Capacity Planner Technical FAQ

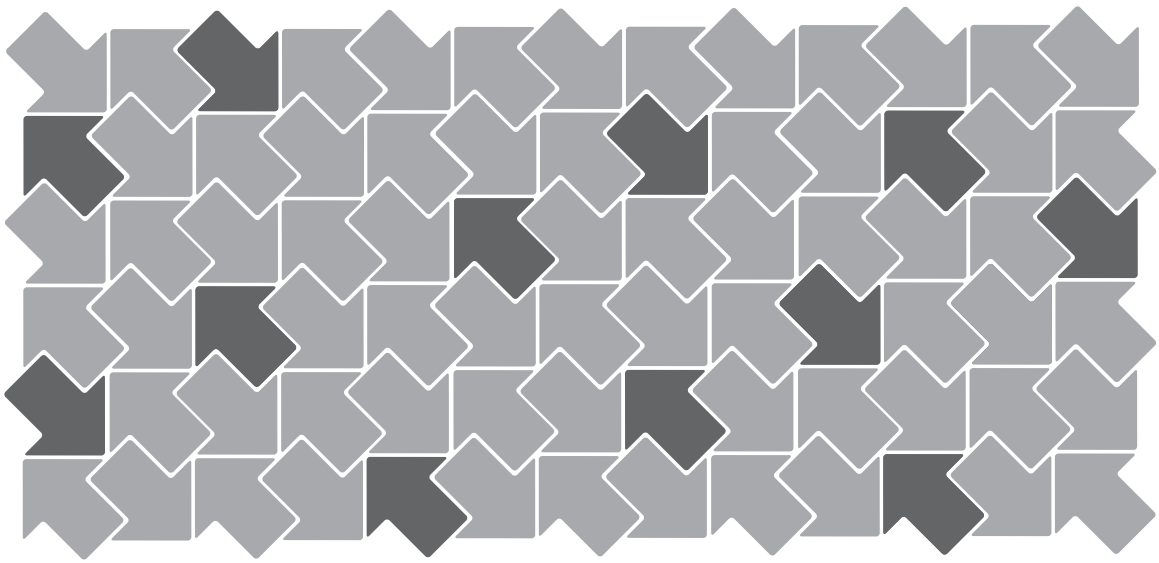

Table of Contents

Requirements

What is the necessary inventory information that you need to conduct an assessment?
For server consolidation and other capacity planning activities, project teams need to know detailed hardware information around four core hardware components: processors, memory, disk, and network interface cards. Detailed intelligence for applications, services, and shares is equally valuable.

What administrative rights are needed to access servers?
On a Windows environment, users enter either global, domain, or individual administrative rights to access remote servers into the Manager. UNIX and Linux systems require root access to access the appropriate data from the software.

What ports need to be open?
We need the following ports open: 135, 137 through 139, and 445. If there are servers behind firewalls that can't be opened or will be challenging to open, it is often easier to install a Collector inside the firewall.

Why is file system cache not considered with VMware recommendations?
The file system cache is not a parameter to be concerned with, but we roll it up in order to give the client an idea of the impact on memory that file system cache access will have. Each virtual windows machine will still operate the same way in regards to file system cache. They will still perform I/O, perhaps paging, and will still store the information in their respective file system cache. This definitely has an impact on the overall system. Certainly, the amount of memory consumed by the physical server is directly related to the amount of file system cache and other memory consumed by individual virtual nodes. The physical server is also affected by how much of the memory has to be populated due to paging. The physical server still reads the date from files and caches it into its system cache. Because Linux is an open system and you never know for certain how it is managing everything, this is not intended to be an estimate of actual system cache usage on the box.

Installation

What are the technical skills needed to install and analyze data?
People installing Capacity Planner should have basic System Administration skills for Microsoft, Linux, or UNIX.

Do I need to install agents?
With Capacity Planner, you do not need to install any agents on the target servers. Capacity Planner is agent-less software that collects across server infrastructure, leveraging existing data sources already on the systems. As a result, Data Collectors only need to be installed to pull the data from the target systems.

Do I need to purchase other software or hardware to run this tool?
Unlike other competing products, Capacity Planner was designed to leverage basic hardware and software requirements. Capacity Planner requires a Windows 2000 or newer operating system on a laptop, desktop, or server with at least 1GHz CPU, 512 MB of RAM, and 250 MB of free disk space.

