EXPANDED STORAGE MANAGEMENT WITH MVS/ESA

Donald R. Deese

Expanded storage is offered as a relatively inexpensive way of using high speed processor storage to minimize I/O operations. Under many circumstances, expanded storage achieves this objective. In some situations, the expanded storage algorithms may not yield the expected benefits. In a few situations, expanded storage can cause severe performance problems. This paper describes the concepts and algorithms used to manage expanded storage with MVS/ESA and highlights major changes between MVS/XA and different releases of MVS/ESA.

1.0 INTRODUCTION

This introduction provides a brief overview of the terms and concepts used by the System Resources Manager (SRM), the Real Storage Manager (RSM), and the Auxiliary Storage Manager (ASM) to manage the use of processor storage. A brief definition of terms is provided so the more detailed explanations have a basic framework:

- **Processor** storage consists of (1) **central** storage and (2) **expanded** storage. Central storage is very high speed storage and is addressed at an individual byte level. Expanded storage is relatively lower speed storage and is addressed only at a page level (4096 bytes).
- **Auxiliary** storage is the storage represented by direct access storage devices (DASD). Auxiliary storage is much lower speed than is expanded storage and is addressed only by I/O operations transferring a page or several pages at a time.
- **Physical swapping** moves an address space out of central storage to either expanded storage or to auxiliary storage.
- **Logical swapping** keeps an address space in central storage, but the SRM adjusts control blocks to indicate that the address space is in a "swapped out" status.
- A **page** consists of 4096 contiguous bytes of program code or data. A page may reside in processor storage or reside in auxiliary storage.
- A **page frame** is a 4096-byte block of central or expanded storage. The page frame may hold a page of program code or data, or the page frame may be unused.
- **Page stealing** steals page frames from the Common area or from swapped-in address spaces.
- **Swap trim** removes page frames from an address space which is to be logically or physically swapped.
- **Migration** moves pages from expanded storage, through central storage, to auxiliary storage. Migration is performed by the **Migrator** component of the RSM.
- **Migration Age** is a single number which is used as an estimate of how long pages have remained in expanded storage before they are transferred to central storage or are migrated to auxiliary storage.

Programs and data may reside in central storage, expanded storage, or auxiliary storage. Not all instructions nor all data are required to be in central storage in order for a program to execute. Much of a program's instructions or its data may be in "virtual" storage and may reside in expanded storage or reside in auxiliary storage, rather than reside in central storage. However, programs can execute instructions only if the instructions reside in central storage and instructions can reference data only if the data resides in central storage.

The main reason for different levels of storage is financial; central storage is very expensive, expanded storage is less expensive, and auxiliary storage is least expensive.

As mentioned above, expanded storage is addressed only in increments of 4096 bytes. Each 4096 bytes is addressed using a block number, and each block of 4096 bytes is termed a "frame" of expanded storage. A frame of expanded storage may hold information which has been transferred from central storage, or the frame may be unused. A page of information (program code or data) may be in a central storage frame or in an expanded storage frame.

2.0 CONTROL OF EXPANDED STORAGE

The RSM moves pages from central storage to expanded storage and from expanded storage to central storage. The RSM maintains an Expanded Storage Table (EST), which describes the status of each frame of expanded storage: whether the expanded storage frame is in use, what type of page resides in the frame (primary working set page, secondary working set page, etc.), various indicator bits to reflect the status of the page, and so forth.

Elements in the EST describe expanded storage frames in ascending order by expanded storage block number (that is, the first element in the EST describes the first expanded storage frame configured to the MVS image).

The RSM maintains a queue of pointers to the EST elements for expanded storage frames which are not assigned and are available. When the RSM wishes to move a page from central storage to expanded storage, an EST pointer is acquired from this queue. When the RSM moves a page from expanded storage to central storage, the EST pointer for the expanded storage frame is added to the bottom of the queue. For both of these actions, the EST is updated to reflect the new status of the expanded storage frame.

Expanded storage frames are used when MVS decides to move pages from central storage to expanded storage. Expanded storage frames are made available for reuse when pages are moved from expanded storage to central storage. Expanded storage frames also are made available when they are freed (for example, when a batch job terminates and frees its processor storage).

Both these actions (using expanded storage frames and making the frames available) occur on a random basis. Consequently, entries in the EST are updated randomly and any particular expanded storage frame may be used or unused at any specific time.

The RSM identifies all expanded storage frames which are assigned to the Common area. The RSM identifies these frames by placing information about the frame on an in-use Expanded Storage Table Element (ESTE) queue for the Common area. Additionally, the RSM identifies all frames which are assigned to each address space, by placing information about the frame on the in-use ESTE queue associated with each address space.

Entries are placed on bottom of the ESTE queues in the order in which frames are assigned. Consequently, the entries on the top of each of the ESTE queues are those frames which have been in expanded storage the longest time (for the particular type of queue), and the frames on the bottom of the queues are those frames which have been in expanded storage the shortest time.

As will be discussed later, the EST and the ESTE queues are important in many decisions the RSM and SRM make about the use of expanded storage (for example, they are important in determining which pages in expanded storage are migrated to auxiliary storage).

2.1 PAGES SENT TO EXPANDED STORAGE

The SRM and RSM distinguish seven categories of pages which are sent to expanded storage or to auxiliary storage: (1) stolen pages, (2) swap trim pages, (3) swap working set pages, (4) hiperspace pages, (5) VIO pages, (6) virtual fetch pages, and (7) requested page out pages. These distinctions are made so that the SRM can control placement of pages and so users can indicate preference about what types of pages are sent to expanded storage versus those sent to auxiliary storage.

- **Stolen pages** are pages which the RSM has removed from the Common area or from an address space because insufficient central storage frames are available to handle page fault resolution or to handle the swap-in of address spaces. The RSM takes central storage pages from active address spaces or from the Common area, sends these pages to expanded storage or to auxiliary storage, and adds the central storage frames to the central storage Available Frame Queue.
- **Swap Trim pages** are pages which the RSM removes from address spaces which are logically or physically swapped out. Swap trim removes pages which are not working set pages from address spaces which have been placed in a "swapped out" status. Pages which are not working set pages are subsequently referred to as "non-working set pages".

  Swap trim occurs (1) when the SRM performs Quiesce processing for certain swaps which are to be physically swapped (the Request/Transition/Storage Shortage swaps), (2) when the SRM performs Quiesce processing for Wait State swaps which are to be logically swapped, and (3) when the SRM needs to replenish the central storage Available Frame Queue from logically swapped address spaces.
- **Swap Working Set pages** are those pages which remain after the Swap Trim process has completed removing pages from an address space. The distinction between "working set pages" and "non-working set pages" depends upon system conditions. As a minimum, the working set consists of all Local System Queue Area (LSQA) pages, all fixed pages, and pages with an Unreferenced Interval Count (UIC) of 0 and the page "referenced" bit turned ON. This minimum working set may be significantly increased if central storage is not constrained.

  The SRM determines whether to target a swap to expanded storage or send it to auxiliary storage based on the specification of the ESCTxxxx keywords in IEAOPTxx (and based on whether abundant expanded storage exists). If the swap is directed to auxiliary storage, the entire working set is swapped out in what is termed a "single stage" swap. If the swap is targeted for expanded storage, a "primary working set" and a "secondary working set" are built. The primary working set pages consist of LSQA pages, fixed pages, and one page from each segment (a segment is one megabyte of memory). The secondary working set pages consist of all pages in the working set which are not primary working set pages.

