z/OS Performance: Managing Processor Storage in an all "Real" Environment

White Paper
December 18, 2002
Version 1.1

Washington Systems Center
Advanced Technical Support

Introduction:

OS/390 R10 and z/OS have support for 64-bit mode real storage environments. OS/390 R10 has the ability to run in either 31-bit mode or 64-bit mode on a zSeries, while installations migrating to z/OS on a zSeries processor are able to run only in a 64-bit mode real storage environment. z/OS 1.2 and later releases provide virtual storage exploitation of the addressing range above 2GB.

The 64-bit mode environment, as contrasted with the 31-bit mode environment, provides only central storage (now called real storage, both terms are used interchangeably in this document) as opposed to a combination of central and expanded storage. Removing expanded storage means the 64-bit mode operating system has a 2-level storage hierarchy; real storage within the processor, and auxiliary storage backed by DASD. The elimination of expanded storage from the processor storage hierarchy and how it impacts the resource management of an MVS system is the main focus of this document.

This document is intended to help installations migrating to an all real environment. The first goal of the paper is to help readers become more familiar with reviewing performance metrics in a z/Architecture environment to ensure optimal performance. The second goal of the paper is to provide information on changes in the configuration and tuning of the paging subsystem will be covered. And lastly information will be discussed which discusses the performance implications of stand-alone dump, and SVC dumps in an “all real” environment.

Editor’s Note: A series of RSM APARs were taken in 3Q2002 to address issues in the Real Storage Manager (RSM) support for z/Architecture. This paper on storage management discusses storage management with the perspective these correcting fixes have been installed. To better understand the support changes introduced by these APARs the reader is directed to the APAR cover pages for APARs; OW55209, OW55729, OW55255, OW55033, OW54399, OW55902, OW54938, OW55190, OW55408, OW55051.

1.0 MVS Performance Metrics

This section will discuss changes in performance metrics to help installations evaluate their processor storage configurations in a 64-bit mode environment.

1.1.0 Storage Service Units and Storage User Coefficient (MSO)

Service Units (SU) are used by the System Resource Manager (SRM) as a basis for resource management. SRM calculates CPU and SRB service based on the task and SRB execution time, the CPU model, and the number of processor's online. Tables describing the service consumed per second of execution time by CPU model are published in the MVS Initialization and Tuning Guide. The values listed for each processor are called SRM constants. The SRM constant is a number derived by IBM product engineering to normalize the speed of a CPU, so a given amount of work would report the same service unit consumption on different processors. This removes the need to redesign your performance parameters each time you upgrade your CPU. Service units can be accumulated for CPU / SRB time, as well as I/O and processor storage. Installations, through the use of installation defined user coefficients, can give additional weight to one type of service relative to another. This allows the installation to specify which type of resource consumption should be emphasized in the calculation of service rates.

Unweighted storage service units are accumulated for a unit of work based on:

```
(Central page frames used) x (CPU service units) x 1/50
```

where 1/50 is a scaling factor designed to bring the storage service component in line with the CPU component. Since CPU service is a factor in the calculation for storage service units, work acquires storage service units only when they are also using CPU time.

The user defined coefficient for weighting service related to processor storage is called MSO (Main Storage Occupancy). The formula for how weighted storage service units are accumulated for a unit if work is:

```
(MSO user coefficient) x (storage service units)
```

RMF reports weighted service units, and weighted service units are the basis for period switching and swap recommendation values. Contrary to the belief of many, the default for the MSO user coefficient is not MSO=3.0. The z/OS default value is actually MSO=1.0. For more than 10 years IBM has been recommending the MSO value be changed to either 0.0000 or 0.0001. An MSO value of 0.0001 will record some measurement of storage used, while 0.0000 will give no information on storage usage. This recommendation has arisen from advances in processor design where the ability to configure a larger amount of processor memory has reduced the need to weight the use of memory as stringently as perhaps was justifiable when MVS was running on a S/370 3033-U with 4MB of memory!

Installations migrating to an all real environment need to be aware of the change in both storage service units and the importance of setting MSO to a smaller number. Storage service units are based on the number of central storage frames, but a workload's storage working set was previously spread between central and expanded storage. In the migration to a 64-bit mode

environment a workload may now accumulate more storage service units as the working set is now backed entirely by central storage.

The change in storage service units accumulated should impact only swappable workloads, and workloads defined with multiple periods. Change in the rate transactions accumulate service units can impact metrics which control swapping (such as the contention index and swap recommendation value), and period switching.

If the rate at which workloads accumulate service units increases, work may transition periods quicker, spending less time in earlier periods. Work in earlier periods generally have access to more resources via either a higher dispatch priority or MPL level. This change in the relative amount of time spent in different periods may impact throughput and responsiveness of work such as Batch, TSO, and USS applications.

After a transition to 64-bit mode it would be worthwhile to review the percentage of transactions ending in first period versus later periods. Check to see if changes in the service units accumulated by a workload have changed the period distribution of ended transactions. If the distribution has changed (and the change is viewed to be unfavorable in terms of the installation's Service Level Agreement) then the duration definitions for the multi-period workloads need to be adjusted to allow more service to be accumulated in the earlier periods, or adjust the MSO user coefficient.

If an installation is running with a large MSO user coefficient and chooses to reduce the impact of central storage on the service rate then a review of the impacted performance groups / service classes needs to be done to ensure work switches periods in the same manner as before, or as is determined to be correct.

Note: IBM has recommended for many years to reduce the MSO user coefficient and reduce the impact of storage on the service rate. A similar recommendation has been made for installations migrating to WLM goal mode. In this case, the recommendation was to reduce the scale of all of the user coefficients to CPU=1.0, SRB=1.0, MSO=0.0001, IOC=0.5. This allows CPU service to be scaled at the same metric used for managing WLM resource groups, and allows the relative contribution of the four different components to remain relatively the same. Installations following this recommendation have to adjust the duration value on any multi-period service class to ensure period transitions are consistent before and after the change to the user coefficients. The user coefficient changes can be done before the migration to goal mode or the migration to 64-bit mode.

Customers running in WLM compatibility mode can find period duration's and the MSO definitions in the IEAIPSxx member of parmlib. Customers running WLM goal mode change the multi-period duration and the MSO definition by updating the WLM Service Definition via the WLM ISPF application.

1.1.1 RMF Measurement Data and Reports

Below is an RMF workload activity report before and after a migration to 64-bit mode.

Workload Activity Report - 31-bit mode - TSO workload by period - MSO=0.1000

TRANSACTIONS		---SERVICE----		--SERVICE RATES--		----STORAGE----	
AVG	2.00	IOC	172134	ABSRPTN	2366	AVG	1311.79
MPL	2.00	CPU	4613K	TRX SERV	2203	TOTAL	2448.98
ENDED	**179929**	**MSO**	**10995K**	TCB	470.0	**CENTRAL**	**1761.96**
		SRB	117620	SRB	12.0	**EXPAND**	**687.02**
		TOT	15898K	RCT	86.1		
		/SEC	4416	IIT	4.2	SHARED	8.55
				HST	0.0		
				APPL %	15.9		

TRANSACTIONS		---SERVICE----		--SERVICE RATES--		----STORAGE----	
AVG	1.07	IOC	217436	ABSRPTN	3164	AVG	1183.70
MPL	1.04	CPU	3780K	TRX SERV	3078	TOTAL	1236.26
ENDED	**18732**	**MSO**	**7876K**	TCB	385.0	**CENTRAL**	**904.13**
		SRB	24790	SRB	2.5	**EXPAND**	**332.13**
		TOT	11898K	RCT	0.9		
		/SEC	3305	IIT	3.8	SHARED	4.95
				HST	0.0		
				APPL %	10.9		

TRANSACTIONS		---SERVICE----		--SERVICE RATES--		----STORAGE----	
AVG	6.94	IOC	2008K	ABSRPTN	4161	AVG	1648.53
MPL	6.94	CPU	26492K	TRX SERV	4160	TOTAL	11436.8
ENDED	**17737**	MSO	74703K	TCB	2698.9	**CENTRAL**	**7346.85**
		SRB	713938	SRB	72.7	**EXPAND**	**4089.98**
		TOT	103917K	RCT	1.9		
		/SEC	28866	IIT	27.2	SHARED	37.82
				HST	0.0		
				APPL %	77.8		

Workload Activity Report - 64-bit mode - TSO workload by period - MSO=0.1000

TRANSACTIONS		---SERVICE----		--SERVICE RATES--		----STORAGE----	
AVG	1.76	IOC	160392	ABSRPTN	3986	AVG	1341.19
MPL	1.72	CPU	5301K	TRX SERV	3909	TOTAL	2309.13
ENDED	**186576**	**MSO**	**19135K**	TCB	540.1	**CENTRAL**	**2309.13**
		SRB	111519	SRB	11.4	**EXPAND**	**0.00**
		TOT	24708K	RCT	71.7		
		/SEC	6863	IIT	3.9	SHARED	7.97
				HST	0.0		
				APPL %	17.4		

TRANSACTIONS		---SERVICE----		--SERVICE RATES--		----STORAGE----	
AVG	1.21	IOC	206859	ABSRPTN	4060	AVG	1343.09
MPL	1.18	CPU	4082K	TRX SERV	3964	TOTAL	1582.89
ENDED	**30563**	**MSO**	**12906K**	TCB	415.9	**CENTRAL**	**1582.89**
		SRB	28162	SRB	2.9	**EXPAND**	**0.00**
		TOT	17224K	RCT	0.7		
		/SEC	4784	IIT	3.8	SHARED	5.47
				HST	0.0		
				APPL %	11.8		

TRANSACTIONS		---SERVICE----		--SERVICE RATES--		----STORAGE----	
AVG	7.24	IOC	2825K	ABSRPTN	4734	AVG	1813.06
MPL	7.24	CPU	22510K	TRX SERV	4734	TOTAL	13120.0
ENDED	**20950**	**MSO**	**97044K**	TCB	2293.2	**CENTRAL**	**13120.0**
		SRB	952530	SRB	97.0	**EXPAND**	**0.00**
		TOT	123332K	RCT	1.2		
		/SEC	34259	IIT	36.7	SHARED	41.19
				HST	0.7		
				APPL %	67.5		

Using these examples of TSO workloads from a production environment some native workload variability is to be expected. First period TSO will tend to have less variability than other periods because the resource consumed in first period is controlled by the duration definition.