What is the appropriate operating system for the Collector and Manager?
Capacity Planner requires Windows 2000 or 2003 with an ASCII language operating system. VMware mandates US English Windows 200x operating system as best practice.

Discovery and Collection

What is the discovery technique?
Capacity Planner uses Active Directory, IP Scanning, DNS queries, and NETBIOS to discover the systems across a network.

What does a discovery collect?
Discovery using Capacity Planner is simply a system count of the environment. There is relatively little information beyond a physical count that is provided. The discovery enumerates the list using the Active Directory, IP Scanning, DNS queries, and NETBIOS options. It does not verify that the machine is online or accessible.

What operating systems does it discover?
Capacity Planner collects from Windows NT 3.51, NT 4.0, 2000, and 2003; Red Hat Linux 8 and 9 and Enterprise Linux (ES/AS/WS) 3 and 4; SUSE Linux 8, 9, 10, and Enterprise Server 9; HP-UX 10.x, 11.0, 11.11, 11.22 (PA-RISC) and 11.23 (Itanium); and Solaris 7, 8, 9, and 10 (Sparc) and 9 and 10 (x86) Operating Environments.

How are discovery and inventory different?
Discovery is a free count of the number of systems in the customer's environment. Inventory is a paid service that collects all the hardware, software, and services inventory data.

What inventory data is collected?
Capacity Planner collects hardware, software, and services inventory data.

What performance data is collected?
Capacity Planner, by default, collects memory, disk, network, and processor data as well as particular application information if it detects that application on a particular system. Multiple systems can be collected at the same time as a configurable parameter.

Can Capacity Planner collect on multiple hardware vendors?
VMware Capacity Planner supports all the hardware vendors.

Will Capacity Planner collect remote servers?
Yes, Capacity Planner is designed as an agent-less tool to pull data from remote servers.

Why do I have server entries with zeros?
Either those servers are not included in your target scope or you have been unable to collect inventory information using your current inventory collection settings.

What performance metrics are collected by Capacity Planner?
Capacity Planner collects from targeted servers 300+ core performance statistics and additional relevant statistics for specific applications. This low-overhead query collects performance metrics from the four main data groups of processor, memory, disk drive, and network utilization. For memory, for example, Capacity Planner collects not only paging data or what is available in bytes but also specific cache information that affects the overall project decision strategy. This data is then correlated with the previously collected inventory data.

Can Capacity Planner collect Unicode language systems?
Yes, Capacity Planner can collect Unicode language systems; however, the Collector and Manager cannot be installed on a Unicode system.

What impact does collection have on my network?
Capacity Planner typically only uses 20,000 bytes during data collection across a network.

What impact does collection have on my servers?
The impact on the target server varies, depending on what is collected. Windows systems traditionally are affected less than 1% utilization. UNIX systems might peak at 5 to 10% utilization during inventory collection cycles.

How long does Capacity Planner take to collect inventory?
This varies depending on a number of factors, such as the Collector system, location of target systems, and network speed. Traditionally, Capacity Planner collects inventory on one system every 20 seconds. Multiple systems can be collected at the same time as a configurable parameter.

How long do we need to collect performance data?
VMware recommends a minimum of three weeks of data collection to start to compile a server profile with at least 1,000 performance samples. In addition, performance collection over three to four weeks provides server trending on a daily, weekly, and monthly basis that could be critical to understanding when servers are peaking. In addition, the longer you collect data, the more valuable it is.

Does Capacity Planner collect application statistics?
Yes, it collects limited application statistics that are written to Microsoft Performance Monitor (PerfMon).

How does Capacity Planner decide which application statistics to collect?
Once Capacity Planner collects inventory information, it can determine which applications and services are running on a systems and request additional counters for those applications or services to be added to the performance collection cycle.

How does Capacity Planner differentiate between a single and dual core processor?
Currently, Capacity Planner's new server recommendations can be adjusted to account for single and dual core processors.

How do you compile an industry averages?
The term industry average could be misleading. We use Industry Average to indicate a combination of all customers that comprise our information warehouse.

