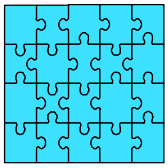
RACF and the Parallel Sysplex

New York RACF User's Group

October 20th, 2009

New York City NY



Russ Hardgrove
RACF Level 2
IBM - z/OS Software Service
Poughkeepsie, NY 12601
hardgrov@us.ibm.com

Objectives:

- Understand the Sysplex Environment
- Implement RACF Sysplex Communication
- Implement RACF Sysplex Data Sharing
- Understand the Recovery Modes available
- Describe the steps to define the Coupling Facility Policy for RACF

- **Explain the purpose of RACF Sysplex Communications and RACF Sysplex Data Sharing.**
- **list the required software and hardware products**
- **Evaluate the functions for applicability in your shop.**
- **Explain how to implement RACF SC and SDS.**

RACF Sysplex Support Objectives

Performance

- Reduce Contention for RACF Database

System Management

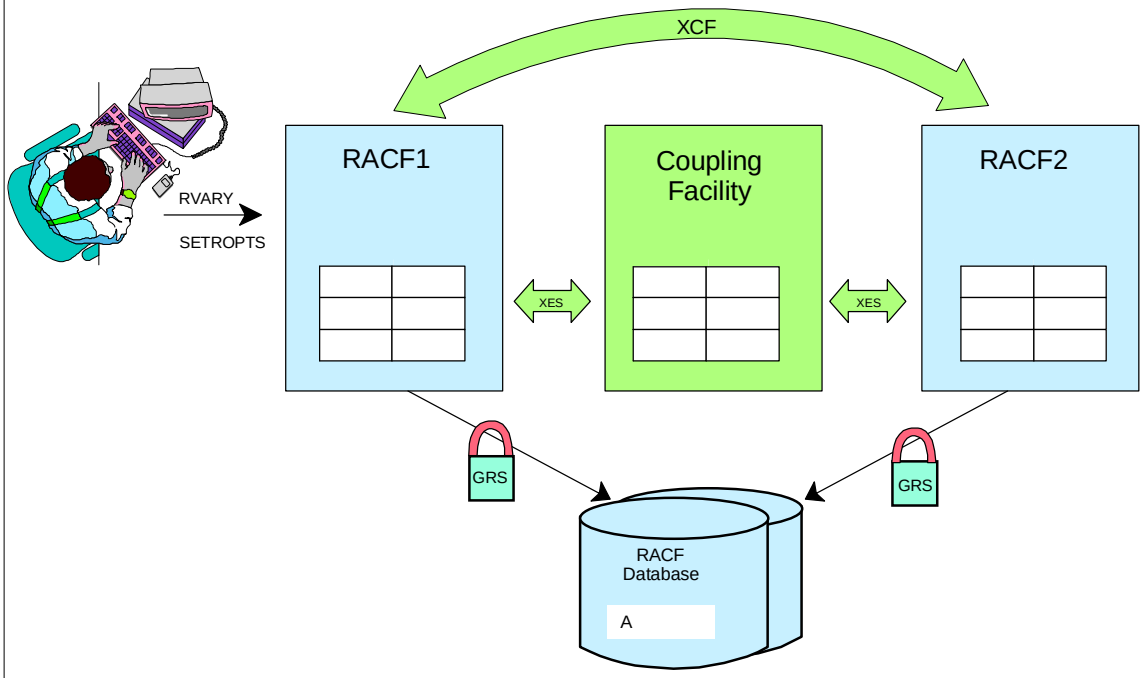
- Provide Single-System Image for Security Administration

Availability

- Propagate RVARY to ALL Systems that Share the RACF Database
- Minimize Sympathy Sickness

- Exploits sysplex facilities to address problems when many systems share a RACF DB.
- RACF (pre-SDS) uses RESERVE/RELEASE
- SETROPTS RACLIST/REFRESH (+ others) and RVARY functions are propagated to all members.
- When in RACF SDS systems that fail no longer can do so when holding a RESERVE.

Overview - How It Works



- When in RACF SC, SETROPTS and RVARY are propagated to all participating systems via XCF.
- When in RACF SDS,he CF as alarge shared cache for the RACF DB. When RACF needs a block, it looks first in the resident data blocks, if not there it looks in the CF cache, and if not there doesn an I/O.
- Serialization is done with GRS instead of hardware RESERVE/RELEASE.

Requirements

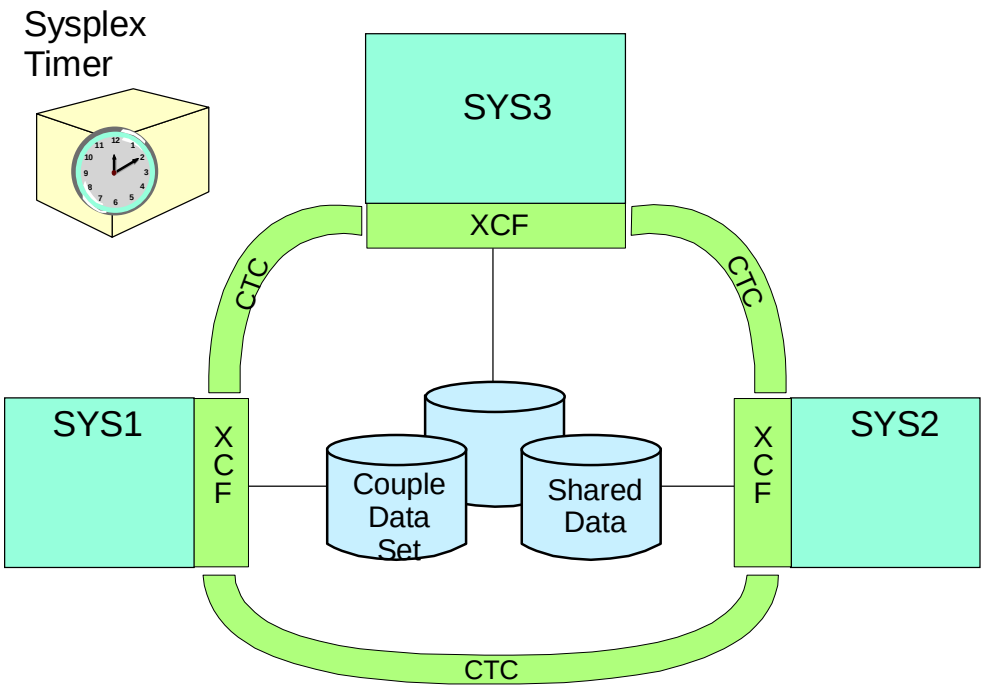
RACF Sysplex Communication

- Sysplex capable via CTC
- all software levels supported

RACF Sysplex Data Sharing

- Parallel Sysplex capable with CFs
- all software levels supported

What Is a Sysplex?



