White Paper

Intel Information Technology

Computer Manufacturing

Server Virtualization

Memory Sizing for Server Virtualization

Intel IT has standardized on 16 gigabytes (GB) of memory for dual-socket virtualization hosts. We based this decision on extensive analysis of maximum physical memory consumption on more than 3,000 servers running non-virtualized workloads in our business computing environment. We found that approximately half of these servers consume 1 GB of memory or less. For workloads of this size, we believe that we can achieve high consolidation ratios of up to 15-20 to 1 using low-cost dual-socket virtualization hosts based on quad-core processors. We expect that this strategy will minimize cost because we avoid paying for unused memory and associated power and cooling.

Sudip Chahal and Todd Glasgow, Intel Corporation

July 2007

IT@Intel

Executive Summary

Intel IT has standardized on 16 gigabytes (GB) of memory for dual-socket virtualization hosts. We based this decision on extensive analysis of maximum physical memory consumption on more than 3,000 servers running non-virtualized workloads in our business computing environment. We found that approximately half of these servers consume 1 GB of memory or less. For workloads of this size, we believe that we can achieve high consolidation ratios of up to 15-20 to 1 using low-cost dual-socket virtualization hosts based on quad-core processors and configured with 16 GB of memory.

We believe we can achieve consolidation ratios of up to 15-20 to 1 using dual-socket virtualization hosts with quad-core processors and 16 GB of memory.

To determine optimum memory size for virtualization host servers, we monitored maximum physical memory consumption on more than 3,000 business systems at Intel and compared the results with data from other companies. We reconciled the actual memory consumption information with available memory configurations and costs to determine optimum memory size for the virtualization hosts.

We found that:

- Memory utilization can vary widely, even within similar workload categories.
- Excluding the upper 25 percent, or top quartile, the servers sampled used an average of 1 GB of memory. Utilization patterns were similar at the other companies.
- 2 GB dual in-line memory modules (DIMMs) represented the best tradeoff of capacity, power, and cost.
- Dual-socket servers with quad-core processors and 16 GB of memory can support 15 to 20 virtual machines (VMs) for most applications that are virtualization candidates in the near- to mid-term.

Our analysis shows the importance of monitoring actual memory use when determining optimum virtualization host memory size. We believe our memory sizing strategy will minimize cost because we avoid paying for unused memory and associated power and cooling.

Contents

Business Challenge

Like many organizations, Intel IT is pursuing server virtualization as part of a broad strategy to increase the efficiency and flexibility of our computing environment. We expect to reduce costs in areas including hardware, technical support, and power and cooling. We also expect to be able to provision server capacity more quickly and flexibly to meet changing business requirements.

Our goal is to achieve high consolidation levels of up to 15-20 to 1. This means that each virtualization host must be able to run many workloads in VMs with good performance.

Determining the optimum standard memory size for these virtualization hosts is a complex issue. In order to determine the total memory required on a virtualization host, we need to know how much memory to allocate for each VM. However, it is difficult to make general assumptions about VM size because we are consolidating a broad range of business computing workloads that may have varying memory requirements. Allocating inadequate VM memory may reduce workload performance, while allocating too much memory is likely to result in underutilization and increased total cost of ownership (TCO).

The consequences of specifying too much memory on virtualization hosts could include:

- Paying for memory that is never used, as well as increased power consumption and cooling
- Specifying a different (and possibly less cost-effective) class of servers in order to accommodate the perceived need for additional memory

Specifying too little memory would also be problematic. Potential consequences include:

- Poor application performance
- Purchasing more servers than necessary, increasing TCO for our environment overall

We realized that we needed a method for accurately determining the optimum virtualization host server memory size, taking into account factors such as workload and VM hypervisor memory requirements, available hardware platforms, memory costs, and our consolidation goals.

Determining Virtualization Host Memory Size

We created a step-by-step plan for determining optimum memory size for the virtualization hosts.

1. Analyze the maximum physical memory actually consumed on business computing servers running workloads that we plan to migrate into VMs.
2. Select a candidate virtualization host server platform.
3. Analyze potential memory configurations on the candidate server platform, including the number of memory slots and the density and cost of available memory modules.
4. Estimate VM hypervisor memory requirements, as well as estimated savings from sharing memory pages between VMs.
5. For each potential memory configuration identified in Step 3, use information from Steps 1 and 4 to determine the consolidation levels we can achieve.
6. Reconcile platform memory configurations with cost information and consolidation goals to determine the optimum virtualization host server memory size.

