Active Memory Sharing
Active Memory Expansion

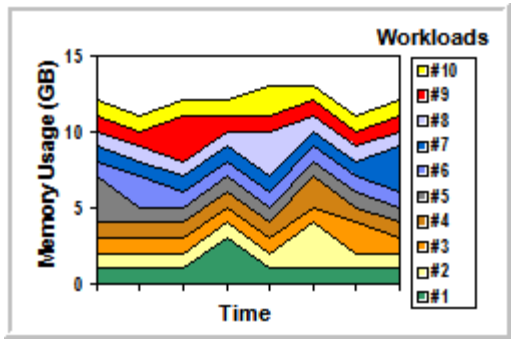
MITEC Session W-210A-3
June 7, 2011

Name: Dave Levites, IBM
Title: Power Systems Architect
email: dlevites@us.ibm.com

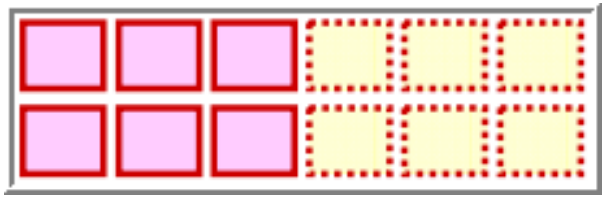
Power your planet.

Advanced Memory Techniques in AIX

Active Memory Sharing



Active Memory Expansion



PowerVM Active Memory Sharing

Agenda

- Overview
- Deployment Considerations
- Performance Guidelines
- Implementation

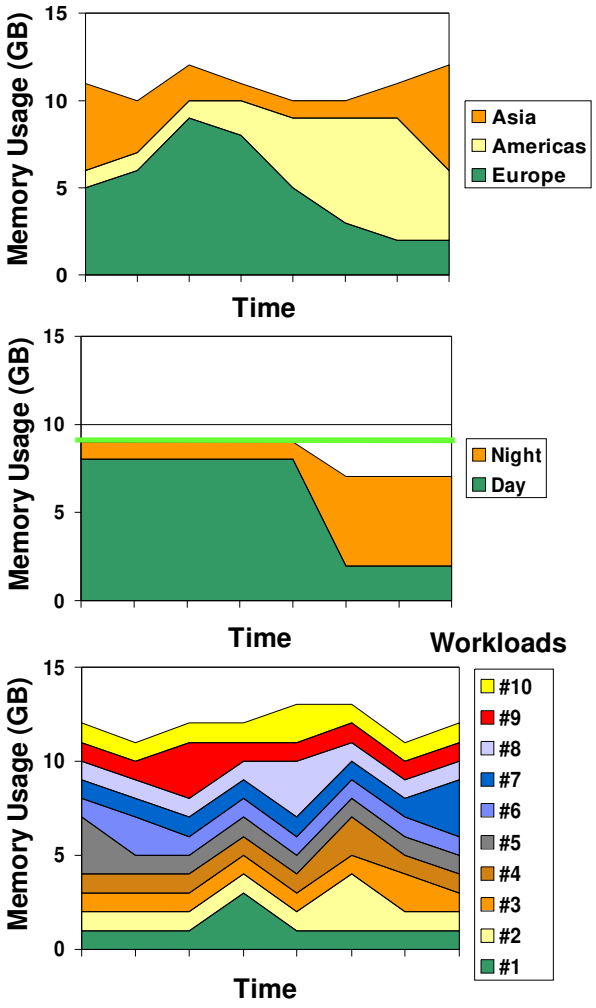
What is Active Memory Sharing?

- **Virtualization of memory similar to processor sharing**
  - A pool of physical memory dynamically allocated by the POWER Hypervisor as needed among AMS client partitions
  - Allows over commitment of memory
  - If needed, Hypervisor paging goes to devices configured on a VIO server
- **For partitions with predictable or low memory requirements**
  - Enables fine-grained sharing of physical memory and automated expansion/contraction of a partition's physical memory footprint
  - Supports OS collaborative memory management to reduce Hypervisor paging

AMS Overview

Increases system memory utilization, reduces costs

- Transparently reallocates physical memory based on partition workload requirements:
  - Designed for LPAR memory workloads that peak at **different times of day**, or
  - For highly-consolidated workloads with **low, sporadic** memory requirements
- AMS uses the POWER Hypervisor and supported operating systems to make this invisible to applications



Power your planet.

Active Memory Sharing Requirements

- PowerVM Enterprise Edition
- Minimum system requirements:
  - POWER6 and POWER7 servers and blades
  - Virtual I/O Server (VIOS) 2.1.1
  - Firmware level: 340_075_039 or 710_043
  - HMC Version 7 Release 3.4 SP 2 or HMC Version 7 Release 7.1
- Minimum operating system requirements:
  - AIX 6.1 TL3
  - IBM i 6.1 plus PTF SI32798
  - SUSE Linux Enterprise Server 11
- Partition Configuration Requirements
  - Must use **shared processors only**
  - **All I/O must be virtualized** through VIOS – no HEA or Fibre Channel supported
  - **4K pages only** – 64K or larger pages are **not** supported

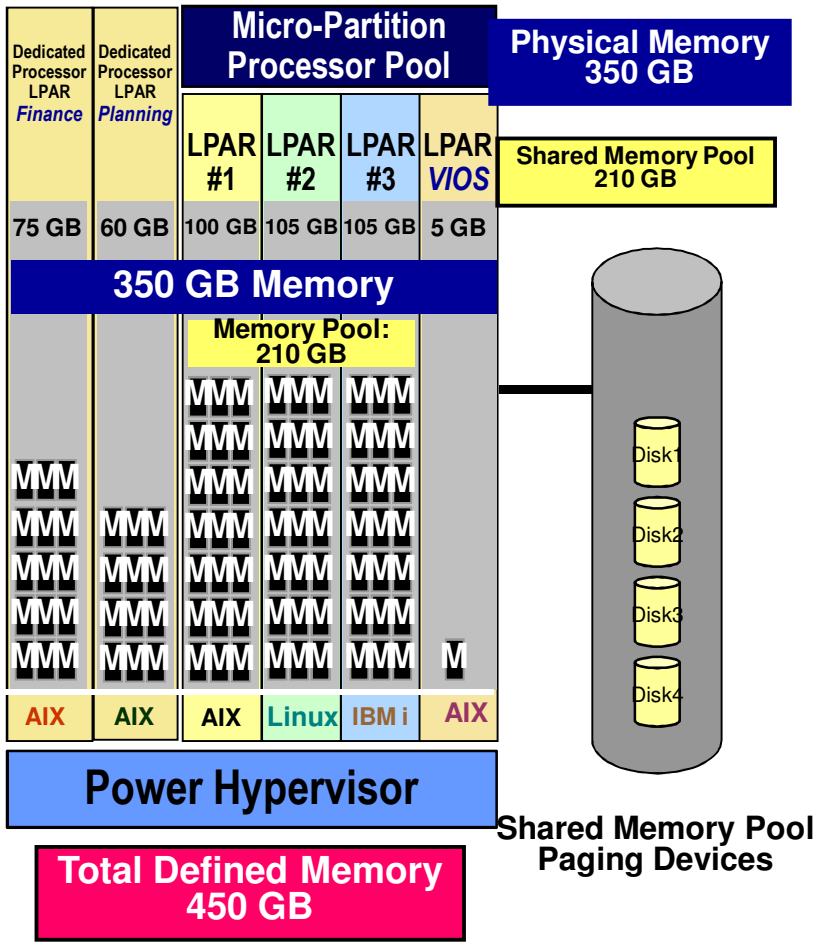
Deployment Considerations

- **Performance**
  - Depends on partition workloads, memory configuration and over-subscription levels
  - Memory latency varies based on shared pool configuration and paging devices
    - More ramp-up latency when moving large amounts of memory
    - Increased system demand can be met by dynamically increasing the shared memory pool
  - High-performance paging devices are recommended to minimize paging impacts
- **Cost Savings**
  - Reducing memory requirements may reduce system cost, depending on workloads and performance requirements
    - AMS may allow creation of more partitions
    - Only actively referenced memory needs to stay resident in an LPAR's footprint
  - AMS can save time and money of system admin who otherwise would be manually reallocating memory

