Memory Virtualization Basics

Asst. Prof. Ashwini Mathur

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/b255d4ea-505a-4f86-8d94-dfa416c7ad29/perf-vsphere-memory_management.pdf

- Before you manage memory resources, you should understand how they are being virtualized and used by ESXi.
- The **VMkernel manages all physical RAM** on the host. The VMkernel dedicates part of this managed physical RAM for its own use. The rest is available for use by virtual machines.
- The **virtual and physical memory space is divided into blocks called pages**. When physical memory is full, the data for virtual pages that are not present in physical memory are stored on disk. Depending on processor architecture, pages are typically 4 KB or 2 MB.

Virtual Machine Memory

▼ Each virtual machine consumes memory based on its configured size, plus additional overhead memory for virtualization

Virtual Machine Memory - Explaination

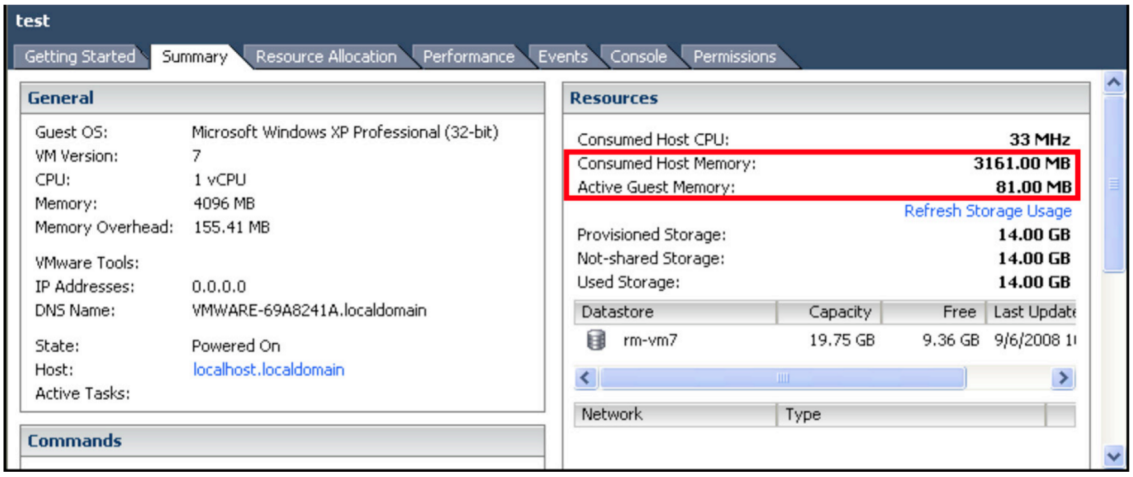


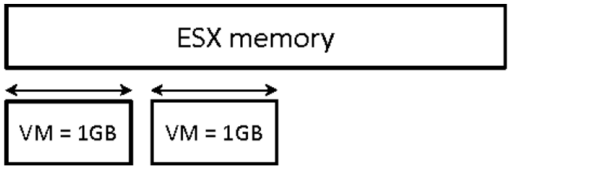
Figure: Host and Guest Memory Usage

1. Consumed Host Memory usage is defined as the amount of host memory that is allocated to the virtual machine, Active Guest.
2. Memory is defined as the amount of guest memory that is currently being used by the guest operating system and its applications.

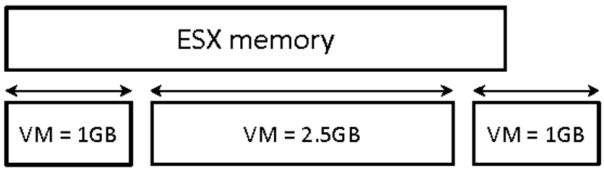
Memory Overcommitment

For each running virtual machine, the system reserves physical RAM for the virtual machine's reservation (if any) and for its virtualization overhead.

Memory Overcommited Explaination



(a)



(b)

Memory Sharing

Memory sharing is a proprietary ESXi technique that can help achieve greater memory density on a host.

Memory sharing relies on the observation that several virtual machines might be running instances of the same guest operating system. These virtual machines might have the same applications or components loaded, or contain common data. In such cases, a host uses a proprietary Transparent Page Sharing (TPS) technique to eliminate redundant copies of memory pages. With memory sharing, a workload running on a virtual machine often consumes less memory than it might when running on physical machines. As a result, higher levels of overcommitment can be supported efficiently. The amount of memory saved by memory sharing depends on whether the workload consists of nearly identical machines which might free up more memor