Physical Memory Consumption on Existing Servers

We began our analysis by examining physical memory consumption on production systems representative of those that we planned to virtualize. We also compared our results with memory consumption data from other companies.

We used internally developed tools to collect utilization data from more than 3,000 business computing servers at Intel. The servers were running non-virtualized workloads on Microsoft Windows*. They included file, e-mail, Web, backup, and database servers.

For each server, we estimated maximum memory consumption as follows:

- We took a snapshot of utilization data every 10 minutes over a five-week period.
- Each snapshot captured key Microsoft Windows performance counters such as minimum memory available.

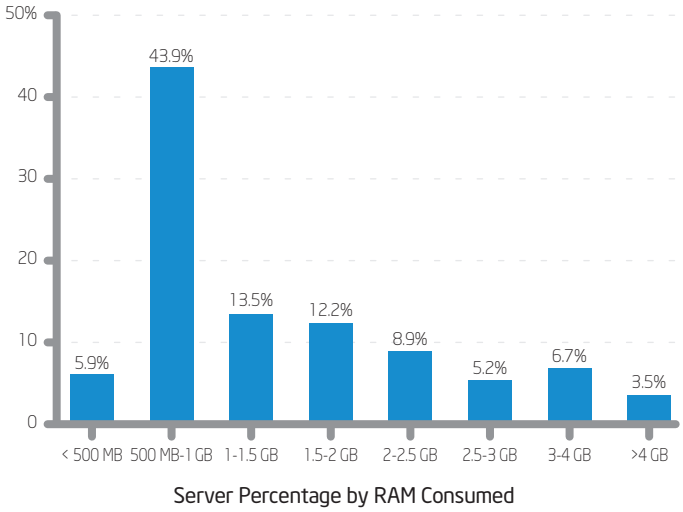


Figure 1. Analysis of memory consumption patterns showed that nearly 50 percent of production servers used 1 GB of memory or less.

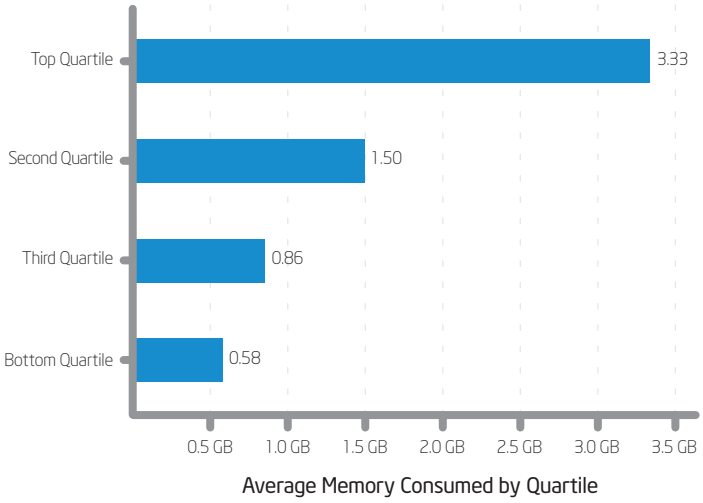


Figure 2. Excluding the top quartile, analysis showed that for 75 percent of production servers, maximum memory consumption averaged about 1 GB.

- Using our inventory database, we obtained information about the total memory installed on each system.
- We used the collected data to calculate maximum memory consumption:
  - For each snapshot, we calculated memory utilization by subtracting the minimum memory available from the total memory installed on the system.
  - We used the largest of all these calculated numbers over the five-week period as our estimate of maximum memory consumption.

Analysis of Memory Consumption

Based on the collected data, we analyzed memory consumption patterns within the sample of servers.

About 50 percent of the approximately 3,000 servers used no more than 1 GB of memory, as shown in Figure 1.

For further analysis, we then divided the servers into four equal groups, or quartiles, based on maximum memory utilization. We calculated the average maximum memory consumption within each quartile, as shown in Figure 2.

When we excluded the top quartile—the 25 percent of servers with the greatest memory consumption—we found that the average among the remaining 75 percent was around 1 GB.

We compared our results with data from other companies. Server memory consumption was similar at these companies, with at least 50 percent of servers utilizing a maximum of 1 GB memory or less.