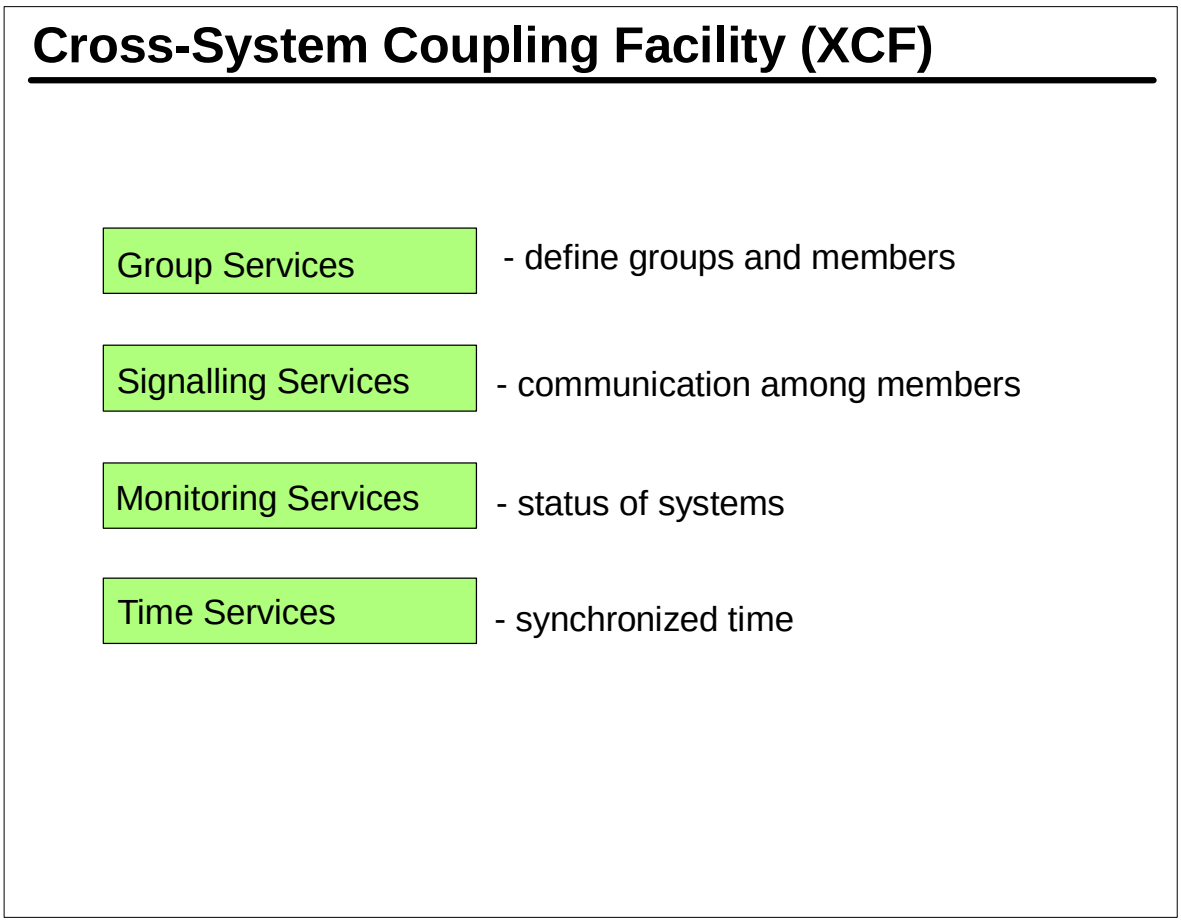
- SYStems comPLEX - collection of cooperating systems behaving as one.
- Coupled together by hardware and software services.
- Communiation via (at a minimum) CTCs.
- Share a common time source (either all on one CEC or have a shared sysplex timer)
- Viewed as a single entity... (i.e. one common master console)

Sysplex Terminology

- Sysplex
- Multisystem Application
- Member
- Group
- Couple Data Set

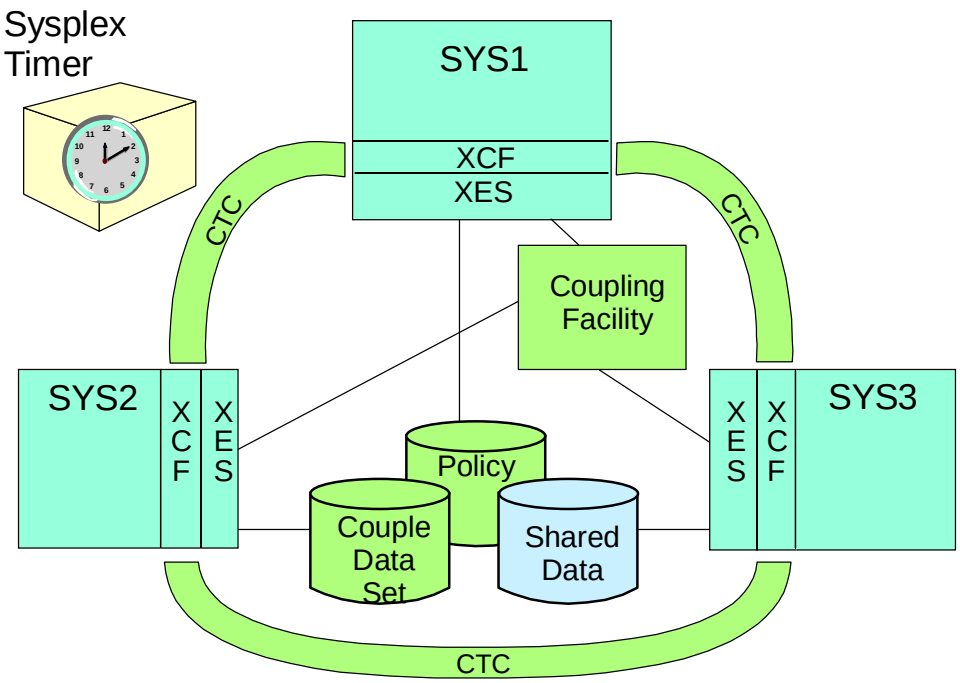
- SYStems comPLEX - collection of cooperating systems behaving as one.
- Multisystem application. An IBM or customer component, subsystem, or application that has various functions distibuted across systems.
- Member is a part of a MSA defined to XCF, it resides on one LPAR and uses XCF to communication with other members.
- Group is a set of related members defined to XCF bu MSA. Known as an XCF group and may have members on multiple systems. May only talk to members in the same XCF group.
- CDS contains sysplex wide data about systems, groups and members utilizing XCF services. Each system in SYSPLEX must have connectivity to the CDS.

Cross-System Coupling Facility (XCF)



- ▸ XCF provides communications services between MVS systems in a plex. XCF services enable MSA, such as GRS or RACF to send signals (messages) among the MVS systems without having to mange the I/O.
- ▸ XCF provides services for MSA that can be grouped:
- **group srvcs** - def grps / mbrs. means to request info regarding mbrs in same XCF grp.
- **signalling srvcs** - method oc comm. between mbrs of same XCF grp.
- **monitoring Srvcs** - allow mbrs to determine it's status and notify other mbrs of changes. Also monitor other mbrs.
- **time srvcs** - provide time synch for events requiring it. (logs, traces etc).

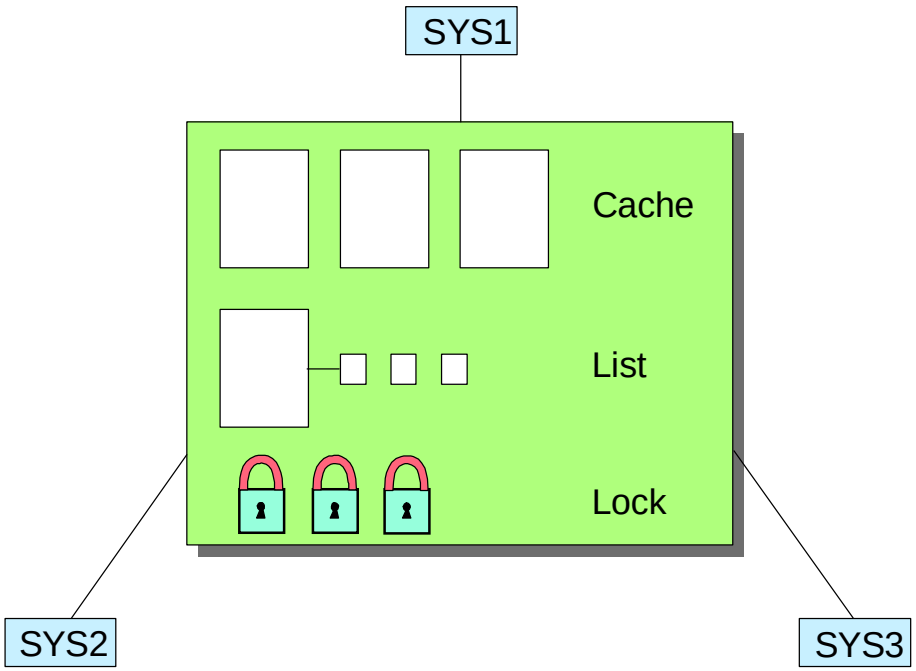
Sysplex with a Coupling Facility



- Picure like before, except a CF (coupling facility) has been added making it a parallel sysplex.
- CFs are connected to CPCs via high speed fiber optic links known as CF channels.
- Cross-system extended Srvcs (XES) is MVS software used to exploit the CF.
- CF working with XCF services provides a high-performance method to share data among many CPCs.
- Parallel Sysplex provides such benefits as:

- Single system image for data access
- Continuous availability
- wokload balancing
- system management

Coupling Facility



Structures - three types:

- Cache - high perf sharing. three types directory-only, store-in o store thru
- List - allows sharing of info in sets of lists or queues.
- Lock - allows users to create set of locks, that when obtained (shr/excl) can be used to serialize resources (including list or cache structures).

