section on Lock structure usage
- Ask Angelo for his SHARE pressie
- Ask Bette Brody for her SHARE stuff
- Check 1.7 intro book, 1.7 PLET

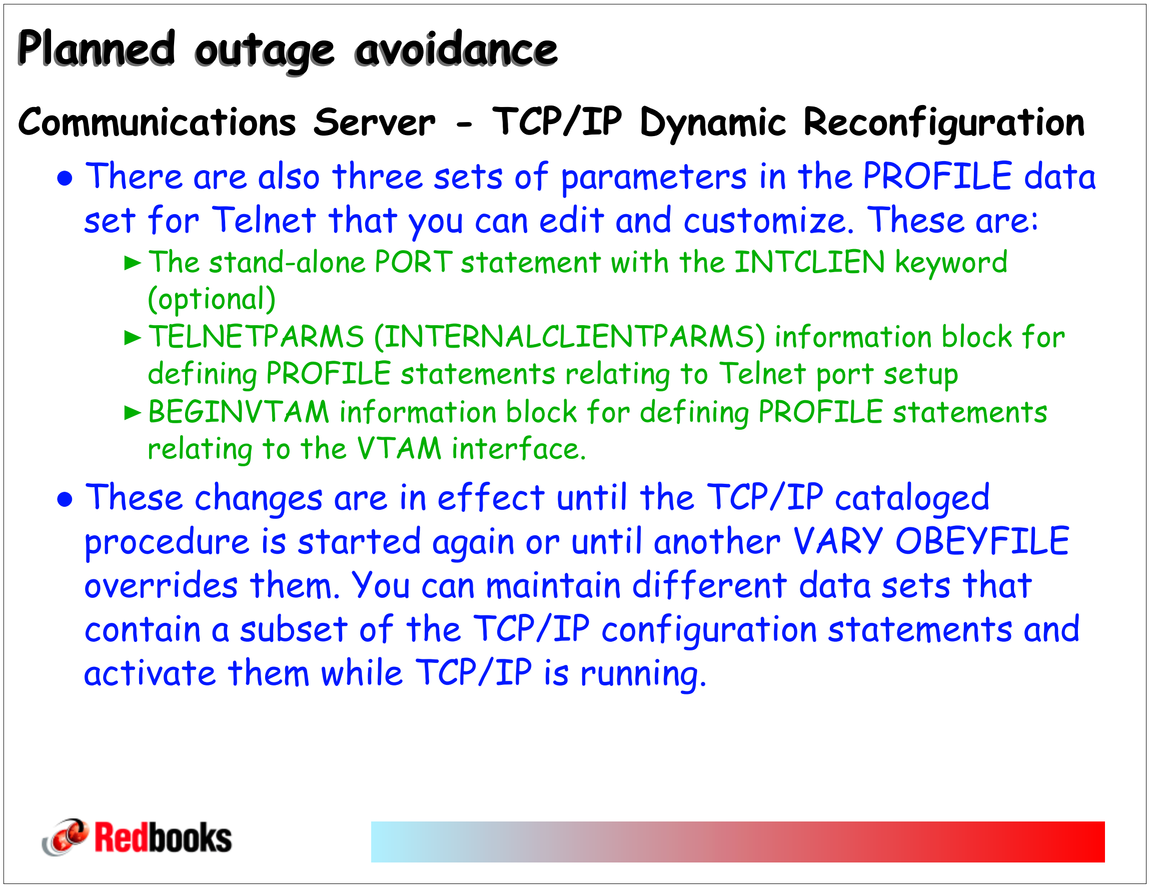


Planned outage avoidance

Communications Server - TCP/IP Dynamic Reconfiguration

- You can use the VARY TCPIP,,OBEYFILE command to dynamically change many of the TCP/IP configuration options established by the PROFILE.TCPIP data set, without stopping and restarting the TCP/IP address space:

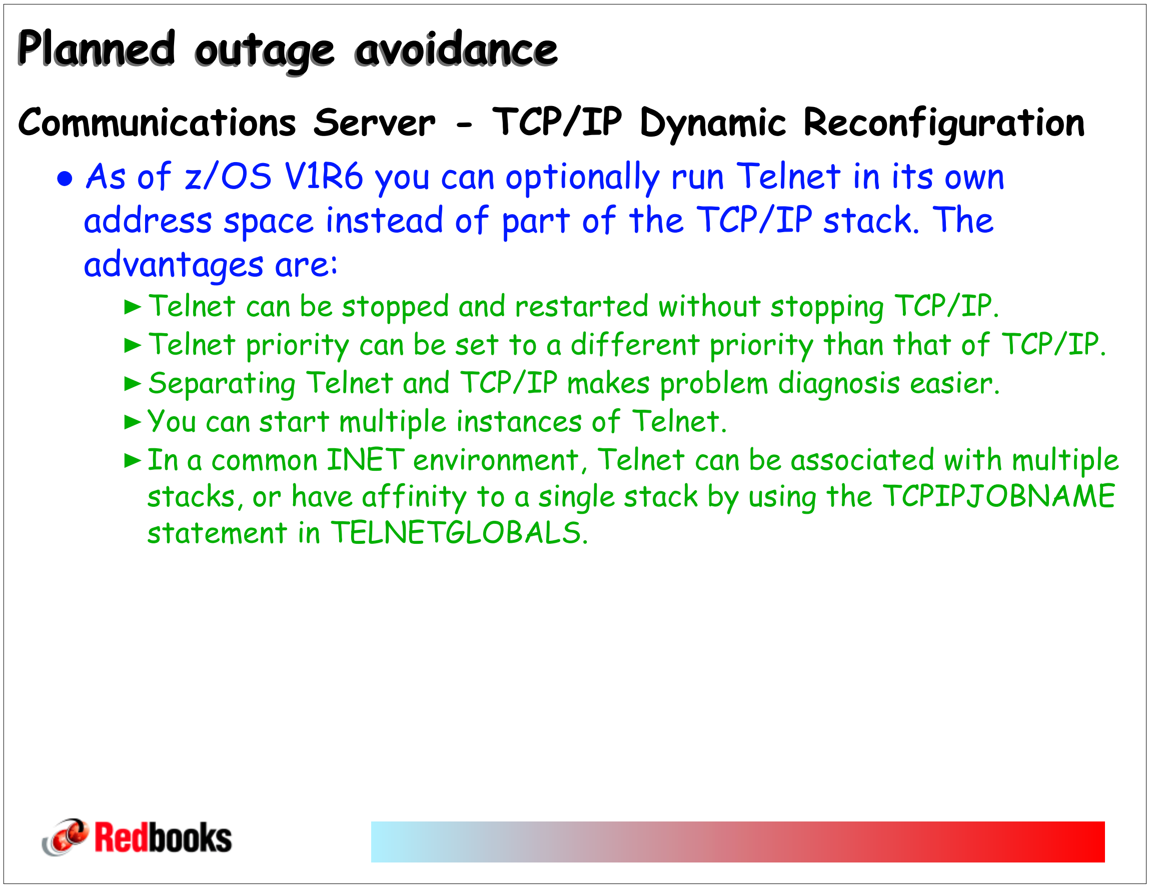
ATMARPSV
ATMLIS
ATMPVC
AUTOLOG
BEGINROUTES, ENDROUTES
BSDROUTINGPARMS
DEVICE and LINK
GATEWAY
HOME
INTERFACE
IPCONFIG
NETACCESS, ENDNETACCESS
NETMONITOR z/OS V1R5
PKTTRACE
PORTRANGE
PRIMARYINTERFACE
SACONFIG
SRCIP z/OS V1R6
TRANSLATE



Planned outage avoidance

Communications Server - TCP/IP Dynamic Reconfiguration

- There are also three sets of parameters in the PROFILE data set for Telnet that you can edit and customize. These are:
  - The stand-alone PORT statement with the INTCLIEN keyword (optional)
  - TELNETPARMS (INTERNALCLIENTPARMS) information block for defining PROFILE statements relating to Telnet port setup
  - BEGINVTAM information block for defining PROFILE statements relating to the VTAM interface.
- These changes are in effect until the TCP/IP cataloged procedure is started again or until another VARY OBEYFILE overrides them. You can maintain different data sets that contain a subset of the TCP/IP configuration statements and activate them while TCP/IP is running.