In the 31-bit mode environment the total TSO ended transaction count was 216,398 for the hour, and 85% of the transactions ended in 1st period. For the 64-bit mode environment the total TSO ended transaction count was 238,089 for the hour and just 78% of the transactions ended in 1st period. The 64-bit mode RMF report still shows a field for expanded storage but of course there is no expanded on the image. Since the MSO user coefficient in this example had already been reduced to 0.1000 the impact of the 64-bit mode migration is noticeable, but not as drastic as may be seen with the MSO value equal to 3.0 or 1.0.

Generally, most installations don't do chargeback accounting on total service units, but rather build chargeback systems based on CPU and SRB service units. Sometimes chargeback systems will also include I/O service. It is still worthwhile to check any chargeback systems to determine if storage service units are used in the system, and if they are, then review the impact, if any, of the migration to 64-bit mode.

1.2.0 System High UIC

The average system-high UIC (unreferenced interval count) is an indicator of central storage contention. RMF, via a PAGING report, shows the minimum, maximum, and average system-high UIC values. The system-high UIC measures how long, in seconds, a page has remained unreferenced in central storage. The system-high UIC is actually an inverse measure of central storage reference. When references to central storage are high, the system-high UIC is low; the number of seconds between references is low, therefore, the page is being referenced frequently. When references to central storage are low, the system-high UIC is high; the number of seconds between references is high, therefore, the page is being referenced infrequently.

The system-high UIC is the highest of the central storage page frame UICs, and the average value is the average over the interval of the highest page UIC. The UIC can be a valuable indicator of storage contention; average system-high UIC is used in both system MPL adjustments and logical swap algorithms. The current system-high UIC is the starting age for stealing frames. The UIC can be misleading because of the method in which it is derived. For example, a single page in central storage can have a UIC of 254, and the next highest page could have a UIC of 50. In such a situation the average system-high UIC would be reported as 254, an accurate but perhaps misleading view of storage.

Other factors can impact the calculation of system-high UIC. If storage isolation has been turned on for an address space or the Common Area, the central storage frames backing these areas are not used when determining the system-high UIC. Storage isolation is enabled by either an IEAIPSxx specification (PWSS and CWSS) or dynamically by WLM. There are times when SRM working set management decides to protect an address space which is paging by turning on central storage isolation again impacting the system-high UIC determination. Logically swapped address spaces are also excluded when determining the system-high UIC.

In 31-bit mode UIC updating occurs approximately every second though it can be a longer interval if central storage contention is low. In 64-bit mode UIC updating is changed to review central storage page reference indicators every 10 seconds, longer if central storage contention is low. The UIC determination is changed with 64-bit mode to accommodate the larger central storage sizes, and to keep the costs of identifying the UIC equal to previous implementations.

Internally the UIC is managed as a value of 0 to 254, though externally the value is adjusted by 10 by RMF to better reflect the duration across which the UIC is determined. For OS/390 R10 and z/OS 1.1, the value in the RMF record for paging statistics (SMF 71) will match the internal values for the minimum (SMF71LIC) and maximum (SMF71HIC) UIC values, so a maximum UIC would be recorded as x'FE' or decimal 254. The internal average UIC value is multiplied by 10 to record the UIC in tenths of a second (SMF71ACA), so a maximum UIC would be recorded as x'9EC' or decimal 2540. For z/OS 1.2 and later all of the UIC fields reported in RMF are adjusted by 10.

Installations using OEM monitoring products should ensure they understand the scale being used when any system-high UIC value is displayed.

Many installations use average system-high UIC as an indicator of central storage contention. In order to get the best feel for storage contention it is recommended to review not just the average system-high UIC, but also the minimum and maximum system high UIC values. Due to some of the issues with the underlying metric other methods of determining central storage "health" may need to be developed to ensure the processor storage provided is sufficient to support the workloads. This next section will discuss some alternatives.

In OS/390 V2R4 new support was added to give more granularity to the central storage picture. Three buckets were created which sorted central storage frames into high, medium, or low impact frames. Frames are assigned to one of the four buckets, and RMF places in the SMF 71 type record the counts for each of the four buckets. The counts are reported for the minimum, maximum, and average frame counts within the interval, as well as available frames in the interval. Available frames counted as part of the UIC updating are smaller than the previous available frame counts because frames kept free to support the MCCAFCTH parameter are not counted in the SMF71CAA/M/X fields. Counts are available for both central storage and expanded storage usage. The data in the fields are relative to each other. What this means is in order to use the counts, the information from the previous bucket(s) needs to be "remembered" and used to calculate the delta's off of the current bucket. These counts are not available in an RMF post-processor report so another reporting tool which post-processes the SMF data is required to obtain the information.

A better picture may emerge by using these buckets to get a feel for how the storage is being referenced by the workloads on the image. By looking at the ratio of the breakdown of frames on the system usage patterns may begin to emerge. As more central storage frames migrate from the low impact category into either the medium or high impact buckets additional processor storage may need to be planned for the processor. Again these buckets are calculated only for frames which are not supporting logical swap users, or storage isolated address spaces.

The SMF 71 fields containing the frame counts with the impact designation are found in:

FIELD	**Comments**
SMF71CLM	Minimum number of low impact frames
SMF71CLX	Maximum number of low impact frames
SMF71CLA	Average number of low impact frames
SMF71CMM	Minimum number of medium impact frames
SMF71CMX	Maximum number of medium impact frames
SMF71CMA	Average number of medium impact frames
SMF71CHM	Minimum number of high impact frames
SMF71CHX	Maximum number of high impact frames
SMF71CHA	Average number of high impact frames
SMF71CAM	Minimum number of available frames
SMF71CAX	Maximum number of available frames
SMF71CAA	Average number of available frames

The data is placed in the different buckets based on the UIC value for the frames. These values are set internally by z/OS and are not a defined external of the operating system. The actual UIC values for the different buckets are subject to change based upon the design enhancements needed for z/OS. These changes may be made via either a new release of z/OS or through the APAR process.

The UIC values used to assign frames to the different buckets are currently set as follows in z/Architecture (based on a UIC scale of 1 to 254) :

- **Low Impact:** Pageable frames unreferenced in 150 seconds or more and not protected (storage isolated) + Available
- **Medium Impact:** Pageable frames unreferenced in 20 seconds or more and not protected + available
- **High Impact:** All pageable frames that are not protected + available.

1.2.1 RMF Measurement Data and Reports - Average System-High UIC

Understanding the impact of storage protection on UIC is critical in understanding the value of the average system-high UIC value. A method to determine the amount of storage isolation is to use an on-line monitor such as RMF Monitor II, and the RMF ASD report. An example of one of these reports is presented below:

```
                    RMF - ASD Address Space State Data          Line 1 of 211

                  MIG=19.1K CPU= 20     UIC= 254 PR=   0    System= WSCM Tota
17:02:22          S C  R  DP   CS   ESF   CS    TAR  X  PIN   ES  TX  SWAP   WSM
JOBNAME  SRVCLASS P L  LS PR    F        TAR   WSS  M   RT   RT  SC    RV    RV

*MASTER* SYSTEM   1 NS    FF 1046    76          0    ---- ----   0     0
CONSOLE  SYSTEM   1 NS    FF  209    24          0  X ---- ----   0     0
WLM      SYSTEM   1 NS    FF  831    79          0  X ---- ----   0     0
ANTMAIN  SYSTEM   1 NS    FF  285    58          0  X ---- ----   0   998
ANTAS000 SYSSTC   1 NS    FE  149    25          0  X ---- ----   0    99
OMVS     SYSTEM   1 NS    FF 1523  1161       2271  X ---- ----   0     0
JESXCF   SYSTEM   1 NS    FF   66    19          0  X ---- ----   0   998
ALLOCAS  SYSTEM   1 NS    FF  241    39          0  X ---- ----   0     0
IOSAS    SYSTEM   1 NS    FF  328   109          0  X ---- ----   0     0
IXGLOGR  SYSTEM   1 NS    FF  761    27          0  X ---- ----   0     0
SMF      SYSTEM   1 NS    FF  198    13          0  X ---- ----   0     0
VLF      SYSSTC   1 NS    FE 8426  3219       8756  X ---- ----   0   998
LLA      SYSSTC   1 NS    FE 1101    12        736  X ---- ----   0   998
```

In the RMF ASD report (Address Space State Data Report), the CS TAR (central storage target) field is interpreted as:

- blank - address space is not monitored
- zero - address space is monitored but not managed
- nonzero - it is monitored and working set managed

CS TAR and TAR WSS (target working set size) are mutually exclusive fields on the ASD report. CS TAR is the working set manager assigned target working set size and the TAR WSS is the target working set size assigned by storage isolation.

