LPAR Performance Considerations for Sharing Resources in a Parallel Sysplex Environment

Introduction

There are several ways you can use LPAR to share resources in a parallel sysplex.

- You can divide a sending processor (CEC) into multiple MVS images.
- If you have multiple MVS images on a CEC, you can share Coupling Facility (CF) links using ESCON multiple image facility (EMIF).
- You can partition the Coupling Facility CEC into multiple CFs.
- You can combine the above options into a configuration which contains MVS images and Coupling Facilities on the same processor, using CF links or the Integrated Coupling Migration Facility (ICMF). (Note: ICMF is not support on zSeries processor)

This flash discusses these various options and the performance implications of each.

Multiple MVS Images on the Sending CEC

If you want to run multiple MVS images on one CEC, you can use LPAR to divide the CEC into logical partitions. The definition of these logical partition allocates system resources (like storage and central processors (CPs)) to each partition. Storage is dedicated to logical partitions. CPs can be assigned to a particular partition (i.e., dedicated) or they can be shared amongst various partitions. Generally, dedicated processor partitions require less LPAR dispatching overhead, resulting in higher internal throughput rate (ITR). However, for fluctuating workloads, sharing processor partitions can provide additional capacity for handling peak demands and result in better external throughput (ETR) and responsiveness.

If the logical partitions are using shared processors, you should look at the ratio of logical processors to physical processors. Larger numbers of logical processors require more LPAR management. As a general guideline, it is recommended the ratio of the total number of logical processors to the total number of physical processors be 2 or less. See the PR/SM Planning Guide (GA22-7123-13) Chapter 5 for additional information.

One of the implications of using shared CPs is a logical processor may complete its allotted time on the physical processor before the request to the CF completes. If this is a synchronous (SYNC) request, the actual request to the CF will complete, but the remainder of the processing will not complete until the next dispatch of the logical processor. The SYNC service time and delays to SYNC requests when no subchannel is available contain only the time the logical processor is actually dispatched on a physical processor. By reporting only the time used when the logical processor is dispatched, the SYNC times truly represent the amount of CPU capacity attributable to SYNC requests.

If an asynchronous (ASYNC) request is being processed when the time slice completes, it will resume with the next dispatch of the logical processor. The ASYNC service time and the ASYNC subchannel queuing time are measured as elapsed time (i.e. they include the time not

dispatched on a physical processor) since they contribute very little to CPU capacity and are used primarily as an indicator of time to complete the function. So ASYNC times in a logical partition which shares CPs will be longer than those observed in a logical partition with dedicated CPs.

Shared processor resources can be divided on either an event-driven or a time-driven basis. The method is selected by setting Wait Completion to either Yes or No (the Wait Completion entry field is displayed after selecting SET Running Time on the LPCTL frame). If you selected a time-driven basis (WAIT COMPLETION set to YES), the logical partitions are considered to be busy 100% of their time slice. Using event-driven dispatching (WAIT COMPLETION set to NO) is recommended because it results in improved responsiveness and the ability to dispatch other work on the physical processor when the logical processor loads a wait state.

MVS images which are not part of the parallel sysplex should not be connected to the coupling facility. If they are connected, some connectivity checking is done on a periodic basis, even if these systems are not using the CF or there are no structures on it. The extent of the connectivity checking was reduced as described in APAR OW15130, but you can avoid it entirely by varying the CF links offline for those partitions which are not part of the parallel sysplex.

Shared (EMIFed) CF Links

If there are multiple MVS images on a CEC, they can share the CF sender links (known as CFS links) from the CEC to the CF. Receiver links (known as CFR links) cannot be shared. When the CFS links are shared, requests from multiple MVS images can occur at the same time, resulting in a "Path Busy" condition. (The number of "Path Busy" conditions encountered is found on the RMF CF SUBCHANNEL ACTIVITY report.

To understand the performance impact of encountering a "Path Busy" condition, it is necessary to back up a bit and look at how requests are sent to the coupling facility. Each MVS image has either 2 subchannels (non-peer mode) or 7 subchannels (peer mode) for each CFS link. MVS gives the request to the subchannel and sends it across the link to the CF. If a subchannel is not available, ASYNC requests are queued; SYNC non-immediate requests are CHANGED to ASYNC requests; and SYNC immediate requests, (like Lock requests), spin until a subchannel becomes available.