Planned outage avoidance

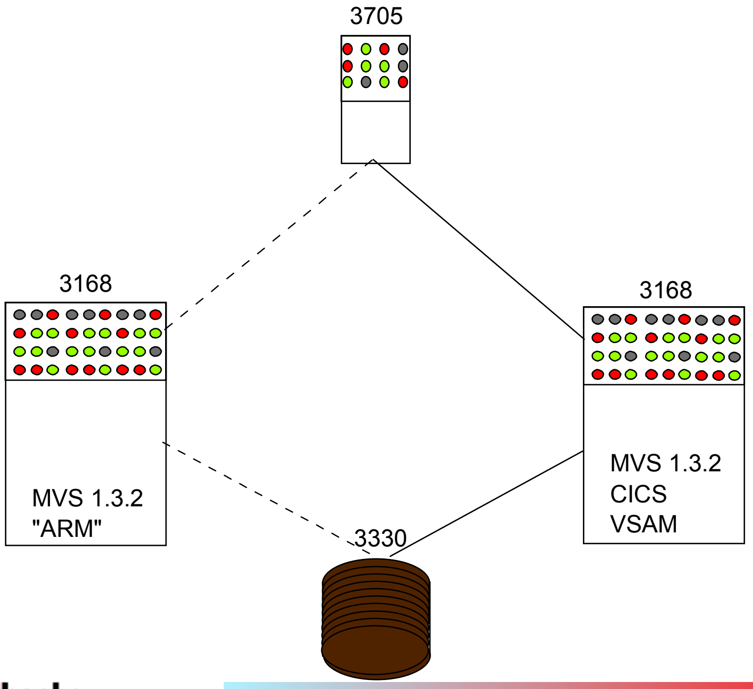
Communications Server - TCP/IP Dynamic Reconfiguration

- As of z/OS V1R6 you can optionally run Telnet in its own address space instead of part of the TCP/IP stack. The advantages are:
  - Telnet can be stopped and restarted without stopping TCP/IP.
  - Telnet priority can be set to a different priority than that of TCP/IP.
  - Separating Telnet and TCP/IP makes problem diagnosis easier.
  - You can start multiple instances of Telnet.
  - In a common INET environment, Telnet can be associated with multiple stacks, or have affinity to a single stack by using the TCPIPJOBNAME statement in TELNETGLOBALS.