A monitored address space is one where implicit block paging has been enabled by SRM. Implicit block paging occurs when blocks of pages, which are contiguous in virtual storage and have the same UIC, are formed at steal time and moved to Auxiliary. The maximum number of pages which can be implicitly block paged is 256 (1MB). When a page fault occurs on a page in the block, the faulted page and other pages in the block in ascending virtual storage order will be loaded subject to available central storage. It is true implicit block paging from auxiliary may waste CPU cycles by preloading pages which are not referenced, but the potential DASD response time savings versus the CPU cost justifies always doing the block paging in from auxiliary.

For a monitored address spaces, SRM will calculate a working set manager recommendation value capable of overriding the swap recommendation value. The working set manager recommendation value measures the value of adding the address space to the current mix of work in the system. Even when the swap recommendation value indicates a specific address space should be swapped in next, the working set manager recommendation value might indicate the address space should be bypassed. In order to ensure no address space is repeatedly bypassed, the system swaps in a TSO/E user 30 seconds after being bypassed. For all other types of address spaces, the system will swap in the address space 10 minutes after being bypassed.

If a monitored address space is paging heavily, SRM may manage its central storage usage. When an address space is managed, SRM imposes a central storage target (implicit dynamic central storage isolation maximum working set) on an address space. Any frames held by the address space over the central storage target are considered preferred steal candidates.

Also in MON II, check the ARD report. Look for any address space with a CR column indicating storage protection. The CR column is an indication of whether WLM is managing the address space as storage critical and/or CPU critical during the reporting interval.

- S - storage critical
- C - CPU critical
- SC - Both storage & CPU critical

RMF - ARD Address Space Resource Data Line 1 of 78

MIG=20.6K CPU= 21 **UIC= 254** PR= 0 System= WSCM Total

17:20:09 JOBNAME	DEV CONN	FF 16M	PRV FF	LSQA CSF	LSQA ESF	X M **CR**	SRM ABS	TCB TIME	CPU TIME	EXCP RATE	SWAP RATE	LPA RT	CSA RT	NVI RT	V&H RT
MASTER	19635	0	816	85	23		0.0	884.1	10805	0.03	0.00	0.0	0.0	0.0	0.0
PCAUTH	0.000	0	0	51	4	X	0.0	0.02	0.07	0.00	0.00	0.0	0.0	0.0	0.0
RASP	0.000	---	---	----	----	X	0.0	0.01	37.59	0.00	0.00	0.0	0.0	0.0	0.0
TRACE	0.117	0	1	51	4	X	0.0	0.01	0.07	0.00	0.00	0.0	0.0	0.0	0.0
DUMPSRV	45.72	0	0	63	6		0.0	12.55	18.85	0.00	0.00	0.0	0.0	0.0	0.0
XCFAS	2114	0	246	1183	13	X	0.0	838.9	942.6	1.27	0.00	0.0	0.0	0.0	0.0

Information on storage critical workloads is also available via RMF Monitor III.

Average system-high UIC values are also reported by RMF on a Paging Activity Report. Review UIC in terms of the minimum, maximum, and average especially if the interval under review is large (15 minutes or more). Storage contention factors can change rapidly and looking at the minimum gives the best picture of storage contention.

```
OPT = IEAOPTPM    MODE = ESAME           CENTRAL STORAGE MOVEMENT RATES
-------------------------------------------------------------------------
   HIGH UIC (AVG) = 2540.0   (MAX) = 2540   (MIN) = 2540
```

In RMF the term MODE=ESAME is a synonym for 64-bit mode of operation.

1.3.0 Available Frame Queue

The available frame queue is a cache of available frames the system uses to satisfy demand for pages in an expeditious manner. In a central / expanded storage implementation there are two available frame queues, one for central storage, and one for expanded storage. When work needs to be pushed out from central storage due to a sudden demand, the work, if eligible for expanded storage, will use frames from the expanded storage available frame queue. In order to maintain the cache of available frames the system keeps two values, an OK threshold, and a LOW threshold. When available frames fall below the LOW threshold the system will steal central storage pages using an LRU algorithm until enough frames are acquired to reach the OK threshold at which time stealing ends. A similar mechanism works for expanded storage.

In 64-bit mode there is only the central storage available frame queue so the total number of pages held in reserve within the processor has been reduced. Installations have the ability to influence both the LOW and OK values governing the depth of the available frame queue through definitions in the IEAOPTxx member of parmlib. This capability exists in both WLM goal mode and compatibility mode. The central storage available frame queue is governed by the IEAOPTxx parameter, MCCAFCTH=(xxxxx,yyyyy) and the expanded storage available frame queue is governed by the IEAOPTxx parameter, MCCAECTH=(xxxxx,yyyyy).

The defaults for these two parameters in ESA/390 mode are MCCAFCTH=(50,100), and MCCAECTH=(150,300). So the OK point for available frames in a 31-bit mode implementation is 400 frames, 100 from central storage and 300 from expanded storage. In order to provide the same amount of available storage in a 64-bit mode implementation the recommendation is to adjust the setting for the LOW and OK thresholds for central storage available frames (MCCAFCTH). It is recommended to set MCCAFCTH=(400,600) in order to provide an equivalent buffer of available frames in a 64-bit mode environment.

1.3.1 RMF Measurement Data and Reports - Available Frame Queue

RMF reports the minimum, maximum and average number of available frames for central and expanded storage. RMF also provides the ability to use RMF trace records to get additional information on the central and expanded storage available frame queue, LOW (RCEAFCLO) and OK (RECAFCOK) levels, as well as the count of times an available frame queue LOW condition (RCEAVQC) was seen in the interval.

These RMF reports are created by inserting RMF statements into the RMF parmlib, ERBRMF02. The insertion of these records will cause RMF to cut SMF type 76 records with the information.

```
TRACE(RCEAECLO,ALL)        /* TRACE AVAIL ESTOR LOW THRESHOLD      */
TRACE(RCEAECOK,ALL)        /*       AVAIL ESTOR OK THRESHOLD       */
TRACE(RCEAFCLO,ALL)        /*       AVAIL CSTOR LOW THRESHOLD      */
TRACE(RCEAFCOK,ALL)        /*       AVAIL CSTOR OK THRESHOLD       */
TRACE(RCEAFC,ALL)          /*       TOTAL OF ALL AVAIL FRAMES      */
TRACE(RCVAFQA,ALL)         /*       AVG AVAIL FRAME COUNT          */
TRACE(RCVAVQC,ALL)         /*       AVQ LOW COUNT                  */
```

The complete list of variables which can be traced by RMF can be found in the *RMF Users Guide*, SC33-7990.

In a 64-bit mode environment you can't trace any of the expanded storage OK and LOW conditions. To format out the records once they have been generated place in the JCL for the RMF Post-Processor a control statement: REPORTS(TRACE).

Below is an example of the type of output generated by the RMF post-processor.

T R A C E A C T I V I T Y

TIME	*		**RCEAFCLO**			**RCEAFCOK**	
MM.SS.TT	*	MINIMUM	AVERAGE	MAXIMUM	MINIMUM	AVERAGE	MAXIMUM
09.00.00	*	83	83.00	83	166	166.00	166
MAXIMUM	*	83	83.00	83	166	166.00	166
MINIMUM	*	83	83.00	83	166	166.00	166
AVERAGE	*	83.00	83.00	83.00			

TIME	*		**RCVAVQC**	
MM.SS.TT	*	MINIMUM	AVERAGE	MAXIMUM
09.00.00	*	0	0.00	0
MAXIMUM	*	0	0.00	0
MINIMUM	*	0	0.00	0
AVERAGE	*	0.00	0.00	0.00

Since the data is in an RMF SMF record any post processing tool should be able to collect and report on the data. Reviewing the available frame counts (minimum, maximum, and average) provides the performance analyst another view of processor storage contention levels and the demand being exerted upon the configuration. Frequent available frame queue low conditions indicate central storage can be considered "full" and additional definition of virtual storage will result in pages being sent to Auxiliary.

Installations may wish to review the count of AVQ low conditions. If they are happening very often then one tuning action which may be considered is to increase the MCCAFCTH thresholds (each sub-parm equally) to provide a larger cache of available frames. This may be especially prudent if the workload's demand for frames is very high such that it empties the available frame queue frequently.

Whenever the available frame queue needs to be replenished stealing (central storage frames moved to auxiliary) will be initiated. The pages being sent to auxiliary will be selected based on an LRU (Least Recently Used) basis, causing the oldest pages in processor storage to be moved to auxiliary. The page movement to auxiliary is not necessarily bad. The condition to watch is the frequency the pages must be brought back into the system via demand paging. Demand page-ins and their impact on performance and throughput are the key contention indicators to review if page movement is being seen from auxiliary. WLM in a goal mode system will be able to detect delay due to demand paging and WLM has the capability to dynamically institute storage working set management to help protect workloads based on their current performance and importance.