In an environment where CFS links are not shared, obtaining a subchannel insures the request will reach the CF with no further delays, barring errors like IFCC or loss of connectivity. If the CFS link is shared, when XES attempts to send a request to the CF, it can encounter a "Path Busy" condition. The performance impact of encountering this condition depends on the type of CF request. If a SYNC request encounters this condition, it will "spin" until the path is available. This "spin" time is included in the SYNC service time; it is not reported separately. If an ASYNC request encounters a “Path Busy” condition, the request is returned and it goes back through the process of obtaining a subchannel. If no subchannel is available, the request could get queued again.

You can obtain an indicator of the number of times this redrive is occurring by looking at the CF Subchannel Activity report. The difference in #REQ TOTAL (the number of requests sent) and

the sum of #REQ SYNC, ASYNC and CHANGED (number of requests completed) is partially due to redrives.

As the number of "Path Busy" conditions increases, the response time of individual requests increases non-linearly. Because each request takes longer to complete, it is more likely the incoming requests will encounter a busy subchannel and these requests may also be delayed or queued. As a guideline, it is recommended no more than 10% of the total requests be delayed for “Path Busy”. If you are approaching this total consider dedicating CFS links or adding additional CFS links.

Multiple Coupling Facilities on Receiving CEC

It is possible to define more than one Coupling Facility on a receiving CEC. For availability reasons, it is not recommended to define multiple production coupling facilities for the same sysplex on the same CEC. However, there are cases where one of the CFs is used for a separate production sysplex or for test purposes. It is important to point out if you define multiple CF LPARs on the same CEC, those CFs are by definition all running the same LIC, thus the same Coupling Facility Control Code (CFCC) and the same CFLEVEL. For example, for a migration scenario you may want one of the LPARs to be CFLEVEL=1 and the other LPAR on the same CEC to be CFLEVEL=2, but it is not possible. If you define multiple CFs on the same CEC, there are a few things you should do to prevent the additional CFs from impacting the performance of your production CF.

If there are multiple CFs on the CF CEC, it is recommended to dedicate CPs to each CF. The CFCC code in the coupling facility partition uses an "active wait" polling algorithm, continuously looking for work. This means the CF partition is continuously busy (if you looked at a SAD display on a CF, you will see it is always 100% busy). The primary motivation for sharing CPs in PR/SM is to allow another partition to be dispatched on this CP when there is no work. With the active wait polling loop this never happens.

If the CF partition shares processors with another CF partition, both will be dispatched for the full duration of their allotted time (as defined by the LPAR WEIGHT and CAP parameters). So a test CF which shares CPs with a production CF can drain valuable resources from the production CF, even if the test CF is doing little or no work. This situation can be controlled via the new Dynamic CF Dispatching. For more information on Dynamic CF Dispatching review:

- WSC Flash W9731A: MVS/ESA Parallel Sysplex Performance: Dynamic CF Dispatching,
- WSC Flash W9846: 9672-R06 Performance: Dynamic CF Dispatch Default set to Enabled

Another new feature, called Dynamic ICF Expansion, is also available to provide additional CF configuration options. For more information on Dynamic ICF Expansion please see:

- WSC Flash W9828: Dynamic ICF Expansion.

To verify the number of CPs assigned to a particular CF, you can look at the SMF 74 subtype 4 records or the RMF CF Usage Summary report will report the number of logical processors defined for this CF and the effective logical processors - the actual number of total processors available to this partition. The effective number of logical processor can change because the

CFs weight does not make a full engine available or the CF partition is defined with dynamic CF dispatching active.

Below is an example of a RMF CF Usage Summary Report (Processor Summary section) which shows the CP assignment. In this case 1 logical processor is assigned, but the LPAR weight is set so the CF is eligible to receive 50% of the processor, hence the effective weight is equal to ½ of the CP.

```
  PROCESSOR SUMMARY
 ------------------------------------------------------------------------------------------------

COUPLING FACILITY         9672    MODEL R06   CFLEVEL 10
AVERAGE CF UTILIZATION  (% BUSY)      7.8   LOGICAL PROCESSORS:   DEFINED  1    EFFECTIVE  0.5
```