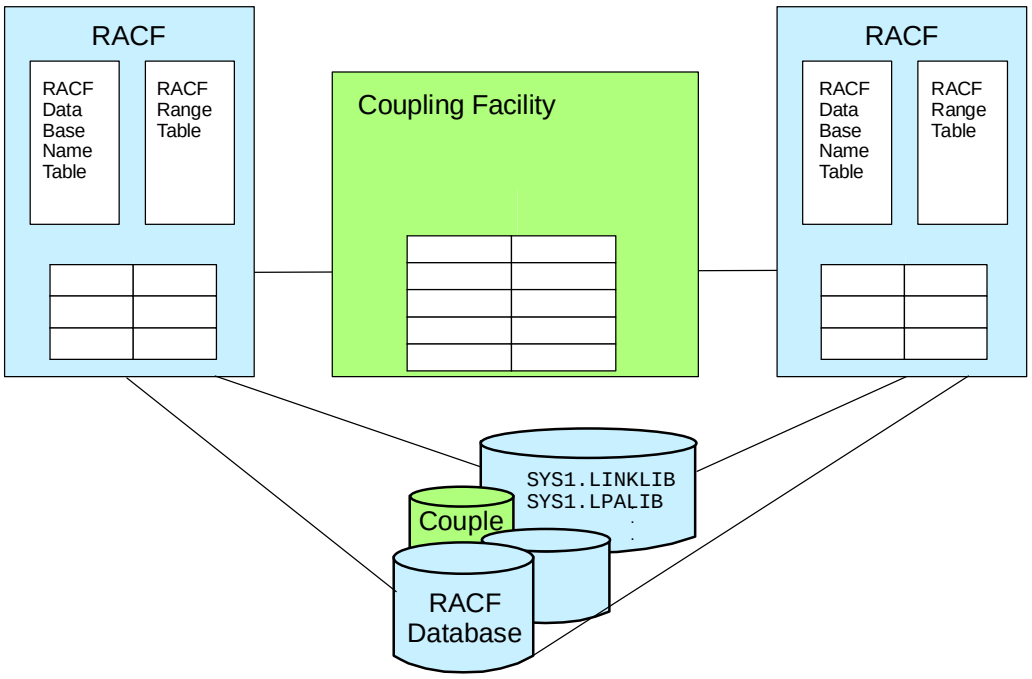
Cross-System Extended Services (XES)

- Connection Services
- Connection Recovery Services
- Application Mainline Services
- Measurement Services
- Dumping Services

XES component of OS/390 or z/OS provides authorized services for using a CF:

- Connection and connection recovery services - used to allocate / define / obtain strcutures.
- Appl mainline srcvs:
  - cache - read / write / delete in structures and do cross-invalidation
  - list - read / write / delete /move list entries
  - lock - obtain / alter / release a lock
- Measurement srvcs - interface for obtaining R/T config and perf data for CF structures
- Dumping srvcs - capture info from CF for inclusion in an SVC dump

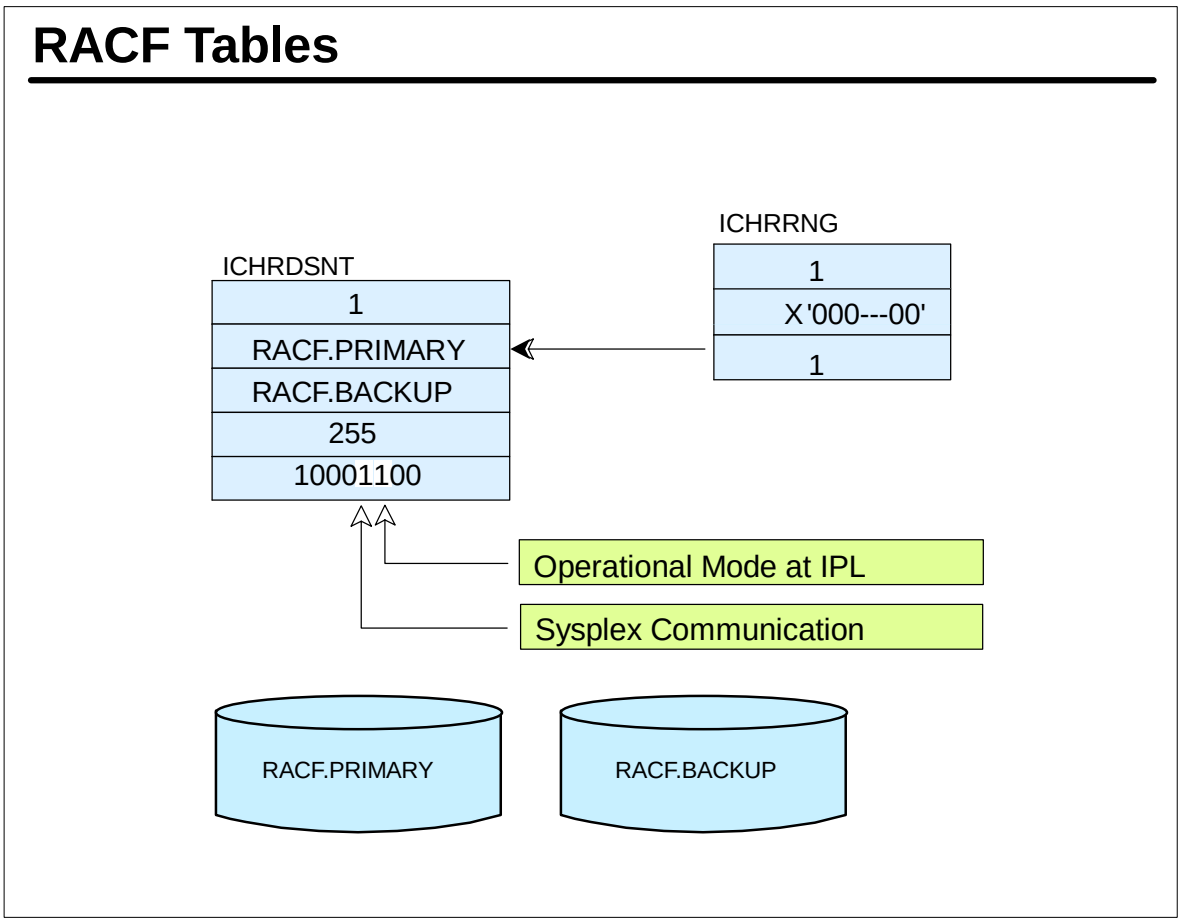
RACF Data Sharing Group



XCF group for RACF SC and SDS is name IRRXCF00 and when implementing RACF SC and SDS these rules apply:

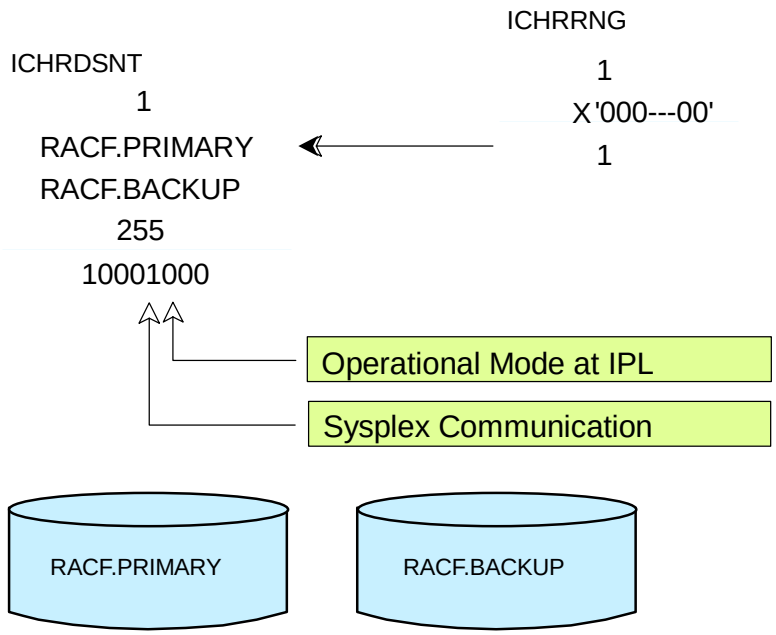
- Can be only one IRRXCF00 in the sysplex
- all members MUST use the same ICHRDSNT and ICHRRNG (if DB split). First system up in plex builds an incore group. Others join and compare theirs - in not equal will use group (ICH555I msg) table.
- RACF DB _cannot__ be shared with a system outside the RACF sharing group. Installastion's responsibility. Corruption will result if violated.