1.3.2 RMF Measurement Data and Reports - Demand Paging

Below is an example of a demand paging report taken from a 64-bit mode system. The report below is broken into two pieces, the page out load and the page in load. The best way to review this is to look first at the load leaving the processor, by reviewing the page out section. This will first tell if the workload being moved is due to swapping or stealing / trim, and gives an indication of what types of virtual pages are being impacted.

Then look at the page in section. This is the section where pain, if any is present, will first show because these are the pages coming in from slow DASD data sets. Identify if the pages are from common, or are related to address spaces and their storage demands. Look to see if the swap load has increased (see the section on effective logical swap if it has), or if the increase is due to demand paging.

	PAGE IN					PAGE OUT			
CATEGORY	SWAP	NON SWAP, BLOCK	NON SWAP, NON BLOCK	TOTAL RATE	% OF TOTL SUM	SWAP	NON SWAP	TOTAL RATE	% OF TOTL SUM
PAGEABLE SYSTEM AREAS (NON VIO)									
LPA		0.52	0.44	0.96	0				
CSA		0.40	0.59	0.99	0		1.49	1.49	1
SUM		0.92	1.03	1.95	1		1.49	1.49	1
ADDRESS SPACES									
HIPERSPACE		0.00		0.00	0		0.00	0.00	0
VIO		0.00		0.00	0		0.00	0.00	0
NON VIO	79.09	94.80	98.13	272.01	99	76.10	181.56	257.66	99
SUM	79.09	94.80	98.13	272.01	99	76.10	181.56	257.66	99
TOTAL SYSTEM									
HIPERSPACE		0.00		0.00	0		0.00	0.00	0
VIO		0.00		0.00	0		0.00	0.00	0
NON VIO	79.09	94.80	100.08	273.96	100	76.10	183.06	259.15	100
SUM	79.09	94.80	100.08	273.96	100	76.10	183.06	259.15	100
SHARED			0.16	0.16			0.41	0.41	

PAGE MOVEMENT WITHIN CENTRAL STORAGE	1.61
AVERAGE NUMBER OF PAGES PER BLOCK	6.3
BLOCKS PER SECOND	15.02
PAGE-IN EVENTS (PAGE FAULT RATE)	115.10

The product of BLOCKS PER SECOND and AVERAGE NUMBER OF PAGES PER BLOCK should be very close to the reported NON SWAP, BLOCK PAGE IN rate. The PAGE FAULT RATE is the sum of NON SWAP,NON BLOCK, NON VIO PAGE IN and BLOCKS PER SECOND. A page-in event is either a single page or a block page transfer. The page fault rate is the best indicator to use to review application delay due to central storage constraint.

Analysts are cautioned to review multiple intervals of data when looking at storage usage patterns. Reviewing the smallest available intervals are best when trying to understand storage reference patterns. Looking at reports which cover an hour or a shift or a day do not provide enough granularity to see the storage dynamics. A useful way of doing this is to use the RMF Overview reports and snap out a key set of indicators across shifts or days. Another method is to

use an RMF Summary report which shows demand paging by interval. Below are a set of OVERVIEW reports showing storage contention indicators across an active period. One line is generated per RMF interval. This data came from an internal benchmark system where RMF was set to a 1 minute interval. Once a specific time frame has been identified online monitors such as RMF Monitor 3 are useful to review storage conditions.

DATE MM/DD	TIME HH.MM.SS	INT MM.SS	TOTPAGRT	DEMPAGNG	PAGEFALT	AVGUIC	MINAVAIL
01/14	15.08.00	01.00	0.016	0.016	0.016	254	397870
01/14	15.09.00	01.00	10.733	1.666	1.666	254	221109
01/14	15.10.00	01.00	60.950	15.250	15.250	254	43341
01/14	15.11.00	01.00	4.183	3.050	3.050	200	0
01/14	15.12.00	01.00	0.083	0.083	0.083	85	0
01/14	15.13.00	00.59	0.066	0.050	0.050	70	0
01/14	15.14.00	00.59	2.550	0.133	0.133	65	0
01/14	15.15.00	01.00	0.033	0.033	0.033	59	0
01/14	15.16.00	00.59	18.666	0.400	0.400	56	0
01/14	15.17.00	01.00	3.150	0.133	0.133	64	0
01/14	15.18.00	01.00	240.310	199.488	0.554	71	0
01/14	15.19.00	01.00	76.183	56.300	56.300	69	0
01/14	15.20.00	01.00	220.135	70.628	33.181	83	0
01/14	15.21.00	00.59	415.877	157.310	5.133	77	0
01/14	15.22.00	00.59	513.217	260.575	85.452	60	0
01/14	15.23.00	01.00	806.716	642.223	126.613	72	0
01/14	15.24.00	00.59	707.891	355.409	328.116	73	0
01/14	15.25.00	00.59	1305.169	605.685	543.513	71	0
01/14	15.26.00	00.59	1277.047	591.712	586.696	71	0
01/14	15.27.00	00.59	885.562	575.202	575.202	71	0
01/14	15.28.00	01.00	1623.283	882.183	447.350	68	0

Field	Description
TOTPAGRT	Total number of pages per second (PIN +POT + SIN +SOT + VIN +VOT + BLP + HOT + HIT) / Interval
DEMPAGNG	Demand Paging per second (PIN +POT) / Interval
PAGEFALT	Page Faults per second (PIN) / Interval
AVGUIC	Average High UIC
MINAVAIL	Minimum number of available central storage frames

PIN - Page In
POT - Page Out
SIN - SWAP In
SOT - SWAP Out
VIN - VIO In
VOT - VIO Out
BLP - Block Pages
HOT - Hiperspace Out
HIT - Hiperspace In

The RMF Overview control statement used to create the above are listed below:

```
//RMFPP2    EXEC PGM=ERBRMFPP,REGION=0M
//MFPINPUT DD   DISP=(SHR,PASS),DSN=*.RMFSORT.SORTOUT
//MFPMSGDS DD   SYSOUT=*
//SYSIN     DD   *
 SYSOUT(O)
 NOSUMMARY
 OVERVIEW(REPORT)
 OVW(TOTPAGRT(TPAGRT))
 OVW(DEMPAGNG(DPAGRT))
 OVW(PAGEFALT(PAGERT))
 OVW(AVGUIC(AVGHUIC))
 OVW(MINAVAIL(CSTORAVM))
/*
```

If the page-ins are address space related (not VIO, and not hiperspace related) then the next step is to review which workloads are being impacted by the demand paging. This is done by looking at the SMF 72 records or an RMF Workload Activity report. Online monitors may also prove very useful for this type of review. Below is a copy of an RMF Workload Activity report showing the paging impacts on specific workloads. The specific report shown is for first period TSO.

```
                          W O R K L O A D   A C T I V I T Y

                                   DATE 01/14/2002          INTERVAL 01.00.000   MODE = GOAL
                          RMF      TIME 15.28.00
                          SERVICE CLASS=TSO_DEF    RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT                              PAGE-IN RATES    ----STORAGE----
 AVG      0.16   ACTUAL                114                              SINGLE    90.9   AVG      584.44
 MPL      0.03   EXECUTION             114                              BLOCK     83.1   TOTAL     14.89
 ENDED      84   QUEUED                  0                              SHARED     0.0   CENTRAL   14.89
 END/S    1.40   R/S AFFINITY            0                              HSP        0.0   EXPAND
 #SWAPS     86   INELIGIBLE              0                              HSP MISS   0.0
 EXCTD       0   CONVERSION              0                              EXP SNGL   0.0   SHARED     0.00
 AVG ENC  0.00   STD DEV               127                              EXP BLK    0.0
 REM ENC  0.00                                                          EXP SHR    0.0
 MS ENC   0.00

 VELOCITY MIGRATION:   I/O MGMT   7.5%    INIT

        ---RESPONSE TIME---  EX   PERF   AVG   --USING%--  ------------ EXECUTION DELAYS %
        HH.MM.SS.TTT         VEL  INDX  ADRSP   CPU   I/O  TOTAL  SWIN  CPU  MPL AUX  AUX
 GOAL                       50.0%                                          PRIV COMM
 ACTUALS
 COF1                        5.1%  9.8    4.1   0.2   0.1    3.8   3.0  0.2  0.2  0.2  0.1
```

This first period TSO workload is seeing delay due to 91 single demand page-ins, and 83 block page-ins. The first page in of a block (page actually faulted on) is counted in SINGLE, and the remaining block pages are counted in BLOCK.

The page-in rates are spread across all of the TSO transactions ended in this period, as well as the TSO transactions who transition through 1st period and end in later periods.

WLM in goal mode will show the impact of demand paging on the transaction response time. Review the execution delay % to see the impact the service class is seeing from the demand

paging activity. In the sample above execution delay is being seen from storage contention. WLM is reporting SWIN delay (swap-in has started but not completed), delay from MPL (work is ready but swap in has not started), and delay from auxiliary paging from both the local page data sets and the common page data set. A description of all of the delay types can be found in the *RMF Report Analysis*, SC28-1950.

What can be seen is even though there is a large amount of demand paging activity it is spread across a large transaction load and so the impacts on end user response times are relatively minor. The TSO workload in the RMF report is not busy and came from a benchmark system and is used primarily to demonstrate the type of WLM delays which can be reported.