- **Hiperspace pages** are pages which are placed into high performance dataspaces, such as Hiperbatch, VSAM Hiperspace buffers, etc.
- **VIO pages** are pages which accessed using the virtual I/O services.
- **Virtual Fetch pages** are pages which are associated with applications (e.g., IMS) which use virtual storage to hold application programs. Virtual Fetch uses VIO paging to access application program modules, rather than using the standard program fetch.
- **Requested Page Out pages** are pages which a user has requested be sent to expanded storage, using the PGLOAD, PGOUT, or PGSER macros.

Before SP4.2, the SRM distinguished between changed and unchanged pages for stolen pages, swap trim pages, and page out pages. With SP4.2, the SRM discontinued this distinction between changed and unchanged pages.

2.2. ESCTxxxx KEYWORDS

Users can influence the SRM's decisions about whether to send pages to expanded storage or to auxiliary storage by using the ESCTxxxx keywords in IEAOPTxx. Most (but not all) of the SRM's logic for determining whether to send pages to expanded storage or to auxiliary storage check the ESCTxxx values.

The purpose of the ESCTxxxx specifications and of the algorithms is to allow users to indicate a preference about whether a specific page type should be sent to expanded storage or to auxiliary storage, based on whether expanded storage is constrained (or whether all processor storage is constrained). If expanded storage (or all processor storage) is constrained, it might be advantageous to send certain page types directly to auxiliary storage; if processor storage is not constrained, then these page types might be sent to expanded storage.

The ability to specify this guidance is important because if pages are sent to expanded storage when processor storage is constrained, pages might have to be migrated to auxiliary storage. As will be described later, migration requires system resources and a high amount of migration can cause serious performance degradation. It is much better to send certain pages to auxiliary storage rather than to expanded storage if migration would result from the pages being in expanded storage.

Even if storage is not constrained when the test is made, it may be desireable to send certain page types directly to auxiliary storage and reserve expanded storage for more important page types or pages associated with higher priority workloads.

The following ESCTxxxx keywords can be used to influence whether the SRM sends pages to expanded storage or to auxiliary storage:

ESCTBDS - Hiperspace pages

ESCTPOx - Requested Page Out pages[1]

ESCTSTx - Stolen pages[1]

ESCTSWTx - Swap Trim pages[1]

ESCTSWWS - Swap Working Set pages

ESCTVF - Virtual Fetch pages

ESCTVIO - Virtual I/O (VIO) pages

The ESCTxxxx keyword specifications are of the form **ESCTxxxx(index) = criterion**, where the "xxxx" defines the specific keyword related to the page type described above, "index" is a value defining the type of workload to which the specification relates, and "criterion" specifies a value against which to compare the Migration Age (and perhaps other indicators of processor storage constraints). These terms are described later in this paper. A simplified form of the SRM algorithms in which ESCTxxxx specifications are used is:

IF
(Migration Age Test Value[2]) greater than criterion
THEN
target page for expanded storage
ELSE
target page for auxiliary storage

The **Migration Age Test Value** used in the previous algorithm differs depending upon the type of page being considered for expanded storage versus auxiliary storage.

- For most page types (Hiperspace pages, Requested Page Out pages, stolen pages, Virtual Fetch pages, and VIO pages), the Migration Age Test Value is the Migration Age. The algorithms simply compare the Migration Age to the criterion specified. If the Migration Age is greater than or equal to the criterion, the SRM

[1] For pre-SP4.2, the last character of the ESCTxxxx keyword indicates whether page guidance applies to a changed or unchanged page: "C" for changed pages and "U" for unchanged pages (for a total of ten ESCTxxxx keywords with pre-SP4.2). The distinction between changed and unchanged pages was dropped with SP4.2 and the last character of the ESCTxxxx keyword is "C" for both changed and unchanged pages (for a total of seven ESCTxxxx keywords with SP4.2).

[2] For simplicity of text, the term "Migration Age Test Value" is used to illustrate the algorithm. This "Migration Age Test Value" may be a combination of the Migration Age, the think time for wait state address space, and the average system high UIC.

will normally[3] send the page to expanded storage. If the Migration Age is less than the criterion, the SRM will normally send the page to auxiliary storage.

- With swap trim pages (the ESCTSWTx keyword), the SRM modifies the test by adding the "Average System High UIC" to the Migration Age. The intent is to assess whether all processor storage is constrained. The Average System High UIC[4] is an indicator of whether central storage is constrained, so adding this value to the Migration Age attempts to assess whether processor storage is constrained.
- With swap working set pages (the ESCTSWWS keyword), the SRM further modifies the test for wait state address spaces by subtracting the address space's think time from the sum of the Migration Age and the Average System High UIC. The address space's think time is computed as the maximum of the previous two wait states (the time a terminal user was "thinking" before hitting ENTER, or an artificial think time of 5 seconds for non-terminal wait address spaces).

  The intent of this refinement is to place less importance on logically swapped wait state users who "think" for a long time. By subtracting their think time from the sum of Migration Age and Average System High UIC, long think times would cause the sum to be less and show more processor storage constraint than if the think time were left out of the algorithm.

An **index** is associated with most keywords (there is no index with the ESCTBDS, ESCTVF, and ESCTVIO keywords for pre-SP4.2). The purpose of the index is to allow the criterion associated with the index to be applied to different categories of workload.

- For pre-SP4.2 systems, the index could be specified only in three categories, defining three types of workload. With SP4.2, the pre-SP4.2 categories are the default categories but users can describe additional categories.

  The index may be a value of 0, 1, or 2 to describe the workload type to which the keyword applies.

  - The criterion associated with an index of "0" applies to pages related to the non-swappable address spaces, the Common area, or privileged address spaces for ESCTPOx (page-out pages) and ESCTSTx (stolen pages) keywords. An index of "0" applies only to privileged address spaces for the ESCTSWTx (swap trim pages) and ESCTSWWS (swap working set pages) keywords. Since non-swappable address spaces and the Common area are not trimmed and swapped, the ESCTSWTx and ESCTSWWS keywords have no meaning for these categories.
  - The criterion associated with an index of "1" applies to all address spaces not included in the "0" or "2" descriptions for the appropriate page type.
  - The criterion associated with an index of "2" applies to TSO address spaces for the ESCTPOx (page-out pages) and ESCTSTx(stolen pages) keywords. Additionally, the criterion for an index of "2" is applied to all wait state address spaces[5] for the ESCTSWTx and ESCTSWWS keywords.
- With SP4.2, the index optionally can be specified as a value ranging from 3 to 99. The additional indices allow criteria to be associated with more detailed levels of workload categories. A user can associate the ESCTxxxx criteria to one or more domains by specifying the index value used for the ESCTxxxx keywords as an operand of the ESCRTABX keyword associated with the domain in IEAIPSxx.

  The SRM will use the criteria specified for ESCTxxxx keywords for all workloads associated with the domain(s) having the index value specified in the ESCTxxxx keywords. Performance group periods are associated with domains using the specifications in IEAIPSxx. Consequently, SP4.2 provides the capability to provide the SRM with expanded storage guidance down to the performance group period level.

The **criterion** associated with each ESCTxxxx specification is used to indicate a preference as to whether the associated page type should go to expanded storage or auxiliary storage. The intent of the

[3] The word "normally" is used throughout the text to describe normal actions in normal operating situations. MVS will take abnormal actions in response to abnormal situations. For example, many abnormal actions will be taken if there is a shortage of fixed storage below 16 megabytes. These actions may include "shutting off" expanded storage, "shutting off" logical swapping, overriding storage isolation, etc. This paper would become large and cumbersome if all abnormal actions were discussed in each situation. Thus, only "normal" actions and important exceptions are typically described.