Utilization and Performance Counters

How are the utilization figures determined?
Utilization and performance counters are determined by collecting multiple samples each hour of each day for each week. The statistics for each hour over a week are correlated together to determine the average, hourly, prime time, and non-prime time utilization for each hour. Average utilization is typically the average utilization during the prime time hours of 7 AM to 6 PM, for example, or for the entire 24-hour period. Capacity Planner also maintains weekly summary statistics that track maximum observed, minimum observed, average, hourly, prime time, non-prime time, and weekend loads. Capacity Planner also maintains a summary for the most recent four weeks of performance statistics on these same criteria. The summary is used to determine peak load for consolidation recommendations. The peak load is determined by evaluating each metric over the most recent four weeks of collection and locating the hour of the day that has the highest average value. It is not the maximum observed value as any statistical analysis eliminates the high and low values from consideration.

An average value is the prime time average. Peak load or "busy hour" is the average value from the hour of the day that reports the highest measurements over a four-week period. It is not the maximum observed value. Capacity Planner tracks the average, prime time, peak, maximum observed, minimum observed, weekend average, non-prime time average, and an average for every hour of the day. Many of these statistics are only viewable through dynamic reports. Clients are often shocked by the low values being reported, and we encourage clients to validate the data against a simple performance monitor collection. Ensure that you run the sample over a long, representative time period. When analyzing capacity requirements, it is never sufficient to collect samples for one hour or one day. It is not appropriate to focus on some hours and not others. It is imperative to identify how loads vary from hour to hour, week to week, and when the peak load occurs. It is not important that a Windows Server hit 100% utilization during one moment in time or that it occasionally hits 30%. Trends over time are required. If you combine four systems together that occasionally hit 30%, it is highly unlikely that the combination of these four systems would yield 100% utilization. The variable nature of the loads would more evenly distribute utilization to have a more consistent value with less variation throughout the day.

What is the difference between peak hour performances versus average utilization?
Peak hour is the one hour in the twenty-four hour period that has the highest average utilization. Average utilization is typically the average utilization during the prime time hours of 7 AM to 7 PM, for

example, or for the entire twenty-four hour period. The difference between these two metrics is significant.

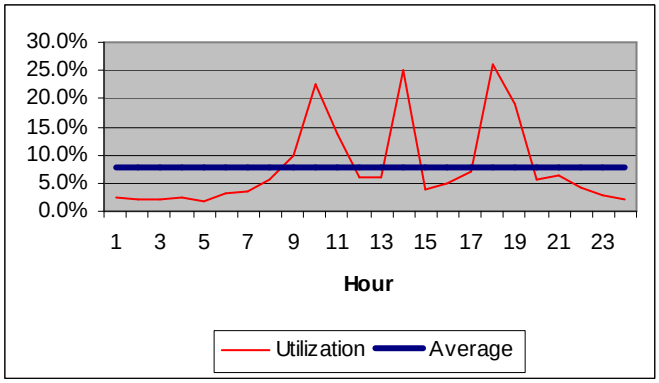

Figure 2: Peak vs. Average Utilization

Assume that servers A, B, C, and D in Figure 1 were Microsoft Exchange servers that were to be consolidated based on average utilization. If this were the case and all peaked to 40 percent utilization in the morning and at the lunch hour, the server would run out of capacity at those critical times.

As another example, assume that the servers in Figure 1 were Citrix servers. On average, most Citrix machines that Capacity Planner has monitored run low CPU utilization, low memory utilization, and extremely low disk utilization. However, the largest variable is network utilization. During peak loads, usually in the morning during sign-on, the network traffic may go from an average of 100,000 bytes per second to millions of bytes per second. If one used average utilization or even prime-time utilization metrics for consolidating a Citrix server, problems would likely arise.

Identifying the peak-hour average provides capacity planners with an upper threshold for consolidating servers. How important is the difference between prime time averages and the peak hour average? Analysis of data in the information warehouse reveals that the peak-hour average is two times higher than the prime-time average for 68 percent of all servers. For 32 percent of all servers, the peak-hour average is five times higher than the prime-time average.

The standard installation samples each server once an hour. Are all servers sampled simultaneously or divided across the time period? Can this be configured?
Servers are sampled in groups of ten servers. They are not all done simultaneously; instead, they are sampled in order until all are sampled. Performance collection is a configurable option within the Manager.

How does Capacity Planner calculate peak hour and other calculations?
Capacity Planner collects data every hour and calculates peak-hour utilization for each one-hour increment in the 24 hour day. After several weeks, it identifies utilization for the busiest hour in the week. Capacity Planner also maintains weekly summary statistics that track maximum observed, minimum observed, average, hourly, prime time, non prime time, and weekend loads. Capacity Planner also maintains a summary for the most recent four weeks of performance statistics on these same criteria. The summary is what is used to determine peak load for consolidation recommendations. Peak load is

determined by finding the hour of the day with the highest sustained load. It is not a measure of maximum observed values.