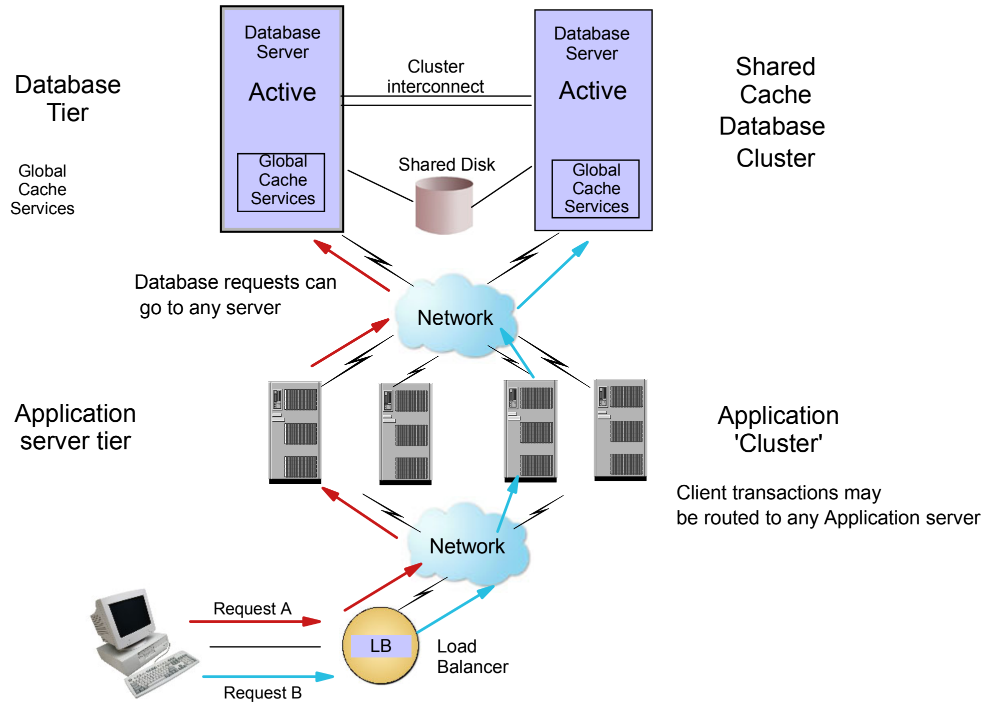
Planned outage avoidance

Migrating to zFS Root file system...

- Made a zFS copy of the sysplex root
- Edit the ROOT statement in the BPXPRMxx member w/ the new sysplex root name, and changed type to ZFS
- Bring down ALL systems but ONE
- Run: <fbpxoinit,filesys=unmountall> to unmount all FSs
- Check w/ <d ovms,f>. All this should show you, after all FSs are unmounted, is a mount w/ the name SYSROOT.
- Run: <set omvs=(xx)> to do the mounts specified in the BPXPRMxx member,which includes the sysplex root
- Make sure all looks good (log into OMVS, check the syslog for error msgs from <set omvs=(xx)> command, run <d omvs,f> to see if all mounts specified in BPXPRMxx are mounted ... etc)
- Then bring all systems up again



Oracle Real Application Cluster (RAC)



