Managing Capacity Using VMware® vCenter™ CapacityIQ

TECHNICAL WHITE PAPER
# Table of Contents

Capacity Management Overview .................................................. 3
CapacityIQ Information Collection ............................................... 3
CapacityIQ Performance Metrics .................................................. 4
CapacityIQ Methodology ............................................................. 4
  - Total Capacity ................................................................. 5
  - Unusable Capacity ............................................................ 5
  - Usable Capacity .............................................................. 5
  - Used Capacity ................................................................. 5
  - Remaining Capacity .......................................................... 6
  - Total Virtual Machine Capacity .......................................... 8
  - Deployed Virtual Machine Capacity ..................................... 8
  - Remaining Virtual Machine Capacity .................................... 8
  - Example of Calculating Average Virtual Machine Capacity .......... 9
CapacityIQ Trending and Forecasting ........................................... 12
Data Resolution and Rollups ....................................................... 14
Data Aggregation and Retention .................................................. 17
  - Weekly Aggregation ......................................................... 17
  - Monthly Aggregation ........................................................ 17
  - Quarterly and Yearly Aggregation ........................................ 18
Capacity Efficiency ................................................................. 19
  - Powered-Off Virtual Machines ............................................ 19
  - Idle Virtual Machines ....................................................... 21
  - Oversized Virtual Machines ............................................... 22
  - Undersized Virtual Machines ............................................. 25
Summary .................................................................................... 29
Managing Capacity Using VMware vCenter CapacityIQ

Capacity Management Overview

Capacity management is a well-established practice in enterprise datacenters. Effective capacity management strives to accomplish two objectives:

• Efficiency: Make efficient use of existing capacity and minimize any waste of computing resources
• Availability: Leave enough capacity headroom to accommodate any spikes in resource demands

These two capacity management objectives often conflict with each other. Rightsizing resource capacity to achieve high efficiency usually comes at the cost of sacrificing capacity availability. Similarly, provisioning an additional capacity buffer tends to reduce capacity efficiency. Consequently, capacity management is a balancing act and tends to become more of an art than a science.

Capacity management in a virtual environment is more challenging for two reasons. First, the physical resources in a virtual environment are shared across multiple virtual machines. Second, VMware® vSphere™ 4 (“vSphere”) has its own advanced features, such as VMware Distributed Resource Scheduler (DRS) and memory optimizations, which manage resource allocations continuously and dynamically. Homegrown solutions based on Excel and scripts tend to ignore these nuances in virtual deployments, and often overestimate capacity needs. VMware vCenter™ CapacityIQ is the purpose-built tool from VMware that takes into account these nuances and helps you address capacity-management challenges in virtual deployments. VMware vCenter CapacityIQ enables users to analyze, forecast, and plan the capacity needs of their virtual datacenter environment. CapacityIQ provides the following key features:

• Dashboard, views and reports for visibility into past, present, and future capacity states
• Views and reports detailing excess capacity from idle, oversized, or powered-off virtual machines for optimizing capacity in existing deployment
• Interactive what-if modeling scenarios to simulate capacity and demand side business needs
• Prediction of the exact timing of potential capacity shortfalls to aid purchasing and provisioning decision making
• Configurable threshold settings that can be customized to the specific IT policies within a datacenter

This document explains the information provided in CapacityIQ Dashboard, Views and Reports. It provides background into the key CapacityIQ concepts and methodology used to provide the details shown in various CapacityIQ screens. This information should help you understand and interpret the capacity data specific to your environment. The document also covers the different configurable settings and how they impact the information shown in CapacityIQ Dashboard, Views and Reports.

CapacityIQ Information Collection

CapacityIQ is deployed per vCenter instance and does not require changes to the vCenter Server setup, such as modifying the logging levels. CapacityIQ collects inventory and performance data from vCenter Server using standard VMware vSphere APIs using port 443 (default) on the vCenter Server. CapacityIQ data collector:

• Collects performance statistics about managed objects in the virtual infrastructure
• Stores the statistics in the CapacityIQ database and builds a performance history for each object
• Collects and stores inventory history for each object, such as a virtual machine move from one host to another

CapacityIQ starts collecting data from vCenter shortly after installation and continues data collection unless you stop CapacityIQ. CapacityIQ collects performance statistics from vCenter every 15 minutes. Depending on vCenter Statistics settings, this may result in 3 (default), 5, or 15 data points for each performance metric that CapacityIQ collects. CapacityIQ averages these data points over an hour, and maintains hourly information in its database for the entire lifetime. In order to provide this hourly granularity, CapacityIQ needs to maintain its own database and cannot rely on the vCenter database. vCenter Server reduces its performance data granularity...
from daily to weekly to monthly as it rolls up performance data. vCenter Server excels at collecting real-time utilization for tuning and performance monitoring, but it does not track capacity changes over time.