Using the above RMF report as a base review the next example of a CF Processor Summary Section. In this example, which can be viewed as the next interval, the Effective time has drop to 0.1 or just 10% of the defined processor. This can happen when the CF partition is set with dynamic CF dispatching. As the request rate to the CF drops, LPAR dynamically adjusts the amount of weight used by the CF. This helps to reduce the impact of the active polling loop on the CEC.

```
  PROCESSOR SUMMARY
 ------------------------------------------------------------------------------------------------

COUPLING FACILITY         9672    MODEL R06   CFLEVEL 10
AVERAGE CF UTILIZATION  (% BUSY)      1.3   LOGICAL PROCESSORS:   DEFINED  1    EFFECTIVE  0.1
```

MVS Images and CF on the Same CEC

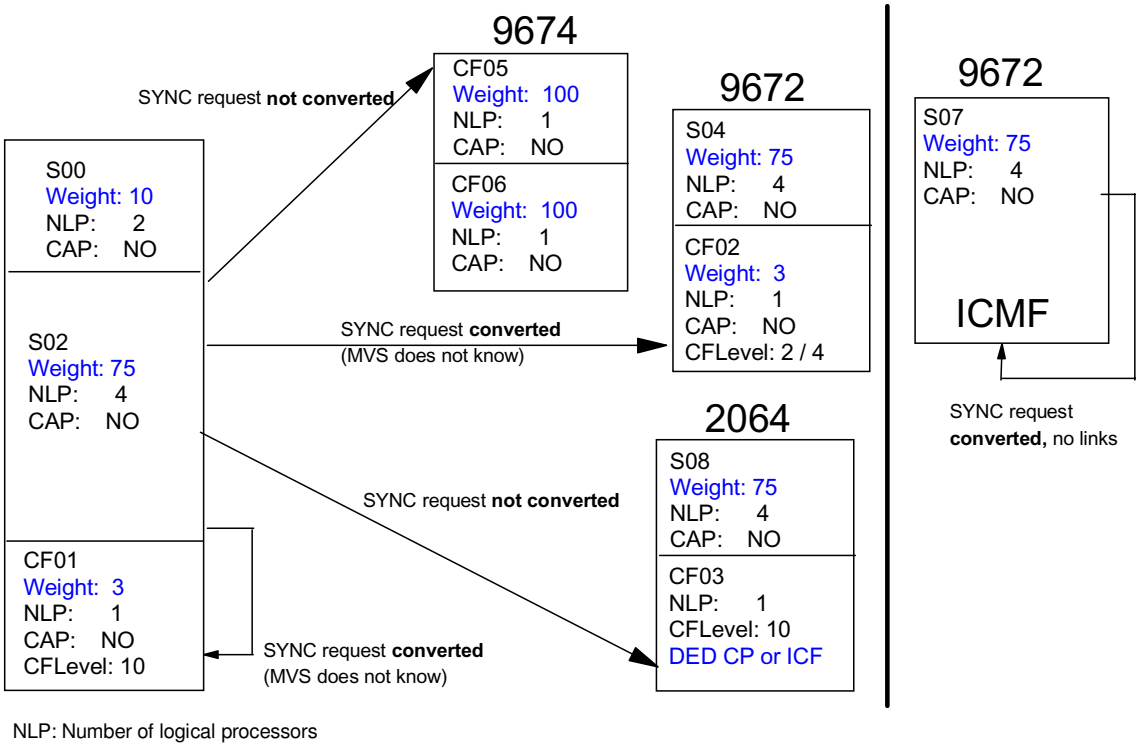
You can configure a CEC with MVS images in some logical partitions and a coupling facility in another partition. The MVS images can share processors with the CF partition. They can be linked to the CF with real CF links or simulated CF links known as ICMF. While the CFS links from the MVS images to the CF can be shared, the CFR links on the CF cannot be shared. This is usually a migration scenario, although it has been suggested as method of obtaining a low cost backup CF. The feasibility of a low cost backup will be discussed later.

While sharing processors between MVS images and the CF partition is technically feasible, there are significant performance implications. Understanding the CF configuration is critical to understanding if the performance impacts apply. If the coupling facility using CFR links shares SCP processors with **ANY** other partition, then in certain situations all request from the MVS image will be treated as ASYNC requests. The conditions are as follows:

- The MVS Image shares an SCP engine with a CF partition, in the same CEC, and there is no ICF defined to the CF partition.
- The MVS Image shares an SCP engine with a CF partition in the same CEC, and there is no ICF defined to the CF partition, and the request is sent to another CF partition, which shares CPs with an MVS image, and there is no ICF in the remote CF partition.