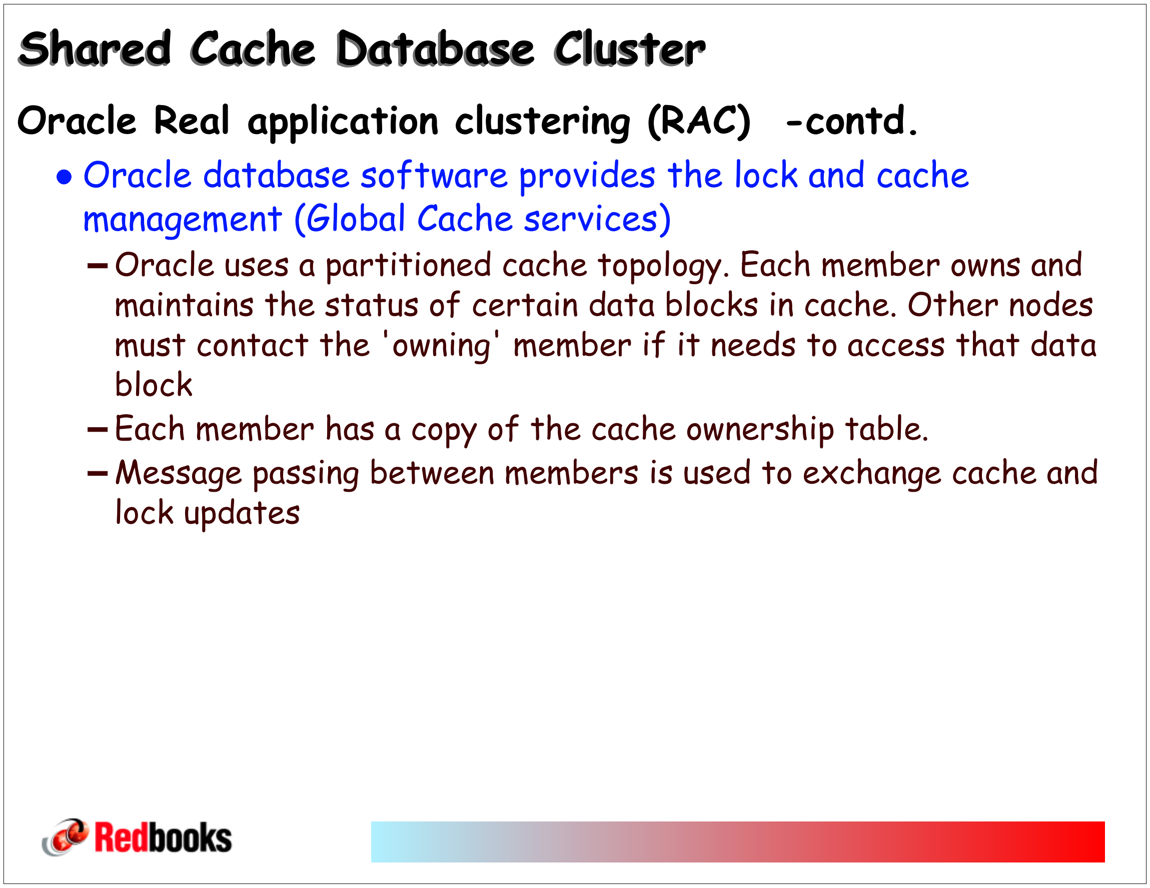
Database requests can go to any server

Client transactions may be routed to any Application server

Shared Cache Database Cluster

Oracle Real Application Clustering (RAC)

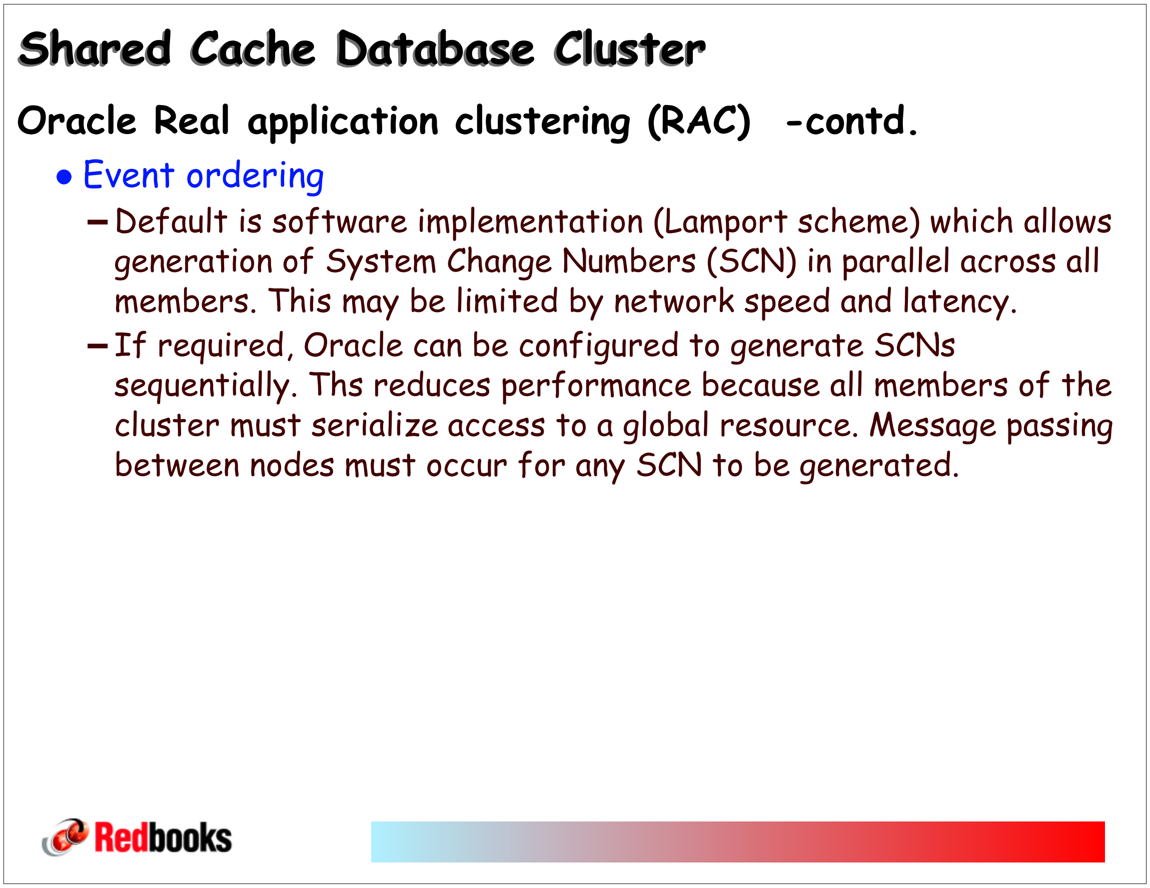
- Oracle clustering software provides :
  - Cluster membership services
  - Failure detection
  - Automation facilities to handle resource failover
- Shared storage is provided by a Storage Area Network (SAN)
- Cluster interconnect may be :
  - LAN
  - Specialized switching H/W (fibre)



