z/OS 1.8 Coupling Facility Resource Management

Performance Enhancements

IBM

Riaz Ahmad
z/OS & Sysplex Technical Support
IBM Washington Systems Center
z/OS 1.8 Coupling Facility Resource Management Performance Enhancements

The ability to have surviving systems recover rapidly for a failed system, and to recover coupling facility (CF) structures after a failure are key elements to achieving high availability in the sysplex environment.

Enhancements to the sysplex subcomponents (XCF/XES) were made in z/OS 1.8 to improve availability and recoverability in a sysplex. This paper describes the issue being addressed, the solution provided, changes in system operation, as well as installation and migration considerations.

The enhancements to the Coupling Facility Resource Manager (CFRM) are intended to improve performance across all products and subsystems using coupling facility resources. Overall, z/OS and Parallel Sysplex availability is improved by shortening the time it takes to recover from failures, especially during periods where sysplex partitioning, structure rebuild, and structure duplexing failover are being processed.

These enhancements provide constraint relief and enable workload growth in the sense that given an existing availability objective, a significant decrease in recovery time allows the sysplex to support more systems, structures, and connectors and still meet the existing objective.

The Coupling Facility Resource Manager active policy defines and maintains status for all coupling facilities, structures, and connectors in the sysplex. The active policy resides in a CFRM Couple Data Set (CDS) accessible from every system in the parallel sysplex. Updates to the policy must be serialized to preserve the integrity of the data. If two systems need to update the policy, the one holding the policy lock can proceed, while the other must wait to obtain the lock before making updates. Contention for the policy lock can therefore cause delay.

The active policy is updated as applications connect to, or disconnect from a structure. Peer application connectors must be informed of these changes as well as other events, and the policy is updated to reflect the status of these notifications for each connector. Actions such as structure rebuild or duplexing failover require multi-phase processes executed in a coordinated fashion throughout the sysplex. These processes update the active policy to record progress and the status of who has started and completed each phase.

In the current implementation, each system must read the policy independently to discover what needs to be done. The policy may need to be read and/or written one or more times as a system performs the necessary processing.
The Issue Being Addressed

When a system fails, CFRM must inform all the surviving connectors about the demise of their peer connectors residing on the failed system. There could be one or more failed connectors for every structure, and there may be surviving connectors on every system in the sysplex. As the number of structures, systems, and connectors increases, the number of connectors needing to be notified will increase.

When a coupling facility fails, all the structure instances in the coupling facility are deemed to have failed. A variety of recovery actions are possible, depending on such things as whether the structure is duplexed and whether user-managed or system-managed rebuild processes apply. But in general, every connector to every failed structure would receive some sort of notification, and most every structure would likely undergo some form of rebuild or duplex fail-over to recover the structure. As the number of structures in the coupling facility increase, and as the number of systems and connectors to those structures increases, the number of connectors to be notified and the number structures needing to be recovered will increase.

The connector notifications as well as the processes used to recover the failed structures all require updates to the CFRM policy. As the number of systems, structures, and connectors increase, so will the accesses to the CFRM policy. Since every system in the sysplex will likely be performing the relevant burst of processing in parallel, the likelihood for policy lock contention and the resulting delay will increase as well. It can take many minutes for the sysplex to complete the failure processing.

The Solution

The policy-based protocols are optionally replaced by message-based protocols which avoid the use of the CFRM CDS as much as possible for those functions that have been bottlenecked by CFRM I/O in the past. The delivery of XES events and the collection of responses required to process these structure and connector events related to structure rebuild and duplexing failover will now use a message-based protocol to greatly reduce serialized access of the CFRM CDS.

In the old policy-based protocol, signals were simply directives telling participant systems to look in the CFRM policy to discover what needed to be done. In the new message-based protocol, signals have content. Thus participant systems can do what needs to be done without reading the policy.

In the old policy-based protocols, participant systems updated the policy to record the status of their processing. In the new message-based protocol, signals are sent to the manager system to indicate the status of their processing. The manager system coordinates these signals, and updates the policy on behalf of all the participants when all systems have responded. Thus “N” write operations are reduced to one. Having one
system perform the write also eliminates much of the policy lock contention. Signalling traffic is slightly increased, but couple data set contention is greatly reduced.

The message–based protocol allows the CFRM processing associated with system failure and CF failure to be accomplished without all the systems trying to access the CFRM CDS at the same time, thereby eliminating the bottleneck on the CFRM CDS thus reducing the elapsed time for recovery.