[4] Note that the *Installation and Tuning Guides/References* describe the test as related to the "system high UIC" which implies the highest UIC in the system when the test is made. The tests actually use the **average** system high UIC. The average system high UIC is computed only when the SRM's Resource Monitor 2 (IRARMRM2) code is executed. The frequency with which the IRARMRM2 routine is executed is based on the number of processors and the speed of the processors. Regardless of the speed or number of processors, the IRARMRM2 routine is executed no more frequently than once every 5 seconds (pre-SP4.2) or once every 2 seconds (with SP4.2). The average system high UIC computed at this time reflects the average of the highest UIC's (excluding storage isolated address spaces) encountered over the previous sampling intervals.

[5] The SP4.2 and SP4.3 *Initialization and Tuning References* are incomplete in their description of the meaning of a "2" index value for the ESCTSWTC and ESCTSWWS keywords. The References state that index "2" describes long, detected, and terminal wait users. APPC wait users are also included in the tests.

specification is to provide guidance to the SRM depending upon whether expanded storage (or processor storage) is constrained. Since Migration Age is one estimate of whether expanded storage is constrained, the criteria are compared with the Migration Age and the Migration Age will usually be the dominating factor in the algorithms.

If the Migration Age is high, the algorithms assume that expanded storage is **not** constrained since pages have remained in expanded storage for a long time without being migrated. If the Migration Age is low, the algorithms assume that expanded storage **is** constrained, since pages are in expanded storage for a short time before being migrated. The Migration Age discussion later in this paper describes how neither of these assumptions may be correct and Migration Age may have no relationship to whether expanded storage is constrained.

The default ESCTxxxx criteria changed significantly with the SP4.2 release.

- The pre-SP4.2 default criteria were very small. The default ESCTxxxx criterion specified for most page types is 100 or lower (the maximum default criterion for stolen pages is only 15). For most installations, the Migration Age never becomes so low that the criteria are applicable. IBM discovered that few installations understood the implications of the ESCTxxxx keywords, and few installations modified the default ESCTxxxx criteria specifications. Consequently, the ESCTxxxx keywords generally were ineffective in providing the SRM with guidance about expanded storage.
- With SP4.2, the defaults were significantly increased (the maximum default criterion for stolen pages increased from 15 to 250 and the maximum default criterion for many other page types increased from 100 to 1200). The intent of the increased defaults was to allow the SRM to be more responsive to any perceived expanded storage constraint.

2.3 OVERRIDING ESCTxxxx

It is important to appreciate that the specifications for the ESCTxxxx keywords simply express a preference to the SRM about whether pages should be sent to expanded storage or to auxiliary storage. The ESCTxxxx specifications are **not** directive in nature. The SRM may decide to send pages to auxiliary storage or to expanded storage and ignore the specifications in the ESCTxxxx keywords, depending upon system conditions.

The primary situations in which the SRM overrides or ignores this guidance are (1) for storage isolated address spaces, (2) when pages are stolen to swap in an address space, (3) when expanded storage is abundant, (4) when expanded storage is tight, and (5) when there is a critical shortage of processor storage. Some of the more common situations are described below:

- **Storage isolated address spaces.** Pages stolen from storage isolated address spaces are normally directed to expanded storage (without regard to the ESCTxxxx criteria[6]), unless the address space is over the maximum working set in central and expanded storage is not considered "tight". The concept of expanded storage being "abundant" or "tight" is discussed later.
- **Storage isolated Common area.** Pages stolen from the Common area if storage isolation has been specified for the Common area are normally directed to expanded storage (without regard to the ESCTxxxx criteria) unless expanded storage has temporarily been "turned off" because of a critical shortage.
- **Storage isolated address spaces over maximum protected working set in central storage**. If an address space is over the maximum protected working set in **central** storage and if expanded storage is considered "tight", the SRM sends the stolen pages to auxiliary storage, regardless of the expanded storage criteria.
- **Special steal -- SRM raised thresholds**. The SRM may raise the processor storage thresholds to accommodate the swap-in of an address space from expanded storage. When this is done, the SRM sets a flag to indicate that this is a special situation. The SRM's page steal processing tests for this flag, and considers the pages stolen to be a "special steal" because SRM has raised thresholds. All pages stolen as a result of this special steal are directed to expanded storage regardless of the ESCTxxxx specifications, so long as there are any expanded storage frames available.

  From one view, this process simply exchanges pages in expanded storage (those being swapped in) with pages in central storage (those being stolen to accommodate the swap-in). However, the ESCTxxxx specifications are ignored for the stolen pages. It is possible that expanded storage could become populated with pages belonging to a workload category which a user wishes to always go to auxiliary storage.
- **VIO pages**. If the job using the Virtual I/O facility is restartable and SYS1.STGINDEX has had no failures, the pages are directed to auxiliary regardless of the ESCTVIO specification.

[6] Note that the MVS/ESA SP4.2 and MVS/ESA SP4.3 *Initialization and Tuning References* are incorrect with regard to the ESCTSTC keyword. Pages stolen from storage-isolated address spaces or from storage-isolated common area are sent to expanded storage **regardless** of whether 32767 is specified as the criteria age. Unfortunately, there is no way to prevent this situation.

- **Expanded storage is abundant**. With SP4.2, the SRM will ignore the ESCTxxxx specifications if expanded storage is abundant, unless 32767 has been specified for the ESCTxxxx keyword. In several situations (described above), the SRM will ignore the ESCTxxxx specification regardless of whether 32767 is specified.

- **Swap Working Set**. If address space is relatively small (if the working set size is less than the migration threshold specified by low value of the MCCAECTH keyword in IEAOPTxx) or if there is plenty of expanded storage, the ESCTSWWS criteria are tested to determine whether to target the swap for expanded storage.

  The secondary working set is sent to expanded storage based on the ESCTSWWS criteria. However, the primary working set will be sent to expanded storage only if there are enough expanded storage frames to hold the primary working set.

  The SRM tests the availability of expanded storage frames just before swapping the primary working set. The primary working set is NOT sent to expanded storage if there are insufficient expanded storage frames at the time of the test, regardless of the ESCTxxxx specification. This means that the secondary working set of an address space could be sent to expanded storage, but the primary working set could be sent to auxiliary storage[7].

  This SRM design might seem strange, since a test could be made to determine whether the primary working set or the complete working set could fit into expanded storage before any working set pages were sent to expanded storage. Indeed, such tests were made with MVS/XA. However, the MVS/ESA processor storage management design is intended to move pages from central storage to expanded storage (or to auxiliary storage) **only** when central storage frames are needed.

  Building of the secondary working set stops whenever enough pages have been added to the central storage Available Frame Queue. Consequently, the RSM may be in the middle of building a secondary working set for an address space when the Available Frame Queue has been replenished.

  Before additional pages must be acquired and added to the secondary working set, central storage frames might be added to the Available Frame Queue from other sources (for example, a different address space could terminate). Consequently, with MVS/ESA, the SRM is unable to determine whether to send the primary working set to expanded storage or to auxiliary storage until it is time to complete the swap out process.

- **Critical shortage of processor storage**. The SRM can temporarily "shut off" expanded storage when the SRM detects that there is a critical shortage of processor storage, when the Migrator is constrained, or when expanded storage is being reconfigured. A flag is set (the MCTESNA indicator) to indicate that expanded storage is not available. All logic about whether to target swaps or to send pages to expanded storage test the MCTESNA indicator. The tests cause swaps or pages to be sent to auxiliary storage if the MCTESNA flag indicates that expanded storage is not available.

2.4 EXPANDED STORAGE THRESHOLDS

The SRM uses two main indicators of whether expanded storage **instantaneously** is constrained or is not constrained. These are the RCEAECLO and RCEAECOK variables, which indicate "low" and "OK" thresholds of available expanded storage frames.