Shared Cache Database Cluster

Oracle Real application clustering (RAC) -contd.

- Oracle database software provides the lock and cache management (Global Cache services)
  - Oracle uses a partitioned cache topology. Each member owns and maintains the status of certain data blocks in cache. Other nodes must contact the 'owning' member if it needs to access that data block
  - Each member has a copy of the cache ownership table.
  - Message passing between members is used to exchange cache and lock updates



Shared Cache Database Cluster

Oracle Real application clustering (RAC) -contd.

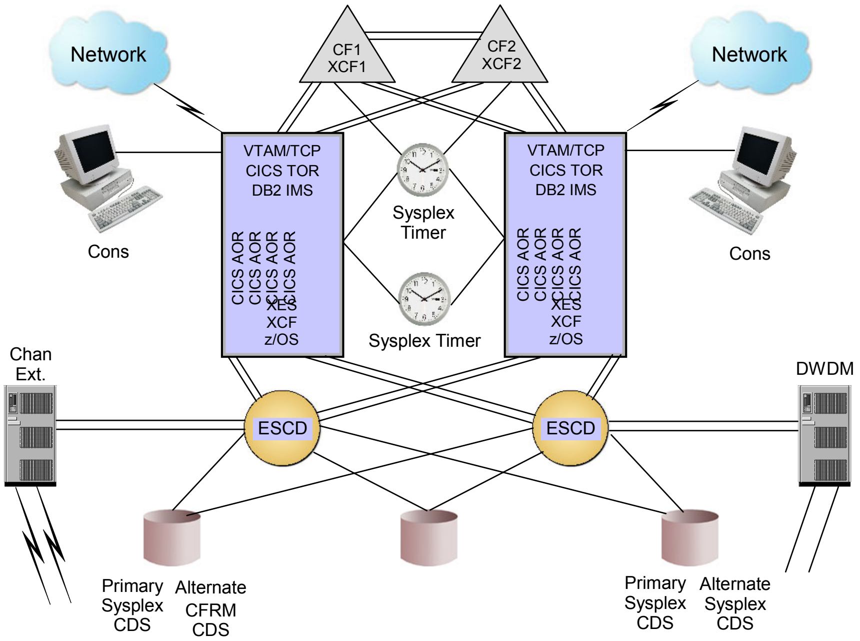
- Event ordering
  - Default is software implementation (Lamport scheme) which allows generation of System Change Numbers (SCN) in parallel across all members. This may be limited by network speed and latency.
  - If required, Oracle can be configured to generate SCNs sequentially. Ths reduces performance because all members of the cluster must serialize access to a global resource. Message passing between nodes must occur for any SCN to be generated.