How is prime time determined?
Capacity Planner uses 7 AM to 6 PM as the standard time period because these are the main hours of operation plus an hour before and after. This range considers the primary hours that part or all of your employees are working. Prime time can be modified easily.

For consolidation, the peak load is considered. Peak load is not the maximum observed value because any statistical analysis eliminates the high and low values from consideration. The peak load is determined by evaluating each metric over the most recent four weeks of collection and locating the hour of the day that has the highest average value. VMware has done quite a bit of research on the difference between 24x7, prime time, and peak loads.

The prime time average is usually slightly higher than the 24x7 average but is occasionally lower due to backups or nightly processes which may skew results. The peak load is three times higher than the prime time average approximately 65% of the time. Peak load is five times higher than the prime time average approximately 35% of the time. VMware is most interested in making conservative recommendations that will work, so we have adopted the use of peak loads so that we have he highest degree of confidence that the combination of loads we recommend will operate effectively.

How do normalized average utilization and average utilization differ?
CPU utilization statistics can be viewed as either rolled-up average utilization or normalized utilization. Normalizing CPU utilization is determined by the megahertz rate times the actual utilization rating while rolled up average utilization is the sum of the individual server averages divided by the number of servers. Normalizing the figure does not skew the utilization when an environment has old processors running on older, slower machines.

Hyperthreading

How does Capacity Planner handle x86 processors that have the hyperthreading feature enabled?
Capacity Planner detects hyperthreading based on the model for the server sent to Capacity Planner. If the chassis model is sent, Capacity Planner checks to see if the hyperthread flag has been turned on for that model in the information warehouse. If the hyperthread flag is on, Capacity Planner ignores half of their CPUs. If they send four, Capacity Planner only enters two into the system, assuming that they did not go through the process to turn hyperthreading off.

Capacity Planner flags a chassis as being hyperthreaded when it is either manually entered or detected by comparing the maximum number of CPUs allowed in the chassis to the number of CPUs being reported. If a chassis can only have four processors, but inventory is sent showing eight processors for that model, Capacity Planner turns on the hyperthread flag for that model.

Application Conflicts

How does Capacity Planner take application conflicts into account?
Application stacking is the only way to create a DLL conflict scenario, and Capacity Planner does not begin with that type of recommendation. Capacity Planner first addresses account issues that are far

more critical than DLL conflicts. Ownership of servers, location of servers, and the environment of the servers (test, development, production) typically present more immediate conflicts. Further, VMware recommends combining like applications. VMware allows you to stack and separate DLLs. This solution also allows you to optimize the utilization of your server resources. If you do not want to virtualize, you still have the option of using Windows 2003 to separate out DLLs and avoid the conflicts that arise in Windows NT primarily and Windows 2000 to a lesser extent. However, this is a more time-intensive process than virtualizing your environment.

Optimizing your environment might also include separating applications from the database so that you can create and manage a larger scale database system on fewer physical machines. You also have to evaluate memory utilization in detail. Just because it appears that you are using all of your memory in Exchange or SQL Servers does not mean that you need to be using that much memory. Without optimizing these environments, you will never truly know what consolidation opportunities exist.

Proposed New Hardware

Does proposed new hardware have VMware automatically included on it?
Proposed new hardware does not automatically have VMware included on it.

How can I help make VMware recommendations?
When leveraging Capacity Planner to run consolidation or virtualization scenarios, users can create a new server with specific hardware specifications and either VMware ESX Server, VMware Server, or standard Microsoft, Linux, or UNIX operating systems. These systems comprise a group of servers to run consolidation scenarios against. The scenarios will allow users to pick their various future platforms and operating system combinations, and then run scenarios to determine how many proposed systems are needed.

How does Capacity Planner determine the native MHz for proposed new hardware?
Capacity Planner reduces the amount of native capacity when multi-processors are involved. The rule of thumb in the industry is to reduce the amount of native capacity available to 80% of the original rating for the second processor. Subsequent processors are multiplied by .8 raised to the power of the slot number of the CPU. So the third CPU contributes .8 raised to the second power multiplied times the rated speed (the third CPU is in slot 2 because slot numbers begin with 0). If this is a 64-bit processor, the percentage is raised to 90% due to improvements in the technology. Therefore, the variables are the number of processors, the rated speed, and the slot number.