When the count of available expanded storage frames is less than the RCEAECLO threshold, the SRM will schedule the Migrator component of the RSM to migrate pages from expanded storage. The Migrator will stop migrating pages when the count of available expanded storage frames is greater than the RCEAECOK threshold. Additionally, the SRM makes decisions about whether to send address spaces to expanded storage or auxiliary storage based on these thresholds.

The RCEAECLO and RCEAECOK variables initially are based on the MCCAECLO and MCCAECOK values. The MCCAECLO and MCCAECOK values are **constants** for pre-SP4.2 and have values of 50 and 100, respectively. The MCCAECLO and MCCAECOK values are **parameters** in IEAOPTxx with SP4.2 (using the MCCAECTH keyword) and have default values of 150 and 300, respectively.

The RCEAECLO and RCEAECOK thresholds are dynamically adjusted based on system conditions.

- The initial RCEAECLO and RCEAECOK values can be adjusted based upon how often expanded storage becomes constrained. The RCEAECLO and RCEAECOK values can be increased when expanded storage is constrained. These values subsequently can be lowered if expanded storage is no longer constrained. The RCEAECLO and RCEAECOK values will not be lowered below their initial values.

  The SRM maintains an indicator which is set ON if all expanded storage frames were allocated (the count of

[7] This condition (primary working set in auxiliary storage and secondary working set in expanded storage) is one of the conditions which would cause purge migration if the Migrator needed to acquire expanded storage frames. Purge migration is described later in this document.

available expanded storage frames was zero). Each SRM second[8], the SRM tests this indicator. A counter is incremented if all expanded storage frames were allocated during the previous SRM second, and reflects a "sample" of the general availability of expanded storage. Additionally, the SRM keeps track of the minimum amount of expanded storage which was available **at the time of the sample** (that is, the minimum available when the SRM took the sample).

The SRM will periodically examine system conditions (whether the processors are over- or under-utilized, whether central storage is over- or under-utilized, whether expanded storage is over- or under-utilized, etc.). Depending upon system conditions, as matched against installation guidance, the SRM will adjust various aspects of its control of the environment. The frequency with which the IRARMRM2 routine is executed is based on the number of processors and the speed of the processors.

These adjustments occur only when the SRM's Resource Monitor 2 (IRARMRM2) code is executed. Regardless of the speed or number of processors, the IRARMRM2 routine is executed no more frequently than once every 5 seconds (pre-SP4.2) or once every 2 seconds (with SP4.2).

During the IRARMRM2 system adjustment processing, SRM determines whether 1000 samples have been taken (that is, 1000 SRM seconds as adjusted for processor speed and number of processors have lapsed since the previous adjustment of expanded storage thresholds). If 1000 samples have been taken, the SRM computes the percent of time that an expanded storage shortage existed. This is simply the count of SRM seconds during which all expanded storage frames were allocated, divided by the number of SRM seconds in the sample.

- If a shortage existed for more than two percent of the previous interval, the SRM will raise the RCEAECLO and RCEAECOK thresholds. By raising the thresholds, the SRM will detect a shortage of expanded storage sooner, start the Migrator to migrate pages, and free expanded storage frames. The thresholds are raised by the percent of samples that an expanded storage shortage existed, times the previous RCEAECLO value (or by four, whichever is larger).

  For example, suppose a shortage existed for 20% of the samples and the previous RCEAECLO and RCEAECOK thresholds were 150 and 300, respectively (the default initial values with SP4.2). The RCEAECLO and RCEAECOK values would be adjusted by 30 (150 * .20). The adjusted RCEAECLO and RCEAECOK values would be 180 and 330, respectively.

- If a shortage existed for less than one percent of the previous interval, the SRM will lower the RCEAECLO and RCEAECOK thresholds. If an expanded storage shortage does not generally exist, there is no need to run the Migrator as often, so the thresholds can be lowered. The thresholds are lowered by one-half the lowest sampled count of available expanded storage (or by four, whichever is larger). As mentioned above, the RCEAECLO and RCEAECOK values will never be lowered below the MCCAECLO and MCCAECOK values.

  For example, suppose a shortage existed for 0.1% of the samples and the previous RCEAECLO and RCEAECOK thresholds were 180 and 330, respectively. Suppose further that when the SRM took samples, the minimum amount of available expanded storage was 1000 frames[9].

  In this example, one-half of the minimum sampled count of available frames would be 500. Subtracting 500 would result in RCEAECLO and RCEAECOK values being less than the MCCAECLO and MCCAECOK thresholds. Consequently, the MCCAECLO and MCCAECOK thresholds would be used. The adjusted RCEAECLO and RCEAECOK values would be 50 and 150, respectively (with pre-SP4.2) or would be 150 and 300, respectively (if the default values were used with SP4.2).

The IRARMRM2 routine which performs the dynamic adjustment of the thresholds is executed every 106 SRM seconds (pre-SP4.2) or every 42 SRM seconds (with SP4.2). Dividing the 1000 samples required to

[8] The "SRM second" is based on the value specified for the RMPTTOM keyword of IEAOPTxx, as adjusted by processor speed. The Initialization and Tuning Guides provide the relationship between the SRM seconds to wall clock time. This number of SRM seconds per second of wall clock typically ranges from 27.6757 (for an IBM 3090 J-series) to 54.0541 for an IBM ES/9000 (Model 900). Depending upon the algorithms involved, the SRM second may be further adjusted by the number of processors.

[9] It may seem strange that the minimum amount of available expanded storage can be greater than zero while there were times when zero expanded storage was available.
This situation can happen because the "sample" of whether expanded storage is short is a sample based on an indicator set on **any** occurrence of zero over the previous sample interval. Expanded storage frames could be freed before the SRM took its sample of the "expanded storage went to zero" indicator.
Consequently, there could have been a temporary shortage of expanded storage between samples but there could have been an abundance of expanded storage frames when the SRM took its sample of the indicator.

make an adjustment by 106 or 42 yields the number of times the IRARMRM2 routine must execute before 1000 samples are taken (the result is 10 for pre-SP4.2 and 24 with SP4.2).

Multiplying by the number of times the IRARMRM2 interval must execute by the minimum execution frequency (5 seconds for pre-SP4.2 and 2 seconds with SP4.2) yields an approximation of the elapsed time between adjustments of the storage thresholds. The dynamic adjustments of the storage thresholds occur about once per 50 seconds for pre-SP4.2 and about once per 48 seconds with SP4.2. The dynamic adjustment algorithms are intended to respond to relatively long-term changes in processor storage requirements.

- The SRM can temporarily raise the RCEAECLO and RCEAECOK thresholds if necessary to make available more processor storage to contain the working set of a swapped-in address space. Once the address space has been swapped into central storage, the thresholds will be lowered by the amounts they were raised.

2.5 EXPANDED STORAGE "TIGHT"

The SRM makes many decisions (particularly those which ignore the ESCTxxxx specifications) based upon whether the SRM considers expanded storage to be "tight" or "abundant" at the time the decision is made. The SRM considers that expanded storage is "tight" when **less** than a certain number of expanded storage frames are available. With SP4.2, the SRM considers that expanded storage is "abundant" when **more** than a certain number of expanded storage frames are available.

The concept of expanded storage being "tight" changed with MVS/ESA SP4.2.

- Prior to SP4.2, the SRM considers expanded storage to be "tight" when the number of available frames is less than three times the MCCAECLO variable, which is a constant with a value of 50. With MVS/XA, the SRM considers expanded storage to be tight if less than 150 expanded storage frames are available.
- With MVS/ESA SP4.2, the SRM considers expanded storage to be "tight" when the number of available frames is less than the RCEAECOK value plus three times the MCCAECOK variable (which is the high value of the MCCAECTH keyword in IEAOPTxx).

