Performance Impacts of Using Shared ICF CPs

Enhancements to the 9672 CMOS family included the introduction of Internal Coupling Facilities, (ICFs). ICFs are Processing Units (PUs) on a 9672, zSeries, or IBM System z processor which are configured as a Coupling Facility engine. The ICF is a lower cost alternative of providing CF capacity. The ICF CP can only be used to perform CF functions and hence is not included in the software licensing charges for the server on which the ICFs reside. Standalone CF's are also still available.

In addition to the ICFs, numerous CF enhancements have been provided which allow more options in the way CF CPU resource can be distributed. Below is a short review of the different enhancements.

- Dynamic CF Dispatching
- Dynamic ICF Expansion
- Enhanced Dynamic ICF Expansion
  1. Support for Shared ICF CPs
  2. Support for CF partition with dedicated and shared CPs

Dynamic CF Dispatching

With the advent of large engines, some installations are defining multiple CFs which share engines. Some of these CFs have very low activity rates, but since each CF is continuously polling, even the CF with little activity looks continuously busy and take it's full share of CP resource.

Dynamic CF Dispatching is a function which allows the installation to limit the impact of CF polling when the CF has a low activity rates and is considered to not be a production CFs. This capability is provided by PR/SM and is available to all IBM CF partitions which share CPs. This support is invoked via a PR/SM command, issued on the CF console, called DYNDISP=ON | OFF, on a partition basis. For CF partitions with dynamic dispatching, (ON or activated), the amount of CP resource the CF partition uses is reduced when there is no activity. As activity increases to the CF partition the amount of resource allocated to the CF will grow limited only by the LPAR weight definitions and the number of logical CPs defined. This function is of most benefit when the CF partition is used as a hot standby or test partition.

There is a performance tradeoff when using DYNDISP=ON. Though the CPU resource used by the backup or test CF is reduced, the responsiveness of the CF partition is also reduced. Therefore sporadic requests to a CF partition configured this way will see CF request service times much greater than CF's without dynamic dispatching turned on. Dynamic dispatching is NOT recommended for production CF's.

Dynamic ICF Expansion

This is support which gives the ability to define dedicated and shared CPs to support CF activity. This capability is available on 9672 G3 processors and above. With this support the dedicated CP was an ICF, and the shared CPs came from the pool of shared general purpose CPs. In this

situation the dedicated ICF will handle most of the requests. Only when the ICF CP is very busy will the expansion take place and the shared CPs start to handle more of the CF activity. Since most of the CF activity is handled by the dedicated ICF the SYNC requests are not converted by PR/SM into ASYNC requests.

There is also a performance tradeoff when using Dynamic ICF Expansion. If expansion should take place and the CF requests start to be handled by the general purpose CPs the amount of processing capacity available to the operating system partitions using the general purpose shared CP pool will be reduced. In this case lower priority work in the operating system partitions may start to see reductions in throughput as capacity is now being used to support the CF Expansion.

Enhanced Dynamic ICF Expansion

This support provides two capabilities. Both will be discussed briefly.

Shared ICF Processors

ICF processors can now be defined as shared in an LPAR partition on the same server which has the potential for an z/OS partition. Prior to this general purpose CPs could be shared, but ICF's could not be shared when an z/OS partition was defined. If the ICFs were on a standalone Coupling Facility the ICFs were always able to be shared.

It was always possible to define a partition as a CF partition and assign it shared engines out of the general purpose pool. However a CF defined in this manner would be using CPs which were covered under the software licensing agreement. Another feature of a CF partition using shared CPs from the general purpose pool is PR/SM will convert all SYNC request from z/OS partitions sharing CPs in the same pool as the CF partition to ASYNC request. This conversion will be done under the covers so to speak and RMF, unaware of the conversion, will continue to report the activity as SYNC requests. The results will be SYNC times will see high response times. Since the request was changed to an ASYNC the sending CP is freed to do other work. So in this case the elongation of SYNC requests does not imply an increase in CPU busy time for the sender CP. However, application response time issues and / or throughput issues may be observed.

Defining an ICF with shared CPs has all of the same performance issues as defining any CF with shared engines. Without enough weight, the CF will appear to be non-responsive to requests. With a CF partition defined using shared CPs from the ICF pool PR/SM will not convert SYNC requests to ASYNC requests. Hence non-responsive CF's will incur increased TCB busy time on the sender side.