All other cases, including a case where the remote target CF is a stand-alone CF, with multiple LPAR partitions, sharing engines, the requests will NOT be converted to ASYNC. It is important to realize the RMF reports will not reflect the SYNC to ASYNC conversion. This

change in the method of CF operation is done without the awareness of MVS. The only symptom of this conversion is increased SYNC service times and high standard deviations. To help understand the rules followed for converting SYNC requests review the following diagram:



This ASYNC conversion is necessary to prevent a lock out situation where a sending MVS image will not give up control because it has issued a SYNC request and the CF cannot get control to process it. This change from SYNC to ASYNC is transparent to the MVS image (and to the person reading RMF reports). It is manifested in increased SYNC service times (which are now really ASYNC requests).

Allocating resources to the CF partition in this configuration also presents some serious performance tradeoffs. As described earlier, the CF partition is running in an "active wait" continuously searching for work, so it will uses as much CPU resource as it can obtain. If you assign a whole processor to the CF partition, either by dedicating a CP or sharing a CP and giving it a weight proportional to the size of one CP, the CF partition will use all of this resource. This may be more resource than you wish to allocate to this function. If on the other hand, you conserve processor resource by assigning a portion of a processor to the CF partition, (either by capping it or giving it a low weight), you will experience degraded response time and lower throughput. This is because the MVS image issuing a request to the CF will have to wait until it is time to dispatch the CF partition, (a maximum of 100 milliseconds if the partition is not capped), before the request can be completed. While this may be acceptable for a test configuration, it will not give satisfactory performance for the production environment.

Giving a low weight to the CF partition may work until contention for CPU resource increases. This is an example of part of the RMF CPU Activity report for an LPAR environment. (Some of the columns have been deleted to get an 80 column example).

```
                P A R T I T I O N   D A T A   R E P O R T

            MVS/ESA                     INTERVAL 000.15.00
            SP5.2.0                     CYCLE 0.500 SECONDS

 MVS PARTITION NAME                 CH1PROD
 NUMBER OF CONFIGURED PARTITIONS          6
 NUMBER OF PHYSICAL PROCESSORS            7
 WAIT COMPLETION                         NO
 DISPATCH INTERVAL                   DYNAMIC

     ---- PARTITION DATA ----     ---   -- AVERAGE PROCESSOR UTILIZATION PERCENT  ---
                                 # OF   -- LOG. PROC --    --- PHYSICAL PROCESSORS
       NAME     STAT  WGT  CAP   LPs    EFFECT.    TOTAL    LPAR MGMT  EFFECTIVE  TOTAL
     PRODAA       A    20   NO     2      4.51      5.09       0.17       1.29      1.46
     PRODBB       A    75   NO     7     47.03     47.34       0.32      47.03     47.34
     CFPART       A     3   NO     1     93.60     93.69       0.01      13.37     13.38
     TESTPART     D
```

Notice the CF partition, even though it has been given a weight of 3, (3/98 or 3% of the entire CEC), is actually using 13.38% of the entire CEC, (or almost an entire CP). This extra processing is possible because the MVS images did not require the extra cycles.

Note however this next example where the MVS LPARs get busier. As the activity in the other partitions increases the CF gets less of the CP resource.

```
    ---- PARTITION DATA ----     ---   -- AVERAGE PROCESSOR UTILIZATION PERCENT ---
                                # OF   -- LOG. PROC --    --- PHYSICAL PROCESSORS --
   NAME     STAT  WGT  CAP    LPs    EFFECT.    TOTAL    LPAR MGMT  EFFECTIVE  TOTAL
PRODAA       A    20   NO      2      6.29      6.51       0.06       1.80      1.86
PRODBB       A    75   NO      7     85.98     86.23       0.25      85.98     86.23
CF1PART      A     3   NO      1     44.55     44.60       0.01       6.36      6.37
TESTPART     D
```

In this case, the CF partition has less than 1/2 of a CP. Response time for requests to the CF will increase and throughput on the MVS partitions will decrease as they wait longer for requests to be satisifed in the CF.