  The RCEAECOK value is initially based on the MCCAECOK variable in IEAOPTxx, which has a default value of 300. With MVS/ESA SP4.2, the SRM considers expanded storage to be tight if less than 1200 expanded storage frames are available[10].

Prior to SP4.2, users experienced situations in which pages were not sent to expanded storage, even though a large amount of expanded storage was available. This situation can arise when the Migration Age becomes low because of a heavy demand for expanded storage, and a large amount of expanded storage subsequently is freed (for example, a large batch job ends). Even though a large amount of expanded storage would be available, the Migration Age would be low and pages might be excluded from expanded storage based on the ESCTxxxx criteria.

Since the Migration Age is based on real time, time must lapse for the Migration Age to increase. While the Migration Age is below the ESCTxxxx criteria specified for different page types, these page types would be excluded from expanded storage even though many expanded storage frames were available.

The SRM designers addressed this problem with SP4.2 by testing whether "abundant" expanded storage exists when making the decision about whether to send page types to expanded storage or auxiliary storage. If abundant expanded storage exists, the SRM normally sends pages to expanded storage, regardless of whether the pages meet the criteria specified in ESCTxxxx.

With MVS/ESA SP4.2, the SRM considers expanded storage to be "abundant" when the number of available frames is greater than the RCEAECOK value plus three times the MCCAECOK variable[11]. During most situations when the SRM is testing the Migration Age against the ESCTxxxx criteria, the SRM adds an additional condition to the test. The additional test determines whether (1) the user doesn't mind that pages be sent to expanded storage and (2) whether "abundant" expanded storage exists. The form of the normal algorithm making the test about whether to target a page or swap to expanded storage or to auxiliary storage is:

```
IF
     (Migration Age Test Value greater than criterion)
OR
          (ESCTxxxx not equal to 32767
     AND
```

[10] With SP4.2, the SRM actually tests the amount of available expanded storage against the RCEAECOK value plus three times the MCCAECOK value. As the text describes, the RCEAECOK value can be increased beyond its default of 300 frames. If this increase should occur, the tests will be based on a value which is greater than 1200 frames. It is simpler to use the default values to illustrate the algorithm.

[11] Notice that RCEAECOK + 3 times the MCCAECOK variable would yield 1200 frames - the same value below which the SRM would consider expanded storage to be tight. With SP4.2, the SRM considers expanded storage to be either tight or abundant, depending upon whether more or less than 1200 expanded storage frames are available.

available expanded is greater than RCEAECOK + (3 * MCCAECOK))

THEN

target page or swap for expanded storage

ELSE

target page or swap for auxiliary storage

The following observations are made about the revised algorithms.

- Users usually can cause the SRM to exclude pages from expanded storage even though "abundant" expanded storage exists by specifying 32767 for the ESCTxxxx keyword. This method works so long as the Migration Age does not exceed 32767. The method does not work if the Migration Age exceeds 32767, since the first part of the OR test would yield a TRUE result[12]. There is no way to exclude pages or swaps from expanded storage if the Migration Age exceeds 32767 unless the relevant PTF has been applied[13].
- Since the initial RCEAECOK value is set based on the MCCAECOK value, the first impression is that pages will be sent to expanded storage whenever the available expanded storage is more than four times the MCCAECOK value. This impression is misleading, because of the dynamic adjustment of the expanded storage thresholds described earlier. The dynamic adjustment of the RCEAECLO and RCEAECOK values could cause the RCEAECOK value to be larger than the MCCAECOK value. This would occur if the SRM considered expanded storage to have been constrained during the previous adjustment intervals, or if the RCEAECLO and RCEAECOK values were temporarily increased to accommodate swap-in processing.

The default specification for MCCAECTH in IEAOPTxx yields a RCEAECOK value of 300. With the default, the minimum available expanded storage to be considered "abundant" would be 1200 frames (300 + 3*300). If the RCEAECOK value had been raised, the minimum could be much larger than 1200 frames.

3.0 MIGRATION

If expanded storage should become full and the RSM wishes to move additional pages to expanded storage, frames must be freed in expanded storage. In order to free expanded storage frames, pages are moved from expanded storage to auxiliary storage. There is no direct path from expanded storage to auxiliary storage. Consequently, the pages must go through central storage to get to auxiliary storage. The process of moving pages from expanded storage through central storage to auxiliary storage is termed "migration" and is controlled by the "Migrator" component of the RSM.

The Migrator does not operate continuously. It is scheduled to run whenever the RSM or SRM detects that there is a shortage of expanded storage frames.

- The Migrator is scheduled by the RSM whenever the RSM detects a shortage during normal page steal processing or detects the shortage during swap trim processing, and the stolen or trimmed page is directed to expanded storage. The RSM detects that there is a shortage of available expanded storage when the RCEAEC count of free expanded storage frames falls below the RCEAECLO threshold.
- The Migrator is scheduled by the SRM during several situations when the SRM detects that insufficient processor storage or insufficient expanded storage exists.
  - The most common situation occurs when the SRM wishes to swap in an address space but the address space will not fit into processor storage. For example, suppose that the SRM wishes to swap in an address space from auxiliary storage, and the address space has 400 pages in the working set. If swapping in the 400 frames would reduce the count of available processor storage frames below the low thresholds, the SRM will make room for the address space. This is accomplished by temporarily raising the central and expanded storage thresholds, and scheduling the Migrator to free expanded storage frames. The SRM will reschedule the swap-in after processor storage has been made available.
  - Another common situation occurs when the SRM's central storage Available Frame Queue replenishment algorithms are invoked. If insufficient expanded storage frames are available to hold the required number of central storage frames, and if any address space or the Common area is storage isolated and is over its maximum

[12] With SP4.2, the *Initialization and Tuning Reference* (pages 198-202) erroneously states "If you specify 32767, the page will never be sent to expanded storage." With SP4.3, the Initialization and Tuning Reference (pages 213-215) partially corrects the SP4.2 documentation error by stating "If you specify 32767, the page will not be sent to expanded storage even when there is enough available." With SP4.3, the Initialization and Tuning Guide (page 3-55) clarifies this feature/constraint by stating that "specifying a criteria age of 32767 does not cause the system to bypass expanded storage when the migration age grows greater than 32767."

[13] Consult APAR OY57191 for a listing of PTFs which prevent pages from being sent to expanded storage when the Migration Age exceeds 32767. This APAR deals with VIO pages, hiperspace pages, virtual fetch pages, page-out pages, and self-steal pages. The expanded storage criteria tests for these page types are made at a **single** location in SRM module IRARMFIP, while the tests for other types of pages are made in **many** locations in various SRM modules. The author has not yet reviewed the code with these PTFs, so it is unclear whether the PTFs apply only to those page types listed above, or whether the PTFs apply to all page types which can be directed to expanded storage.

protected working set, the SRM will schedule the Migrator to implement purge migration.

Purge migration will migrate the pages from the storage-isolated address space(s) or Common area, so that either (1) the shortage of expanded storage is eliminated or (2) all pages above the maximum working set of the address space(s) or Common area are migrated.

When the Migrator runs, it has a "migration quota" which it is trying to fill. This is simply the number of expanded storage frames which must be made available to alleviate the shortage. The quota of frames is the RCEAECOK threshold minus the RCEAECLO threshold. This quota differs depending upon whether the MVS version is SP4.2 or prior versions, and differs depending upon whether the Migrator was scheduled by the RSM or by the SRM.