Active Memory Sharing Configuration

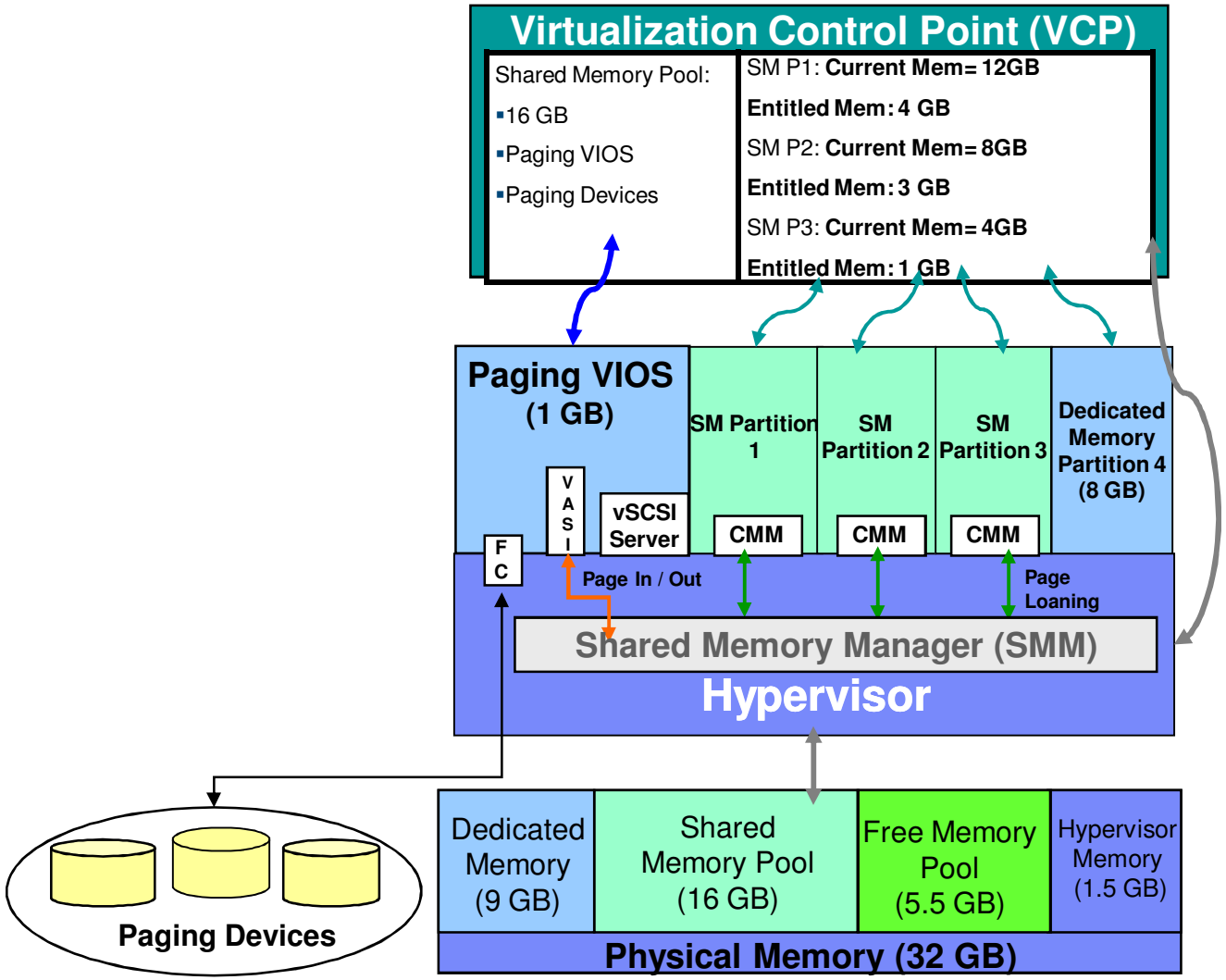
- **Shared Memory Pool**
  - Specify desired and maximum pool size
  - Assign paging devices and paging VIOS
    - Single or Redundant Paging VIOSs
  - Dynamically change pool size as needed
- **Shared Memory Partition**
  - **Partition Attributes**
    - *Min, Max, Desired* **Logical** Memory
    - **I/O Entitled Memory**: maximum amount of physical memory available for I/O mapping.
    - **Memory Weight**: partition's priority to get physical pages
    - **Paging VIOSs**: single or redundant; primary and secondary paging VIOS (optional)
  - **Switch between dedicated and shared memory mode**
    - Need to power down partition to change
  - **DLPAR memory operations** change logical memory
  - **Partition Mobility support**: among AMS capable systems

Note: Partition memory will **never exceed** *desired* value, regardless of load



Power your planet.