We then analyzed a subset of the servers by application type. In several common categories, a significant proportion of servers used less than 1 GB memory. These included:

- Nearly 50 percent of file servers
- Nearly 50 percent of Web servers
- More than 50 percent of backup servers
- About 30 percent of e-mail servers
- Nearly 25 percent of database servers

Target Platform

We assumed that we would use Quad-Core Intel® Xeon® processor-based servers as our target virtualization host platforms. Previous Intel IT and industry testing had shown that these dual-socket quad-core servers, with a total of eight cores, deliver good performance with low TCO.

Potential Memory Configurations

We assessed memory configuration options for our target virtualization servers, as shown in Table 1. The most common dual-socket servers have eight DIMM slots, though some have 12 or 16 slots. At the time we conducted our analysis, four sizes of memory module were widely available for these slots: 512 MB, 1 GB, 2 GB, and 4 GB. The cost per GB was similar for the 512 MB, 1 GB, and 2 GB modules. However, for the 4 GB modules, the cost per GB was several times higher.

We determined that 2 GB DIMMs represented the best trade-off of power, cost, and capacity. Using 2 GB DIMMs, we could configure servers with 16 GB, 24 GB, or 32 GB of memory depending on the number of slots, with relatively low power consumption and at much lower cost than using 4 GB DIMMs. We expect to periodically reassess memory options.

Hypervisor Memory Assumptions

When estimating total server memory requirements, we need to consider the memory required by the VM hypervisor in addition to any other overhead requirements. On the other hand, there will be savings due to memory sharing across VMs because we expect to run multiple copies of the same OS and applications in different VMs on the same server. For simplicity, we assumed that the additional memory required for the hypervisor would be offset by the hypervisor's ability to share common OS and application pages across VMs.

Achievable Consolidation Levels

For the purpose of this memory sizing exercise, we calculated the consolidation levels that we could achieve on our hosting servers

Table 1. Typical Memory Options for Dual-Socket Servers

Dual In-line Memory Modules (DIMMs):				
Available Sizes	512 MB	1 GB	2 GB	4 GB
Price per GB	~$300	~$250	~$275	~$1000
DIMM Slots:				
8	4 GB	8 GB	16 GB	32 GB
12	6 GB	12 GB	24 GB	48 GB
16	8 GB	16 GB	32 GB	64 GB

Note: Prices shown are accurate as of April 2007.

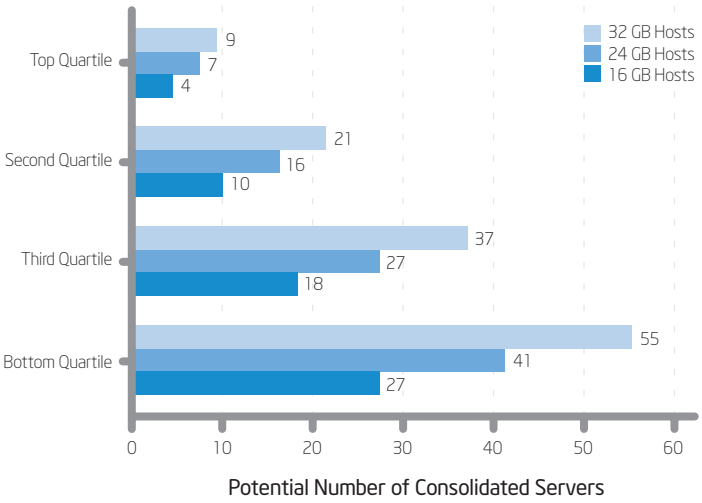


Figure 3. Potential consolidation levels by quartile, based solely on memory constraints. We calculated average consolidation ratios for each quartile by maximum memory consumption.

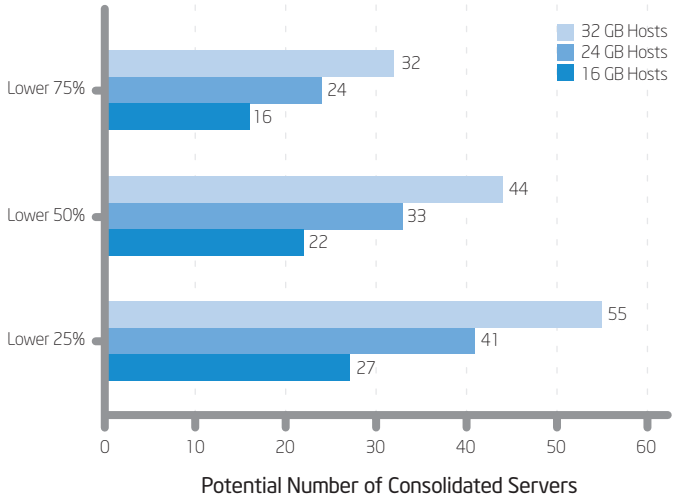


Figure 4. Potential consolidation levels for less-utilized pools of servers, based solely on memory constraints. We calculated average consolidation ratios for the lowest 75 percent, lowest 50 percent, and lowest 25 percent of servers by maximum memory consumption.

