
Rule WLM361: Non-paging DASD I/O activity caused significant delays

Finding: Non-paging DASD I/O activity by the service class was a significant cause of the service class missing its performance goal.

This finding applies service class periods with an average response or percentile response goal, and to service classes with execution velocity goals only if non-paging DASD I/O using and I/O delay are included in the execution velocity calculation.

Impact: This finding can have a LOW IMPACT, MEDIUM IMPACT, or HIGH IMPACT, depending upon the amount of non-paging DASD I/O activity and the delay to the service class caused by the non-paging DASD I/O activity.

Logic flow: The following rules cause this rule to be invoked:

Rule WLM101: Service Class did not achieve average response goal

Rule WLM102: Service Class did not achieve percentile response goal

Rule WLM103: Service Class did not achieve execution velocity goal

Discussion: When CPExpert detects that a service class did not achieve its performance goal, CPExpert analyzes the basic causes (see the discussion in the above predecessor rule). One of the possible causes of delay is that the service class was delayed because of non-paging DASD I/O activity.

The SRM collects I/O using and delay information beginning with OS/390 Release 3. These delays are collected regardless of whether the performance goal is a response goal or an execution velocity goal.

Non-paging DASD using and I/O delays can be a part of the computation of execution velocity beginning with OS/390 Release 3. However, the I/O activity is included only if the Workload Manager has been instructed to include I/O using and I/O delay in the calculation of execution velocity. If I/O using and I/O delay are not included in the calculation of execution velocity, the I/O using and delay information has no relevance to the goal achievement¹.

The non-paging DASD I/O using and delay information is reported in SMF Type 72 records for each service class period. CPExpert analyzes the non-

¹The I/O using and I/O delay can, of course, have a drastic effect on the actual performance of the service class periods with an execution velocity goal. However, if the I/O activity is not included in the Workload Manager's assessment of goal achievement for execution velocity, no action would be taken based on I/O using or I/O delays.

paging DASD I/O using (field R723CIOU) and I/O Delay (field R723CIOD) for service classes missing their performance goal. CPEXpert produces Rule WLM361 when the percent I/O Using or I/O Delay caused by non-paging DASD I/O is greater than the **WLMSIG** guidance variable in USOURCE(WLMGUIDE), and the service class period has a *response goal* specified. CPEXpert produces Rule WLM361 when the percent or I/O Delay caused by non-paging DASD I/O is greater than the **WLMSIG** guidance variable in USOURCE(WLMGUIDE), and the service class period has an *execution velocity goal* specified.

After producing Rule WLM361, CPEXpert analyzes several possible causes of non-paging DASD I/O delay and reports the result in subsequent rules.

The following example illustrates the output from Rule WLM361:

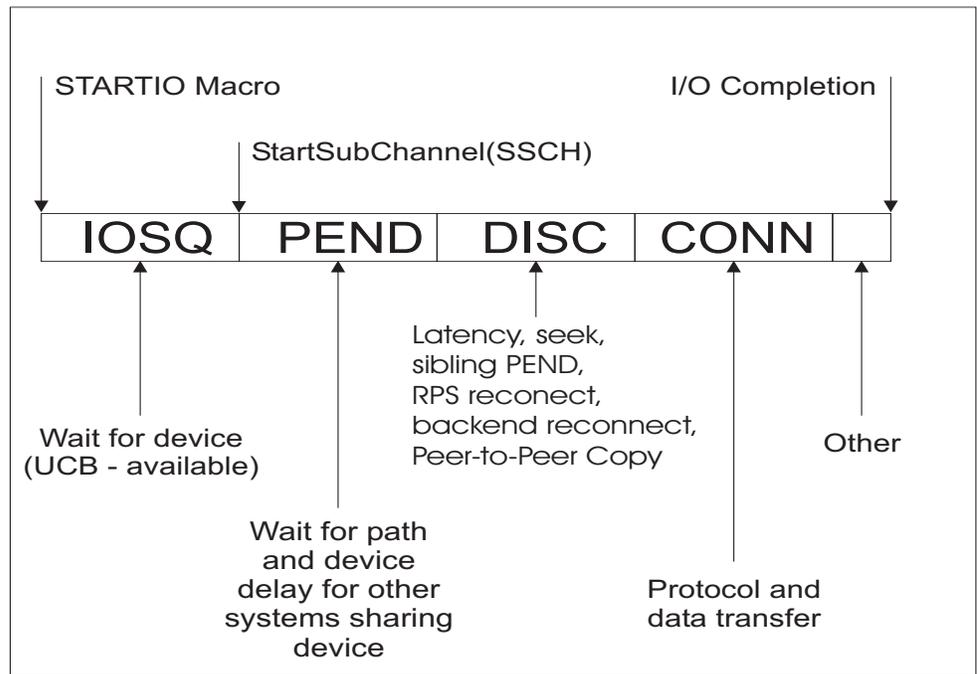
```

RULE WLM361: NON-PAGING DASD I/O EXPERIENCED SIGNIFICANT DELAYS

  BATPHI (Period 1): A significant part of the delay to the service
  class was caused by non-paging DASD I/O activity. The below data shows
  intervals when non-paging DASD I/O operations experienced significant
  activity. The percentages are computed as a function of the EXECUTION
  samples on the local system (the percentages are adjusted to eliminate
  IDLE time, to reflect the effect when the service class was actually
  executing).

  MEASUREMENT INTERVAL      AVG DASD   PCT    ---AVERAGE DASD I/O TIMES---
                             I/O RATE  DELAY  RESP IOSQ WAIT  DISC  CONN
0:30- 0:45,31JUL2003      1,068    49.9  0.007 0.004 0.001 0.000 0.003
0:45- 1:00,31JUL2003         692    50.9  0.008 0.005 0.000 0.000 0.003
1:03- 1:15,31JUL2003         906    51.1  0.008 0.004 0.000 0.001 0.002
1:15- 1:30,31JUL2003      1,013    48.3  0.007 0.003 0.001 0.001 0.002
1:30- 1:45,31JUL2003      1,056    45.7  0.006 0.003 0.001 0.001 0.002
1:45- 2:00,31JUL2003         976    48.0  2.509 0.003 0.001 0.001 2.504
  
```

From a high-level view, there are four key measures of DASD performance: IOS Queue (IOSQ) time, pending (PEND) time, disconnect (DISC) time, and connect (CONN) time. The following figure illustrates these four measures and another potential element of DASD I/O time, titled "Other":



C **IOSQ time.** IOSQ time is the time from the issuance of a STARTIO macro until the StartSubChannel (SSCH) instruction is issued. After the STARTIO macro is issued, the software determines whether the device is busy with *this system*; that is, whether there is an available Unit Control Block (UCB) for the device. If the device is not busy with *this system* (a UCB is available), the SSCH instruction is issued. However, if the device is busy with *this system*, the I/O request is queued. Thus, IOSQ time always means that the device is unable to handle additional requests from *this system*. (The emphasis on "this system" is explained in the below discussion of PEND time.)

This discussion of IOSQ time does not always apply to Parallel Access Volumes (PAVs)². With PAV devices, MVS creates multiple UCBs for each device, depending on how many "alias devices" have been defined. The multiple UCBs allow multiple active concurrent I/Os on a given device when the I/O requests originate from the same system³. Using PAVs can dramatically improve I/O performance by nearly eliminating IOSQ.

Beginning with OS/390 Version 2 Release 4, IOSQ time for service class periods is available in SMF Type 72 records as field R723CIOT.

²PAV devices are available with Enterprise System Storage (ESS). With PAV devices, a "base device" address is defined, and a UCB is associated with this base address. "Alias device" addresses can be defined and UCBs are associated with the alias device addresses.

³Multiple Allegiance allows multiple active concurrent I/O operations on a given device when the I/O requests originate from different systems.