RACF Tables



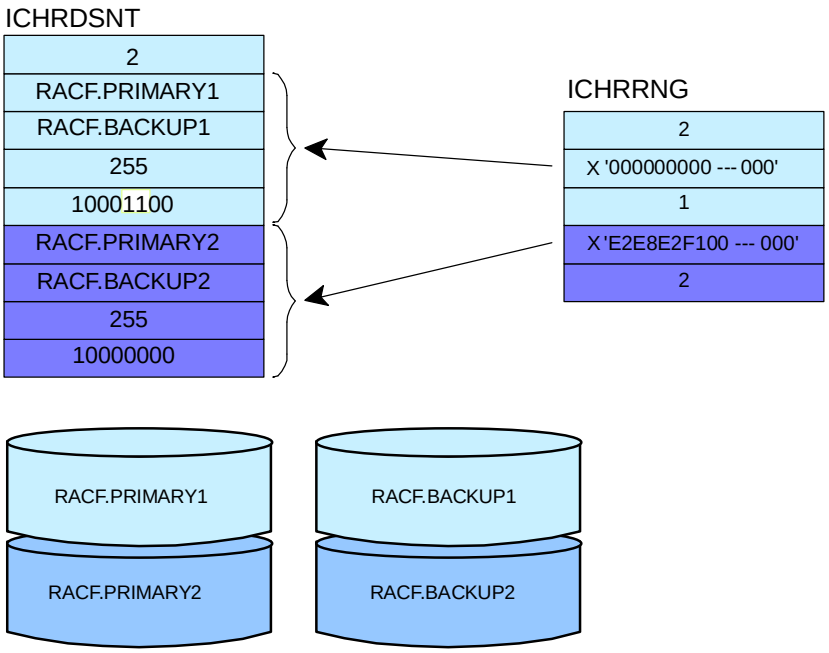
- RACF SDS enabled by two bits in ICHRDSNT.
- when bit x'_8' is set, RACF is enabled for SC, the SETROPTS/RVARY (those eligible) are propagated.
- when x'_4' bit is set RACF is enabled for SDS at IPL. Then uses the CF for structures.
- are in flag bits. refer to RACF SPG for detailed explanations of all bit settings.
- job in SAMPLIB (in mbr RACTABLE)has an example.

RACF Tables



- When in RACF SC (x'_8' it in flag byte of ICHRDSNT) causes propagation of certain SETROPTS and RVARY command.
- Use of CF not required for this capability.

Multiple RACF Databases



- When DB is split it is necessary for the previous bits to be in the first pair of DBs flag byte.

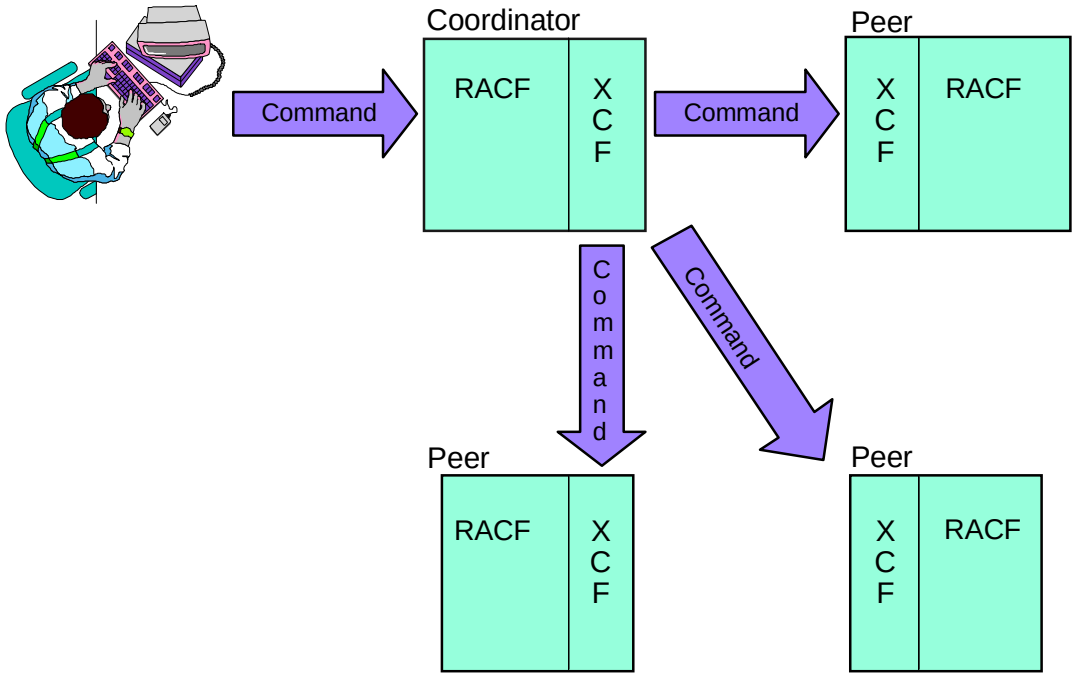
Command Propagation

RVARY
- ACTIVE
- INACTIVE
- SWITCH
- DATASHARE
- NODATASHARE

SETROPTS
- RACLIST
- RACLIST REFRESH
- NORACLIST
- GLOBAL
- GLOBAL REFRESH
- GENERIC REFRESH
- WHEN(PROGRAM)
- WHEN(PROGRAM) REFRESH

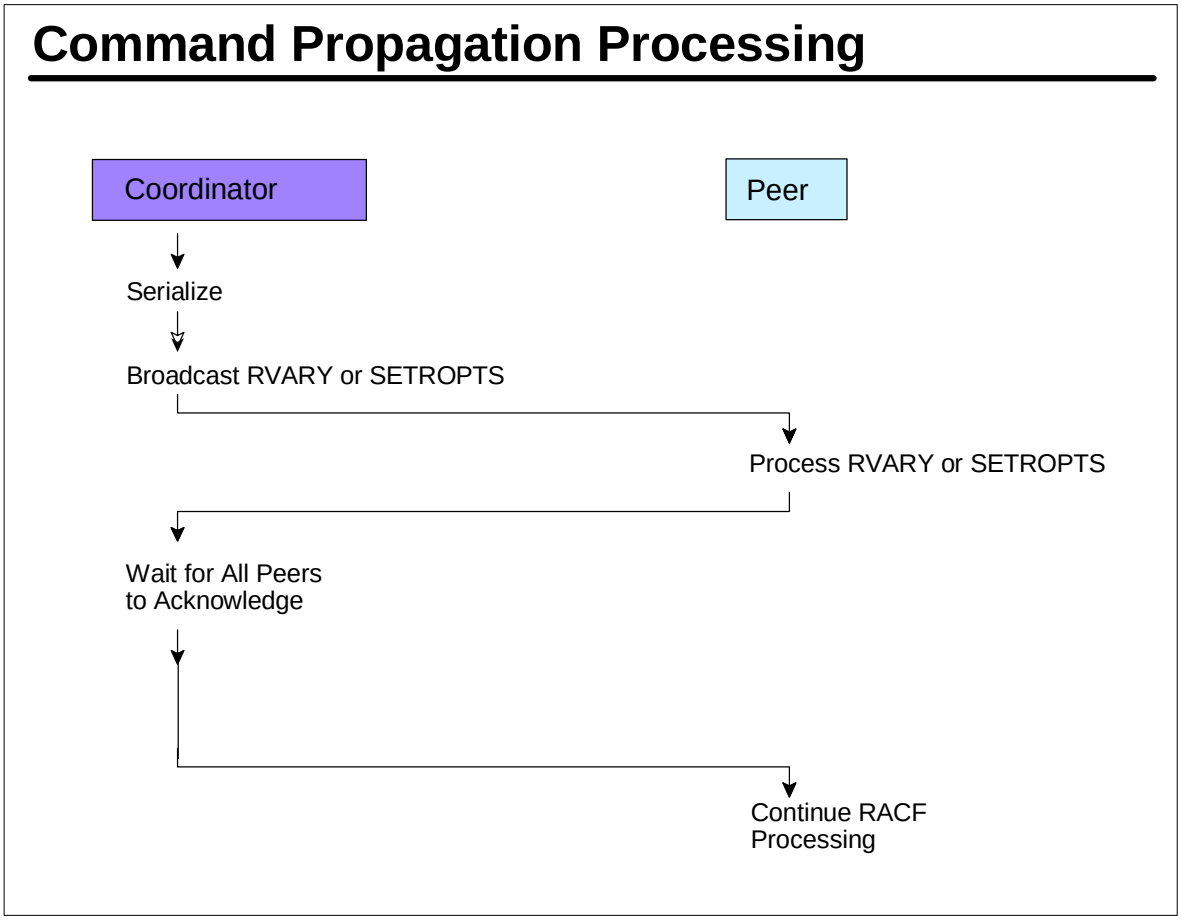
- Pre RACF SC capability most (but not all) comands were communicated to all systems sharing a DB.
- Above were exceptions.
- With RACF SC, using CF messaging services the above commands get propagated to all members. NO CF not required.
- THe x'-8' bit in flag byte of ICHRDSNT accomplishes this.