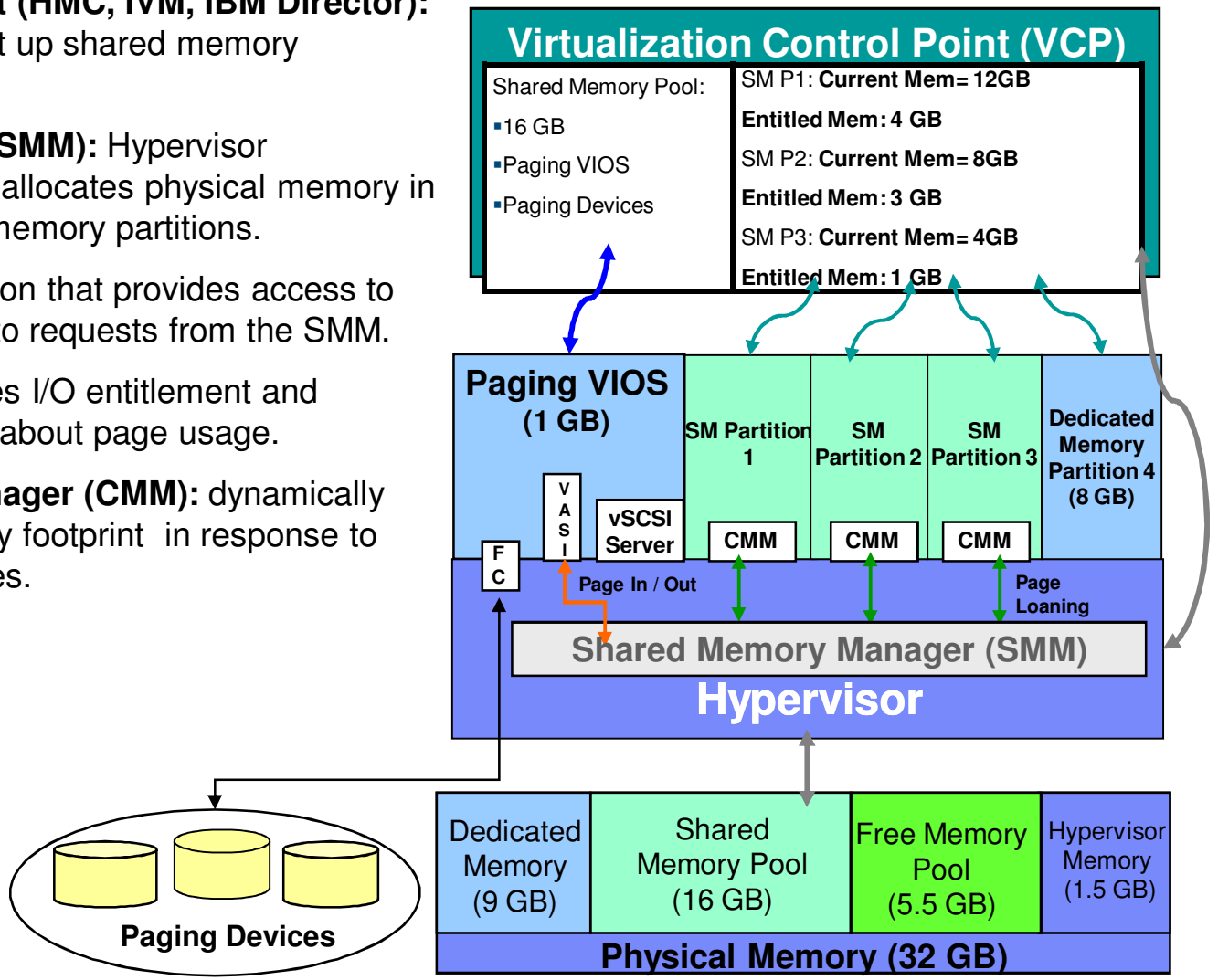
Active Memory Sharing Key Concepts



Power your planet.

Active Memory Sharing Key Concepts (cont.)

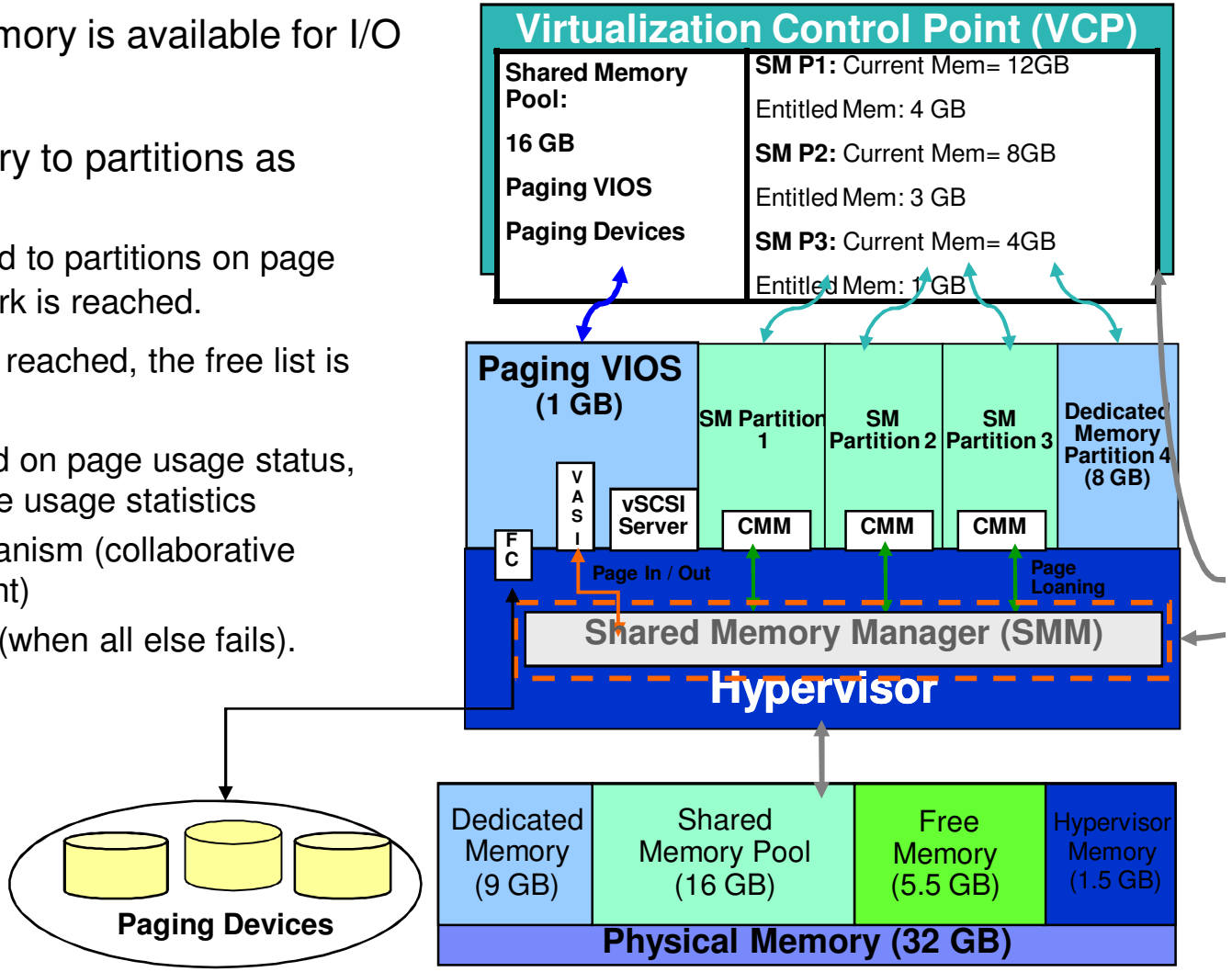
- **Virtualization Control Point (HMC, IVM, IBM Director):** provides user interface to set up shared memory environment.
- **Shared Memory Manager (SMM):** Hypervisor component that dynamically allocates physical memory in the pool among the shared memory partitions.
- **Paging VIOS:** A VIOS partition that provides access to paging devices in response to requests from the SMM.
- **Operating System:** manages I/O entitlement and provides hints to Hypervisor about page usage.
- **Collaborative Memory Manager (CMM):** dynamically changes a partition's memory footprint in response to Hypervisor requests for pages.



Power your planet.