Again RMF Overview reports can be very useful to get a feel for the impact of paging on a workload across a longer interval. Such an approach is very useful after identifying the specific workloads which are paging. Below is an overview report looking at the TSO 1st period and the single and block paging impacts. Report also shows the paging impact on the ops_high service class. The ops_high service class in this benchmark generated data contained the Websphere address space. Again the demand paging impact was high but it was spread across all of the web transactions and again the delay to end user response time was not enough to cause the workload to miss the WLM defined performance objective.

DATE MM/DD	TIME HH.MM.SS	INT MM.SS	STSO1SNG	STSO1BLK	SOPHSING	SOPHBLCK
01/14	15.09.00	01.00	0.0	0.0	0.0	0.0
01/14	15.10.00	01.00	0.0	0.0	0.0	0.0
01/14	15.11.00	01.00	0.0	0.0	0.0	0.0
01/14	15.12.00	01.00	0.0	0.0	0.0	0.0
01/14	15.13.00	00.59	0.0	0.0	0.0	0.0
01/14	15.14.00	00.59	0.0	0.0	0.0	0.0
01/14	15.15.00	01.00	0.0	0.0	0.0	0.0
01/14	15.16.00	00.59	0.0	0.0	0.0	0.0
01/14	15.17.00	01.00	0.0	0.0	0.0	0.0
01/14	15.18.00	00.59	0.0	0.0	0.0	0.0
01/14	15.19.00	01.00	0.0	0.0	4.5	1.5
01/14	15.20.00	01.00	0.0	0.0	2.2	0.9
01/14	15.21.00	00.59	0.0	0.0	0.1	0.2
01/14	15.22.00	00.59	0.0	0.0	5.6	2.1
01/14	15.23.00	01.00	9.1	1.1	7.2	4.4
01/14	15.24.00	00.59	25.2	134.1	20.6	21.4
01/14	15.25.00	00.59	9.5	4.8	35.0	42.9
01/14	15.26.00	00.59	2.1	0.0	33.2	48.3
01/14	15.27.00	00.59	102.0	47.4	38.6	12.2
01/14	15.28.00	01.00	90.9	83.1	32.8	11.7

Field	Description
STSO1SNG	Service class TSO, 1st period, single page ins
STSO1BLK	Service class TSO, 1st period, block page ins
SOPHSING	Service class ops_high, single page ins
SOPHBLCK	Service class ops_high, block page ins

1.4.0 Fixed Frames below 16MB

As faster processors have been introduced an increased need arose for more than 2GB of central storage. Additional storage is needed to allow workloads to take advantage of the increased processing power, and to support newer workloads which have significant demand for central storage. The response to these requirements was the introduction of z/Architecture and it's support for 64-bit mode real and virtual addressing.

One area still needing management is the pressure on using real storage backed below the 16MB line. Of paramount importance is the impact and pressure on the use of fixed frames below the 16MB line. Workloads causing stress on the availability of fixed frames below the 16MB line can cause significant virtual storage shortages which result in the SRM reducing multiprogramming levels and impacting the overall throughput of the system.

Currently the largest demand for fixed frames below the 16MB line is the storage used by SQA (which by definition is fixed) and by page fixed LSQA. Other users of fixed frames below the line are generally application dependent and installations should monitor the usage of this storage, especially as workloads are consolidated or grow to take advantage of the large zSeries processors.

For instance program storage is pagefixed while I/O is being done to fetch the program. For RMODE 24 programs storage with the LOC=(BELOW,BELOW) attribute is used because there might be RMODE 24 programs expecting their program storage to be in 24 bit addressable real storage when fixed. This means although the program fetch I/O does not require real storage below 16MB, it gets it anyway in order to allow subsequent fixes to be below 16MB.

So the fetching of RMODE 24 programs does create a demand for real storage below 16MB. For any one program, it is a short term demand, but it drives page movement within central storage, since these below 16MB frames need to be exchanged for above 16MB frames due to other fetches needing the below 16MB frames.

Designation of V=R Areas

The system allocates virtual equals real (V=R) regions upon request by those programs which cannot tolerate dynamic relocation. Such a region is allocated contiguously from a predefined area of central storage and is non-pageable. Programs in this region will run in dynamic address translation (DAT) mode, although real and virtual addresses are equivalent.

Jobsteps can request to run in a V=R region by using the ADDRSPC=REAL keyword on the EXEC or JOB JCL statement. Installations should review their use of any V=R designations. It is possible to determine if any work has a V=R designation by reviewing the SMF 30 records. Field SMF30SFL has bit 0 set to 1 whenever a V=R jobstep runs. By reviewing this field it is possible to determine if the ADDRSPC=REAL designation is being used. Use of V=R limits the ability to exploit the entire range of real storage, as the storage needed for the job step is placed in central storage below the 16MB line.

The amount of storage allocated to V=R jobs is specified by the REAL parameter in the IEASYSxx parmlib member. The default for REAL= is 76, or 76K. If you are sure you have no requirement for running V=R jobs then specifying REAL=0 will disallow the use of the ADDRSPC=REAL keyword in any JCL.

Another pressure on the fixed storage below the 16MB line comes from EXCP users. EXCP processing will use SQA control blocks to hold a virtual EXCP request's translated channel program. It is possible for users to generate EXCP requests at such a rate the threshold provided by the system for these blocks is exceeded. At this point EXCP processing will acquire additional blocks from SP230 (private below the 16MB line) in the address space of the TCB which opened the data set. An example of this type of user is a large DFSORT job which is sorting a large number of records with a small blocksize.

It is possible to exhaust all of the available storage below 16MB depending upon the number of concurrent users generating this type of EXCP load. The result of this over-commitment would be either an ABEND804 or ABEND878. This failure can be corrected by decreasing storage below 16MB used by the application opening the data set, by increasing the size of the private area below 16MB for all address spaces, by reducing the size of the channel programs to be translated, or by reducing the number of open data sets in use by programs in the address space.

For DFSORT users one method of reducing the size of the channel programs required is to increase the data set blocksize thereby providing better performance by setting blocksize equal to half of the DASD track size. DFSORT's copy function can be used to copy the input data set to a larger blocksize data set before sorting the file.

1.4.1 RMF Measurement Data and Reports

The easiest method of reviewing the area below 16MB is to look at the RMF Paging Activity Report under the Frame and Slot counts. It gives information on fixed frames below the line. Look at the minimums as well as the average, and review periodically. Remember there are a fixed number of these frames (4096) which can be used in an MVS system.

FRAME AND SLOT COUNTS

CENTRAL STORAGE

(181 SAMPLES)	MIN	MAX	AVG
AVAILABLE	1076303	1501313	1380529
SQA	15,582	15,655	15,613
LPA	5,278	5,278	5,278
CSA	11,699	11,881	11,749
LSQA	36,326	38,023	37,128
REGIONS+SWA	540,280	965,361	661,049
TOTAL FRAMES	2113241	2113241	2113241

FIXED FRAMES

NUCLEUS	1,895	1,895	1,895
SQA	13,021	13,094	13,051
LPA	152	153	152
CSA	611	705	617
LSQA	24,316	26,015	25,096
REGIONS+SWA	9,699	23,963	14,568
BELOW 16 MEG	665	1,016	733
BETWEEN 16M-2G	39,329	43,062	40,707
TOTAL FRAMES	50,744	65,089	55,381

You can get a better understanding of the common storage map by using an RMF virtual storage report. In this report SQA is broken out from ESQA to better identify the fixed frame impact on storage below 16MB. This report is created with a REPORTS(VSTOR) RMF control card.

```
      STATIC STORAGE MAP                                                ALLOCATED
 AREA         ADDRESS   SIZE              ------------ BELOW 16M --------------
EPVT          37B00000  1157M                MIN              MAX              AVG
ECSA           6D3E000   782M          SQA  2512K 05.24.19  2620K 04.59.59  2576K
EMLPA          6D3D000     4K          CSA  1572K 05.00.50  1596K 05.15.59  1587K
EFLPA          6D3A000    12K
EPLPA          4267000  42.8M          ALLOCATED CSA BY KEY
ESQA           172F000  43.2M           0   1220K 04.59.59  1240K 05.15.59  1232K
ENUC           1000000  7356K           1     52K 05.00.50    56K 04.59.59    54K
----- 16 MEG BOUNDARY ------           2     36K 04.59.59    36K 04.59.59    36K
NUCLEUS         FCC000   208K           3      0K 04.59.59     0K              0K
SQA            D71000  2412K           4      0K 04.59.59     0K              0K
PLPA           C1E000  1356K           5     12K 04.59.59    12K 04.59.59    12K
FLPA           C14000    40K           6    136K 04.59.59   136K 04.59.59   136K
MLPA                0     0K           7     16K 04.59.59    16K 04.59.59    16K
CSA            900000  3152K          8-F   100K 04.59.59   100K 04.59.59   100K
PRIVATE          1000  9212K
PSA                 0     4K          SQA EXPANSION INTO CSA
```

1.5.0 Swapping

In order to use central storage more effectively and reduce processor and channel subsystem overhead, the SRM logical swap function attempts to prevent the automatic physical swapping of address spaces. Unlike a physically swapped address space, where the LSQA, fixed frames, and recently referenced frames are placed in expanded storage or on auxiliary storage, SRM keeps a logically swapped address space's frames in central storage.