based solely on memory constraints. We recognize that in production environments, other factors independent of memory considerations can impose constraints on the consolidation levels. These factors include performance and risk management considerations.

Because we assumed no net impact of the hypervisor on overall memory requirements, the calculations were simple. We based them on:

- The average of maximum memory consumption within each of the four quartiles in our sample of more than 3,000 servers.
- The three candidate target memory configurations: 16 GB, 24 GB, and 32 GB.

We divided each of our candidate memory configurations by each of the quartile maximum memory utilizations to obtain a range of consolidation scenarios, as shown in Figure 3. For example, to determine how many of the servers in the third quartile we could consolidate into one 16 GB virtualization host, we divided 16 GB by 0.86 MB, the average maximum memory consumption for servers in this quartile. This resulted in a ratio of approximately 18 to 1.

We then calculated potential consolidation levels among larger pools of the servers in the lower three quartiles, as shown in Figure 4. We calculated the average potential consolidation level among the lowest 75 percent of servers by maximum memory consumption, and among the lowest 50 percent. Figure 4 also shows the data for the lowest quartile for comparison.

These less-utilized servers are obvious consolidation candidates because of the potential efficiencies. We estimated the following potential consolidation levels, based on memory requirements only:

- Approximately 25 to 1 for the lowest 25 percent of our server sample
- Approximately 20 to 1 for the lowest 50 percent
- Approximately 15 to 1 for the lowest 75 percent

Consolidation Strategy

Based on our analysis, we believe we can achieve very good consolidation ratios using cost-effective dual-socket virtualization hosts configured with Quad-Core Intel Xeon processors and 16 GB of memory.

In the near term, we are aiming to consolidate servers similar to those in the lower two quartiles of our 3,000-server sample. When we virtualize these workloads, we anticipate that each VM will require between 0.6 and 1.5 GB of memory; many will require 1 GB or less.

By specifying servers with 16 GB memory, we expect to be able to achieve high consolidation levels of up 15-20 to 1.

We are in the early stages of implementing our consolidation strategy. Preliminary results have validated our assessment; all indications are that we will be able to achieve the anticipated consolidation ratios.

Over time, we expect to fine-tune our consolidation strategy. For example, we may choose to consolidate several server workloads of the same type, such as databases, into a single larger workload and then virtualize the consolidated workload. This approach would reduce the number of VMs in our environment, but increase the memory required per VM. We expect that this would not materially change overall virtualization host memory requirements.

Conclusion

Intel IT has standardized on 16 GB of memory for dual-socket virtualization host servers, based on extensive analysis of maximum physical memory consumption on more than 3,000 servers running non-virtualized workloads. We found that approximately half of these servers consume 1 GB of memory or less. For workloads of this size, our experience to date indicates that we can achieve consolidation ratios of up to 15-20 to 1 using low-cost dual-socket virtualization hosts based on quad-core processors.

Our analysis indicates the importance of carefully monitoring memory consumption on the systems to be virtualized in order to size virtualization host memory appropriately. Our monitoring showed that for many servers, actual memory use is modest. By sizing virtualization host memory based on measurements of actual memory consumption, we believe we can maximize utilization of virtualization host server memory and avoid paying for unused memory and associated data center power and cooling.

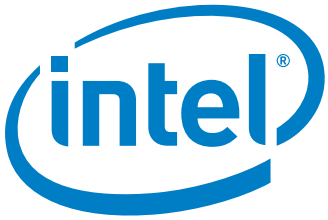
Authors

Sudip Chahal is a compute and storage architect with Intel Information Technology.

Todd Glasgow is a service manager with Intel Information Technology.

Acronyms

DIMM	dual in-line memory module	TCO	total cost of ownership
GB	gigabyte	VM	virtual machine



www.intel.com/IT

Printed in USA
0707/SEP/RDA/PDF

♻ Please Recycle
ITAI Number: 07-1206w