Platforms

Is VMware comparing like with like when comparing Intel and AMD based platforms? If people believe the AMD CPU is more powerful, is their a way to apply a correction factor for AMD platforms?
AMD Opteron is not more powerful. It is a different technology that provides better throughput under certain circumstances. In an Intel platform, all the processors make memory requests through a common physical interface. AMD Opteron has a separate physical interface for each processor. The bottleneck to memory has been removed, and the speed to the memory is faster as well. This is the big difference. It is also a 64-bit platform, but that alone does not offer an advantage over an Intel 64-bit platform. There are

two things to consider. First, are you comparing a 32-bit to a 64-bit platform? Second, are you comparing Intel to AMD Opteron?

The only way you see significant improvements in throughput is in a memory intensive application. Most applications are not total random memory access intensive, only about 2%. If a system is paging due to limitations on the file system cache, 64 bit may solve the problem by allowing a much larger file system cache. If a system is paging because every reference to memory is for new data that has not been read into memory before, this is a disk access and memory speed issue, primarily disk access. If you have a system that is constantly pulling data from memory but not requiring disk access, this favors the AMD Opteron. This typically means large database servers. When we combine a number of systems in a virtualized environment, this creates a memory intensive situation on the ESX Server overall and favors AMD Opteron. There is no way to apply an algorithm to the statistics that would show lower processor utilization. The processor will not run at a lower utilization, but the throughput and response times will be better. Because it cannot be represented statistically, the technology and the ramifications are explained to clients. Essentially, the statistics point out that running an AMD Opteron at 65% utilization and Intel at 50% utilization provides greater throughput with AMD Opteron.

Groups

I have a small network environment. Do I need server groups?
Not necessarily. You can create an All Servers Group and move all servers into that group. However, you might want to consider disaster recovery or your use of other applications.

How are the groups created?
Capacity Planner has four grouping categories: Departments, Environments, Locations, and Applications. These groups are created by the partner or customer. Groups can be created or imported on both the Manager and Dashboard.

For all consolidation projects, servers and applications with similar attributes need to be assessed to determine what approach to consolidation should be used. Groups of servers with similar attributes are defined and registered within the Manager or Dashboard. Typically, Capacity Planner scenarios are used to select source systems for consolidation analysis because these provide the best overall control.

Capacity Planner can be configured with a unique list of server groups that can represent any grouping criteria you decide on. The servers in these groups are themselves cross-selectable using the built-in Capacity Planner predefined basic subgroups.

These subgroups are typical subdivisions that you normally need, but their contents are still user defined typically as follows:

- Environment – Production, development, test, and so on
- Department – Sales, IT support, marketing, and so on
- Domain – Usually added automatically by data collection
- Function – Major application or server role, for example, IIS, DC, EXCH
- Location – Names of your customer's sites

Transferring Data

Are there other ways to transfer data besides HTTPS?
HTTPS is the default transfer mechanism, but you have the ability to transfer data through FTP, email, or burning the data to a CD.

Can the data be exported?
Users can export data from either the Manager or Dashboard to a CSV file.

Can I import a list of servers?
Yes, you can import a list of servers as a comma-delimited file instead of discovering all the servers in the environment. At the same time as importing host names, it is extremely useful if you include other information, such as location, function, environment, and department, to simplify the later grouping of servers according to use.

What is the file format for the server import file?
The import file should be in the CSV format. The naming convention is "DomainName,SystemName".

Reports

Can you create custom reports?
Dashboard includes Dynamic Reports that permits users to create customizable reports and export them to CSV files for further analysis.

Anomalies

How do you know if an anomaly is based on hardware or software?
Anomalies are determined by the particular class and metric reported from PerfMon and UNIX. These are based on either a hardware or software performance metric.

How are alerts and anomalies different?
Anomalies indicate performance that is significantly different from the industry performance averages of like servers provided by the information warehouse. Capacity Planner uses three standard deviations to determine what significantly different performance is. The performance difference indicated by the anomaly can be either good or bad. The real-world, real-time performance comparison that anomalies provide can be leveraged to configure similar machines if good or to identify the problem and proactively address the problem before you have performance issues if bad. Alerts indicate performance that exceeds vendor-provided thresholds.