Shared ICF and shared general purpose CPs are separate "pools" of physical resources.
Processing weights for partitions using shared ICFs are totaled and managed separately from the total weights of partitions using general purpose shared CPs. If the PR/SM dispatch interval is dynamic then the calculation of dynamic run time is done by pool. If the dispatch interval is set by the installation then this time will be applied to both pools.

A caution about shared ICFs is in order. Because a function is provided does not mean an installation should employ this function. An installation needs to clearly understand the performance impacts of sharing ICF resources. No right or wrong configuration exists. Tradeoffs will need to be made when using shared CF CPs . Installations need to review their requirements for availability, connectivity, cost, and performance when making these decisions.

CF Partitions with dedicated and shared CPs

Provides support to allow a CF partition to contain both dedicated and shared CPs. This is also sometimes called an L-shaped LPAR. This support is provided to more efficiently use ICF resources across production, test, and backup partitions. The shared CPs defined in this manner can be defined as coming from either a pool of shared ICF CPs or from a pool of shared general purpose CPs but not from both.

As with Dynamic ICF expansion the majority of the request will be handled by the dedicated ICF. Only when the dedicated ICF becomes very busy will ICF expansion to either of the shared pools occur.

Configuring for Performance vs Configuring for Function

Care needs to be exercised when configuring coupling facilities to ensure the configuration meets the installations performance requirements. Certain configuration options may cause a coupling facility to become non-responsive. Let's review some of the alternatives which have been seen lately which are causing issues.

1. CF partition has dynamic dispatch turned on. This is recommended for a test partition or a hot backup type of CF. Typically a hot backup CF will have no structures in it and so the fact no CP resources are given to the CF partition means no performance problems are generally seen.

   However a test partition may be different. If the testing is of a sporadic, low volume nature the CF will generally be viewed as non-responsive. The result will be both SYNC and ASYNC service times will be very high, as will the standard deviations of these response times. Since the activity rate is low the impact is also generally low. The one condition to watch though is the sending CP on SYNC requests will start to see increase TCB busy times. Several things can fall out of this:

   a.) A test partition may now start to use more of it's weight. If there is another production LPAR in the same Server and it is using more than it's allotted weight the production LPAR may start to see a reduction in throughput. So low priority work in the production LPAR may be impacted by the CF testing in the test partition.

   b.) While executing SYNC commands a test partition may exceed the LPAR time slice on the sending CP side. LPAR will not interrupt the SYNC operation and the amount of time to have the SYNC request complete may be far in excess of the defined LPAR time slice. LPAR in this case may adjust the actual amount of processor dispatch time to be equal to the dispatch interval. The result may be an RMF report which has LPAR effective dispatch time greater than the physical dispatch time. The result of this is RMF will show '****' as LPAR management time. The SMF 70 records, the basis for the

RMF report, will also show the differences in the effective and total dispatch times. Effective dispatch time is recorded in SMF70EDT and total dispatch time is recorded in SMF70PDT. Any installation reporting program using these fields will need to be aware of the reporting issues discussed above. In order to eliminate this reporting situation it is necessary to provide a more responsive CF environment.

c.) Configurations with non-responsive coupling facilities should not be used for benchmarking or performance analysis. The CPU per Transaction number will be inflated for workloads using test CFs with dynamic dispatching turned on. Performance testing needs to be done in a coupling facility which has consistent access to more CF CP resource.

2. An ICF CP is shared among two or more partitions, and dynamic dispatch is turned off for all the partitions. The most common configuration is two coupling facility partitions sharing a single ICF engine, and the weights are set to give each partition equal weight. In this case the ICF CP is unavailable 50% of the time. This is a non-responsive coupling facility. SYNC service times in excess of 3000 microseconds are not uncommon, and ASYNC response times can exceed 10,000 microseconds.
   The options here are to:

   a.) Live with it. Understand the CPU costs from the sender side is a function of the SYNC request rate. Also there is a response time cost. Somewhere behind every SYNC and ASYNC request is an application which cares to some degree about the requests finishing in a timely manner.

   b.) Turn dynamic dispatching on in all but one of the CF partitions and make the others lightly used test or hot back-ups. In this manner you can designate a specific CF to be the production CF with all the structures, has dynamic dispatching turned off, and in essence get access to all the ICF CP resources. If the test partitions become active they will start to use more of their share, and the production CF will start to see increased service times and high standard deviations.

   c.) Define one partition as the production partition and all the others as test or hot backup partitions and change the weights so the production partition gets 90% of the weight. Dynamic dispatch is turned off for the production CF and turned on for all other partitions. Now if the test or backup partitions start to take more CF CP resource the production partition will be more isolated.