**CapacityIQ Performance Metrics**

In a vSphere deployment, determining exact capacity requirements for a particular virtual machine can be challenging. A typical vSphere environment is very dynamic, with multiple virtual machines sharing the same physical resources. The virtual machines use up resources based on configured values as well as those based on resource allocation policies determined by shares, reservations and limits. These resources can also be changed dynamically without the need to reboot virtual machines. vSphere allows you to overcommit resources so that the sum of configured values across all virtual machines on a vSphere host can be more than the physical resources available. The actual resource utilization for a virtual machine should include any overhead due to virtualization as well. When multiple virtual machines have to contend for the same physical resources, vSphere uses advanced optimization techniques to share, throttle, or reclaim resources using various techniques. vSphere maintains a variety of performance counters to track this information. In order to accurately reflect the capacity needs of a virtual machine, CapacityIQ goes beyond simply tracking CPU or memory used. CapacityIQ defines special resource utilization metrics for CPU and memory:

- **CPU Demand**: CPU demand takes into account a virtual machine’s CPU usage and its CPU ready and CPU wait times to derive the CPU capacity demand for the virtual machine.
- **Memory Demand**: Memory demand considers memory consumed by a virtual machine, which reflects any ballooning, page sharing and memory swapping for the virtual machine, as well as memory overheads due to virtualization, to derive the memory capacity demand for the virtual machine.

Furthermore, CapacityIQ converts the CPU usage (percent) values into corresponding MHz used based on the processor speeds in order to normalize for different processors.

CapacityIQ captures this information for various objects in the vCenter inventory and aggregates them to datacenter or other appropriate levels as needed. These resource utilization numbers are displayed in the Dashboard under the Resources section.

**CapacityIQ Methodology**

CapacityIQ analyzes capacity at various levels of abstraction. CapacityIQ treats capacity for physical resources and virtual machine capacity in different ways. For physical capacity, CapacityIQ captures utilization from the five core physical resources:

- CPU — cycles of processor
- Memory — bytes of RAM
- Disk Space — amount of space on disk (shared and local)
- Disk I/O bandwidth — available throughput of storage I/O channels
- Network I/O bandwidth — available throughput of network links

In the previous section, we already discussed how CapacityIQ uses these metrics to calculate derived metrics that accurately reflect capacity demand in the vSphere environment. In addition, CapacityIQ also takes into account capacity reserved for failover, virtualization overhead or any capacity headroom you may need. To understand this further, let’s look at how CapacityIQ works with physical capacity (see Figure 1).
Managing Capacity Using VMware vCenter CapacityIQ

Total Capacity
Total capacity is the total amount of physical resources provisioned in the hosts and clusters that make up the virtual datacenter.

Unusable Capacity
Unusable capacity is capacity that will not be available for use by virtual machines due to one of the following reasons:

- HA Reservation: This is the capacity needed to meet VMware High Availability (VMware HA) failover commitments when host failures occur. CapacityIQ removes the host with the largest amount of memory to account for one host failover capacity.
- ESX Overhead: This is the small amount of capacity needed to run VMware ESX® 4 and its services.
- Capacity Buffer: This is the capacity that is set aside by the user as a buffer against sporadic spikes in demand. For example, it is a typical practice in most IT environments to leave 15–20 percent headroom on CPU and memory as well as storage (for snapshot). CapacityIQ allows users to configure this headroom. CapacityIQ allows you to customize this value using the Capacity Buffer Limits global settings to suit your datacenter policies.

Usable Capacity
Usable capacity is capacity that is available for use by virtual machines:

\[ \text{Usable Capacity} = \text{Total Capacity} - \text{Unusable Capacity} \]

Used Capacity
This is the aggregate capacity used by all the virtual machines running in the virtual datacenter.
Remaining Capacity

This is the total capacity available for use by existing or new virtual machines.

The CapacityIQ Dashboard provides visibility into the current capacity utilization in your datacenter. CapacityIQ analyzes the performance statistics for objects managed in the VMware infrastructure, such as ESX hosts, clusters and virtual machines, to help you understand the current state of capacity in your virtual environment.

CapacityIQ displays information about physical capacity for your environment in the dashboard view. Figure 2 shows the total physical capacity available. Figure 3 shows the total usable physical capacity after the unusable capacity is taken out.

