z/OS Availability: Blocked Workload Support

This document highlights new function delivered in z/OS (R) V1.9 and rolled back to z/OS 1.7 and z/OS 1.8 via APAR OA17735. This support was inspired by experiences with DB2 but it applies to any implementation with TCBs or SRBs running in multiple address spaces and/or enclaves sharing internal or external resources. The support addresses a potential threat to application availability. The threat to availability occurs when an address space or enclave is CPU starved while holding a resource(s) needed by other transactions. CPU starvation means the address space or enclave is not high enough in dispatch priority to run for significant periods of time. What is a significant period of time is relative to the importance of the workloads and the turnaround criteria set for the workload. The address space or enclave holding the resource is often defined as of low importance and able to be postponed in favor of more important work. The transactions needing the resource(s) may often be among the most important in the system but becomes stalled and times out due to unavailability of the resource(s).

The demand for CPU causing the starvation is usually a much higher peak than the monthly peak. In most cases, it has been a new, record peak. There have been many cases of complete outage of an application due to the inability to identify the root cause of the problem, namely low priority work holding a resource needed by high importance work. The threat to availability of the application is very significant because it is a latent threat, inherent in one of the z/OS strengths, specifically the ability to honor goals and the importance of those goals. A well-designed WLM policy may inadvertently allow service levels to degrade when lower importance work is CPU starved while holding resource(s).

History of MVS and z/OS support for blocked work

z/OS provides several capabilities to recognize contention for resources and promote TCBs or SRBs to resolve the contention. This is done for suspend locks such as the local lock and it is done for TCBs holding a resource via ENQ which causes contention. The GRS component informs WLM/SRM when there is contention for a resource. WLM/SRM promotes the holder of the resource to a dispatch priority set high enough to ensure the holder is able to execute.

New IEAOPTxx Function Provided

The new function called blocked workload support provides the means to promote CPU starved address spaces to the ENQHOLD priority for very small amounts of CPU time. The system must be running at 100% utilization for the support to take effect. This means there must be zero MVS wait time. Blocked workload support does not apply to work running on specialty engines such as zIIPs and zAAPs. It is expected zIIPs and zAAPs are not run to 100% busy for any significant period of time and therefore has no need for blocked workload support.

The blocked workload support provides installation control over the amount of CPU blocked workloads can receive. There are two IEAOPTxx parameters introduced with this support; BLWLINTHD and BLWLTRPCT.

BLWLINTHD

Specifies the threshold time interval for which a blocked address space or enclave must wait before being considered for promotion. If the CPU utilization of a system is at 100% (no MVS wait time), workloads with low importance (low dispatch priority) might not be dispatched. This may lead to problems if the low priority work holds a resource required by high priority workloads. Therefore, if an address space or enclave has ready-to-run work units (TCBs or SRBs) but does not get CPU service for the specified time interval, it will be temporarily promoted to a higher dispatch priority. Swapped out address spaces are not considered for promotion.

The syntax is:

BLWLINTHD=option

Value Range: 1-65535 seconds
Default Value: 20 seconds

BLWLTRPCT

Specifies how much of the CPU capacity is to be used to promote blocked workloads. This parameter does not influence the amount of CPU service a single, blocked address space or enclave is given. Instead, this parameter influences how many different dispatchable units can be promoted at the same point in time. If the value specified with this parameter is not large enough, blocked workloads might need to wait longer than the time interval defined by BLWLINTHD. The specification represents tenths of 1 percent of total LPAR CP resource allowed for blocked workloads.

The syntax is:

BLWLTRPCT=option

Value Range: 0-200 (up to 20%. 0% indicates blocked workload support is not enabled.)
Default Value: 5 or 0.5%

Description

For a z9 EC processor, the promotion allows approximately 480 microseconds of CPU time per interval to one blocked TCB. In a 1 minute interval this is only 0.0008% of a logical processor. This is a small amount of CPU capacity to release a resource but debugging experiences with DB2 latches indicates 1-3 bursts of this size should free most instances of a latch. One can apply more CPU to a blocked workload by reducing BLWLINTHD; a value of 20 would provide 1.44 milliseconds of promotion time per minute on a z9 EC processor. The default value for BLWLTRPCT of 0.5% of all CP resources of the LPAR allows a large number of blocked TCBs and SRBs to be promoted. Configured zAAPs and zIIPs are not considered in generating the amount of CPU to be used by this function. A z9 EC LPAR with 2 CPs of weight would require 1250 TCBs/SRBs to be blocked every time the blocked workload function executes to consume the 0.5% of CPU allowed by the default values. It is unlikely for a system to have this many TCBs/SRBs blocked by CPU in any reasonably managed z/OS system.

Let's examine a more probable CPU starvation example. Assume an LPAR is on a z9 EC processor and is defined with 2 logical processors. The LPAR is at 100% CPU utilization, with 20 address spaces classified as either discretionary or importance level 5, and this work receives no CPU for 20 seconds. Further assume each address space has 1 TCB ready to run. SRM will not use all the resource allowed by BLWLTRPCT (0.5% of the LPAR) unless there are sufficient blocked TCBs/SRBs to justify the use of the promotions. If the blocked interval time (BLWLINTHD) is set to 20 seconds, WLM will promote each of the 20 TCBs for 480 microseconds every 20 seconds which represents 0.024% of the processor resource to improve availability. This is calculated as:

((20 TCBs * 480us) / (1000000us * 20 secs * 2CPs) = .00024

this small amount of CPU allocation is intended to improve availability; the work promoted may be productive work albeit low priority.