- For pre-SP4.2, the initial low value (RCEAECLO) is 50 frames and the initial OK value (RCEAECOK) is 100 frames. For pre-SP4.2, the Migrator has a quota of at least 50 frames (100 - 50) which it will attempt to migrate.
- With SP4.2, the initial values are based on the MCCAECTH keyword in IEAOPTxx. The MCCAECTH has a default specification of (150,300). Thus, the default low value (RCEAECLO) is 150 frames and the OK value (RCEAECOK) is 300 frames. With SP4.2, the Migrator has a quota of at least 150 frames (300 - 150) which it will attempt to migrate.
- The initial RCEAECLO and RCEAECOK values can be adjusted based upon how often expanded storage becomes constrained. The RCEAECLO and RCEAECOK values can be increased when expanded storage is constrained (so the Migrator will run earlier). These values subsequently can be lowered if expanded storage is no longer constrained (so the Migrator does not run as soon). The RCEAECLO and RCEAECOK values will not be lowered below their initial values. The migration quota will never be less than the amount computed based on the default values.
- The quota which the Migrator is attempting to fill can be larger than the default. This situation occurs if the SRM has scheduled the migration to accommodate the processor storage requirements of an address space which the SRM is attempting to swap in. The SRM can raise the RCEAECLO and RCEAECOK thresholds if necessary to make available more processor storage to contain the working set of a swapped-in address space.

It is important to appreciate that the Migrator does not migrate pages at a steady rate as might be reported by RMF, but migrates in "bursts" of pages. The "burst" will be the quota which the Migrator is attempting to fill. As the previous paragraphs discussed, the Migrator will have a quota of either 50 or 150 pages (or more), depending upon the version of MVS and whether the Migrator was scheduled by the RSM or the SRM. This number of pages will be migrated during one "burst" of migration, even though RMF might report an average migration rate of only 10 pages per second (for example).

3.1 TYPES OF MIGRATION

There are three basic types of migration: (1) reconfiguration migration, (2) purge migration, and (3) least recently used (LRU) migration.

- **Reconfiguration migration** takes place whenever expanded storage is to be reconfigured offline. The reconfiguration migration process migrates all frames within the range of the expanded storage frames to be reconfigured offline.

  Additionally, if any primary or secondary working set pages of an address space are located in the range of expanded storage to be reconfigured, all pages not in the working set of the address space and the entire secondary working set pages are migrated. This is done regardless of where the pages are located in expanded storage.

  Further, if any primary working set pages are located in the range of expanded storage to be reconfigured, the address space is migrated (SRM is notified to update its control blocks and the SRM then initiates the migration of the address space).

- **Purge migration** attempts to select pages for migration when the pages belong to (1) address spaces which were deferred for migration, (2) address spaces which have been placed on auxiliary storage but which have pages in expanded storage, and (3) address spaces which are storage isolated and are over their maximum protected working set.

  - When address spaces are migrated, their primary working set must swapped into central storage. Depending upon the size of the primary working set, the swap-in may fail because there may be insufficient central storage to accommodate the primary working set. If this should occur, the address space will be deferred for migration.

    The SRM will raise the storage thresholds so that central storage frames will be made available, and will reschedule the migration swap. The address space is marked as having been deferred for migration, and the SRM will allow the Migrator to attempt to migrate expanded storage frames from other sources. Migration for the address space will be attempted later (when more central storage

frames are available and the swap-in of the address will be successful).

- As described earlier, the primary working set of an address could be swapped to auxiliary storage, even though the secondary working set had been placed in expanded storage. The SRM will identify these address spaces as candidates for purge migration. The Migrator will preferentially migrate all pages in expanded storage associated with the address space, before attempting LRU migration.
- If storage isolation is applied to address spaces or to the Common area, the SRM tests whether the address space or Common area is over its maximum protected working set **in processor storage**. The tests are made by both the Page Steal algorithms (IRARMPR5) and the logical swap Available Frame Queue replenishment algorithms (IRARMMS2).

  The Migrator is scheduled by these algorithms to perform purge migration of the pages over the maximum protected working set. The Migrator will preferentially migrate pages from the address space(s) or Common area, up to the quota which the Migrator is attempting to fill but not more than the number of pages over the maximum protected working set.

When address spaces or the Common area are selected for purge migration, pages belonging to the address space are migrated based on the order in which the pages appear in the in-use ESTE queue (IUEQ) for the address space or Common area. The IUEQs are in LRU order, so the oldest pages belonging to the address space or Common area are first selected for migration.

The concept behind the purge migration algorithms is sound: basically, the pages selected for purge migration are those pages which should not be in processor storage if processor storage is needed for other purposes. Unfortunately, as will be discussed later, purge migration can significantly distort the SRM's assessment of whether expanded storage is constrained and may cause the ESCTxxxx specifications to be meaningless.

- **LRU migration** is the process by which the Migrator attempts to migrate "old" pages. "Old" pages are pages which were in expanded storage during the last scan of the Expanded Storage Table (EST). LRU migration is not performed if the Migrator's quota of pages was met by purge migration.

  The Migrator scans the EST to determine which pages are "old" and thus can be migrated. The Migrator's scan of the EST does not restart from the beginning of the EST each time the Migrator is started. Rather, the Migrator keeps track of the last EST element examined and scans the EST from that point on.

  As the Migrator scans the EST searching for pages to migrate, it makes its decision about whether to select a page for migration based on the status of the ESTOLD bit in the EST. The ESTOLD bit initially is set OFF when a page is moved to expanded storage. When the Migrator finds an entry in the EST which has the ESTOLD bit set OFF, it sets the ESTOLD bit ON and continues the scan.

  After completing the scan of the EST, the Migrator begins a new scan of the EST. For this subsequent scan, the Migrator knows that any EST element with the ESTOLD bit set on represents an "old" page (that is, a page which was present during the previous scan). The Migrator selects the "old" page as a candidate for migration.

  Note that the selected page is not necessarily the "oldest" page in expanded storage. The page simply has been in expanded storage the last time the Migrator scanned the EST. The Migrator does not attempt to migrate all pages in expanded storage based on the length of time the pages have been in expanded storage. The Migrator simply selects the first "old" page it encounters in the EST based on the status of the ESTOLD bit.

  Further LRU processing by the Migrator depends upon whether the page belongs to the Common area, is a Virtual Fetch page, or belongs to an address space.

  - **Common area pages**. If the page belongs to the Common area, the Common area in-use Expanded Storage Table Element (ESTE) queue is selected for processing. This queue is in least recently used (LRU) order, and pages on the queue are migrated based on their age in expanded storage (that is, the Common area pages which have been in expanded storage the longest are migrated first).

    Migration of Common area pages continues until the migration quota has been met or until the page selected by scanning the EST is migrated. If the page selected by scanning the EST is migrated and the migration quota has not been met, the Migrator continues to scan the EST searching for a frame with the ESTOLD bit ON.
  - **Virtual Fetch pages**. If the page is a Virtual Fetch page, only the selected page is migrated. If the migration quota has not been met, the Migrator continues to scan the EST searching for a frame with the ESTOLD bit ON.

- **Address space page**. If the page belongs to an address space, the in-use ESTE queue for the address space is selected for processing. The ESTE queue for each address space is in least recently used order, and pages on the queue are migrated based on their age in expanded storage.

  Note that there may be pages for other address spaces in expanded storage which have been in expanded storage longer than the selected address space. The selected address space simply has the characteristic of having pages which have were in expanded storage the last time the Migrator scanned the EST. While the pages within an address space are migrated in LRU order, the least recently used algorithm does not apply to all address spaces in expanded storage.

  Migration of pages associated with the selected address space continues until the migration quota has been met or until the page selected by scanning the EST is migrated. If the page selected by scanning the EST is migrated and the migration quota has not been met, the Migrator continues to scan the EST searching for a frame with the ESTOLD bit ON.

  Pages not in the working set and secondary working set pages can be migrated as they are encountered in the ESTE queue for the address space. However, if the RSM encounters a primary working set page in the ESTE queue for the address space before encountering the page selected by scanning the EST, the RSM issues a SYSEVENT to notify the SRM that the **address space** is being selected for migration.