Figure 2. Total Physical Capacity.

The used capacity can be obtained in two ways:

- Total Host Resources Used perspective
- Total Usable Host Resources Used perspective

Figure 4 shows the total physical capacity used. Similarly, Figure 5 shows the used portion of the total usable physical capacity.
The remaining capacity can be obtained from the Total Usable Host Resources Remaining perspective as shown in Figure 6.

**NOTE:** CapacityIQ also has the remaining perspective for Total Host Resources (Figure 7) — remember this does not take into account the reserved capacity.
Managing Capacity Using VMware vCenter CapacityIQ

Figure 7. Remaining Total Physical Capacity.

CapacityIQ translates physical capacity into an abstract measure called virtual machine capacity. Translating physical capacity to units of virtual machine capacity provides a practical way of tracking and planning capacity in a virtual datacenter.

**Total Virtual Machine Capacity**

Total virtual machine capacity is the sum of deployed and remaining virtual machine capacity.

\[
\text{Total virtual machine count capacity} = \text{deployed virtual machine capacity} + \text{remaining virtual machine capacity}
\]

**Deployed Virtual Machine Capacity**

Deployed virtual machine capacity is the number of virtual machines already deployed in the datacenter or cluster. Deployed virtual machines include virtual machines regardless of whether they have been powered on or off.

**Remaining Virtual Machine Capacity**

Remaining virtual machine capacity is a measure of the number of new virtual machines that can be deployed.

First we present the Dashboard screens that provide information about virtual machine capacity. The Deployed perspective in the Objects section (Figure 8) shows the average number of deployed virtual machines, hosts and clusters in your virtual environment.

Figure 8. Deployed Virtual Machine Capacity.
The Remaining perspective in Objects section (Figure 9) shows the remaining capacity in terms of number of deployed virtual machines, hosts and clusters.

Figure 9. Remaining Virtual Machine Capacity.

Remaining virtual machine capacity value is based on assumptions that CapacityIQ makes about capacity demands of new or future virtual machines you expect to deploy. To estimate the number of remaining virtual machines, CapacityIQ applies the notion of an average virtual machine. CapacityIQ assumes that any new virtual machine you add to your environment will represent this average virtual machine, and hence is a better reflection of your capacity utilization requirements. This approach will result in realistic estimates of remaining virtual machines as long as the population selected resembles the population of new virtual machines expected to be deployed.

CapacityIQ uses the concept of a virtual machine population as the basis for calculating the average virtual machine. A virtual machine population is a set of virtual machines with known or given capacity demand patterns. This set of virtual machines is considered representative of the new virtual machines you expect to deploy. Any new virtual machines you plan to deploy will likely resemble this current population of virtual machines. Typically this population includes a mix of small, medium, and large virtual machines. CapacityIQ uses weighted averages of resource demands for each of the machines in this population to compute the average virtual machine. Note that the average virtual machine is calculated based on deployed virtual machines rather than powered-on virtual machines.

Example of Calculating Average Virtual Machine Capacity

Consider the following virtual machine population composed of 30 virtual machines.

- The population includes virtual machines of three types, which are small, medium, and large.
- 9 of the virtual machines are small, which is 30 percent of the population.
- 15 of the virtual machines are medium, which is 50 percent of the population.
- 6 of the virtual machines are large, which is 20 percent of the population.

Using the virtual machine population described in this example, here is how CapacityIQ calculates the Average Virtual Machine CPU Demand in this cluster.

\[
\text{Average VM CPU Demand} = 0.3 \times (\text{CPU Demand Small VM}) + 0.5 \times (\text{CPU Demand Medium VM}) + 0.2 \times (\text{CPU Demand Large VM})
\]

Here is how CapacityIQ calculates the Average Virtual Machine Memory Demand in this cluster.

\[
\text{Average VM Memory Demand} = 0.3 \times (\text{Memory Demand Small VM}) + 0.5 \times (\text{Memory Demand Medium VM}) + 0.2 \times (\text{Memory Demand Large VM})
\]
The result of Average VM CPU Demand and Average VM Memory Demand represent the average virtual machine profile for this cluster.

![Virtual Machine Population Diagram](image)

**Average Virtual Machine**

\[
\text{CPU Demand}_{\text{avg. VM}} = 0.3 \times \text{CPU Demand}_{\text{Small VM}} + 0.5 \times \text{CPU Demand}_{\text{Medium VM}} + 0.2 \times \text{CPU Demand}_{\text{Large VM}}
\]

\[
\text{Memory Demand}_{\text{avg. VM}} = 0.3 \times \text{Memory Demand}_{\text{Small VM}} + 0.5 \times \text{Memory Demand}_{\text{Medium VM}} + 0.2 \times \text{Memory Demand}_{\text{Large VM}}
\]

Figure 10. Average Virtual Machine Calculation Example.