Address spaces swapped for wait states (for example, TSO/E terminal waits) are eligible to be logically swapped out whenever the think time associated with the address space is less than the system threshold value. SRM adjusts this threshold value according to the demand for central storage. SRM uses the average system high UIC to measure this demand for central storage. As the demand for central storage increases, SRM reduces the system threshold value, as the demand decreases, SRM increases the system threshold value. This threshold value is determined by WLM in a goal mode system, and can be set in the IEAOPTxx parmlib member for a compatibility mode system.

In an all real environment the efficiency of effective logical swap becomes more important. As the effectiveness of logical swap decreases the amount of swap traffic to the local page data sets will increase. More importantly swappable workloads may start to see delays as they await swap-in processing. RMF SMF 71 records give information on swapping.

SWAP data sets are eliminated in OS/390 R10 for both 31-bit mode and 64-bit mode installations. Swapping controls work the same in this environment except the load is carried by the local page data sets.

Installations migrating to 64-bit mode should review the number of transition and request swaps. In 31-bit mode a workload undergoing a transition or request swap may be swapped into expanded storage. In 64-bit mode, and before z/OS 1.2, a workload's transition or request swaps are swapped via the auxiliary paging subsystem. This activity will place a greater dependency on the paging subsystem. In z/OS 1.2 transition swaps are done in real storage. New swap reason code 18 indicates an in-real swap has occurred, and is reported by both RMF and SDSF.

A transition swap may occur when the status of an address space changes from swappable to nonswappable. A transition swap will prevent the job step from improperly using reconfigurable storage. After the subsequent swap-in, frames are allocated from preferred storage and the address space is non-swappable. For example, the system performs a transition swap out before a nonswappable program or V=R step gets control. The ability to request a transition swap is also under program control via use of the SYSEVENT TRANSWAP service. SYSEVENT DONTSWAP / OKSWAP sequences are used to make work non-swappable for short periods of time, generally less than 1 minute, and do not cause a transition swap to occur. The associated LSQA and private area pages are not necessarily put into preferred storage for this type of swap sequence since the workload will be non-swappable for a short period of time.

A request swap occurs whenever a processor storage element is configured offline to allow movement of work out of the area designated to be removed. In z/OS 1.2 changes were made to perform an in-real swap instead.

In order to eliminate the impact of transition swaps, installations should review their use of the IEASYSxx parameters RSU and REAL. Installations should set RSU=0 and REAL=0 if they have no requirement for reconfiguring storage or a need to run a V=R job. This setting will cause all SYSEVENT TRANSWAPs to be treated as a NOP. In 31-bit mode having an RSU=0 and REAL=0 meant there was no non-preferred storage in the system and a SYSEVENT TRANSWAP did not require physical swap-in. In 64-bit mode, segment tables are managed as non-preferred storage, regardless of the RSU and REAL settings. This means in 64-bit mode there can be conditions when RSM will need to do a REQSWAP to manage the segment tables. Any request swap or transition swap (via SYSEVENT or other method) is carried by the local page data sets until z/OS 1.2 which provides the in-real swap support.

1.5.1 RMF Measurement Data and Reports

Run an RMF post-processor report with the control cards REPORTS(PAGING) to generate reports like those below. This report will show the swap activity to auxiliary and will give an indication of the swap type. By tracking the logical swap effective percentages an installation will be able to further identify the contention for central storage, and the load which is being sent to the paging subsystem

```
OPT = IEAOPT00    MODE = ESAME              S W A P   P L A C E M E N T
                          *------ AUX STORAGE -------* *---LOGICAL SWAP--*
                          AUX                 AUX STOR
                          STOR      AUX STOR     VIA                  LOG SWAP
                          TOTAL     DIRECT    TRANSITION LOG SWAP   EFFECTIVE
TERMINAL        CT      237,108          0           0          0    237,108     237,108
INPUT/OUTPUT    RT        65.86       0.00        0.00       0.00      65.86       65.86
WAIT             %        96.2%       0.0%        0.0%       0.0%     100.0%      100.0%
LONG            CT        4,319          0           0          0      4,319       4,319
WAIT            RT         1.20       0.00        0.00       0.00       1.20        1.20
                 %         1.8%       0.0%        0.0%       0.0%     100.0%      100.0%
DETECTED        CT        4,559          0           0          0      4,559       4,559
WAIT            RT         1.27       0.00        0.00       0.00       1.27        1.27
                 %         1.8%       0.0%        0.0%       0.0%     100.0%      100.0%
UNILATERAL      CT          560          0           0          0        560         560
                RT         0.16       0.00        0.00       0.00       0.16        0.16
                 %         0.2%       0.0%        0.0%       0.0%     100.0%      100.0%
TRANSITION      CT            3          3           3          0          0           0
TO NON-         RT         0.00       0.00        0.00       0.00       0.00        0.00
SWAPPABLE        %         0.0%     100.0%      100.0%       0.0%       0.0%        0.0%
TOTAL           CT      246,549          3           3          0    246,546     246,546
                RT        68.49       0.00        0.00       0.00      68.49       68.49
                 %       100.0%       0.0%      100.0%       0.0%     100.0%      100.0%
```

1.6.0 SQA Definitions

A RAS APAR has been developed which changes the thresholds at which the SRM detects a common storage shortage below the line. APAR OW50225 increases the point where SRM issues the IRA100E and IRA101E messages. IRA100E signals SRM is stopping all new address spaces create functions, such as START, MOUNT, or LOGON commands, due to a critical shortage of common below the line. The IRA101E message signals the system will reject LOGON, MOUNT, and START commands until the shortage is relieved, and will fail GETMAIN requests from jobs requesting more SQA than is available. The changes are made to provide earlier notice so detection mechanisms into the shortage can be initiated sooner, allowing installations the ability to avoid an outage due to lack of common below the line.

Message	Reason	Before OW50225	After OW50225
IRA100E	SQA Shortage	8 4K pages (32K)	128 4K pages (512K)
IRA101E	Critical SQA Shortage	4 4K pages (16K)	64 4K pages (256K)

Some customers have voiced concerns the thresholds introduced by this RAS APAR were too aggressive for their environments. As a result APAR OW50225 was PE'ed (PTF in Error) and the default thresholds were changed to reduce the levels at which the IRA100E and IRA101E messages were issued. Also enhancements have been made to allow customers to further customize these values as needed prior to IPL by zapping the high and low thresholds. Customers are urged to review the APAR text for APARs OW50225, and OW54022.

Message	Reason	After OW54022
IRA100E	SQA Shortage	64 4K pages (256K)
IRA101E	Critical SQA Shortage	32 4K pages (128K)

Changes in the hardware configuration, for both processors and I/O definitions, can change the amount of SQA which must be defined. This includes adding new I/O devices, altering the amount of expanded storage, adding additional CPs, or changing the architectural mode. Sometimes the SQAs definition must be changed due to additional SQA requirements, such as new devices. Other times it's an effect of "rounding" the internal allocation of SQA to a 64K boundary. Or it may be a change in internal SQA usage patterns.

During IPL in 31-bit mode there is an initial amount of ESQA set aside for NIP use by MVS. During NIP, RSM will calculate the amount of ESQA needed to hold the expanded storage page tables, which will then be rounded to a 64K boundary. The storage itself is not used till later in the IPL. This amount of storage, in addition to any SQA specification made in IEASYSxx, is then set aside to ensure sufficient ESQA exists to support the IPL.

Some installations intentionally define a very small IEASYSxx SQA specification, and rely on SQA to overflow into CSA. This ability is not available during the entire IPL process, and storage shortages may result early in the IPL process if the specification is too small. The UCBs and other I/O related control blocks must reside in SQA. The system does not internally adjust the amount of allocated SQA based on the I/O configuration. Storage for these devices is part of the SQA specification in IEASYSxx. If a very small amount is specified for ESQA in IEASYSxx the IPL may fail. To correct the failure you must increase the amount of ESQA that is defined.

Increasing, reducing, or eliminating expanded storage, which always occurs when moving to 64-bit mode, can result in an ESQA shortage. The system may be using part of the excess ESQA added during rounding of the reserved amount to a 64K boundary in order to successfully IPL. When the amount of expanded storage is changed the excess could change from 63K to 0K, and this could cause an IPL failure.

Changes may be required during a migration to an all real environment. In 64-bit mode there is no expanded storage and so no ESQA set aside is done by RSM. There can be system conditions where a system during IPL is using the ESQA reserved for the expanded storage page tables for other purposes. It is not until later in the IPL sequence when the expanded storage page tables are moved from the MASTER address space into the area previously reserved for them in ESQA. So the period where the ESQA is set aside and the pages are actually copied the reserved ESQA can be used to satisfy other ESQA requests during IPL. This double use of the ESQA most often happened when the installation IPL'ed with an inappropriate ESQA parameter in IEASYSxx (the ESQA operand was set to 0 or very low).

If this previously undetected double use of the ESQA set aside for expanded storage control blocks is happening on a 31-bit mode system then migration to a 64-bit mode system will need more ESQA than is provided for in the current IEASYSxx definition. The absence of the expanded storage set aside may cause the IPL to fail with an ABEND 878. The recommendation is to set the ESQA operand correctly in IEASYSxx to meet the needs of the IPL. Minor over definition of ESQA will hardly be noticed in current systems and this over definition insures the system will have sufficient ESQA to operate.