**Assigning Manager System**

As is the case for many sysplex processes, the system selected to be the manager system for message-based processing is generally the winner of a race. Such races can occur, for example, when systems IPL into the sysplex, when systems are removed from the sysplex, or when a message-based capable CFRM CDS is brought into use. When truly a race condition exists, the winner will tend to be systems that are good at getting to the CDS quickly (a very desirable attribute for the manager system).

There is a limited ability to influence the choice of manager system during planned reconfigurations by carefully controlling the order in which systems IPL into the sysplex, or which system handles removal of the manager system from the sysplex, or which system brings a CFRM CDS into use. For unplanned changes, and even for some planned changes, such orchestration is not plausible.

For some installations, concerns about workload imbalance, processor power, or other factors may motivate the desire to explicitly control the selection of manager system. This can be done, but only after the fact. Once a system becomes the manager, it stays the manager until it is either removed from the sysplex or message-based processing is stopped. If the manager system is removed from the sysplex, a surviving system will take over the role of manager system. The SETXCF STOP,MSGBASED command issued from any system which has access to the CFRM CDS will cause the sysplex to revert to policy-based processing. There is no manager system for policy-based processing. Issuing the SETXCF START,MSGBASED command on a system which has access to the CFRM CDS will cause the sysplex to start using message-based protocols. The system processing the SETXCF command will become the manager system. In order to control selection of the manager system, operational procedures may need to be updated and/or system automation provided.

**How to determine Manager System**

The output of the DISPLAY XCF,STR command indicates whether message-based protocols are being used, and if so, which system is the manager system. For the detailed structure display (shown below), there is also a line to indicate which protocol is being used for that particular structure. A structure could be using policy-based protocols even though the sysplex is capable of message-based protocols. The last line of the output indicates which protocol is being used by the sysplex and which system is the Manager System.
New message IXC548I is issued when a system becomes Manager System. The message IXC549I identifies the manager system if the sysplex is using message-based protocols.

```
D XCF,STR,STRNM=J2CKPT1
IXC360I 17.44.46 DISPLAY XCF 668
STRNAME: J2CKPT1
STATUS: ALLOCATED
EVENT MANAGEMENT: MESSAGE-BASED
TYPE: SERIALIZED LIST
POLICY INFORMATION:
  POLICY SIZE : 15000 K
  POLICY INITSIZE: N/A
  POLICY MINSIZE : 0 K
  FULLTHRESHOLD : 80
  ALLOWAUTOALT : NO
  REBUILD PERCENT: N/A
  DUPLEX : DISABLED
  ALLOWREALLOCATE: YES
  PREFERENCE LIST: CF1 CF2
  ENFORCEORDER : NO
  EXCLUSION LIST IS EMPTY

ACTIVE STRUCTURE
----------------
  ALLOCATION TIME: 10/26/2006 13:45:08
  CFNAME : CF1
  COUPLING FACILITY: 002084.IBM.02.000000023A6A
  PARTITION: 05 CPCID: 00
  ACTUAL SIZE : 15616 K
  STORAGE INCREMENT SIZE: 256 K
  ENTRIES: IN-USE: 2878 TOTAL: 3485, 82% FULL
  ELEMENTS: IN-USE: 2878 TOTAL: 3466, 83% FULL
  LOCKS: TOTAL: 2
  PHYSICAL VERSION: BF9CAA60 ED59F90A
  LOGICAL VERSION: BB632BC3 36CC995C
  SYSTEM-MANAGED PROCESS LEVEL: 8
  XCF GRPNAME : IXCLO006
  DISPOSITION : KEEP
  ACCESS TIME : NOLIMIT
  MAX CONNECTIONS: 32
  # CONNECTIONS : 4

  CONNECTION NAME ID VERSION SYSNAME JOBNAME ASID STATE
  ---------- ----- ------- ------ ------- -------
  JES2_SYSA 01 00010074 SYSA JES2 001C ACTIVE
  JES2_SYSB 02 0002008D SYSB JES2 001C ACTIVE
  JES2_SYSC 03 0003005A SYSC JES2 001C ACTIVE
  JES2_SYSD 04 0004007F SYSD JES2 001C ACTIVE

  DIAGNOSTIC INFORMATION: STRNUM: 00000006 STRSEQ: 00000000
  MANAGER SYSTEM ID: 0100047C
  NAME/MGR #QUEUED 1STQESN LASTQESN CMPESN NOTIFYESN
  SYSB 00000000 00000000 00000000 00000008 00000008

  EVENT MANAGEMENT: MESSAGE-BASED MANAGER SYSTEM NAME: SYSA
```
Migration

There are no hardware or application software dependencies for this function. All systems in the sysplex where this function is being deployed must be at z/OS V1R8 or later.

