z/OS Performance: Capture Ratio Considerations for z/OS and IBM System z Processors - V2

This updated flash discusses newly provided support which may reduce the z/OS capture ratio on the most powerful IBM System z processors. z/OS APAR OA18452 was taken to address a problem where the SRM timing routines were running too frequently causing a lower z/OS capture ratio. The fixes provided by this APAR are now available. Correcting PTFs are being provided for z/OS 1.7 (UA32973), for z/OS 1.7.1 (UA32972), and for z/OS 1.8 (UA32971). Installations running on z/OS releases prior to 1.7 should continue to use the circumventions originally developed and documented later in this flash.

SRM Timing Routine:

Many of the SRM functions are invoked by the SRM Timer DIE (Disabled Interrupt Exit). These routines are intended to run after a fixed amount of work executes on a CP regardless of the speed of the processor. This means a CP which is twice as fast will execute this code twice as often. SRM does this to maintain its responsiveness to changes in the system. The CPU time used to execute these functions is uncaptured time. On z/OS systems prior to z/OS 1.7 the SRM timer is rounded to units of 1.024 milliseconds. On a zSeries z990, the SRM interval is rounded up to 2.048 milliseconds. On an IBM System z9 EC processor the timer interval is rounded down to 1.024 milliseconds. Therefore the System z9 with full capacity CPs, (processor 2094-701 through 2094-754) schedules 100% more timer interrupts than a zSeries z990 which lowers the capture ratio. In z/OS 1.7 the defect occurred because the SRM timer precision changed to microseconds while the setting of the timer interval remained in units of 1.024 milliseconds.

The extra timer interrupts introduced due to rounding do not increase the SRM CPU time in direct proportion to the interrupt rate but they do produce unnecessary overhead. On non-production LPARs the extra interrupts are particularly of low value. The SRM Timer DIE is being driven approximately 40% more often than the SRM design intended due to the rounding done for the z9 and the imprecision of units. APAR OA18452 has been taken to address this problem.

SRM Support for z/OS 1.7 and Later Releases:

While researching the issue of lower than expected z/OS capture ratios additional improvements in SRM processing were identified. These improvements are aimed at reducing the uncaptured time in a z/OS system. The SRM design was developed in the 1970s and over time portions of the design have become less than optimal. Areas which impacted the design include how much faster processors have become and the increase in the number of address spaces when summed for all logical partitions. Several functions processed in the SRM DIE (e.g. calculation of MTTW dispatch priority for the discretionary service class) do not need to be processed as often as the current implementation allows. There are other functions (e.g. page replacement algorithm) which are time sensitive and still need to be processed in a timely fashion on a production system.

The fixes for APAR OA18452 provide design improvements to change the frequency for some less critical SRM functions to reduce uncaptured time. Some SRM processing was done continually but is only relevant when the system is at 100% utilization. This processing is improved to execute only when all logical processors are busy. The improvement in capture ratio may be slightly greater for LPARs running less than 100% busy.

The RMPTTOM parameter is specified in the IEAOPTxx member of SYS1.PARMLIB. This parameter is the SRM invocation interval constant and is used to influence the interval used by the SRM Timer DIE. The specified real-time interval is adjusted by relative processor speed to become SRM time in order to ensure consistent SRM control across various processors. Increasing this value reduces the frequency of execution of the SRM DIE. With the application of the fix for APAR OA18452 the default value for RMPTTOM is being changed to 3000. The recommendation though is to not code the value at all in the IEAOPTxx member of a production LPAR. There may still be valid reasons to use a value greater than the default for certain types of LPARs (see below).

Circumvention for z/OS 1.6 and earlier releases

The IEAOPTxx parameter RMPTTOM can still be used to circumvent this problem on z/OS releases where a fix is not being provided. In z/OS releases without a fix provided by OA18452 the default value is 1000 (msec). If a customer is using the default value then changing this value to 2000 will negate the problem described by the APAR. It is recommended all customers running on zSeries and System z processors rated at a per CP speed greater than 100 MIPS and running on z/OS 1.6 or earlier releases update their RMPTTOM setting to 2000.

Impact of the Change

Once the APAR fix has been applied, or the circumvention implemented for releases without a fix, the installation's ability to save CPU time will depend upon different configuration factors. The amount of uncaptured time is a function of:

- The SRM interval (influenced by RMPTTOM)
- Number of Address Spaces in the LPAR
- Number of LPARs
- Utilization of the logical CPs

At a fixed SRM interval logical partitions with a large number of address spaces (e.g. >800) will have more processor time consumed by SRM than a logical partition with fewer address spaces and hence greater opportunity for CPU savings. Since each z/OS LPAR has an SRM timer function, an environment with many LPARs will have more timer interrupts per second and will correspondingly have a greater opportunity for CPU savings.

The PTFs for OA18452 were evaluated for effectiveness in both the z/OS 1.7 and z/OS 1.8 environments. The tests were done using two different LSPR workloads; WASDB (WebSphere Application Server using a DB2 database), and OLTP-T (traditional IMS on-line workload with VSAM and OSAM databases). The WASDB workload requires approximately 45 address spaces on a 32 way z9 EC with a single LPAR, resulting in little advantage to the capture ratio after application of the PTFs. The savings were about 6 MIPS and about 0.1% ITR.

The OLTP-T workload requires approximately 1600 address spaces on a 32 way z9 EC with a single LPAR. The OLTP-T workload demonstrated significant improvement. Approximately half the SRM processing was eliminated which resulted in about a 60 MIPS saving and a 1% ITR improvement.

The methodology for both the WASDB and OLTP-T workloads is to run very close to 90% processor utilization. Since the processors were not saturated, a small part of the savings is due to eliminating processing with no great value unless the system is 100% busy.

Opportunity to Further Reduce Uncaptured Time via Tuning

Installations have always had the opportunity to reduce the impact of SRM CPU usage by changing the RMPTTOM default. With any tuning exercise, the installation is responsible to understand how these tuning changes apply to their environment.

For most environments with the fix for APAR OA18452 applied the new RMPTTOM default of 3000 should be adequate. With or without the fix, installations may choose to experiment with larger RMPTTOM values for production LPARs running on processors with a per CP speed above 100 MIPS. With the APAR fix larger values like 5000 may have little influence on SRM CPU time. For any LPAR, but especially small LPARs (less than 150 MIPS) or non-production LPARs, which may not need SRM period switch or can live with less precise period switches, numbers up to 20,000 or larger may be used. Customers must do the analysis to ensure high values do not impact responsiveness and system efficiency. Small changes in RMPTTOM should be tried with analysis of the effect of each change. An installation may decide higher values may be acceptable due to the nature and importance of the workload on the LPAR.

Values above 20,000 for a z9 may not provide much benefit since the SRM cost is reduced considerably at 20,000 and one probably does not want to risk poor resource management. A setting of 20,000 allows an SRM timer interrupt every 28 milliseconds or about every 17 million instructions (of a single CP) on a z9. One would use lower limits for z990 and z900 based on the respective processing power to keep the timer interrupt interval at about 28 milliseconds. Values of 15,000 for the z990 and 10,000 for the z900 may be specified to enable SRM to manage resources at the same 28 millisecond interval (approximately.)

Special Notice:

The information contained in this document has not been submitted to any formal IBM test and is distributed on an “as is” basis without any warranty either expressed or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer’s ability to evaluate and integrate them into the customer’s operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Performance data contained in this document was determined in a controlled environment; therefore the results which may be obtained in other operating environments may vary significantly. No commitment as to your ability to obtain comparable results in any way is intended or made by this release of information.