RVARY and SETROPTS Command Propagation



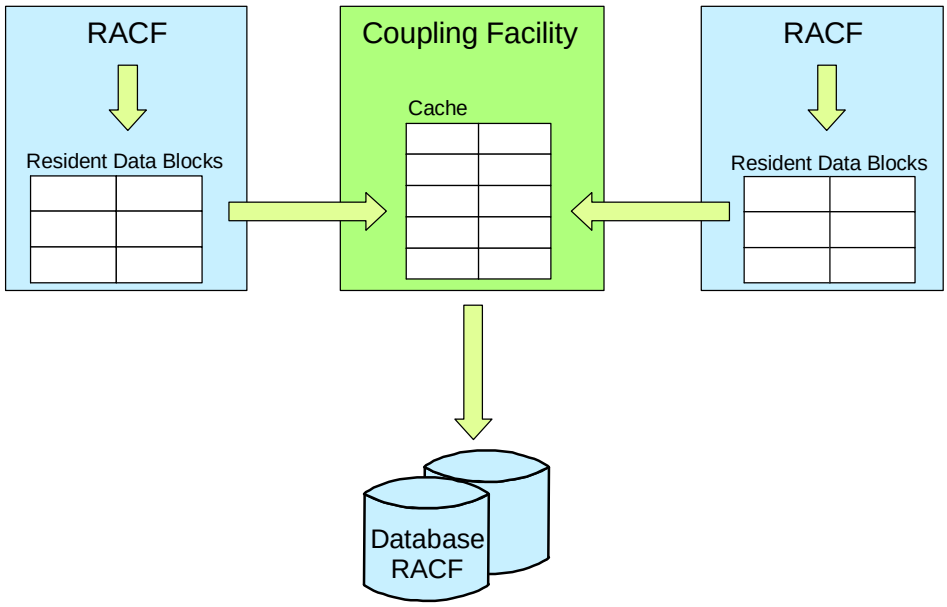
- Here we see a SETROPTS or RVARY entered on 1 system and being propagated to all members of the RACF DSG.
- System where entered called coordinator. Others are referred to as peers.
- For command propagation to occur.
  - ICHRDSNT must have x'_8' bit on
  - system must have joined RACF DSG - IRRXCF00
- CF not required

Command Propagation Processing

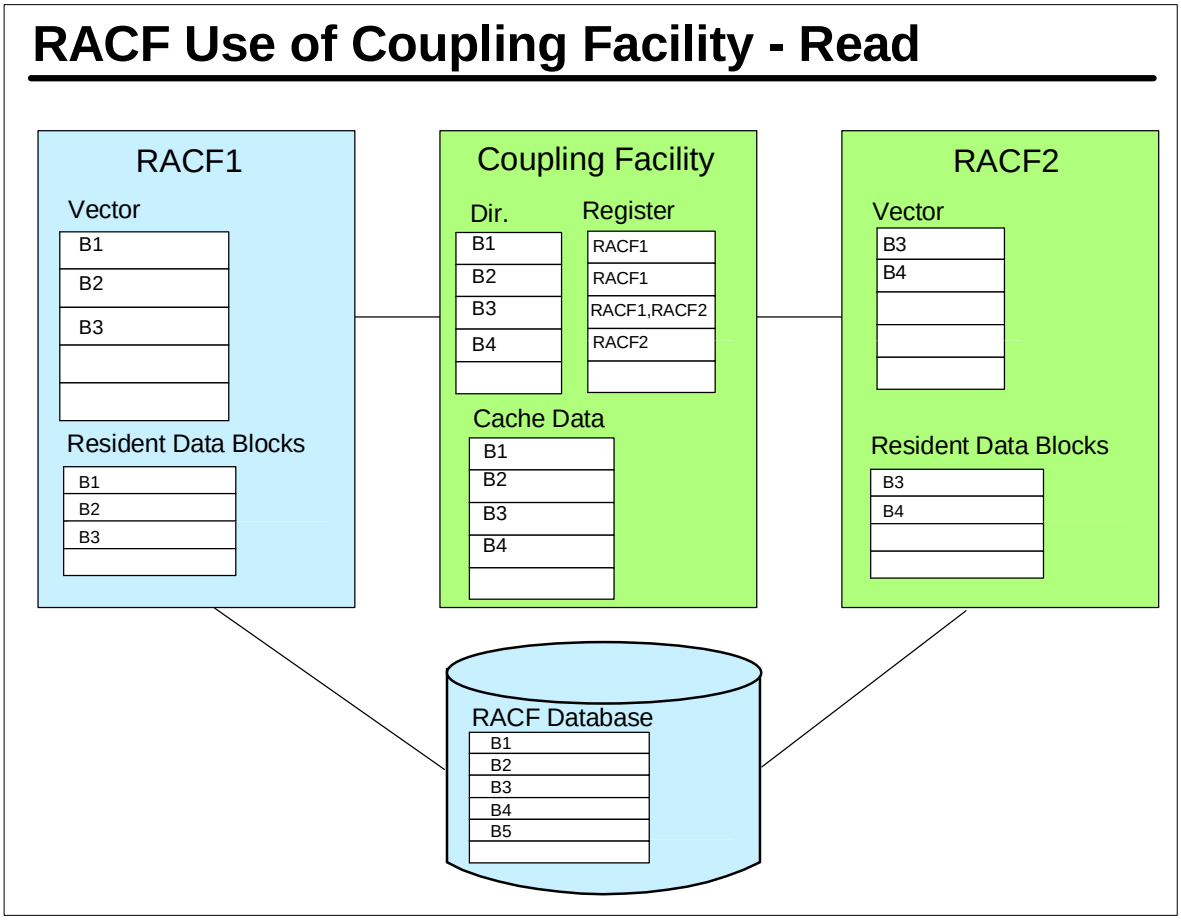


- commands propagated to all mbrs in a coordinated / syncronized manner. also insures security decisons are using same profiles for all members of plex.
- coordination done by enqueue (to prevent competing commands)
- once enq held, command is broadcast to all members.
- once all acknowledgements (done!) received, message sent to peers to continue processing.
- insure GRS handles SYSZRACF/SYSZRAC2 scope=systems correctly. IBM GRS, nothing required (insure not EXCL'd). MIM must be told explicity to do so.

Conceptual View of Cache

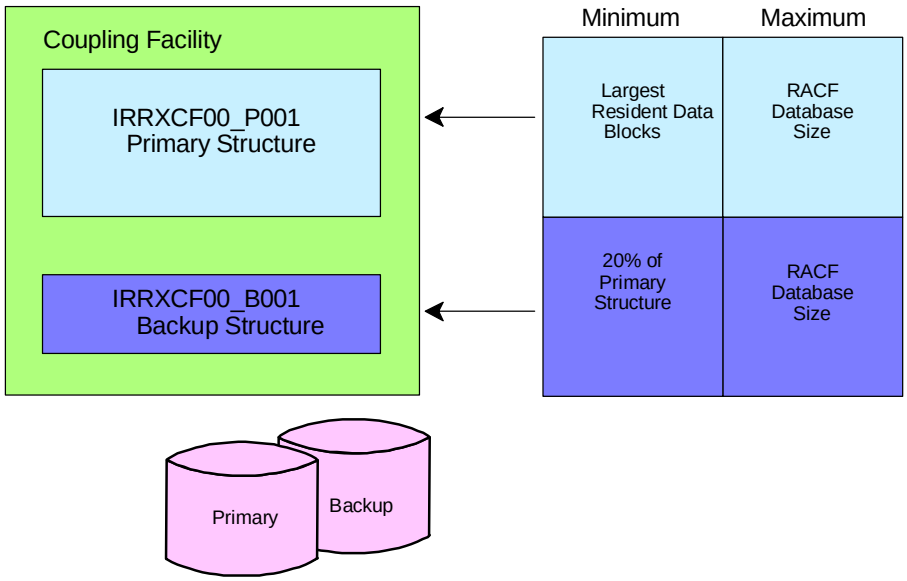


- to read a block, RACF first looks in resident data blks.
- If block is there, RACF uses it.
- If block is not found, RACF next looks in CF, and if there use it.
- if block is not in CF, then an I/O must be done.
- Simple concept, right? However how about updates fron another system? Serialization?