It is doubtful DB2 or any other large virtual storage region user would be more or less successful if ESQA is over-defined, but the operating system supporting the entire environment may be much happier. There will be savings in ESQA due to the removal of expanded storage, and the movement of the central storage page table to an RSM owned Data Space. Installation's should assume 8MB of ESQA are saved for every 1GB of processor storage on the processor.

1.6.1 RMF Reports and Measurements - SQA thresholds

An RMF virtual storage report can be used to review the allocation of SQA and CSA below the lines to determine if the changes introduced by the RAS APAR will impact the system, and will govern if additional tuning should be undertaken to ensure sufficient reserves for common storage. Use RMF Control card REPORTS(VSTOR) to get the report listed below.

```
                     ---------- BELOW 16M ----------------
                       MIN              MAX              AVG
CSA
 FREE PAGES (BYTES)  1160K 10.25.59  1288K 10.00.49  1269K
 LARGEST FREE BLOCK  1156K 10.26.10  1276K 10.11.10  1255K
 ALLOCATED AREA SIZE 1864K 10.00.10  1996K 10.26.10  1876K
SQA
 FREE PAGES (BYTES)     0K 10.00.00    24K 10.21.19     4K
 LARGEST FREE BLOCK     0K 10.00.00     8K 10.21.19     2K
 ALLOCATED AREA SIZE 2412K 10.00.00  2412K 10.00.00  2412K
MAXIMUM POSSIBLE USER REGION -  9012K BELOW AND 1086M ABOVE
```

1.7.0 Virtual Storage Memory Limits

In z/OS 1.2 exploitation of 64-bit virtual storage is enabled. The virtual storage above 2GBs is organized as memory objects. Memory objects are created via use of the IARV64 service. Any 31-bit assembler program can use the support to obtain, store and manipulate data in the virtual storage above 2GB. Programs continue to be loaded and run in the first 2GB of storage (RMODE=31). The support for memory objects provides significant data caching capability to subsystems, and exploitation of virtual 64-bit support will require additional capacity planning to ensure enough real storage is provided to support the virtual storage definitions and / or the auxiliary paging subsystem can support an increased virtual definition.

There is no practical limit to the amount of virtual addresses an address space can request above 2GB. Instead limits are placed on the amount of useable virtual storage above 2GB which can be used by an address space at one time. This ability to limit the allocation of virtual storage above 2GBs is through the use of the MEMLIMIT support.

The limit placed on the use of virtual storage above 2GB is 0 unless one of the following is specified:

- New SMF MEMLIMIT parameter. This provides an installation control over virtual above 2GB. It is specified in the SMFPRMxx parmlib member.
- New MEMLIMIT keyword is specified in the JCL. Any JCL with MEMLIMIT specified will fail on an ESA mode system.
- REGION=0 specified on jobcard.
- IEFUSI exit sets a limit.

For more information on memory objects review the manual *z/OS MVS Programming: Extended Addressability Guide*, Document Number SA22-7614. Installations running z/OS 1.2 or later should review their current methods of limiting virtual storage to incorporate the use of virtual storage above 2GB.

2.0 Auxiliary Paging Data Sets

With the introduction of 64-bit mode addressing, and the elimination of expanded storage, installations should review the resources needed to support the auxiliary (AUX) paging subsystem. Even if currently there is little or no activity to the local page data sets, a migration to 64-bit mode should ensure the system has a robust paging subsystem. The auxiliary subsystem will now be the only resource available to support a workload disruption to central storage. This section will review the current recommended "best practices" for the auxiliary paging system. These recommendations are valid for both 31-bit and 64-bit mode systems.

RSM passes paging requests (read and write) to ASM for processing. RSM can request one or more pages per request. In response ASM builds CCWs to allow a "burst size" of data to be transferred to a local page data set. The maximum amount of data in the burst size for a 3380 or 3390 DASD device is 30 pages. The request is considered completed when the entire burst size is transferred. ASM will start the CCWs by adding the CCWs to the end of a running channel program, resuming a suspended channel program, or starting a new channel program. ASM will then notify RSM when processing for a page is complete.

When writing pages to a local page data set ASM selects a target data set to be used. The selection process ASM uses is to treat the set of eligible local data sets as a circular queue. ASM remembers the last data set used for writing a burst and evaluates the next data set on the queue. ASM will select a "good" performing data set. Goodness is determined by reviewing the average service time being seen by the page data set being evaluated relative to the overall average service time of all of the local page data sets. If no available slot space exists in any of the "good" performing local data sets then any local page data set will be used. Other restrictions apply regarding VIO pages and which local page data sets can contain VIO pages.

A performance benefit available to ASM is the use of the contiguous slot algorithm which allows the burst size to be transferred together to a cylinder. This allows the I/O access time to read or write a burst to be reduced. The search duration will vary based on the number of available slots so fewer slots may mean a longer search. If there are no contiguous slots available to hold the burst size then the pages will be allocated into individual slots. Based on empirical evidence, the contiguous slot algorithm is rendered ineffective whenever the page data sets are more than 30% allocated. The contiguous slot algorithm is not turned off when the local's slots are more than 30% allocated. Instead the likelihood of finding the required contiguous slots fell off when the data set reached this utilization level.

The benefit of contiguous slot in today's I/O subsystems with RAID DASD probably relates more to the impact on the cache controller than the impact on physical disk layout. High performance I/O is based on cache performance and paging I/O performance analysis would be no different. In a control unit like the 2105 (Shark), all I/O is cached. Define extent commands to bypass cache or inhibit cache load are still recognized but primarily their use is to influence the LRU algorithms used to manage the control unit. Note: prior to z/OS 1.3, ASM always specified the bypass cache attribute. Starting with z/OS 1.3, newer devices, such as the 2105, will not have the bypass cache attribute specified. Understanding page data set performance means reviewing performance of the cache controller.

Reading pages from the local page data sets is also limited by the maximum burst size. ASM does provide special handling for reading a single page to resolve a page fault, by using a special burst size of 1 slot.

1.2 Page Data Set Allocation - Size and Number

In a paging situation throughput from the Auxiliary paging subsystem is the key, and throughput is obtained by having a high level of parallel access to Auxiliary Storage. So after determining the amount of auxiliary storage needed, providing this capacity by more data sets is preferable to fewer, very large, page data sets. The minimum recommended number of local page data sets is four. The locals should be placed on high performance DASD, and configuring them for maximum parallel access, (on multiple volumes, across multiple control units, on well performing channel paths).

It is also recommended to allocate locals with roughly the same size. Having some large locals and small locals will work but may not provide optimal results. Use of the locals is generally round robin, except for specific poor performance situation, and because of this selection method issues can arise in relation to the contiguous slot allocation. Since ASM directs paging to each of the data sets equally, smaller locals will tend to fill before larger locals. Since the allocation level of the page data set is not considered (only the availability of slots is used) there may be situations where the large locals can still easily support the contiguous slot allocations but the smaller data sets have exceeded the 30% slots used guideline. In this case the paging load directed to the smaller local will allocate slots on an individual basis, increasing the search times and reducing the responsiveness of the I/O operation.

Installations should plan to not let local page data sets exceed 70% slot utilization. If the number of slots in use goes above 70% address space creation stops, SVC dump processing terminates, and other system slowdowns may be experienced. When reviewing the utilization of local page data sets ensure there are enough free slots to support the requirements of SVC dump processing. The number of slots required will be governed by the MAXSPACE keyword on the CHNGDUMP operator command. The default for the MAXSPACE parameter is 500 MB. Another guideline is to provide enough slots to allow a runaway address space to create a 2GB space without hitting the 70% allocated slots mark.

Because ASM reviews the service time of local page data sets when selecting a page data set the installation may choose to put other user data on the volume (including page data sets from other images). If other data becomes active to such an extent it impacts device service time ASM will dynamically move away from this data set to other, better performing data sets. ASM will check once in a while to see if the previously shunned local page data set should be reconsidered for use. Understand, any page fault being resolved from this local means the amount of time to read in this page can be impacted by the activity caused by the other user data.

Placing other data on the volume accessed from the same image will impact the suspend/resume channel protocol used by ASM for local page data sets. The performance benefits of suspend/resume channel programming may be very small given current processor speeds and the

sophistication of current cache controllers. When placing multiple page data sets on a volume shared by multiple systems it is important to ensure the owning system's name is a part of the page data set's name. This will help ensure a system does not inadvertently use another system's page data set, because they will have unique data set name.

If the installation places multiple local page data sets on a specific volume or places user data on the volume the use of parallel access volumes is recommended. This allows concurrent access to the volume as long as the activity is to different extents. In order to use PAVs, prior to z/OS 1.3, static PAVs would be required to be used because ASM does all of its device queuing internally. Since IOS queue length is the trigger for dynamic PAV's this function would not work for local page data sets.