  The SRM may decide to migrate the address space or may tell the RSM to continue to scan the EST searching for a frame with the ESTOLD bit ON. Migration of address spaces is described later.

3.2 MIGRATION AGE

The Migration Age is a key control mechanism in determining whether the SRM will direct pages to expanded storage or auxiliary storage. The Migration Age is used to estimate whether expanded storage is **generally** constrained.

As mentioned above, when the Migrator completes its scan of the EST searching for old pages to migrate, it "wraps" the EST and begins a new scan of the EST. The Migrator counts the "wrap" by incrementing the RCEWRAPS variable. The SRM IRARMPR1 routine in the IRARMSTM module is invoked each second of real time. Among other responsibilities, the IRARMPR1 routine keeps a count of the number of times it has been invoked since the Migrator began the current scan of the EST. This count is maintained in the MCVMGCNT variable. The SRM uses the value in the MCVMGCNT to calculate the Migration Age.

- When the IRARMPR1 routine is invoked, the routine tests whether the RCEWRAPS variable has changed and thus detects that the Migrator has begun a new scan of the EST. If the Migrator has begun a new scan of the EST, the SRM computes the Migration Age based on the count in the MCVMGCNT variable accumulated during the previous scan of the EST, and sets the MCVMGCNT variable to zero.
- If the Migrator has not begun a new scan of the EST (that is, the RCEWRAPS variable has not changed), the SRM calculates the current Migration Age, based on the current value of the MCVMGCNT variable. The current Migration Age is the maximum of the previously computed Migration Age (based on the previous scan of the EST) and the currently computed Migration Age (based on the current scan of the EST).
- The computed Migration Age is simply 1.5 times the count of the number of times that the IRARMPR1 routine has been invoked since the last time the Migrator moved through the EST. Since the IRARMPR1 routine is invoked each second, a Migration Age of 1800 (for example) represents 1200 seconds of elapsed time during the Migrator's previous or current scan of the EST.

  Why is the Migration Age computed as 1.5 times the Migrator's scan of the EST? Consider the process by which the Migrator sets the ESTOLD bit to indicate that a page will be considered "old" on the next scan of the EST. If a page were moved to expanded storage immediately ahead of the Migrator's scan, that page would immediately be flagged as an "old" page, and would be eligible for migration on the next scan. In this case, the page could stay in expanded storage exactly one scan of the EST before the Migrator selected it as a candidate for migration.

  However, if a page were moved to expanded storage immediately behind the Migrator's scan, that page would not be flagged as an "old" page until the **next** scan of the EST. In this case, the page could stay in expanded storage for two scans of the EST before the Migrator selected it for migration.

  On average, a page will stay in expanded storage for 1.5 scans of the EST before being selected as a candidate for migration. This is why the Migration Age is computed as 1.5 times the number of seconds to scan the EST. The Migration Age is a statistical value which is computed as the **estimated** average time "old" pages (those pages with the ESTOLD bit set) have

remained in expanded before being considered eligible for migration.

It is important to appreciate that the Migration Age does not provide a measure of how long a page has remained in expanded storage. The Migration Age is not associated with any **particular** page nor is it even a measure of how long **any** page in expanded storage might have remained in expanded storage. The Migration Age is simply a measure of how long it takes the Migrator to scan the Expanded Storage Table.

The Migration Age can be quite deceptive since the Migration Age does **not** provide information about the **number** of pages which have remained in expanded storage at the Migration Age. Expanded storage frames are freed without migrating in a variety of situations (the page is moved from expanded storage to central storage during normal page fault processing, job steps complete, TSO users log off, a FREEMAIN is issued, etc). As expanded storage frames are freed, these frames become available and migration is not required to make the expanded storage frames available.

Since migration is not required to make the frames available, a small number of pages might remain in expanded storage for a long time. The Migration Age could become quite high, even though only a small number of "old" pages were in expanded storage.

This situation is particularly pervasive if there is a high rate of expanded frames being freed without migration, there is a high page movement rate between expanded storage and central storage, or most migration is accomplished by purge migration. In these cases, the Migrator might not invoked frequently, there might be a relatively small number of frames eligible for migration during a scan of the EST, or the LRU migration algorithms might not even be invoked. The "old" pages could have been in expanded storage significantly longer or significantly shorter than the Migration Age.

- The "old" pages could have been in expanded storage for up to 33% longer than the Migration Age (for example, if the Migration Age were 1800, all "old" pages could have been in expanded storage for as much as 2400 seconds). This situation could occur if a few pages were moved to expanded storage and the EST elements (ESTE) for these expanded storage frames were immediately behind the Migrator's scan of the EST. (The 33% additional time represents two scans of the EST, while the Migration Age represents 150% of the time to scan the EST.)
- The "old" pages could have been in expanded storage for a very short time (e.g., less than a second) even though the Migration Age might be very high. For example, suppose that the Migrator were invoked infrequently (because of page movement to central, because expanded frames were freed, etc.). The Migration Age could become high. Once the Migrator was invoked, it is possible that it could find no "old" frames remaining between its current position in the EST and the end of the EST. It is possible that the pages in the remaining frames had just been placed into expanded storage. These frames would be marked as "old" frames by turning ON the ESTOLD bit.

  As the Migrator wrapped the EST and began another scan, it could discover that all expanded storage frames up to its previous scan contained "new" pages. The Migrator would set the ESTOLD bit ON for these frames, but the Migrator would not encounter "old" frames until it encountered those frames it had just marked as "old" during the previous scan. These frames would have been in expanded storage for a very short time, yet would be considered as "old" frames and selected for migration. Thus, the Migration Age could be high (for example, it could be 7000) while frames which had been in expanded storage for less than a second were selected for migration.

  The Migrator might reach the end of the EST before filling its quota of pages to migrate. In this case, the Migration Age would sharply decrease (for example, it might immediately decrease from 7000 to 20). On the other hand, the Migrator might fill its quota and not be again invoked for some considerable elapsed time. Consequently, the Migration Age might not so sharply decrease.
- Purge migration can seriously distort the meaning of the Migration Age. Purge migration migrates pages without implementing the EST scan processing. Since the EST scan processing is not performed, "old" pages are not selected for migration. It is quite possible that purge migration could make expanded storage frames available without migrating any "old" pages. One consequence of this could be that the Migration Age would increase, even though only a few "old" pages might be in expanded storage. No statistics are available to determine whether purge migration is being performed versus LRU migration being performed.

These examples illustrate situations in which the Migration Age has no relationship to the length of time pages have remained in expanded storage.

When evaluating performance problems, it is important to realize that the Migration Age is intended to simply be one indicator of whether expanded storage is constrained. To appreciate whether expanded storage really is constrained, both the Migration Age and the number of expanded storage frames at that Migration Age must be assessed.

Unfortunately, the ESCTxxxx keywords use only the Migration Age as the metric indicating whether expanded

storage is constrained. Since the Migration Age is based on LRU migration, the specifications associated with these keywords may be ineffective in many environments.

3.3 ORDER OF MIGRATING PAGES

In understanding the migration algorithms, it is interesting to consider the order in which pages for an address space are sent to expanded storage. For any address space, stolen pages are sent to expanded storage while the address space is active, swap trim pages are sent to expanded storage while the address space is logically swapped or while being trimmed for physical swap, secondary working set pages are sent to expanded storage when the address space is physically swapped to expanded storage, and finally the primary working set pages are sent to expanded storage.