Shared Memory Manager in the Hypervisor

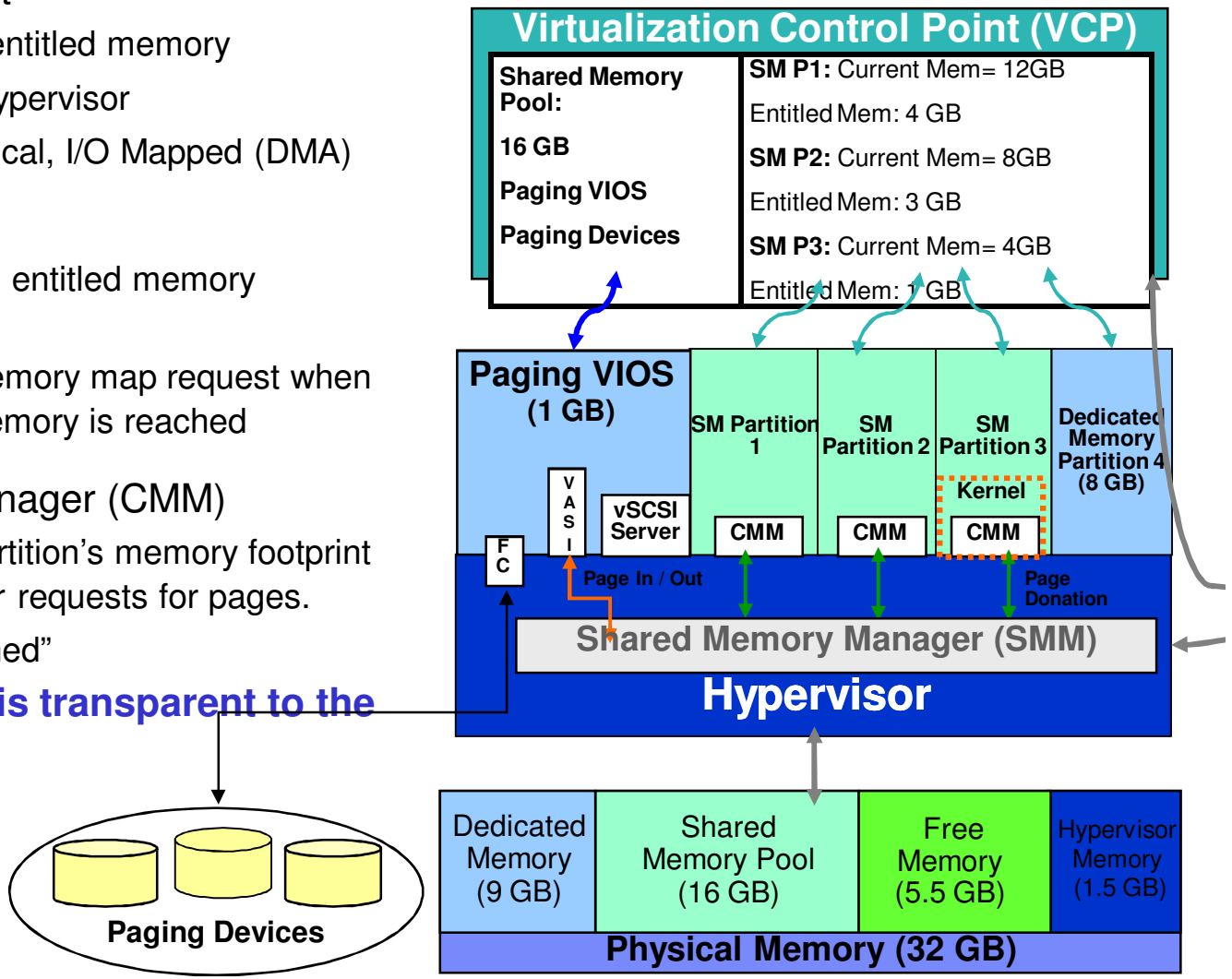
- Guarantees physical memory is available for I/O mappings
- Allocates physical memory to partitions as needed
  - Memory frames assigned to partitions on page faults until low water mark is reached.
  - When low water mark is reached, the free list is replenished through:
    - **Page stealing** based on page usage status, memory weight, page usage statistics
    - **Page loaning** mechanism (collaborative memory management)
    - **Hypervisor paging** (when all else fails).



Power your planet.

Operating System Support

- Operating System Support
  - Manages partition's I/O entitled memory
  - Provides page hints to hypervisor
    - Unused, Active, Critical, I/O Mapped (DMA)
- Device Driver Support
  - Communicates minimum entitled memory requirements to kernel
  - Handles failure of I/O memory map request when partition's I/O entitled memory is reached
- Collaborative Memory Manager (CMM)
  - Dynamically changes partition's memory footprint in response to hypervisor requests for pages.
    - Mark pages as "loaned"
- **Active Memory Sharing is transparent to the application layer**



Power your planet.

Deployment Strategy

- Use dedicated memory partition values as a reference point
- **Base:** Shared Memory Pool's size **equals** the aggregate physical memory in dedicated client partitions
  - The pool has enough physical memory to cover the clients' peak needs concurrently; no over-commitment
- **Logical Overcommit:** Pool size **less than** aggregate desired by all clients, but **sufficient** to cover estimated peak workloads occurring at different times
  - Frequently changing workloads could impact latency; additional memory can be added to the pool to meet response time criteria
- **Physical Overcommit:** Pool size **less than** aggregate desired by all clients, and **insufficient** to cover concurrent demands
  - **Hypervisor paging** results
  - **Pool size** can be changed **dynamically**

Paging Device Monitoring

- Paging devices need to be tuned
  - Look for read, write response time
  - Queue depth
  - Add more disks to array
- No loaning
  - No need to tune AIX paging
- Default loaning
  - Need to tune both (PHYP will have more paging than AIX)
- Aggressive loaning
  - Need to tune both (AIX will have more paging than PHYP)

Live Partition Mobility and Paging Devices

- Paging devices can only be assigned to one shared memory pool at a time
  - You cannot assign the same paging device to a shared memory pool on more than one system at the same time.
- Prior to migrating a shared memory partition, ensure that the destination system has an available paging device for the mobile LPAR to use
  - The paging device must be equal to or greater than the partition's maximum logical memory setting

Summary & Conclusions

- ✓ IBM PowerVM AMS technology takes virtualization to the next level by optimizing memory utilization and automating memory allocation to meet the changing demands of workloads.
- ✓ As a consequence, AMS not only improves memory utilization, but also reduces administration cost as it automates memory provisioning.
- ✓ In many cases, these benefits will give a boost to data centers and other large and midrange IT organizations, improving their total cost of ownership.
- ✓ Key to proper deployment of AMS is choosing appropriate workloads that have the right variability over time and which do not demand high, sustained memory residency.
- ✓ Generally AMS performance compares favorably with dedicated memory environments.