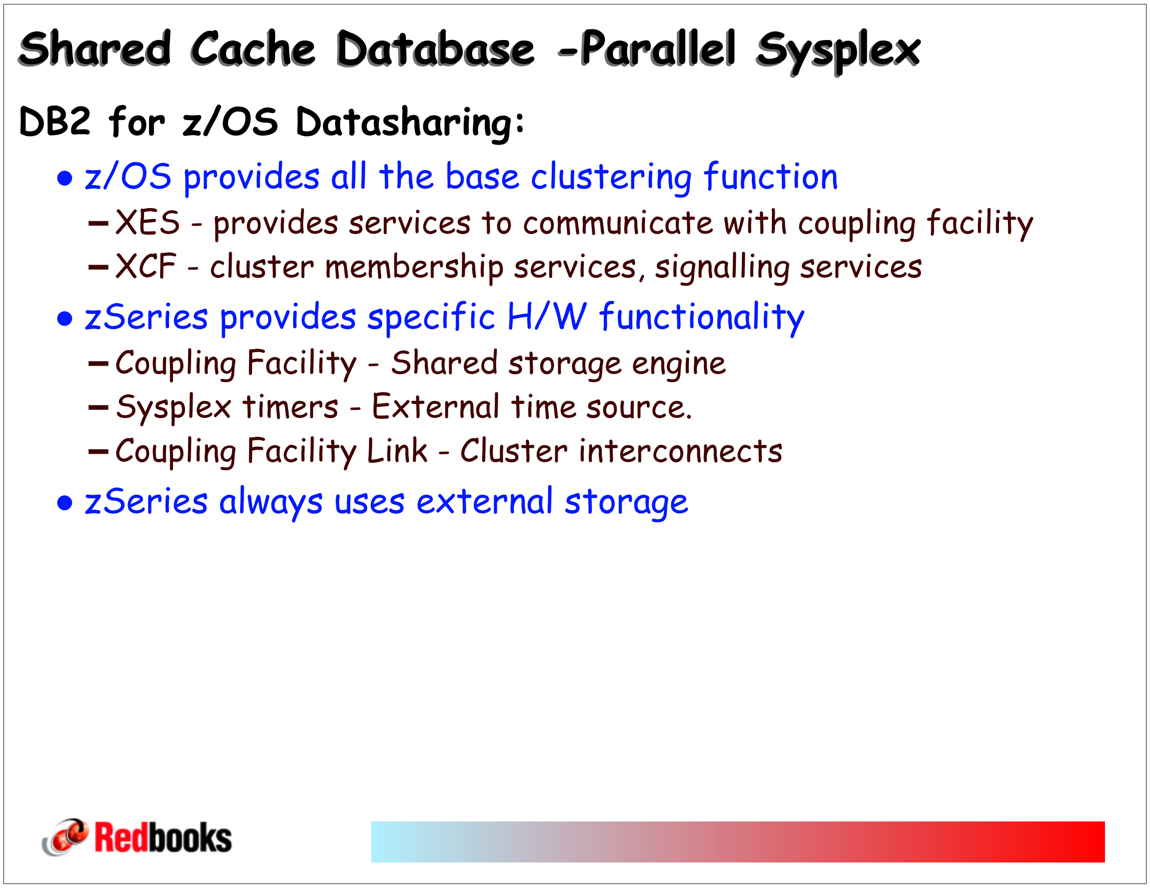
Shared Cache Database Cluster

Oracle Real application clustering (RAC) -contd.

- Oracle Database software provides the member restart and recovery services
  - If a member fails then Oracle must rebuild its Global Resource Directory to reassign ownership of data blocks to surviving members. It must also assess what in flight transactions exist and work out what data blocks need recovery actions taken against them. During this time the entire database is unavailable.

Parallel Sysplex components





Shared Cache Database -Parallel Sysplex

DB2 for z/OS Datasharing:

- z/OS provides all the base clustering function
  - XES - provides services to communicate with coupling facility
  - XCF - cluster membership services, signalling services
- zSeries provides specific H/W functionality
  - Coupling Facility - Shared storage engine
  - Sysplex timers - External time source.
  - Coupling Facility Link - Cluster interconnects
- zSeries always uses external storage

Redbooks