   Generally the difference between alternative b and alternative c is alternative b is usually chosen when the other CF partitions are hot-standbys. In this case when the hot standby is called into action because of a failure the increased request activity will cause the partition to grow it's weight, and hence it's access to resources, up to it's full share. Since the configuration is now in a failure scenario the degradation of response time for both the production and new activated hot standby may be acceptable.

   Alternative c is generally chosen when the other partition is a test partition and the bulk

of the ICF CP resource is always deemed to be more important to give to the production side. In this situation the test CF will always see poor CF service times. The impact to the sender side CPU resource needs to be reviewed.

Sample RMF Data

Below are representative samples of RMF data which will highlight the above discussion.

Example: Shared CPs

```
COUPLING FACILITY MODEL 009672      VERSION R06      CFLEVEL   6
AVERAGE CF UTILIZATION  (% BUSY) 1.4   LOGICAL PROCESSORS: DEFINED  1  EFFECTIVE   0.5
```

This CF partition is defined with 1 CP. However the CP is defined as shared. This can be determined by looking at the LOGICAL PROCESSOR: EFFECTIVE value. In this case EFFECTIVE is 0.5, the production CF is only getting 1/2 of a CP. Any CF request which arrives when this CFs time slice is not active will wait. This is why you see large average service times and high standard deviations.

```
STRUCTURE NAME = ISGLOCK            TYPE = LOCK
           # REQ    -------------- REQUESTS -------------
 SYSTEM    TOTAL              #     % OF  -SERV TIME(MIC)-
 NAME      AVG/SEC            REQ    ALL     AVG   STD_DEV
 WSC11       940K   SYNC     940K   100%   742.6   3415.1
           522.1    ASYNC      0    0.0%     0.0      0.0
                    CHNGD      0    0.0%  INCLUDED IN ASYNC
```

In this situation the installation decided to turn on dynamic dispatching in the other CF partition sharing this 9672. When this happened SYNC times dropped to under 50 microseconds on this partition, while the other partition saw SYNC times grow to >5000 microseconds.

This situation is demonstrated on a standalone 9672 R06 coupling facility. The same situation can be seen with a shared ICF configuration.

Example: Dynamic Dispatch

```
Interval 09:00.00
COUPLING FACILITY MODEL 009672      VERSION R06      CFLEVEL   6
AVERAGE CF UTILIZATION  (% BUSY) 30.4   LOGICAL PROCESSORS: DEFINED  1  EFFECTIVE   0.0

Interval 09:30.00
COUPLING FACILITY MODEL 009672      VERSION R06      CFLEVEL   6
AVERAGE CF UTILIZATION  (% BUSY) 10.1   LOGICAL PROCESSORS: DEFINED  1  EFFECTIVE   0.1

Interval 10:00.00
COUPLING FACILITY MODEL 009672      VERSION R06      CFLEVEL   6
AVERAGE CF UTILIZATION  (% BUSY)  2.4   LOGICAL PROCESSORS: DEFINED  1  EFFECTIVE   0.0
```

In the above case you can see dynamic dispatch is active by looking at the EFFECTIVE field. The amount of CF CP capacity which is allocated during the interval is reported. Variations in the EFFECTIVE field shows the CF partition is gaining and losing weight in response to the activity rate. The CF Utilization value is the utilization in terms of the effective amount of CF CP resource, not the defined CF CP resource.

Another way to determine if dynamic dispatching is turned on is to look at the weight allowed to the CF partition. If a partition is allowed 50% of 1 processor then the LOGICAL PROCESSOR: DEFINED field will be 1 and the EFFECTIVE field should be 0.5. If the EFFECTIVE field is less than this than dynamic dispatching is turned on.
Likewise the opposite situation can be determined. If a CF partition is defined as 1 processor with a weight of 50% and the EFFECTIVE field is reported as greater than 0.5 then the other CF partition sharing the processor most likely has dynamic dispatching turned on, and this partition, using more than it's fair share weight, does not have dynamic dispatching turned on.

These situations are possible with either a standalone CF or an ICF configuration.

Example: RMF Reports with Shared ICFs

Another feature of ICFs is the impact they have on RMF reports. RMF reports on the Partition Data Report the total number of configured PUs. So if a processor is a 9672-X77, with an ICF defined RMF will report based on an 8-way processor, even though the general purpose shared pool is made up of 7 CPs. The interpretation of these reports will force the reader to rebase the partition utilization to the actual number of shared CPs in the appropriate pool.