In z/OS 1.3, ASM has written code to work with WLM to enable dynamic PAV management for volumes with local page data sets. ASM signals to WLM the minimum number of PAVs needed to support the number of page data sets on a volume. This minimum is used to protect the paging subsystem during periods of low paging activity by directing WLM to not reassign the aliases even though the PAVs may appear to be currently underutilized. The WLM / ASM support for PAVs in z/OS 1.3 also helps provided increased system availability for certain events such as SVC dumps which may cause a burst of paging activity which locks out page fault resolution for critical workloads. The z/OS 1.3 ASM support for PAVs will allow the system to prevent page-outs from blocking page-in operations against a specific local page data set.

2.1.2 RMF Measurement Data and Reports

Use an RMF paging space report to get information on the allocation of slots on the paging data sets. Use RMF control card REPORTS(PAGESP).

P A G E / S W A P D A T A S E T A C T I V I T Y

NUMBER OF SAMPLES = 1,800

PAGE SPACE TYPE	VOLUME SERIAL	DEV NUM	DEVICE TYPE	SLOTS ALLOC	PAGE DATA SET USAGE ---- SLOTS USED --- MIN	MAX	AVG	BAD SLOTS	% IN USE	PAGE TRANS TIME
PLPA	MCATPD	3868	33903	90000	11355	11355	11355	0	0.00	0.000
COMMON	MCATPD	3868	33903	126000	32	32	32	0	0.00	0.000
LOCAL	PAGPD1	3E66	33903	599400	10410	10440	10422	0	0.00	0.000
LOCAL	PAGPD2	3E67	33903	599400	10709	10734	10715	0	0.00	0.000
LOCAL	PAGPD3	3E68	33903	599400	10723	10743	10731	0	0.00	0.000
LOCAL	PAGPD4	3EA1	33903	599400	10674	10690	10682	0	0.00	0.000
LOCAL	PAGPD5	3E69	33903	599400	10746	10776	10762	0	0.06	0.006
LOCAL	PAGPD6	3E40	33903	600300	10761	10778	10768	0	0.00	0.000

3.0 Dumping Support

3.1.1 SADUMP Introduction

Migration to z/Architecture requires a much larger real storage to be dumped over a ESA/390 implementation. Additional planning is required to allow SADUMP to complete in a timely manner and with minimal operator intervention. It is also important to plan for the additional DASD resources which will be needed to allow SADUMP to capture a complete dump.

Define stand-alone dump (SADUMP) data sets as multi-volume dump groups. SADUMP will write concurrently to all volumes defined to the dump data set.

Define multiple dump data sets each with multiple volumes. Each SADUMP data set can extend across 16 volumes. As an initial starting point IBM recommends the definition of 3 dump data sets, with each data set having 5 volumes defined. The 5 volumes should be dedicated 3390-3s, (Use of 3390-9 is also possible as long as only the first 65,536 tracks are allocated). Use the AMDSADDD REXX utility to format the SADUMP data sets, being sure to specify a BLKSIZE of 24960, with a logical record length of 4160. Once a multi-volume dump data set is formatted it is important the installation not move the volume as special control information is written to allow SADUMP to correctly identify the volumes. If dedicated volumes are not used then it is critical the volumes used in a dump set do not contain a page data set from the system being dumped.

The actual amount of DASD resources required is a function of the real memory size of the image being dumped and the amount of virtual storage on the page data sets. The variable in dumping space requirements for SADUMP will be the amount of virtual storage on the AUX page data sets. For a 3390-3 device there are 50,085 tracks per volume. SADUMP will write 12 records per track, with each record representing one 4K page. A dedicated 3390-3 device will be able to dump approximately 2.46GB, and a custom volume using all 65K tracks would be able to dump approximately 3.2GB. For a 12GB region, using 3390-3 approximately 6 devices will be needed; five devices to support the real memory requirements, and additional space to support the virtual storage requirements. If a custom volume is created it would take 5 volumes; four volumes for the real memory and an additional volume for the virtual storage.

The performance of SADUMP will be directly tied to the performance of the I/O configuration defined to support SADUMP. Standard DASD tuning practices should be followed when identifying the DASD resources for SADUMP. Where possible the volumes should be on control units with adequate channel capacity and across different channel path groups, and the volumes should be located across multiple LCUs, and should not share the same RAID Ranks.

The amount of time it takes to complete a SADUMP is dependent upon the amount of concurrent write operations, which is directly dependent upon the I/O configuration. If the initial starting point of 3 data sets with 5 volumes each does not provide sufficient speed then additional volumes should be added to increase data transfer concurrence. For planning purposes a well configured 3390-3 logical volume should be able to transfer between 35-40 MB per second.

The capabilities of the control unit will also impact the storage requirements of SADUMP. Control units which perform compression will have the ability to reduce the physical space required to contain the dump output. The actual compression ratio will be a function of the data which needs to be dumped.

3.1.2 SADUMP Options:

In general the defaults for SADUMP should be used with the exception of the DDSPROMPT keyword. Use DDSPROMT=YES to allow SADUMP to prompt the operator for additional data set names while DDSPROMPT=NO will allow SADUMP to run without operator intervention. Operational considerations, which vary by customer, will dictate the level of operator intervention desired.

MINASID=PHYSIN is not as meaningful as in previous architectures because address spaces are ordered to get a more effective partial dump. Dumping is done in a priority order with the storage for ASIDs 1-4 dumped, then selected address spaces, and any address spaces executing at the time of the dump. All swapped in address spaces will be dumped before swapped out ones. Messages are issues which indicate when swapped-in spaces are being dumped and when swapped-out spaces being dumped. Depending upon the situation it is possible to terminate the dump when processing the virtual storage of the swapped out address spaces begins as this information is often less valuable. Because of these changes in dump processing taking the default for MINASID (MINASID=ALL) is acceptable. If the dump is terminated before it completes it is important to generate a CPU external interrupt to ensure a clean stop is done which allows IPCS to process the dump.

3.1.3 SADUMP Additional Information

Additional information on the performance on SADUMP can be found on the web. ReviewWSC Flash 10143 which can be found at the following URL:
http://www-1.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/Flash10143

This document, prepared by the zSeries performance team reviews performance runs for different configurations of SADUMP. It is important to know this article discusses the use of stripes for SADUMP. The use of the stripes terminology really refers to the implementation of multi-volume dump groups. Another good source on SADUMP can be found in product manual *z/OS MVS Diagnosis: Tools and Service Aids* (GA22-7589).

And finally information is available to get output performance statistics from a dump by using the IPCS command VERBX SADMPMSG 'STATS'. This IPCS command gives a whole raft of things, some interesting only to a standalone dump developer and others of general interest, including how fast the data transfer rate was when taking the dump.

Customers should also be aware of the PUTDOC tool which can be used to send documentation to IBM. This tool has the ability to automatically break apart very large files, like SADUMP data sets, automatically. For more information on the PUTDOC tool see the web site at the following URL: http://techsupport.services.ibm.com/server/nav/zSeries/putdoc/putdoc.html

3.2.1 SVCDUMP Introduction

A migration to z/Architecture will also require some review of the SVC dump capabilities. Unlike SADUMP, additional real storage by itself will not cause changes in the dumping requirements for SVC Dump. Exploitation of virtual storage through use of memory objects would increase the requirements for dumping. As installations exploit memory object they should review their performance and capacity requirements for SVC Dump.

The zSeries performance team also provided performance information on SVC Dump which was published as a flash. This information can also be found on the web at the following URL: http://www-1.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/Flash10182

In this document when references are made to stripes the intent is to use the VSAM striping support.

Editor's Note: Several APARs have also been take in the areas of both SADUMP and SVC Dump. Installations should ensure they are have reviewed appropriate and applied maintenance relating to dumping services.

Below are some important APARs to review:
- SADUMP: OW56411, OW56405, OW56867
- SVCDUMP: OW56817

Summary:

The migration to 64-bit mode will take some planning and evaluation of the current 31-bit mode system. After the migration performance analysts will need to rethink how they review processor storage. The good news is the 64-bit mode architecture has no distinction of storage being either central or expanded. Now the entire processor storage range can be used to support the workloads.

Migration to 64-bit mode is not expected to cause any requirement for additional processor storage. The amount of processor storage installations should plan for would be the current amount of central and expanded storage on the 31-bit mode system.

Additional processor storage may be needed or desired if the current system is seeing demand paging delay impacting workloads in 31-bit mode, or if additional workload (new or growth) requires additional storage. Likewise, database manager changes (increasing buffers, additional databases, use of dataspaces) may require additional processor storage in a 64-bit mode system.

The most immediate benefit will be the ability to unlock the previously available frames in expanded storage to support the current workloads. Utilization of the expanded storage was dependent upon the installation exploiting data in memory techniques, or it was limited to supporting the page movement to and from central storage. Now all of the configured processor storage is "real" storage and the CPU costs of accommodating central storage shortage with expanded storage are now avoided. Most importantly, 64-bit mode allows installation to continue to launch new applications with ever larger storage demands and have them supported by this very large, all "real" environment.

Special Notices

A.0 Addendum

A.1 Summary of Changes

The following is a list of changes which have been made to this document based on version number.

Version 1.1

Change	Page
Added a Cover Page with Version number.	cover
Added Editor's Note on list of HIPER RSM APARs	1
Updated the UIC field attributes for SMF71LIC / HIC	7
Documents the UIC buckets for impact frames in the SMF 71 record	9
Updated recommendation for MCCCAFCT post the RSM Hiper APAR Activity	12
Added Chapter 3.0 on Dumping Support	29-30