The Average Deployed Virtual Machine Profile — Summary view shows the average virtual machine profile for your environment in the datacenter.

![Dashboard View](image)

**Average Deployed Virtual Machine Profile - Summary**

- CPU Demand: 1.37 GHz
- Memory Consumed: 244.6 MB
- Virtual Disk Size: 1 GB
- Network I/O Usage: 303.9 Kbps
- Disk I/O Usage: 940.6 KBps

Figure 11. Average Deployed Virtual Machine Profile — Summary View.

CapacityIQ uses this average virtual machine profile to calculate the remaining virtual machine capacity or the number of new virtual machines that you can deploy for this inventory object.
The remaining capacity is the additional number of average virtual machines that can be deployed with the capacity that is available after the new virtual machines are sized. For example, if the remaining CPU demand value would allow you to create 15 new virtual machines but the remaining virtual machines’ memory value is 10, you can create only 10 more virtual machines because you do not have enough memory capacity for more.

The capacity demand profile of the average virtual machine is important for understanding the concept of remaining virtual machines.

CapacityIQ first calculates remaining virtual machines based on each resource individually and then determines the remaining virtual machines based on the limiting resource, as illustrated in Figure 12. CapacityIQ calculates remaining virtual machine capacity using remaining physical capacity divided by average virtual machine demand for CPU and memory.

\[
\text{Remaining VMs CPU} = \frac{\text{Remaining CPU Capacity}}{\text{CPU Demand for Average VM}}
\]

\[
\text{Remaining VMs Memory} = \frac{\text{Remaining Memory Capacity}}{\text{Memory Demand for Average VM}}
\]

Whichever of these two values is smaller determines how many more virtual machines can be deployed.

**NOTE:** CapacityIQ 1.0 considers only CPU and memory resources in estimating remaining virtual machines.

For example, consider the following example scenario:

- A cluster has remaining CPU capacity of 6GHz and remaining memory capacity of 9GB.
- The average virtual machine has 1.5GHz of CPU demand and 4GB of memory demand, on average over time — that is, the assumption is that new virtual machines expected to be deployed on this cluster will have a CPU demand of 1.5GHz and memory demand of 4GB on average.
- Given the above, the remaining virtual machines in this cluster is two — that is, the minimum of four virtual machines based on CPU alone (6GHz/1.5GHz) and two virtual machines based on memory alone (9GB/4GB).

We will discuss CapacityIQ trending and forecasting in the next section.
**CapacityIQ Trending and Forecasting**

The CapacityIQ Dashboard also provides information on how much time is remaining before the specific inventory object you run is out of capacity. CapacityIQ analyzes the current deployment trends for the objects, and forecasts this time based on the total physical resources used in your environment. CapacityIQ projects time remaining for each of the physical resources individually based on the profile of average deployed virtual machine in the datacenter. The time remaining displayed for the object is the minimum of all the times displayed for the individual resources, because that particular physical resource is likely to be used up first.

Time remaining is a measure associated with a virtual datacenter or a given cluster. Time remaining is estimated for both physical capacity and virtual machine capacity. For physical capacity, time remaining is based on forecasting future capacity usage of the physical resource (for example, CPU) to a future point at which the forecasted usage is equal to usable capacity. For virtual machine capacity, time remaining corresponds to the point in time in which the remaining virtual machine capacity is zero — that is, time remaining for virtual machine capacity is a measure of how much time remains before no additional virtual machines can be added to a virtual datacenter or a cluster.

CapacityIQ fits historical capacity usage data into trends to help visualize capacity usage patterns over time to forecast capacity use and availability. Trends are derived by fitting historical data into a trend curve. The trend curve is projected into the future to yield a forecast, as illustrated in Figure 13. The forecast horizon is the number of future intervals into which CapacityIQ projects the trend.