Below is a representative RMF report. Some fields in the report may have been altered to enable the report to fit the paper.

```
MVS PARTITION NAME              WSC11
NUMBER OF CONFIGURED PARTITIONS          5
NUMBER OF PHYSICAL PROCESSORS           10
                    CP                   8
                    ICF                  2
                                         0
WAIT COMPLETION                         NO
DISPATCH INTERVAL                  DYNAMIC

- PARTITION DATA --           ---- AVERAGE PROCESSOR UTILIZATION PERCENTAGES ----
                             PROCESSOR  LOGICAL PROCESSORS  - PHYSICAL PROCESSORS -
  NAME    WEIGHTS  NUM  TYPE EFFECTIVE    TOTAL  LPAR MGMT   EFFECTIVE     TOTAL
CF2         100      2  ICF    99.08      99.14      0.01       19.82     19.83
PROD         70      8  CP     14.08      14.32      0.19       11.27     11.46
TEST2        10      2  CP     11.33      10.05      ****        2.27      2.01
TEST3        10      2  CP     10.56       9.43      ****        2.11      1.89
CF3         280      1  ICF     0.39       0.57      0.02        0.04      0.06
*PHYSICAL*                                           0.61                  0.61
                                                     ------     ------    ------
  TOTAL                                              0.83       35.50     35.85
```

In the system represented above all partitions are active and no capping is being used.
The important information comes from the logical partition dispatch time which is listed below

```
             PARTITION PROCESSOR DATA --
            ----DISPATCH TIME DATA-----
            EFFECTIVE        TOTAL
CF2         00.29.56.702   00.29.58.698
PROD        00.16.57.892   00.16.59.946
TEST2       00.03.23.744   00.03.13.904
TEST3       00.03.10.215   00.02.59.396
CF3         00.00.03.079   00.00.05.347
            ------------   ------------
TOTAL       00.53.31.002   00.53.17.291
```

The PHYSICAL PROCESSOR TOTAL field is based on available capacity in the interval for 10 processors, or 10 * 15 minute interval = 150 minutes. Hence CF2 used approximately 30 minutes out of 150 which is why the TOTAL PHYSICAL PROCESSOR for CF2 is about 19%. If you were looking only at the ICF pool, the actual value would be 2 ICFs * 15 minute interval = 30 minutes. This would then represent CF2 as using 98% of the ICF capacity.

Likewise the PROD partition is listed as using 11.46% of the configured 10 way. In reality it can only use 8 of the configured PUs. So it's actual utilization should be based upon 8 CPs * 15 minute interval = 120 minutes. This would then state PROD is using approximately 17 minutes out of 120 minutes or approximately 14% busy.

By reviewing the weights for CF3 it can be seen this CF partition is using a shared ICF, and it's weight allows it to use 70% of 1 ICF. However this CF partition must have DYNDISP=ON since the partition is not using it's total weight. This indicates the active polling by the CF has been reduced. The CF3 partition is using 0.5% of the 10-way, but actually is using only the ICF pool, and so is using approximately 0.5 * 10 = 5 across 2 ICFs = 2.5% of the ICF capacity.

What can also be seen with this data is the impact of a non-responsive CF and the impact of dynamic dispatching. TEST2 and TEST3 are using CF3. Since the CF has very little access to CF CP resource the SYNC requests to CF3 are taking a long time to complete. In fact the LPAR time slice is being exceeded. Hence PR/SM is manipulating the total dispatch time. As a result the LOGICAL EFFECTIVE dispatch time is greater than the LOGICAL TOTAL dispatch time and as a result the LPAR MGMT time goes negative and RMF reports it as "****".

Summary:

Installations sharing CF CP resources, or using dynamic dispatching, need to understand the performance penalties which may occur from these configurations. Though shared ICFs are possible such a configuration may not provide the CF performance required for their applications. Non-responsive CFs will cause increased CPU busy time on the sender side and may cause response time issues for applications using data in the coupling facilities.

Though this article discusses ICFs the performance impacts are the same if the CF is internal or external. Customers migrating to ICFs with the shared CP capability need to be aware of the performance tradeoffs which are present when using this capability. Functions such as dynamic dispatching and ICF expansion may be used to build CF configurations which meet both the functional and performance requirements of the sysplex.