The promotion time is sensitive to the speed of the processor. The promotion allowance for a z990 is about 650 microseconds compared to 480 microseconds for a z9 EC. This is the standard SRM adjustment for values such as microseconds of time per service unit. If unsure of the relationship of a processor to a z9 EC the following WEB site may be helpful:

https://www-304.ibm.com/servers/resourcelink/lib03060.nsf/pages/srmindex?OpenDocument

RMF support for blocked workloads

In support of blocked workload, RMF has enhanced SMF records and provided extensions to 2 reports; CPU Activity and Workload Activity. SMF 70 and SMF 72 extensions are not described here. The additions to the records can be found in the SMF book for z/OS R1.9. The RMF report extensions are documented in the RMF APAR for the rollback. The RMF APAR OA18244 provided PTFs are UA90375 and UA90377.

RMF CPU Activity Report

The Postprocessor CPU Activity report provides a new section with information about blocked workloads.

```
BLOCKED WORKLOAD ANALYSIS
  BLWLTRPCT (%)   0.5  PROMOTE RATE: DEFINED   14  WAITERS FOR PROMOTE: AVG   0.010
  BLWLINTHD        60                USED (%)   4                       PEAK      1
```

PROMOTE RATE: DEFINED - Number of blocked work units which may be promoted in their dispatching priority per second. This value is derived from IEAOPTxx parameter BLWLTRPCT.

PROMOTE RATE: USED (%) - The utilization of the defined promote rate during the reporting interval. This is calculated per RMF cycle and averaged for the whole RMF interval. It demonstrates how many trickles were actually given away (in percent of the allowed maximum) for the RMF interval.

WAITERS FOR PROMOTE - Average number of TCBs/SRBs and enclaves found blocked during the interval and not promoted according to IEAOPTxx parameter BLWLINTHD.

WAITERS FOR PROMOTE - PEAK – the maximum number of TCBs/SRBs and enclaves found blocked and not promoted during the interval according to the IEAOPTxx parameter BLWLINTHD. The AVG value might be quite low although there were considerable peaks of blocked workload. Thus, the peak value is listed as well.

As long as WAITERS FOR PROMOTE is greater than 0, the system has work being blocked longer than the BLWLINTHD setting. In such a case it may be advisable to increase BLWLTRPCT. If there are still problems with blocked work holding resources for too long even though there are no waiters seen in the RMF data then decreasing the BLWLINTHD setting may be advisable.

RMF Workload Activity Report

The RMF Postprocessor Workload Activity report will provide the CPU time transactions in a service or report class were running at a promoted dispatching priority.

```
                          SERVICE CLASS=BATLOW      PERIOD=1 IMPORTANCE=5
---SERVICE---   SERVICE TIME  ---APPL %---  --PROMOTED--  ----STORAGE----
IOC     60439K  CPU 11303.13  CP    294.32  BLK    7.498  AVG    17647.61
CPU    664890K  SRB  176.656  AAPCP   0.00  ENQ    0.000  TOTAL  359251.4
MSO         0   RCT    0.715  IIPCP   2.68  CRM    0.000  SHARED   208.28
SRB     10392K  IIT   73.298                LCK   46.300
TOT    735720K  HST    0.013  AAP      N/A  SUP    0.000  -PAGE-IN RATES-
/SEC   204367   AAP      N/A  IIP    26.62                SINGLE       0.0
                IIP  958.208                              BLOCK        0.0
```

This new field is called PROMOTED.

Recommendations

It is important to review the new blocked workload information in RMF. If blocked workloads are seen frequently for moderate to long intervals there may be negative effects on the running of the system. CPU starvation of workloads which hold critical system resources such as DB2 latches should be avoided. It is recommend a review be done of the WLM goals and importance settings of services classes receiving CPU through the blocked workload support to ensure the amount of processor delay experienced by the work in these service classes is appropriate.

It is also strongly recommended for this function to be enabled for all production systems. Given the very small amounts of CPU time provided to blocked work by this function, and the positive effects it can have, installations are also encouraged to consider enabling the blocked workload support on all LPARs.

Change in Blocked Workload Defaults

In z/OS 1.9 the default settings for the blocked workload support has the function enabled for use. For z/OS 1.7 and z/OS 1.8 the PTFs shipped in OA17735 delivering the original support for these releases specified a default for BLWLTRPCT of 0% thereby disabling the function.

APAR OA22443 is being provided for all supported releases (including z/OS 1.9) which will provide PTFs for all supported releases which will change the defaults for the blocked workload support. The defaults for all supported releases will be changed as follows:

BLWLINTHD=20
BLWLTRPCT=5

These values allow a promotion every 20 seconds for a blocked address space or enclave providing 1.44 milliseconds of promotion time in one minute and allow a total of 0.5% of CPU resource in the extreme case of there being 417 blocked TCBs/SRBs in each 20 second interval on the hypothetical 2-way z9 EC processor.

APAR OA44526 has made additional changes to the blocked workload support. This APAR has lowered the amount of time which a blocked workload can wait before promotion from five seconds to one second. For certain LPARs where the workload in the LPAR is all critical, high importance online waiting 5 seconds for this support may be too long. In these all online environments having a more aggressive blocked workload threshold is important. In these types of environments a setting of 2 for BLWLINTHD may be more appropriate.

Once APARs OA22443 and OA44526 are available, all installations who have the blocked workload support active (BLWLTRPCT>0) should evaluate changing their IEAOPTxx member to specify a setting of BLWLINTHD=5. Waiting 20 seconds to receive this availability support is too long for current processing environments.