CapacityIQ uses standard forecasting/curve fit techniques for trending — linear, polynomial, exponential, logarithmic, and power. You can choose the forecast method from the Capacity and Trend Analysis view in Global Settings. The forecasting method you choose should closely reflect the actual and future deployment in your datacenter. The quality of the capacity analysis is directly related to how well the trend curve fits the reality. Using more data points for this regression analysis is a way to improve the fit. Standard statistical best practices suggest using 12 data points for a good start, hence CapacityIQ defaults to 12 intervals for its trending and forecasting. As you collect more data, you can choose to increase this number of intervals in the Capacity and Trend Analysis view in Global Settings.
Note that CapacityIQ focuses on managing the capacity on a longer term and not the day-to-day resource utilization. Near–real time spikes do not impact capacity planning as much as they do problem response. To this end, CapacityIQ also provides the flexibility to use different time intervals — daily, weekly, monthly, and quarterly. Depending on the interval you choose, this provides an indication about the amount of data CapacityIQ should collect for high-confidence capacity analysis. Table 1 summarizes this:
Managing Capacity Using VMware vCenter CapacityIQ

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>MINIMUM DATA COLLECTION PERIOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>12 days (~2 weeks)</td>
</tr>
<tr>
<td>Weekly</td>
<td>12 weeks (~3 months)</td>
</tr>
<tr>
<td>Monthly</td>
<td>12 months (1 year)</td>
</tr>
<tr>
<td>Quarterly</td>
<td>12 quarters (3 years)</td>
</tr>
</tbody>
</table>

Table 1. CapacityIQ Minimum Data Collection Period.

By default, the CapacityIQ Dashboard uses the weekly interval and hence the need to deploy for three months for more accurate predictions. You can change the time interval used from the Dashboards, Views and Reports Settings (Figure 16). For example, during initial proof-of-concepts, you can change to a daily interval and get useful information from CapacityIQ after two weeks. However, once again, remember that capacity management is a long-term initiative.

![Figure 16. Dashboards, Views and Reports Settings.](image)

Data Resolution and Rollups

CapacityIQ performs capacity estimates and forecasts based on historical capacity usage data for virtual machines and hosts. For the purpose of trending and forecasting, CapacityIQ uses rollups of capacity usage over time. A rollup is an average value across a given time period. By default, CapacityIQ maintains historical capacity usage information in a fundamental unit called a business week. A business week is a 24x7 matrix, totaling 168 hours, in which each cell represents one hour in a week. Each hour represents the average capacity used during that hour of the business week. Figure 17 illustrates a daily rollup of CPU capacity usage for Monday — the daily rollup is a single number representing the average of the per-hour averages of CPU usage across the 24 hours of Monday. Likewise, the weekly rollup will be a single number representing the average across all 7 days in the week. CapacityIQ provides rollups based on daily, weekly, monthly, quarterly, and yearly time periods.
Additionally, CapacityIQ provides a way to limit which hours in a day are used to calculate rollups using the concept of business hours. Business hours specify the hours in a given day of the week to be used in calculating rollup. For example, if business hours were set to 8:00 a.m.–6:00 p.m. for Monday, the daily rollup for Monday would be based on capacity usage only during the hours between 8:00 a.m. and 6:00 p.m. on Monday. Business hours enable capacity to be planned based on capacity usage only during hours that represent expected, future capacity use.

For example, consider an organization in which virtual machines run workloads only during working hours — 7:00 a.m. –7:00 p.m. (12 hours a day), during which each virtual machine has an average CPU capacity usage of 1.1GHz. Virtual machines have near-zero capacity usage during non-working hours (7:00 p.m. –7:00 a.m.). In planning capacity for these virtual machines, considering all 24 hours in a day in the daily rollup, average capacity usage for each virtual machine would be 0.55GHz (averaged CPU usage over 24 hours). Consequently, capacity would be planned on the assumption that each virtual machine needs 0.55GHz of CPU — this underestimates the true capacity demands during working hours. If business hours of 7:00 a.m. –7:00 p.m. are used, capacity estimates would be based on each virtual machine needing 1.1GHz of CPU, reflecting the actual capacity demands that must be accounted for in planning capacity.
The Utilization Average Calculation Settings (Figure 19) enables you to select the business hours in your environment.

Figure 19. Utilization Average Calculation Settings.
Data Aggregation and Retention

CapacityIQ aggregates historical capacity usage data in order to reduce the amount of information stored. CapacityIQ uses a progressive aggregation scheme in which the degree of aggregation depends on the historical time frame of the data relative to the current time. CapacityIQ aggregates historical data by month, week, quarter (3 months), and year.

Weekly Aggregation

Figure 20 illustrates the aggregation scheme by week. CapacityIQ maintains data for each hour in a week in the 24x7 hour business week (see Data Resolution and Rollups section). The retention period specifies how many past weeks of data is retained in CapacityIQ’s database. For example, if the retention period for weekly aggregation is set to 12 weeks, CapacityIQ retains, at any point in time, one business week of data for each of the last 12 weeks — a total of 12 business weeks of data. As each week passes, data for the oldest business week is purged and data for the most recent business week is added.