- The page steal process steals pages from active address spaces based on the UIC of the page. The stolen pages normally are sent to expanded storage if they meet the criteria specified in ESCTxxxx, if the pages are stolen from storage isolated address spaces, or (with SP4.2) if abundant expanded storage is available. If the RSM encounters contiguous virtual storage pages with the same UIC, the RSM blocks these pages and sends them to expanded storage as a block.

  Prior to SP4.2, a maximum of 10 pages would be stolen from any active address space. This restriction has been removed with SP4.2 page stealing. With SP4.2, the RSM will be directed to steal as many pages from an address space as necessary to replenish the Logical Swap Available Frame Queue, at a given UIC "steal criteria"[14]. This change in logic is intended to take advantage of the performance benefits of block paging.

- The swap trim process steals pages not in the working set from logically swapped address spaces or from address spaces which are being trimmed for a physical swap. The RSM normally will be directed to trim as many pages not in the working set pages from an address space as necessary to replenish the Logical Swap Available Frame Queue. The swap trim pages can be directed to expanded storage or to auxiliary storage, based on the ESCTxxxx criteria, or (with SP4.2) if abundant expanded storage is available. If the swap working set pages are targeted for expanded storage, the UIC comprising the working set is increased by two. Consequently, the working set of address spaces targeted for expanded storage may be larger than the working set of an address space directed to auxiliary storage.

  The swap trim process will trim an address space down to a maximum of the large job working set size (default of 512 frames) or the protected working set (for storage isolated address spaces), and to a minimum of 96 frames.

- The working set of an address space which is directed to expanded storage will be broken into a "primary" and a "secondary" working set. The **primary** working set consists of all LSQA pages, all fixed pages (these are converted to disabled reference or DREF pages), and one page from each segment of virtual storage used by the address space (a segment is one megabyte of virtual storage). The **secondary** working set consists of all other pages in the working set.

  For a typical TSO address space, the primary working set would be relatively small (perhaps 20-30 pages) while the secondary working set might be 80-90 pages.

  The address space will be targeted for expanded storage based on the ESCTxxxx criteria, or (with SP4.2) if abundant expanded storage is available. Additionally, the SRM will ensure that the address space will actually fit into available expanded storage. The SRM will send the address space to auxiliary storage even though the ESCTxxxx criteria might have been met if the address space will not fit into available expanded storage.

  The secondary working set will be sent to expanded storage in blocks, as central storage frames are needed. The primary working set will be sent to expanded storage, as a group, after the secondary working set has been sent to expanded storage.

To summarize, stolen pages are sent to expanded storage, followed by non-working set pages,

followed by secondary working set pages, followed by primary working set pages.

The order of migrating pages associated with an address space is based on the ESTE queue. Since the ESTE queue for each address space is in LRU order, the oldest pages are migrated first. The oldest pages are stolen pages, followed by non-working set pages, followed by secondary working set pages. Pages in these categories **can** be migrated a page at a time. In practice, a block of pages normally is migrated since the Migrator is attempting to fill a quota.

When the RSM encounters a primary working set page, it issues a SYSEVENT to notify the SRM that the **address space** is being selected for migration. The

[14] As described in the companion paper "Central Storage Management with MVS/ESA," the UIC "steal criteria" can encompass a range of UIC values. This range is possible because the UIC steal criteria is not lowered by 1 for each loop through the address spaces and common area one UIC at a time. Rather, the UIC steal criteria is decremented by an algorithm applied to the current steal criteria.
For example, if the current steal criteria were 200, the steal criteria algorithm would yield a new steal criteria of 159. This algorithm has the advantage of quickly removing old pages from central storage, while minimizing calls to the RSM steal algorithms.

SRM then makes a decision about whether to migrate the address space. Unless the address space has become ready to execute, the SRM normally decides to migrate the address space.

3.4 MIGRATING ADDRESS SPACES

Migrating an address space is a far more serious process than simply migrating non-working set pages or secondary working set pages. This is because the **ENTIRE** primary working set must be migrated.

In order to migrate the primary working set, the address space must be swapped into central storage, and then swapped out to auxiliary storage. The SRM goes through most of the processing associated with normal swap-in/swap-out when an address space is migrated. This processing includes determining whether the address space will fit into central storage.

Note that there was a shortage of central storage when migration began. This statement is true because pages were being sent from central storage to expanded storage and these pages were sent only because there was a shortage of central storage frames. The shortage of central storage frames triggered a movement to expanded storage and the shortage of expanded storage frames triggered migration.

The RSM reserves a relatively small amount of central storage frames to accommodate the migration of **pages** in expanded storage. However, this reserve cannot always accommodate the migration of an **address space** (in fact, the reserve cannot always accommodate the normal migration of pages if a large demand is encountered).

If the SRM cannot fit the address space into central storage, the SRM signals the RSM to make available more central storage. The migration of the address space is deferred until more central storage is available. Since more central storage must be made available and expanded storage is constrained, the SRM will temporarily direct all page movement to auxiliary storage. This has the effect of temporarily "shutting off" expanded storage, and is considered a temporary "crisis" situation.

Note that the swap-out of address spaces will be directed to auxiliary storage rather than to expanded storage during this "crisis" situation. Further, (depending upon the urgency of the central storage shortage) logical swapping may be turned off. These actions can temporarily result in extremely poor performance (particularly to on-line systems). However, the alternative (continuing to send pages to expanded storage when expanded storage was constrained) would be worse!

The amount of central storage reserved by the RSM for migration is a function of the RCEAECLO value, and the Migrator's quota of pages to migrate is the difference between RCEAECOK and RCEAECLO. These expanded storage thresholds are based on the MCCAECLO and MCCAECOK values.

With SP4.2, the MCCAECLO and MCCAECOK values can be adjusted using the low and high values of the MCCAECTH keyword in IEAOPTxx. Additionally, the central storage thresholds are based on the MCCAFCLO and MCCAFCOK values. With SP4.2, the MCCAFCLO and MCCAFCOK values can be adjusted using the low and high values of the MCCAFCTH keyword in IEAOPTxx.

One way to reduce the performance impact of migrating address spaces is to raise the default processor storage control values. Raising the central storage default thresholds will signal a shortage of central storage before available central storage gets quite as low. Raising the expanded storage default thresholds will signal a shortage of expanded storage before available expanded storage gets as low. Both these actions will facilitate migration of address spaces. Of course, altering the defaults should be done only if you have detected a performance problem caused by migrating address spaces.

4.0 SUMMARY

This paper has explained the basic techniques and algorithms used by the SRM to manage expanded storage: the type of pages sent to expanded storage, how a user expresses preference that a page be sent to expanded storage versus auxiliary storage, situations when the SRM overrides or ignores the user guidance, expanded storage thresholds which are used in many of the SRM's decisions, migration, and migration age. A companion paper ("Central Storage Management with MVS/ESA") explains the techniques used by the SRM to manage central storage.

ACKNOWLEDGEMENT

The author would like to thank Tom Beretvas (Beretvas Performance Consultants) for his many contributions to this paper, and for his constructive criticism of the draft document.

REFERENCES

MVS/ESA SP4.2 Initialization and Tuning Guide, IBM document GC28-1634-3

MVS/ESA SP4.2 Initialization and Tuning Reference, IBM document GC28-1635-3

MVS/ESA SP4.3 Initialization and Tuning Guide, IBM document GC28-1634-4

MVS/ESA SP4.3 Initialization and Tuning Reference, IBM document GC28-1635-4

MVS/ESA SP2.2 Working Set Management and Block Paging Presentation Guide, IBM document GG66-3204

MVS/ESA Component Diagnosis and Logic: System Resources Manager, IBM document LY28-1592

MVS/XA SP2.2 Source Code, IBM Microfiche LJB2-9573

MVS/ESA SP4.2 Source Code, IBM Microfiche LJB2-9605