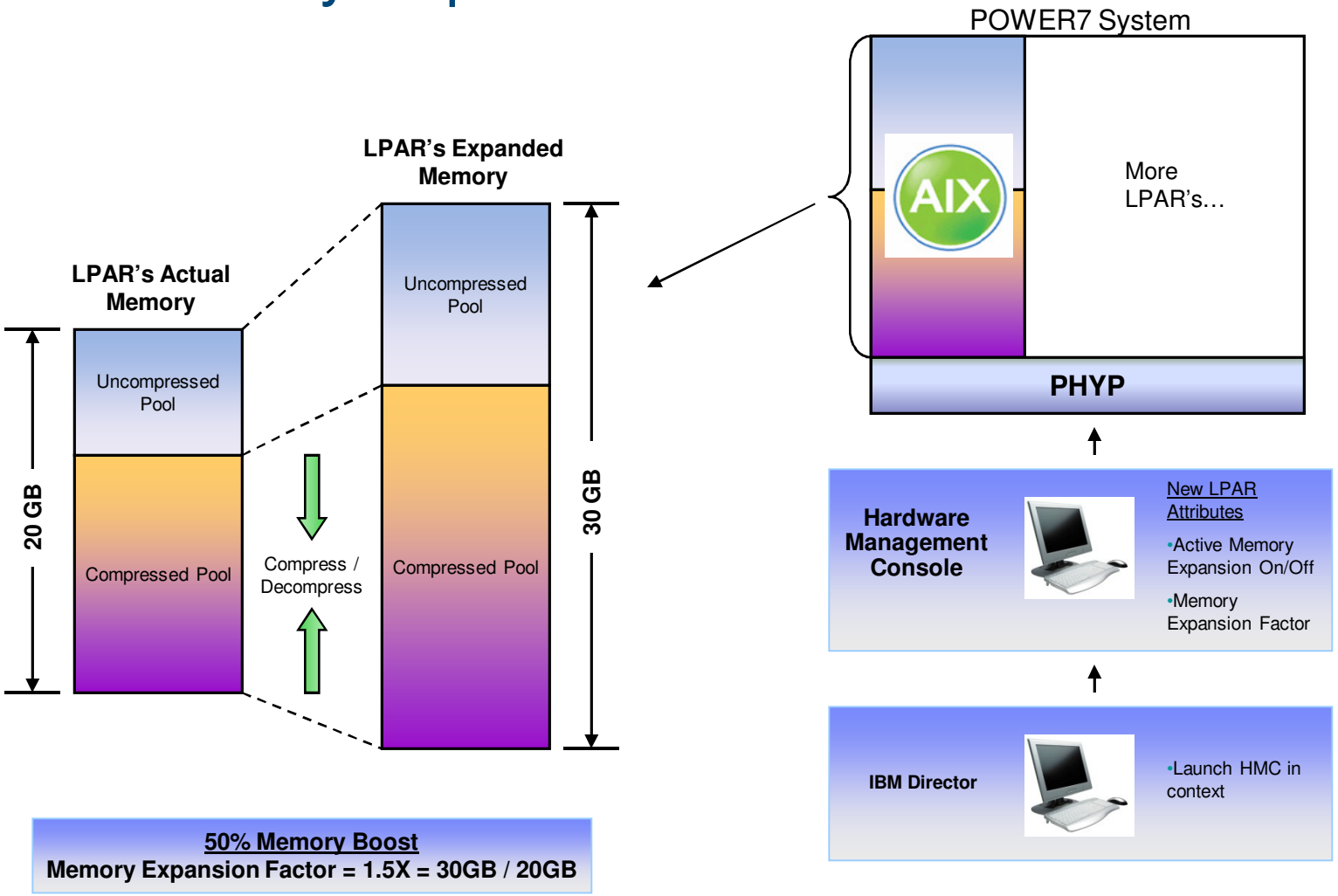
PowerVM Active Memory Expansion

Active Memory Expansion (AME) for POWER7

Expand memory capacity through in-memory data compression

- **AME, a chargeable feature, is managed by the Hypervisor and the AIX operating system**
- **Memory compression is transparent to applications**
- **Active Memory Expansion requires some additional CPU for compression and decompression**
- **AME is configurable on a per-LPAR basis and supports dedicated memory as well as Active Memory Sharing**

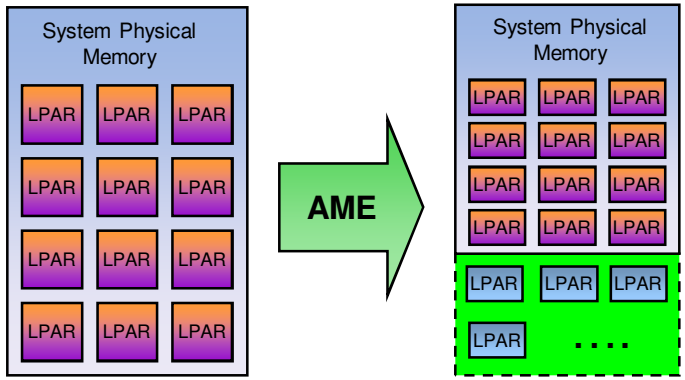
Active Memory Expansion - Overview



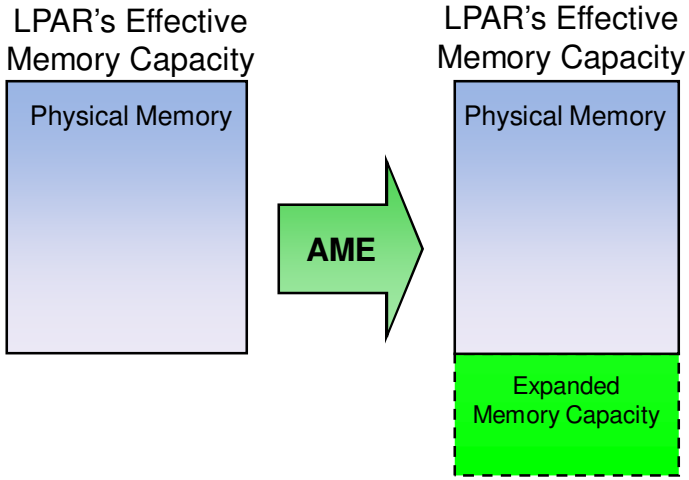
Power your planet.

The Value of Active Memory Expansion

- **Enables more LPAR's per server**
  - Reduces the memory requirements for existing LPARs so more can be created using the same resources

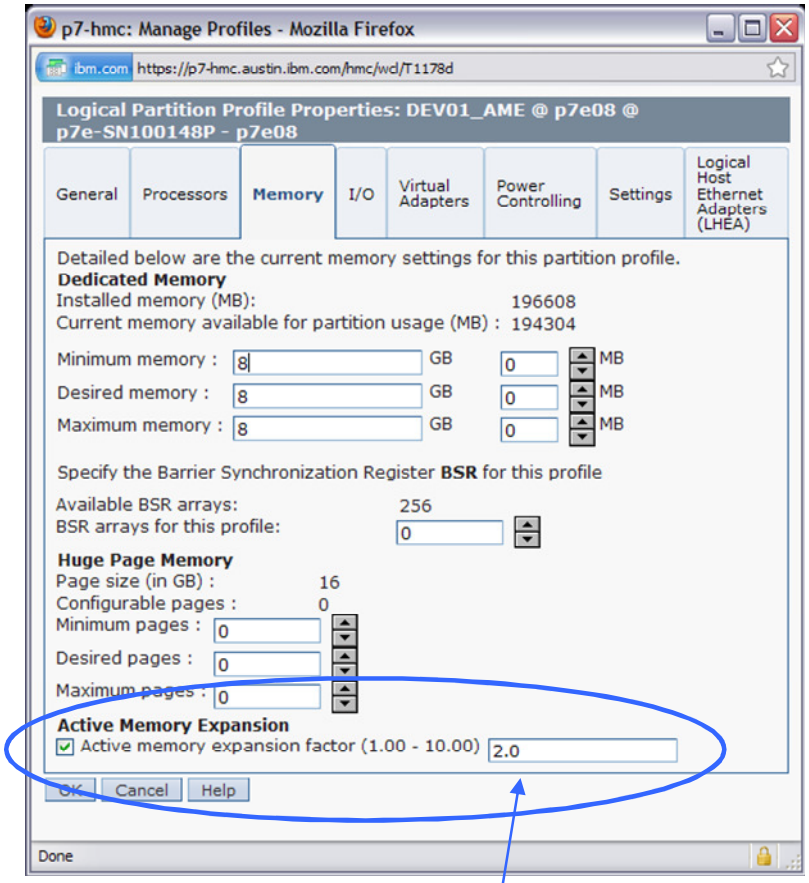


- **Increases the throughput of existing LPARs**
  - More throughput achieved by increasing the effective memory capacity



Active Memory Expansion Configuration

- HW and SW minimum requirements
  - POWER7 system
  - HMC: V7R7.1.0.0
  - FW: 7.1
  - AIX 6.1 TL4 SP2 or later
- License Requirements
  - AME requires a special activation code
  - Activation code is similar to PowerVM activation codes
  - 60-day trial activation available
    - Go to the CoD website to obtain enablement
- Configuration
  - New AME LPAR attribute (memory expansion factor)
  - Available in the memory tab of a LPAR's profile configuration in HMC



Memory Expansion Factor

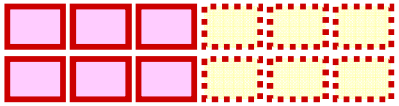
(1.0 – 10.0)

Power your planet.