![Weekly Aggregation Diagram](image)

**Figure 20. Weekly Aggregation.**

Monthly Aggregation

Figure 21 illustrates the aggregation scheme by month. CapacityIQ aggregates data across all the business weeks in a month into a single business week representing the entire month. For example, as shown in Figure 21, capacity usage average across every Monday, 8:00 a.m.–9:00 a.m. in a given month is averaged into a single number representing capacity usage for Monday, 8:00 a.m. – 9:00 a.m. for the single business week representing the month.

The retention period for monthly aggregation specifies how many months worth of business week summaries is retained by CapacityIQ. For example, if the retention period for monthly aggregation is set to six months, CapacityIQ will retain, at any point in time, one business week for each of the last six months. As each month passes, the business week for the oldest month is purged and data for the most recent month is added.
Quarterly and Yearly Aggregation

CapacityIQ performs similar aggregation by quarter and year. A quarterly aggregation is a single business week representing a given quarter and is based on averages across the three business weeks representing each of the three months in the quarter. A yearly aggregation is a single business week representing a given year and is based on averages across the three business weeks representing each of the four quarters in the year. Figure 22 illustrates the aggregation scheme by quarter and by year.

Figure 21. Monthly Aggregation.

Figure 22. Quarterly and Yearly Aggregation.
Similar to weekly and monthly aggregations, the quarterly and yearly retention periods specify how many quarters’ and years’ worth of business week summaries are retained by CapacityIQ.

The CapacityIQ data retention is configurable from the Data Retention global settings.

![Figure 23: Data Retention Settings.](image)

**Capacity Efficiency**

CapacityIQ can be used to optimize the use of existing capacity in the virtual infrastructure. The ease of deploying virtual machines can result in virtual machine sprawl and inefficient use of virtual capacity.

CapacityIQ includes several views that can provide insights into the virtual machines in your datacenter and clusters:

- Powered-off virtual machines
- Idle virtual machines
- Oversized virtual machines
- Undersized virtual machines

In the next sections, we will discuss how CapacityIQ identifies these virtual machines.

**Powered-Off Virtual Machines**

CapacityIQ identifies virtual machines that have been placed in a powered-off state for a significant portion of time. The motivation behind this analysis is to identify virtual machines that are rarely or never used. Such virtual machines are candidates for decommissioning and represent an opportunity to reclaim unused capacity.

CapacityIQ identifies powered-off virtual machines based on a threshold setting that’s configurable from CapacityIQ global settings. A virtual machine is identified as powered-off in a given interval if the total time it is powered-off (% time powered-off) is greater than a given % time powered-off threshold. The Powered-Off Virtual Machines – List view shows the result for the specific inventory object selected.
Examples
The examples below illustrate the concepts above.

**Example 1**
- A virtual machine is powered-off 50% of the time over the last four weeks.
- Assume that the % time powered-off threshold is set at 60%.
- This virtual machine will not be identified as a powered-off virtual machine, since its % time powered-off (50%) < % time powered-off threshold (60%).

**Example 2**
- A virtual machine is powered-off 70% of the time over the last four weeks.
- Assume that the % time powered-off threshold is set at 60%.
- This virtual machine will be identified as a powered-off virtual machine, since its % time powered-off (70%) > % time powered-off threshold (60%).
Idle Virtual Machines

CapacityIQ identifies virtual machines that show an insignificant level of capacity usage on one or more resources most of the time. The motivation behind this analysis is to identify virtual machines that have practically fallen out of use for any number of reasons—for example, users are no longer running workloads within the virtual machine. Such virtual machines are candidates for decommissioning and represent an opportunity to reclaim unused capacity.

Figure 25 illustrates how CapacityIQ identifies idle virtual machines. CapacityIQ identifies idle virtual machines based on capacity usage of the following resources:

- CPU
- Disk I/O
- Network I/O

For each of these resources, you can set the utilization threshold from the CapacityIQ Global Settings. You can configure CapacityIQ to identify idle virtual machines based on any one or all of these thresholds. For a given time interval, CapacityIQ calculates the total time during which capacity usage is below a given usage less than threshold. If the percentage of this total idle time is more than % time idle threshold, then the virtual machine is identified as idle for that particular resource.

![Diagram of Idle Virtual Machines](image)

Virtual Machine Idle if (% Time Idle) > (% Time Idle Threshold)
Examples

The examples below illustrate the concepts above.