If the CF is not given enough processor resource, more serious problems can occur. If the request to the CF takes more than 300 milliseconds, the request will be timed out and reported to MVS. These show up in MVS LOGREC as Interface Control Checks, (IFCC's). They are functionally harmless as MVS will respond to this by reissuing the request. The result however is yet further elongated response time for the original request. If continuous time-outs span a time of about 3.5 seconds, MVS will consider the path to the CF to be broken and seek other paths. If all the paths to the CF are broken, MVS will declare a loss of connectivity to the CF and start recovery actions.

To avoid these more serious problems and get a somewhat reasonable response time in this shared CP environment:

- **DO NOT CAP** the CF partition.
- Limit the number of logical CPs defined to the CF to the minimum needed to meet the needs of the CF.
- Choose a weight for the CF partition based on the anticipated CP requirements of the coupling facility. If these requirements are less than one physical CP; as a general guideline, set the weight of the CF partition so the coupling facility logical processor has at least 50% or more of a PHYSICAL CP resource. If you have less than a full PHYSICAL CP allocated to the coupling facility logical processor, this will result in some elongation of response time. This will have a greater impact on some CF functions (like locking) than on others.
- Use the Dynamic CF Dispatching function, (DYNDISP), to control the test or hot standby CF partition.

ICMF (Not supported on zSeries processors)
One variation of this configuration is a test or migration system which has all the MVS images and the CF defined on one CEC. In this case, ICMF would be an alternative to CF links. The dispatching enhancements now available with this facility no longer require the CF to poll for work. In this case, it is more viable to run shared CPs. There will still be longer response times because of shared resources but the installation can manage the capacity of the CF partition with ICMF by its weight more effectively because the CF does not exhaust its CP allocation with an active wait polling loop. In this environment, all SYNC requests are changed to ASYNC requests for both shared and dedicated CPs

Hot Standby Configuration
A parallel sysplex not only has the advantage of continuous availability but a customer can select the level of continuous availability best suited to his environment and financial considerations.

One of the options a customer can choose is a variation of the configuration described above, namely a BACKUP CF which is an LPAR on a sending CEC having a low weight, real CF links and no defined structures. Another alternative to this is to have the CF defined with an appropriate weight and use dynamic CF dispatching. With a hot standby configuration the customer has one or more separate CFs which hold all of the defined structures. Although this quiesced backup CF has some obvious financial attractiveness, customers implementing this strategy should consider the following points and cautions.

XCF signaling must always be enabled from every system of a sysplex to every other system If you are using CF structures for XCF signaling, and you do not have XCF CTCs configured, loss of your one and only coupling facility also results in loss of signaling capability. Under these circumstances, your backup CF (the one with no structures allocated) may not be able to take over for the failed CF.

If you have specified CONFAIL(YES) and ISOLATETIME, your systems may be placed in a non-restartable wait state if the primary coupling facility fails or connectivity to it is lost and the ISOLATETIME interval expires. GRS Ring disruptions may result if the GRSCNFxx TOLINT interval or the COUPLExx INTERVAL expire before signaling is restored.

If you plan to have a configuration with a backup CF similar to the one described above, you must have alternate signaling paths. IBM recommends CTCs as the alternate signaling mechanism.

If you opt for this configuration, be advised in the event the primary CF is lost, the backup CF partition must be given a reasonable LPAR weight for it to function properly. This implies either you have spare capacity on this CEC to be used in case of emergency, or you are willing to deactivate other partition(s) to provide sufficient cycles to the CF. If this is not the case, you will see seriously degraded performance of all subsystems using the coupling facility.

If your primary CF should fail, you need operational procedures to insure the LPAR weight of the backup CF is increased as soon as possible. This operational step can be avoided with the use of dynamic CF dispatching as LPAR will automatically detect the increased load to the CF and will dynamically increase the CFs weight. Make sure the CF LPAR is not capped. If the backup continues to operate with a very low LPAR weight, it may take a very long time to rebuild and reallocate the structures from the primary CF. This can impact applications dependent upon CF structures. You will have to increase the weight of the backup CF partition to a size which will accommodate the CF functions which will be rebuilt there. At a minimum, set the weight of the CF partition so the coupling facility logical processor has 50% or more of a PHYSICAL CP resource. If you have less than a full PHYSICAL CP allocated to the coupling facility logical processor, this will result in some elongation of response time. You will have to give more weight to the CF partition for functions with stringent response time requirements, such as IRLM locking.

Special Notices