Active Memory Expansion & Active Memory Sharing

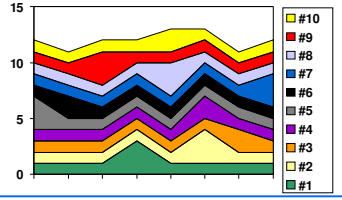
Active Memory Expansion

- Effectively gives more memory capacity to the partition using compression / decompression of some contents in true memory
- AIX partitions only



Active Memory Sharing

- Moves memory from one partition to another
- Best fit when partitions need memory at different times
- AIX, IBM i, and Linux partitions



Active Memory Expansion PLUS Active Memory Sharing

- AME would act to supplement memory removed by AMS, thus avoiding Hypervisor paging
- Considerations
  - Only AIX partitions using Active Memory Expansion
  - Active Memory Expansion value is dependent upon compressibility of data and available CPU resource

Power your planet.

AME – A Closer Look

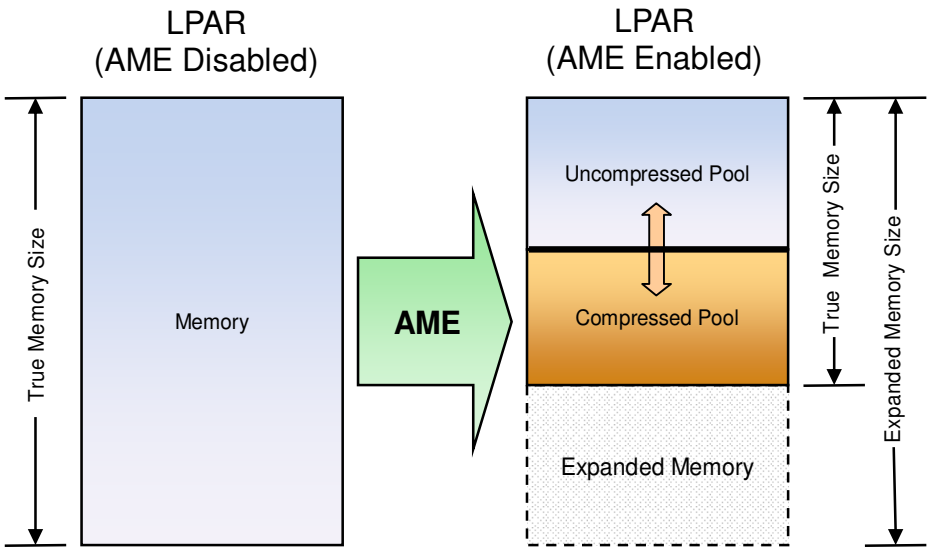
Unique Attributes of the Active Memory Expansion Environment

- Compressible memory
  - Only **unpinned**, **working storage** pages can be compressed (4K and 64K pages *)
  - Workloads that consist mostly of **pinned** or **file cache** pages are **not** good candidates for AME
- Expanded Memory View
  - Tools and API's that report memory usage will report *expanded* memory usage
  - mem0 in ODM is the exception: it will always show the true memory size of the LPAR
- Page Size
  - For improved performance, 64K pages are disabled by default for AME LPAR's

* 64K pages can be manually enabled via the vmo vmm_mpsize_support restricted tunable, but this is not recommended because of increase CPU usage

AME – A Closer Look

- Compressed pool management
  - Managed at the block level
  - Number of blocks required to compress a 4 KB page varies with the compressibility of the data
- Compressed pool size
  - Dynamically changes with the workload
  - Initially, the compressed pool will be empty
  - It expands and shrinks based on load and memory usage
- Compressed pool exhaustion
  - Pages from uncompressed pool will be paged out to paging space as needed
  - Compressed pages will NOT be paged out to paging space



Memory Expansion Factor

- **Memory Expansion Factor**
  - Controls the size of expanded memory
  - An indication to the operating system on how much memory to compress
  - The OS will compress in-memory data if possible until the specified amount of expanded memory has been achieved
  - A memory expansion factor of 2.0 means the expanded memory size for a LPAR is twice the size of the LPAR's true memory size
- **Memory Expansion Factor is NOT the compression ratio for a workload**
  - Compression ratio is an indication of ***how well*** a workload's data compresses
  - When choosing a memory expansion factor for a workload, the compression ratio should be considered
  - The memory **expansion factor** can **never be greater** than the **compression ratio** for a workload
  - Some amount of a workload's memory will always be uncompressed

Example Configurations

LPAR's True Memory Size	LPAR's Memory Expansion Factor	LPAR's Expanded Memory Size
10GB	1.5	15GB
20GB	1.5	30GB
30GB	1.5	45GB
10GB	2.0	20GB
20GB	2.0	40GB
30GB	2.0	60GB

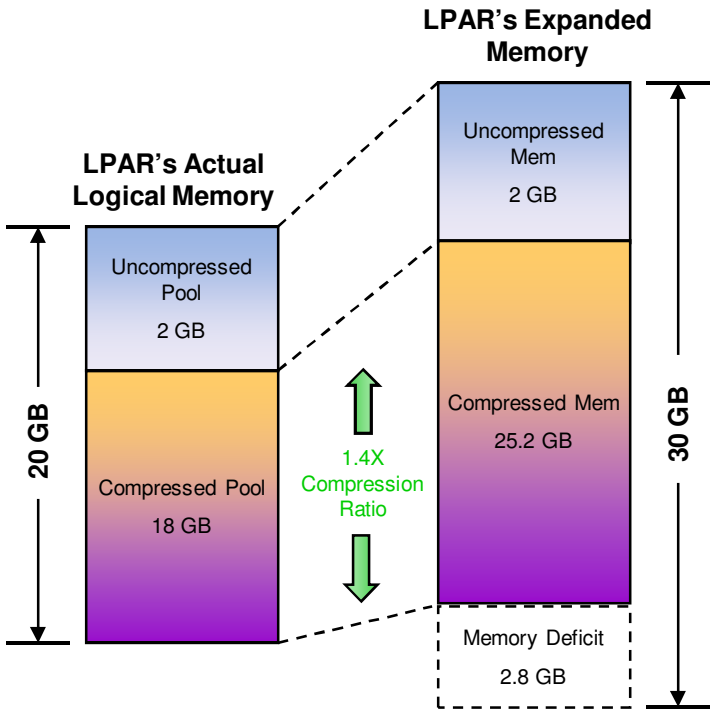
LPAR's Memory Expansion Factor	Minimum Compression Ratio
1.25	1.28
1.5	1.56
1.75	1.84
2.0	2.11
2.25	2.39
2.5	2.67