**Example 1**
- A virtual machine’s CPU usage is around 0.05GHz, 80% of the time in the last four weeks.
- Assume that the usage less than threshold for CPU is set at 0.10GHz.
- Assume the % time idle threshold is set at 90%.
- This virtual machine will not be identified as an idle virtual machine due to CPU. This is because its % time idle for CPU (80%) < % time idle threshold for CPU (90%).

**Example 2**
- A virtual machine’s CPU usage is around 0.05GHz, 95% of the time in the last four weeks.
- Assume that the usage less than threshold for CPU is set at 0.10GHz.
- Assume the % time idle threshold is set at 90%.
- This virtual machine will be identified as an idle virtual machine due to CPU. This is because its % time idle for CPU (95%) < % time idle threshold (90%).

**Oversized Virtual Machines**

CapacityIQ identifies virtual machines that consistently use less capacity than their configured capacity. The motivation behind this analysis is to identify virtual machines that can be right-sized down to a smaller configured capacity. For example, a virtual machine configured with 2 vCPU and 4GB memory that is consistently using only 1 of its 2 vCPUs and only 10% of its 4GB of configured memory might be considered an oversized virtual machine. For virtual machines identified as oversized, CapacityIQ also calculates a recommended CPU and memory configuration such that right-sizing the virtual machine to the recommended configuration will result in the virtual machine being no longer underutilized.
CapacityIQ identifies the oversized virtual machines in two steps:

1. Identify virtual machines that are underutilized relative to their configured capacity.
2. Calculate a recommended configured capacity for each identified underutilized virtual machine.

Like powered-off and idle virtual machines, CapacityIQ uses configurable utilization thresholds to identify an oversized virtual machine. However, unlike powered-off and idle virtual machines, CapacityIQ uses a two-level analysis to mark a virtual machine as oversized or undersized. Each of these two levels has a separate threshold, which can be set from the CapacityIQ Global Settings. For the given time interval, CapacityIQ first calculates if a physical resource (for example, CPU, memory) is underutilized based on the configurable utilization less than threshold. CapacityIQ 1.0 identifies oversized virtual machines based capacity usage of the following resources:

- CPU
- Memory

Next, it calculates the degree to which a virtual machine is oversized. In simple words, the degree oversized is the ratio of the area shaded blue to the area shaded grey. A virtual machine is identified as oversized if its degree oversized is greater than the % oversized threshold.

The idea behind this two-level approach is to factor both (1) the amount of time a virtual machine spends in an underutilized state, as well as (2) the extent to which the virtual machine is underutilizing its configured capacity. This is illustrated in Figure 28. The figure shows resource utilization of two virtual machines — one in blue and other in green. For this example, we assume that both virtual machines are underutilized for the same amount of time in a given interval. However, based on the % oversized threshold, the virtual machine depicted in blue may not be identified as oversized.
Once a virtual machine has been identified as oversized, CapacityIQ uses a proprietary technique to calculate a recommended configured capacity such that the virtual machine will no longer be underutilized if right-sized to the recommended configuration. The Oversized Virtual Machines – List shows the final result of this analysis.
Managing Capacity Using VMware vCenter CapacityIQ

**Examples**

The examples below illustrate the concepts above.

**Example 1**
- A virtual machine’s configured capacity is 2 vCPU. The virtual machine is deployed in a cluster in which each host has a 3GHz CPU — therefore the virtual machine’s configured capacity of 2 vCPU translates to $2 \times 3GHz = 6GHz$ of physical CPU capacity.
- The virtual machine uses 0.5GHz of CPU, 95% of the time in the last four weeks.
- Assume that the utilization less than threshold for CPU is set at 50% of configured capacity — in other words, the utilization less than threshold is 50% of 6GHz ($= 3GHz$).
- Assume the % oversized threshold is set at 70%.
- The virtual machine will be identified as underutilized on CPU — because its degree oversized on CPU is $((3GHz - 0.5GHz) \times (0.95 \times 4 \text{ weeks})) / (3GHz \times 4 \text{ weeks}) (= 79\%) > \% \text{ oversized threshold} (= 70\%)$.
- CapacityIQ will calculate a recommended vCPU configuration for this virtual machine such that its utilization relative to the new configuration will be $> 50\%$.