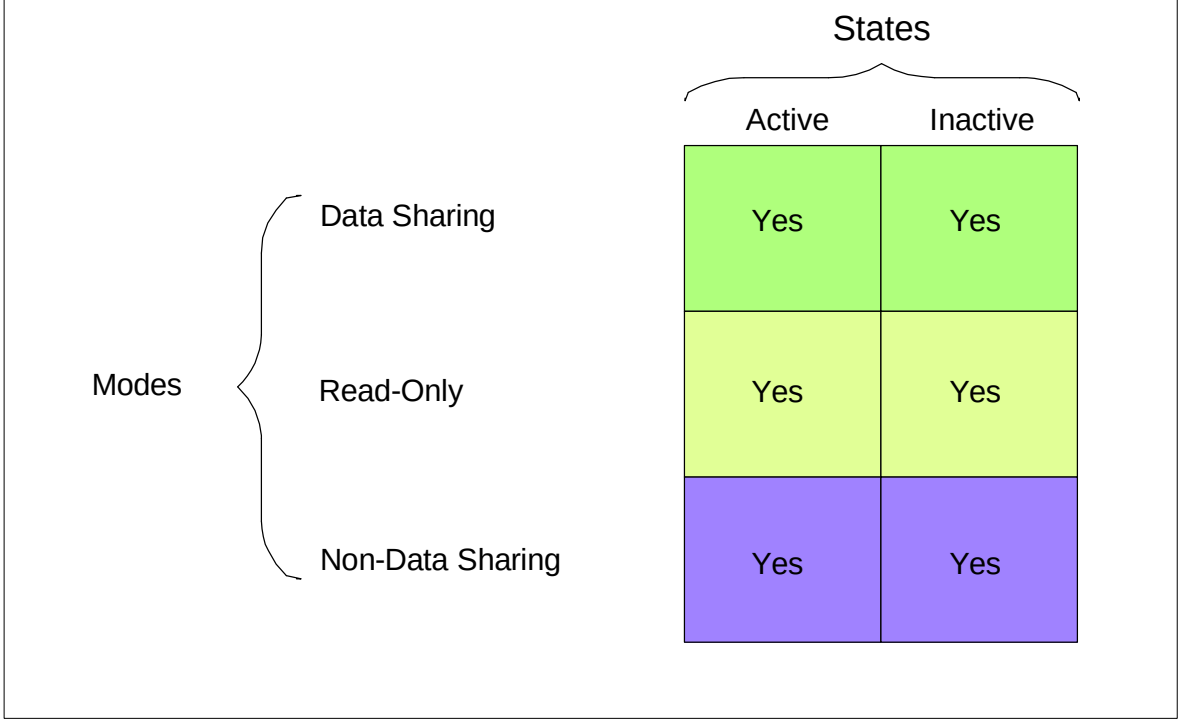
- RACF uses the CF to greatly speed up read operations.
- without RACF SDS, invalidation is done for an entire type of block such as all level1 indices or all profile blocks. A great benefit of RACF SDS is that blocks in the resident data blocks are invalidated on a block by block basis. This improves the efficiency of resident data blocks. There is a validity vector that is a string of bits that are used to indicate whether a block is valid or invalid.
- The CF as a directory that points to each block in the cache. A register keeps track if which RACFs have a copy of a block in their resident data blocks.
- If a block isn't in a system's resident data blocks, then RACF looks in the CF. Only if not there is an I/O required.

Size of Coupling Facility Cache



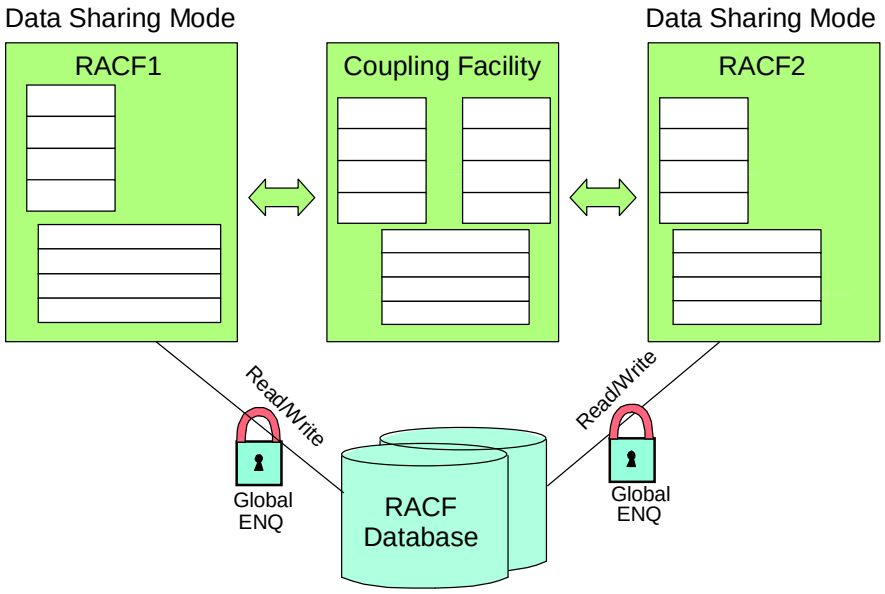
- THere must be a structure for each peice of the RACF DB. Example a two way split (2 primaries / backups) must mean total of 4 structures defined.
- Miminum size (primary) is the size of the largest local buffer (resident data blocks) of any of the RACFs in the DSG, but not less than 50 blocks. Max size is the size of the RACFDB.
- Minumum size of backup RACF DB structures is 20% of the primary DB's, but not less than 10 blocks. Max is gan the size of the RACF DB.
- The more the cache the better the probability of finding a block in the cache. Later we well look at a formula to calculate a suggested minumum size for the RACF cache.

Operational States and Modes



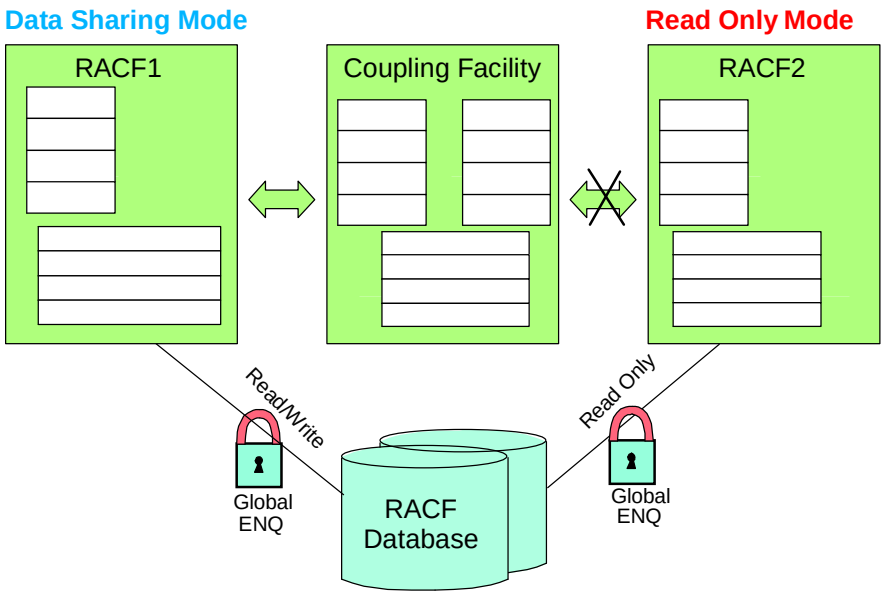
- With RACF SDS there are three operational modes.
  1) datasharing
  2) read-only and
  3) non-data sharing.
- These modes have meaning only for systems that use RACF SDS and do not apply to systems NOT configured for RACF SDS.
- The term mode applies only to DS, RO and NDS sharing operational modes.
- The term STATE applies to whether RACF is active or inactive. RACF can be in any of these modes/states.
- For a system that has RACF confugured for DS, DS is the normal mode of operation. The RO and NDS modes are for recovery from problems associated with the CF.