C PEND time. PEND time is the time from the issuance of the StartSubChannel (SSCH) instruction until the device is selected by the control unit and physical positioning commands (such as seek and set sector) are transferred to the device. With modern fixed block architecture (FBA) devices, the PEND time ends when the physical positioning commands are presented to the *logical volume control block* within the control unit. The PEND time is caused by queuing for the path (wait for channel, wait for director port, wait for control unit, wait for device, or wait for “other” reasons)⁴.

The PEND time can be caused by the device being busy from *another system*. In this case, the system issuing the STARTIO macro (*this system*) would have no knowledge that the device was busy with another system. Rather, if a UCB were available for the device, the SSCH would be issued. However, the device could not necessarily be selected (unless multiple allegiance were available), since the device would be busy from another system. Additionally, PEND time could accumulate even with PAV devices if the access were to an extent that was busy with another I/O operation from *this system*.

PEND time for service class periods is available in SMF Type 72 records (field R723CIWT⁵).

C DISC time. DISC means that there is some delay that is often (but not always) associated with a mechanical movement during which the device disconnects from the control unit.

With legacy systems (e.g., 3380 drives attached to 3990-2 control units), the DISC time of most concern was associated with seek (arm movement) and rotational position sensing (time waiting for the disk platter to rotate to the location where desired data resides). Considerable performance improvement efforts were directed at reducing the seek activity and reducing the rotational position sensing (RPS)⁶ delays for the legacy systems. These two mechanical delays still exist for most modern

⁴PEND time is significantly reduced with FICON channels. FICON channels can have multiple I/O operations concurrently active, which reduces the potential PEND time caused by channel busy. There is no port busy time with FICON switches, and control unit time is significantly reduced. This statement regarding PEND time is not necessarily correct if a large number (more than 5) I/O operations are concurrently executing on a FICON channel. Dr. H. Pat Artis and Mr. Robert Ross have presented the results of research indicating that performance degrades significantly when more than 5 I/O operations are concurrently active on a FICON channel (see “Understanding FICON Channel Path Metrics” at www.perfassoc.com).

⁵While the SMF documentation described R723CIWT as “queue time + pending time, the “queue time” refers to queuing for controller, rather than IOSQ. This meaning has been confirmed by IBM SRM/WLM developers and by RMF developers. IOSQ time was added in OS/390 Version 2 Release 4 by the SMF Type 72 field R723CIOT.

⁶RPS delays were caused by a path not being available when the required data came under a device read head. Since a path was not available, the data could not be read and another rotation of the platter was experienced until the data again came under the device read head. Multiple rotations might be required, depending on the busy level of the path.

*redundant array of independent disks (RAID)*⁷ systems, but their impact can not be directly reduced with normal methods.

With modern disks, data is cached into Actuator Level Buffers (ALBs), that contain data read from a track on the disk platter. Using ALBs eliminated the RPS delays, since required data is read into the device buffer during a single rotation and stored until a path is available to transfer the data.

Additionally, data is cached into increasingly large cache on the controller. For a read operation, desired data often is found in the cache. Write operations normally end as the data to be written is placed in the cache; and the storage processor writes the data to the device asynchronous with other activity (as a “back end” staging operation).

Consequently, DISC time for modern systems is a result of *cache read miss* operations, potentially back-end staging delay for write operations, peer-to-peer remote copy (PPRC) operations, and other miscellaneous reasons⁸. DISC time often can be very small with adequate cache. For example, there would be zero disconnect time for a cache read hit (the record was found in the cache).

DISC time for service class periods is available in SMF Type 72 records (field R723CIDT).

C CONN time. CONN time includes the data transfer time, but also includes protocol exchange⁹ (or “hand shaking”) between the various components at several stages of the I/O operation.

For devices attached to paths that include parallel channels and ECON channels, the data transfer time is simply the number of bytes transferred divided by the transfer speed. This is because a parallel channel or ESCON channel can have only one data transfer operation in execution at one time.

For devices attached to paths that include FICON channels, the algorithm is more complicated. This primarily is because a FICON channel can perform multiple data transfer (read and write) operations at one time. The data packets for multiple read or write operations are interleaved (or

⁷ An array is an ordered collection of physical devices (disk drive modules) that are used to define logical volumes or devices.