AME Details – A Closer Look

Modifying the memory expansion factor

- The expansion factor is set in the LPAR **profile** when the partition is created
- It can be changed *dynamically* via **DLPAR** once the partition is active
- Changes to the memory expansion factor appear as **DLPAR memory** changes to the applications and users in the LPAR
- When the LPAR's **true** memory size is changed dynamically, the expanded memory size is ***updated*** by a corresponding amount based on the memory expansion factor
- AME cannot be turned on or off without **shutting down** and **rebooting** the partition

Active Memory Expansion – Memory Deficits



Memory Expansion Factor = 1.5X = 30GB / 20GB

- **Memory Deficit**
  - Workload's data does not compress small enough to achieve desired expanded memory size
  - Results in a "hole" in expanded memory
- **Identifying a Memory Deficit**
  - Performance tools will report the size of a LPAR's memory deficit
    - **vmstat, lparstat, svmon, amepat**
  - Some tools report memory deficit in terms of **memory size** (e.g. bytes)
  - Other tools report memory deficit in terms of **reduced expansion factor**
- **Resolving a memory deficit**
  - The memory expansion factor can be reduced (thus reducing the expanded memory size of the LPAR)
  - To keep the expanded memory size the same, reduce the memory expansion factor and increase the amount of memory configured to the LPAR

Planning Considerations

- Workload's Compressibility
  - Better compression = better expansion
  - Most data tends to compress very well, but compressed data objects won't compress and further
- Workload's Type of Memory Usage
  - Cached file data won't be compressed
  - A file server not a great candidates for Active Memory Expansion
- Workload's Pinned Memory Usage
  - Pinned memory will not be compressed by Active Memory Expansion (ie, AIX V7 kernel)
  - Workloads that have a large pinned memory footprint are not good candidates for Active Memory Expansion

The AME Planning Tool can analyze a workload for all of these considerations.

Active Memory Expansion – Client Deployment Steps

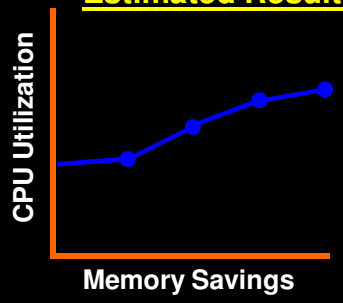
1 AME Planning Tool

A. Run AME planning tool on existing workload

B. Tool calculates compressibility of workload data and estimates potential CPU utilization increase due to AME

C. Tool provides initial recommendations on AME configuration

Estimated Results



2 AME Trial

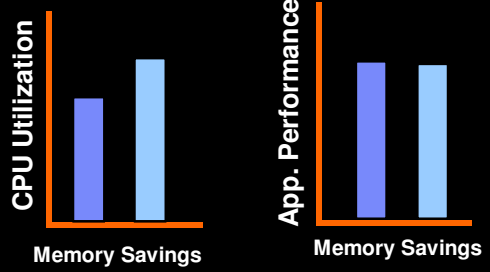
A. Activate AME (possible trial activation)

B. Configure LPAR based on recommendations from AME planning tool

C. Use performance tools to monitor workload in AME environment

D. Tune AME settings based on actual performance results

Actual Results

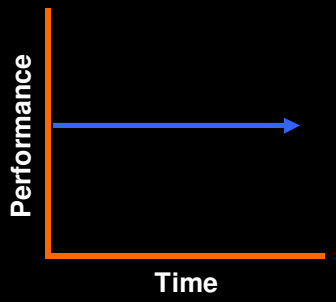


3 Deploy into Production

A. Permanently activate AME (if using trial activation)

B. Deploy workload into production

C. Continue to monitor workload using performance tools



No Application or Middleware Changes Required

Power your planet.

Active Memory Expansion Planning Tool (amepat)

- AME Planning Tool is available with AIX 6.1 TL4 SP2
- Planning Tool can be run on POWER4 through POWER7 systems
  - For example, the planning tool can be run on a POWER5 system to plan for moving a workload from a POWER5 system to a POWER7 system
- AME Planning Tool should be run in conjunction with the workload being monitored
  - AME Planning Tool will analyze the running workload
  - AME Planning Tool should ideally be run during the workload's peak utilization time
- AME Planning Tool is run in two phases
  - 1st Phase (On-Line Phase): Monitors the workload and writes metrics to recording file
  - 2nd Phase (Off-Line Phase): Generates a report from the recording file
- AME Planning Tool can be launched from the AIX command-line or from smit
  - Command-line Example:
    - Monitor a workload for 120 minutes and log recording information to ame.recfile:
    - Generate a report from the recording file, ame.recfile:

AME Planning Tool – Sample Report (Page 1)

```
Command Invoked                 : amepat 120
Date/Time of invocation         : Wed Oct  7 15:50:19 CDT 2009
Total Monitored time            : 2 hours
Total Samples Collected         : 10

System Configuration:
---------------------
Partition Name                          : p7e08
Processor Implementation Mode           : POWER7
Number Of Logical CPUs                  : 8
Processor Entitled Capacity             : 2.00
True Memory                             : 8.00 GB
SMT Threads                             : 4
Shared Processor Mode                   : Enabled
Active Memory Sharing                   : Disabled
Active Memory Expansion                 : Disabled

System Resource Statistics:                 Min             Average           Max
---------------------------             -----------     -----------     -----------
CPU Util (Phys. Processors)              0.81 [ 41%]     0.83 [ 41%]     0.84 [ 42%]
Virtual Memory Size (MB)                 6230 [ 76%]     6230 [ 76%]     6230 [ 76%]
True Memory In-Use (MB)                  6404 [ 78%]     6404 [ 78%]     6404 [ 78%]
Pinned Memory (MB)                        719 [  9%]      719 [  9%]      719 [  9%]
File Cache Size (MB)                      158 [  2%]      158 [  2%]      158 [  2%]
```


Power your planet.

AME Planning Tool – Sample Report (Page 2)

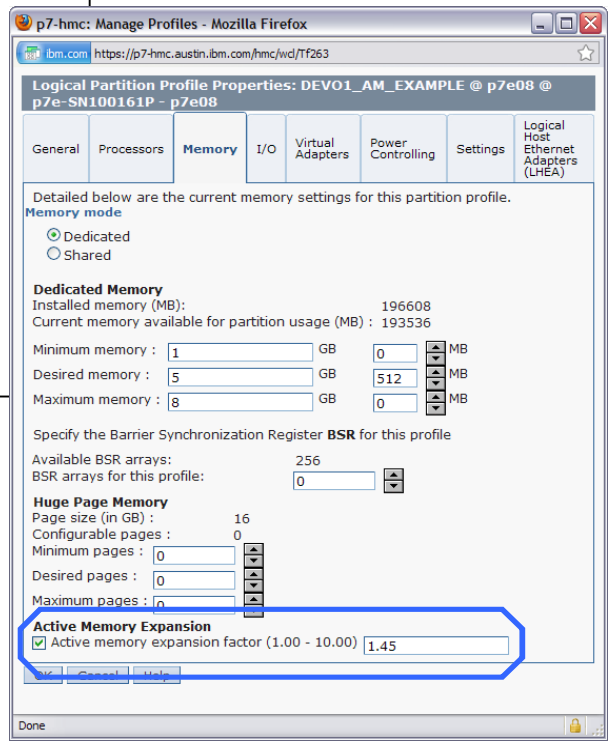
```
Active Memory Expansion Modeled Statistics:
-------------------------------------------
Modeled Expanded Memory Size :   8.00 GB
Average Compression Ratio    :   6.81

Expansion      Modeled True         Modeled           CPU Usage
 Factor         Memory Size       Memory Gain          Estimate
---------    --------------    ------------------    -----------
  1.19       6.75 GB              1.25 GB [ 19%]     0.00 [  0%]
  1.28       6.25 GB              1.75 GB [ 28%]     0.20 [  5%]
  1.39       5.75 GB              2.25 GB [ 39%]     0.35 [  9%]
  1.45       5.50 GB              2.50 GB [ 45%]     0.58 [ 15%]
  1.60       5.00 GB              3.00 GB [ 60%]     1.46 [ 73%]

Active Memory Expansion Recommendation:
---------------------------------------
The recommended AME configuration for this workload is to configure the LPAR
with a memory size of 5.50 GB and to configure a memory expansion factor
of 1.45.  This will result in a memory gain of 45% from the LPAR's current
memory size.  With this configuration, the estimated CPU usage due to AME is
approximately 0.58 physical processors, and the estimated overall peak CPU
resource required for the LPAR is 1.42 physical processors.
```