Data Sharing Mode



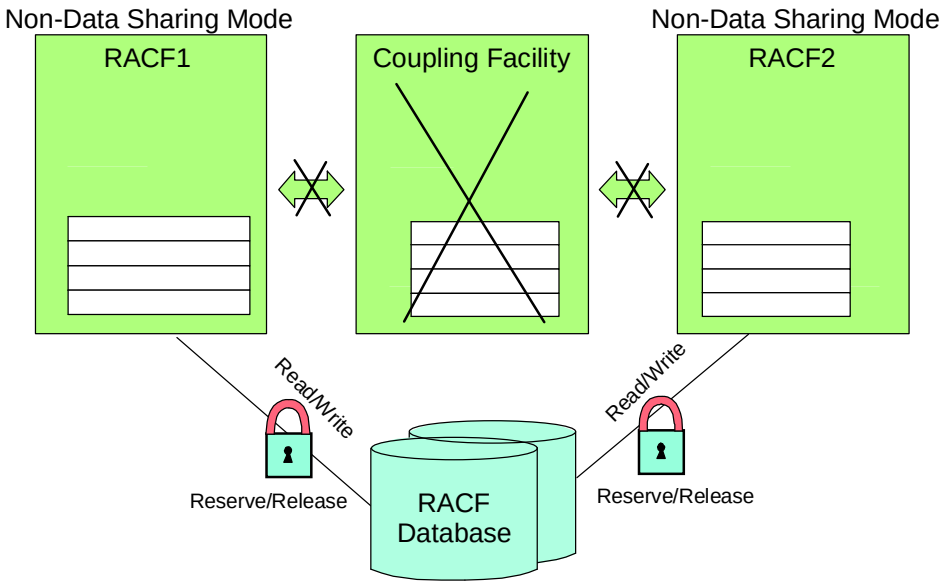
- Data sharing is the normal mode of operation when RACF is enabled for DS by setting the x'_8' and x'_4' bits in the flag byte of ICHRDSNT.
- When the system is IPLed, RACF attempts to connect to the CF structures for the RACF datasets. If no problems are encountered then RACF enters data sharing mode. Each member in the DSG goes through the same process,
- When operating in DS mode:
  - RACF exploits the CF to cache the RACF DB.
  - RACF uses GRS global enqueues to serialzie access to the RACF DB

Data Sharing and Read-Only Modes



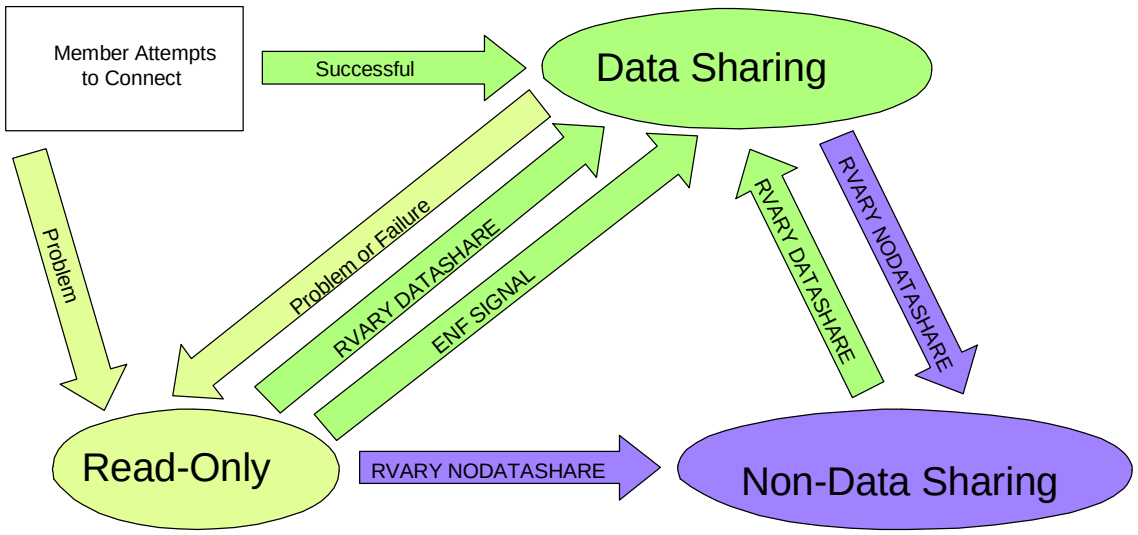
- There is NO command to cause RACF to enter RO mode. RO mode is entered when RACF cannot connect to the CFcache structure. Some reasons could be:
  - CF isn't available
  - CF channel is down.
  - Structure size in the MVS policy is less than required size.
  - CF structure constraints such that there is insufficient storage to allocate the structure.
  - Structure is currently in a rebuild state.
- In some situations RACF will automaticaly redrive the connection when problem is cleared up. It knows to do this when it receives ENF (event notification facility) signal that the CF is now available.
- GRS is used for serialization in RO mode. While in RO mode RACF can function, but it cannot update the RACF DB. However, it can do reads and statistical updates to the DB.
- Note: READONLY can apply to only ONE LPAR, while others

Non-Data Sharing Modes



- Non datasharing mode is another recovery mode
- NDS mode can be entered via the RVARY NODATASHARE command on one of the RACF mbrs of the RACF DSG. This command (when reposnded to) is then propagated to all mbrs so they all enter NDS mode.
- If one is in NDS mode then all must also be. It isn't possible to have 1 system in NDS mode while others in the DSG are in DS or RO mode. This is due to the difference in serialization (NDS uses RESERVE/RELEASE).
- Perhaps the main reason to issue RVARY NODATASHARE is to place the DSG in that mode due to one (or more mbrs) being in RO mode and a critical update cannot occur. Once the update is done and /or the reason for the RO mode switch (CF failure?) is past the RVARY DATASHARE can be issued. IF the CF is still not available applicable members will enter RO mode.

How Do Modes Change?



- **Summary of the ways RACF can move from one operational mode to another.**
- **DS mode is normal mode of operations for systems configured for RACF DS. RO ad NDS modes are considered for recovery.**
- **At IPL, all mbrs of the RCF DSG attempt to connect to the structures in the CF. (remember the x'_8' and x'_4' bits in the flag byte of ICHRDSNT. When a mbrs connects it enters DS.**
- **If a mbr cannot connect it enters RO mode. In this mode RACF monitors ENF signals and if pissible will redrive the connect to enter DS mode.**
- **Remember, there is NO command to cause RACF to enter RO mode. RO mode is enter only when RACF has a problem connecting to cache structures.**
- **Another way to move from RO mode to DS mode is to isue the RVARY DATASHARE casuing RACF to attempt a connect to the CF.**
- **Note that NDS mode is entered by RVARY NODATASHARE CMD. RVARY DATASHARE is used to move from NDS (or RO) mode to DS mode.**
- **Your operations staff needs to understand these operational modes and the appropriate use of RVARY DS / NDS.**

New topics: 1

z/OS V1R10 has support to help STOP Database corruption - see old SUG APAR OW52482

uses GXFACILI profiles. See 1.10 SPG section title

4.10.1.3 Guarding against data corruption resulting from incorrect database sharing

- Support was added due to an increasing number of caes where customers inadvertantly (and incorrectly) shared a DB inside and outside a CF.
- all cause by a simple snafu as regards 1 bit in ICHRDSNT.

New topics: 2

Better cross lpar cache invalidation - started in z/OS V1R2
- uses (only) SYSPLEX Comm

w/o this support CROSS lpar VLF cache purge is possible
big. with it, it is more 'surgical'..