⁸ Artis has described a “sibling PEND” condition that results from collisions within the physical disk subsystem of RAID devices. See “Sibling PEND: Like a Wheel within a Wheel,” www.cmg.org/cmgap/int449.pdf.

⁹ Note that the protocol exchange occurs at multiple points in the normal I/O operation, even though it is shown only once in this exhibit.

multiplexed) in the FICON link. CONN time for an individual I/O begins with the first frame of data transferred and ends last frame of data transfer, even though data for other I/O operations might be transferred concurrently on the link. Consequently, if multiple data packets (representing data for multiple read or write operations) are interleaved on the FICON link, the elapsed time for any particular I/O operation can be elongated¹⁰ when compared with the elapsed time of the same I/O operation on an ESCON channel.

CONN time for service class periods is available in SMF Type 72 records (field R723CICT).

C **OTHER time.** There are at least two other potential I/O delays for DASD: (1) waiting for the I/O completion interrupt to be serviced by a processor and (2) waiting for the I/O interrupt to be serviced by a domain under PR/SM. Neither potential I/O delay is expected to be of the magnitude of the four "standard" I/O delays. However, they can be significant in special circumstances.

C Multi-processor configurations can use any processor to service an I/O interrupt. However, when a processor services an I/O interrupt, the processor's high-speed cache storage is no longer valid when control is returned to the interrupted task. Consequently, many of the processor's high-performance design features may be nullified.

A hardware feature allows processors to be disabled for I/O interrupts. With this method, only a small number (perhaps only one) processor is enabled for interrupt processing. Only this processor will have its high-speed cache storage disturbed by the task-switching required for interrupt processing, and only this processor will periodically have its high-performance design features nullified. The disadvantage to this approach is that an interrupt may occur while the processor is busy servicing a previous interrupt.

If an interrupt is pending and no processor is enabled to service the interrupt, the interrupt must wait until a processor is available. This time should be insignificant, unless the system is processing a significantly large number of I/O operations. If the system is processing a large number of I/O operations (or if the I/O is particularly time-sensitive), the interrupt pending delay could pose performance problems.

¹⁰The relative speed of a FICON channel is much higher than that of an ESCON channel. Consequently, the elapsed time of any particular I/O operation should be less on a FICON channel than on an ESCON channel, even if there are multiple I/O operations interleaving data. This statement regarding elapsed time is not necessarily correct if a large number (more than 5) I/O operations are concurrently executing on a FICON channel. Dr. H. Pat Artis and Mr. Robert Ross have presented the results of research indicating that performance degrades significantly when more than 5 I/O operations are concurrently active on a FICON channel (see "Understanding FICON Channel Path Metrics" at www.perfassoc.com).

After the processor completes processing for an I/O interrupt, it issues a Test Pending Interrupt (TPI) instruction to determine whether there are any interrupts pending. If an I/O interrupt is pending, the processor proceeds to service that interrupt.

The CPENABLE keyword in the IEAOPTxx member of SYS1.PARMLIB is used to specify the percent of I/O interrupts detected by the TPI instruction, compared with all I/O interrupts. When the percent exceeds the high threshold of the CPENABLE keyword, MVS enables another processor to handle pending I/O interrupts. If the percent falls below the low threshold of the CPENABLE keyword, MVS will disable a processor (to the point that only one processor is enabled). IBM's recommended setting for the CPENABLE keyword differs, depending on the level of processor.

- C MVS environments running under as a guest under VM or in a logical partition (LPAR) under PR/SM are subject to I/O interrupt delays. These delays can occur if another guest (for VM) or another domain is in its dispatch interval when the I/O interrupt completion is posted. The I/O interrupt remains pending until the guest or domain is dispatched. These delays have been estimated to be far more significant than might otherwise be expected.

OTHER time for service class periods is not available in SMF Type 72 records.

Suggestion: There are no suggestions associated with this finding. Subsequent rules will be produced to provide suggestions, depending on where delays occur.