Only systems at z/OS V1R8 or later can use a CFRM CDS formatted to support message-based protocols. Once such a CDS is switched into use, down level systems will not be able to join the sysplex. Since a sysplex-wide outage is required to fall back to a CFRM CDS that does not support message-based protocols, it is advisable to delay using a message-based capable CFRM CDS until no system in the sysplex will need to run z/OS V1R7 or earlier.

Use the Couple Data Set format utility IXCL1DSU to format two new message-based capable CFRM CDS by adding the following statement as input:

```
ITEM NAME(MSGBASED) NUMBER(1)
```

Note this specification enables message-based support and also enables support for system-managed rebuild and system-managed duplexing as well.

Finally, switch the new CFRM CDS in. Remember to update COUPLExx member with the new CDS names!

Activating Message-Based processing

If an existing sysplex’s CFRM CDSes were formatted for message-based protocols, the sysplex will automatically activate the message-based capability when a message-based capable CDS becomes the primary CFRM CDS. From that point on the message-based capable CFRM CDS remembers whether message-based or policy-based processing is to be used. The protocol is retained across subsequent CDS switch processing that promotes an alternate CFRM CDS to become primary. When a new instance of the sysplex is IPLed with a message-based capable CFRM CDS as the primary the sysplex will use whatever protocol is recorded in the CDS.

If a message-based capable CFRM CDS is in use by the sysplex, the SETXCF command can be used to start or stop message-based processing in the sysplex. If stopped, the message-based CFRM CDS will indicate policy-based processing is to be used. This setting persists until the SETXCF START,MSGBASED command is issued to have the sysplex use message-based protocols.

After the message-based protocol is selected for CFRM as a whole, an individual structure may continue to use policy-based processing until in flight events are
completely processed. Structure displays (D XCF,STR,STRNAME=strname) will continue to show structure using the policy-based protocol until the next event occurs for that structure, which may be quite some time. So even if the sysplex is using message-based protocols, a particular structure could be using either protocol. A structure used for XCF Signalling always uses policy-based protocols.

**Fallback**

If for some reason you want to revert to policy-based protocols, issue the SETXCF STOP,MSGBASED command from any system in the sysplex with access to the CFRM CDS. The sysplex will *immediately revert* to the old policy-based protocols. The CFRM CDS is updated to indicate policy-based protocols are to be used. In particular, this means if the sysplex were to be re-initialized with this CFRM CDS, the sysplex would use policy-based protocols. After stopping the use of message-based protocols, the policy-based protocols will continue to be used until explicitly started again through use of the SETXCF START,MSGBASED command.

*Because of the SETXCF STOP,MSGBASED fallback capability, there should be no need to revert to a non-MSGBASED CDS, because the installation can revert to the policy-based processing without changing CDSes.*

CDS version checking will not allow down level systems (z/OS V1R7 or lower) to use a primary CFRM CDS formatted for message-based processing. Thus a CFRM CDS formatted for message-based processing cannot become a primary CDS for a sysplex while there is a down level system in the sysplex. Nor will a down level system be able to join the sysplex once a CFRM CDS supporting message-based processing becomes the primary. To enable a down level system to join the sysplex, one would have to get the sysplex to use a CFRM CDS that is *not* formatted to support message-based protocols. But once the sysplex starts using a CFRM CDS formatted with support for message-based protocols, it cannot revert to a CDS formatted without this support because ACOUPLE processing will reject it as unsuitable. The only way to revert to such a CDS would require a sysplex-wide IPL. Do not bring a message-based capable CDS into use until you are certain there is no need to IPL a down level system into the sysplex.
New and Changed Messages

Following new messages are introduced as part of this function:

**IXC547I** is issued to indicate whether the SETXCF STOP (or START) MSGBASED command completed or failed. For failures, the message explains why the request failed.

**IXC548I** issued when the sysplex changes CFRM protocols or changes the manager system when using message-based protocols. The system that updates the CFRM policy to record the change issues the message.

**IXC549I** is issued to indicate the current CFRM protocol being used. Each system issues this message when CFRM initialization is done (usually during IPL).

Following messages are updated as part of this function:

**IXC358I** was updated to report data provided for the CFRM CDS which may be formatted to support MSGBASED.

**IXC359I** reports summary information for structures and was updated to report the CFRM protocol in use by the sysplex and the manager system if using message-based protocols.

**IXC360I** reports detail information for a structure and was updated to report whether message-based or policy-based protocols are being used for the structure. It also reports the CFRM protocol in use by the sysplex and the manager system if using message-based protocols. There are also some diagnostic data if message based protocols are being used allowing IBM service personnel to obtain information about the state of message-based processing for the structure.