- Picture DB shared between LPAR1 and LPAR2. Admin connectes USERX to a new
- group. They do this command from lpar1.

before this support

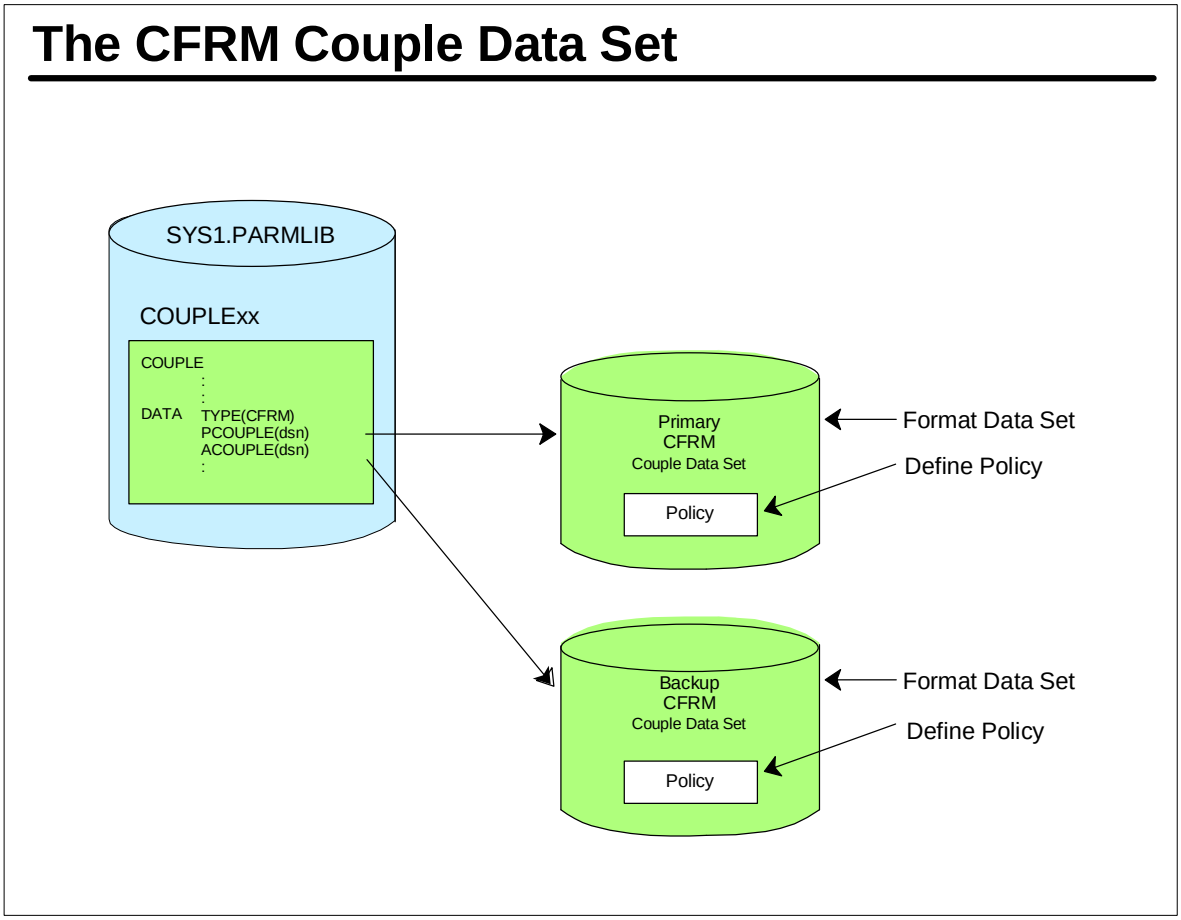
- USERX's vlf cache entry deleted on LPAR1. LPAR2's entire IRRACEE vlf cache
- is purged,

after this support

- On both lpars, only USERX's acee in VLF is deleted.

- This is a good performance improvement

The CFRM Couple Data Set



- The coupling facility policy resides in the coupling facility resource manager (CFRM) couple data set.
- That dataset is pointed to by the COUPLExx member of SYS1.PARMLIB. It contains:
- PCOUPLE(dsn) specifies primary CFRM cds.
- ACOUPLE(dsn) specifies alternate CFRM cds
- Utilize the IXCLIDSU to format a CFRM cds (if one doesn't already exist). The add the policy via the IXCMIAPU utility

Formatting The CFRM Couple Data Set

```
//FORMAT       JOB     MSGCLASS=H,NOTIFY=JJONES,MSGLEVEL=(1,1)
//STEP         EXEC    PGM=IXCL1DSU
//STEPLIB      DD      DSN=SYS1.MIGLIB,DISP=SHR
//SYSPRINT     DD      SYSOUT=  *
//SYSIN        DD *
//
//


  DEFINEDS     SYSPLEX(PLEX1)
               DSN(CFRM_couple_dsn)  VOLSER(vvvvvv)
               CATALOG


  DATA TYPE(CFRM)
          ITEM  NAME(POLICY)  NUMBER(3)
          ITEM  NAME(CF)  NUMBER(2)
          ITEM  NAME(STR)  NUMBER(6)
          ITEM  NAME(CONNECT)  NUMBER(6)


/*
```

- Before we can define the CF policy, we need (possibly) to format the CFRM dataset.
- We use the IXCL1DSU utility to format it specifying such info as:
  - Max number of policies.
  - Max number of CFs
  - Max number of structures
  - Max numbr of mbrs that connect to the CF
- Once the CFRM dataset has ben formatted it's ready for policy definitions.

- For additional info, refer to MVS/ESA PLANNING: SYSPLEX MANAGEMENT, GC28-1449

Coupling Facility Storage Calculation

Suggested Minimum Starting Point:

Primary Structure Size = (RDB x 4K) + (.1 x RDB x N x 4K)

Backup Structure Size = (.2 x Primary Structure Size)

Where: RDB = Largest Number of Resident Data Blocks
N = Number of Systems in Sysplex

Example For 16-Way Sysplex

Primary Structure Size = (255 x 4K) + (26 x 16 x 4K)
= 2684K

Backup Structure Size = (.2 x Primary Structure Size)
= 537K

- Minimum size for primary cache size is 50 blocks or the largest number of resident data blocks of any member of the plex, whichever is larger.
- Minimum size for backup cache is 10 blocks or 20% of the primary cache, whichever is larger.
- A recommended mimimum starting point is the size of the largest resident data locks plus 10% of the sum of the resident data blocks of all the systems in the DSG. This formula is based upon the assumption that most of the cache is being accessed by all the systems and some data (10%) is being accessed by only one of the systems at any point in time.
- As actual experience is gained, this starting point will probably be adjusted upwards as needed for performance as as allowed b-y CF storage constraints.

Defining The CFRM Couple Data Set

```
//DEFPOL        JOB         MSGCLASS=H=JJONES,MSGLEVEL(1,1)
//             EXEC        PGM=IXCMIAPU
//SYSPRINT      DD          SYSOUT=*
//SYSIN         DD          *
DATA  TYPE(CFRM) REPORT(YES)
DEFINE  POLICY  NAME(POL1)  REPLACE(YES)
   CF  NAME(FACIL01)    TYPE(009674))      MFG(IBM)     PLANT(PK)
                        SEQUENCE(0000040021)       PARTITION(1)
                        CPCID(00)     SIDE(1)   DUMPSPACE(2000)
   CF  NAME(FACIL02)    TYPE(009674))      MFG(IBM)     PLANT(PK)
                        SEQUENCE(0000040022)         PARTITION(1)
                        CPCID(00)  SIDE(1)  DUMPSPACE(2000)
   STRUCTURE  NAME(IRRXCF00_P001)
              SIZE(2688)  PREFLIST(FACIL01)
   STRUCTURE  NAME(IRRXCF00_B001)
              SIZE(538)  PREFLIST(FACIL02)
   /*
```

- After the CFRM cds is formatted (or if it pre-existed) we then use the CFRM policy utility (IXCMIAPU) to define our cache structures.
- One structure is required for each RACF dataset. For example a shop with 2 primaries and 2 backups would require 4 structures.
- The example shown here has just 1 policy. Several policies can exist in the CFRM cds, but only one can be active at a time.
- Code structure statement for each structure. Note the naming convention.

- IRRXCF00_P001 cache structure for first primary
- IRRXCF00_B001 cache structure for first backup
- IRRXCF00_P002 cache structure for second primary
- IRRXCF00_B002 cache structure for second backup

Unit Summary

- The Sysplex Environment
- RACF Sysplex Communication
- RACF Sysplex Data Sharing
- Recovery Modes
- Defining the Coupling Facility Policy for RACF

- With sysplex communications, RACF will propagate the RVARY and certain SETROPTS commands to all members of the sysplex.
- Installations that have a CF can implement RACF SDS to cache the RACF database.
- RACF Remote Sharing Facility (RRSF) can be used to ease the migration to sysplex, as it allows to logicaly share a RACF database as we move systems one at a time to the parallel sysplex environment.
- Systems that cannot connect to the CF cannot share a RACF DB with systems that are using RACF SDS. However by using RRSF, these separate systems can use a RACF DB kept in sync with the sysplex's RACF DB.