Active Memory Expansion Configuration Steps

1. Run AME Planning Tool

```
Active Memory Expansion Modeled Statistics:
-------------------------------------------
Modeled Expanded Memory Size :   8.00 GB
Average Compression Ratio    :   6.81

Expansion      Modeled True         Modeled            CPU Usage
 Factor         Memory Size       Memory Gain          Estimate
---------    --------------    -----------------    -----------
  1.19       6.75 GB              1.25 GB [ 19%]     0.00 [  0%]
  1.28       6.25 GB              1.75 GB [ 28%]     0.20 [  5%]
  1.39       5.75 GB              2.25 GB [ 39%]     0.35 [  9%]
  1.45       5.50 GB              2.50 GB [ 45%]     0.58 [ 15%]
  1.60       5.00 GB              3.00 GB [ 60%]     1.46 [ 73%]

Active Memory Expansion Recommendation:
---------------------------------------
The recommended AME configuration for this workload is to configure the LPAR
with a memory size of 5.50 GB and to configure a memory expansion factor
of 1.45.  This will result in a memory gain of 45% from the LPAR's current
memory size.  With this configuration, the estimated CPU usage due to AME is
approximately 0.58 physical processors, and the estimated overall peak CPU
resource required for the LPAR is 1.42 physical processors.
```

2. Enable AME for LPAR



3. Go!

Power your planet.

AME Performance Tools Cheat Sheet

- Help size/plan a workload for deployment with AME
  - amepat
- Get AME configuration for LPAR
  - lparstat –i
- Monitor AME activity for a LPAR
  - vmstat –c
  - lparstat –c
  - topas
  - amepat
- Detailed information on compressed memory usage
  - svmon –O summary=ame,unit=auto

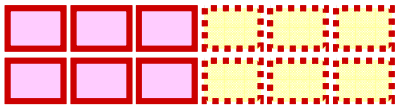
AME Performance Analysis: lparstat

```
lparstat -c 1


System configuration: type=Dedicated mode=Capped mmode=Ded-E smt=On
                      lcpu=8 mem=8704MB tmem=4352MB

%user  %sys  %wait  %idle  %xcpu   dxm
-----  -----  ------  ------  ------  ------
 34.3  36.2    4.3   25.2  12.0      0
 39.1  37.6    0.2   23.1  12.5      0
 23.8  40.6    3.0   32.6  13.9      0
 32.7  41.9    0.0   25.4  14.4      0
 35.3  40.2    3.4   21.1  14.1      0
 37.2  39.9    2.4   20.4  14.0      0
 39.0  39.3    1.2   20.4  13.2      0
```

Percentage of CPU consumed for memory compression; This is a percentage of the overall CPU capacity in the LPAR.

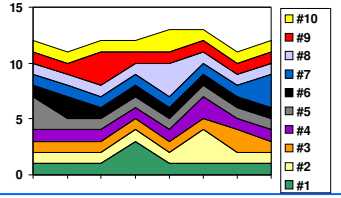
Active Memory Expansion & Active Memory Sharing

- Active Memory Expansion
- Effectively gives more memory capacity to the partition using compression / decompression of the contents in true memory
- AIX partitions only



Active Memory Sharing

- Moves memory from one partition to another
- Best fit when one partition is not busy when another partition is busy
- AXI, IBM i, and Linux partitions



Active Memory Expansion Plus Active Memory Sharing

- Supported, potentially a very nice option
- Considerations
  - Only AIX partitions using Active Memory Expansion
  - Active Memory Expansion value is dependent upon compressibility of data and available CPU resource

Additional Information:

- Hands-on Videos for AIX, Power 7, PowerVM
  http://www.ibm.com/developerworks/wikis/display/wikiptype/movies

- IBM Redbooks:
  AMS: http://www.redbooks.ibm.com/abstracts/redp4470.html?Open
  AME: http://www.redbooks.ibm.com/abstracts/sg247590.html?Open

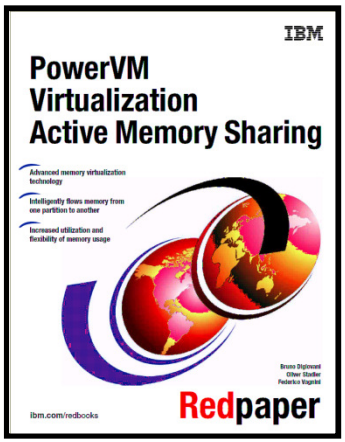
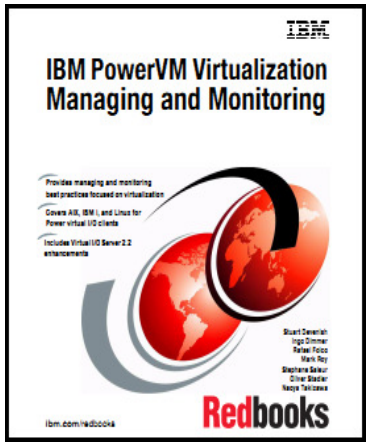


Table of contents
Chapter 1. Overview
Chapter 2. Detailed architecture
Chapter 3. Planning for Active Memory Sharing
Chapter 4. Configuring and managing
Chapter 5. Monitoring



Section 16.4:
Monitoring AME

Active Memory Sharing Documentation

- IBM PowerVM Active Memory Sharing Performance
  ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/pow03017usen/POW03017USEN.PDF
- Using Active Memory Sharing on SLES11
  http://www.ibm.com/developerworks/wikis/display/LinuxP/Using+Active+Memory+Sharing+on+SLES11
- PowerVM portal on IBM web site
  http://www-03.ibm.com/systems/power/software/virtualization/
- PowerVM information roadmap
  http://publib.boulder.ibm.com/eserver/roadmap_powervm.html
- *PowerVM Virtualization Active Memory Sharing* Redpaper
  http://www.redbooks.ibm.com/abstracts/redp4470.html?Open

Active Memory Expansion Documentation

- IBM PowerVM Active Memory Sharing Performance
  ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/pow03017usen/POW03017USEN.PDF
- Active Memory Expansion: Overview and Usage Guide
  http://www-03.ibm.com/systems/power/hardware/whitepapers/am_exp.html
- IBM Developers Works Wiki
  https://www.ibm.com/developerworks/wikis/display/WikiPtype/IBM+Active+Memory+Expansion
- Performance Whitepaper
  http://www-03.ibm.com/systems/power/hardware/whitepapers/am_exp_perf.html

Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

Special notices (cont.)

Thank you for your interest in IBM Power Systems

Power your planet.