**Example 2**
- Consider the same virtual machine in the above example — a 2 vCPU virtual machine, deployed on a cluster with 3GHz CPU hosts.
- The virtual machine uses 0.5GHz of CPU, 30% of the time in the last four weeks.
- Assume that the utilization less than threshold for CPU is set at 50% of configured capacity — that is, the utilization less than threshold is 70% of 6GHz ($= 3GHz$).
- Assume the % oversized threshold is set at 70%.
- The virtual machine will not be identified as underutilized on CPU — because its degree oversized on CPU is $((3GHz - 0.5GHz) \times (0.3 \times 4 \text{ weeks})) / (3GHz \times 4 \text{ weeks}) (= 25\%) < \% \text{ oversized threshold} (= 70\%)$.

**Undersized Virtual Machines**

CapacityIQ identifies virtual machines that consistently use nearly all of their configured capacity. The motivation behind this analysis is to identify virtual machines that might need to be right-sized up to larger configured capacity so that the workloads running within the virtual machine can get sufficient capacity. For example, a virtual machine configured with 1 vCPU and 1GB memory that is consistently using 95% of its 1 vCPU of CPU capacity and 95% of its 1GB of configured memory might be considered an undersized virtual machine.

To identify undersized virtual machines, CapacityIQ uses similar two-level analysis as oversized virtual machines.
Unlike the oversized virtual machines list, however, CapacityIQ does not provide a recommended CPU and memory configuration for undersized virtual machines. Each of these two levels has a separate threshold, which can be set from the CapacityIQ Global Settings. For the given time interval, CapacityIQ first calculates if a physical resource (for instance, CPU, memory) is overutilized based on the configurable utilization more than threshold. CapacityIQ 1.0 identifies undersized virtual machines based capacity usage of the following resources:

- CPU
- Memory

Next, it calculates the degree to which a virtual machine is undersized. In simple words, the degree undersized is the ratio of the area shaded blue to the area shaded grey. A virtual machine is identified as undersized if its degree undersized is greater than the % undersized threshold.
As explained under the oversized virtual machines section, the idea behind this two-level approach is to factor both (1) the amount of time a virtual machine spends in a stressed state, as well as (2) the extent to which the virtual machine is stressed relative to its configured capacity. This is illustrated in Figure 33. The figure shows resource utilization of two virtual machines—one in blue and other in green. For this example, we assume that both virtual machines are overutilized for the same amount of time in a given interval. However, based on the % undersized threshold, the virtual machine depicted in blue may not be identified as undersized.

![Degree Undersized](image)

Virtual Machine is Undersized if (Degree Undersized) > (% Undersized Threshold)

The Undersized Virtual Machines – List shows the final result of this analysis.

**Examples**

The examples below illustrate the concepts above.

**Example 1**

- A virtual machine’s configured capacity is 1 vCPU. The virtual machine is deployed in a cluster in which each host has a 3GHz CPU—therefore the virtual machine’s configured capacity of 1 vCPU translates to 1 x 3GHz = 3GHz of physical CPU capacity.
- The virtual machine uses 2.4GHz of CPU, 50% of the time in the last four weeks.
- Assume that the utilization target more than threshold for CPU is set at 60% of configured capacity—in other words, the utilization more than threshold is 60% of 3GHz (= 1.8GHz).
- Assume the % undersized threshold is set at 30%.
• The virtual machine’s degree undersized for CPU will be \( \frac{(2.4\text{GHz} - 1.8\text{GHz}) \times (0.50 \times 4 \text{ weeks})}{(3\text{GHz} - 1.8\text{GHz}) \times 4 \text{ weeks}} = 25\% \). Since the degree undersized (25\%) > % undersized threshold (30\%), this virtual machine will be identified as undersized.

**Example 2**

• Consider the same virtual machine in the above example—a 2 vCPU virtual machine, deployed on a cluster with 3GHz CPU hosts.

• The virtual machine uses 2.4GHz of CPU, 50\% of the time in the last four weeks.

• Assume that the utilization target less than threshold for CPU is set at 70\% of configured capacity—in other words, the utilization target less than threshold is 70\% of 3GHz (= 2.1GHz).

• Assume the % undersized threshold is set at 25\%.

• The virtual machine’s degree of stress on CPU will be \( \frac{(2.4\text{GHz} - 2.1\text{GHz}) \times (0.50 \times 4 \text{ weeks})}{(3\text{GHz} - 2.1\text{GHz}) \times 4 \text{ weeks}} = 17\% \). Since the degree of stress < % oversized threshold (25\%), this virtual machine will not be identified as oversized.

CapacityIQ analyzes capacity efficiency over a time interval. This time interval is defined by the Non-trend views setting in CapacityIQ Global Settings (Figure 35). Depending on how a virtual machine uses physical resources, a virtual machine can be marked as both oversized and undersized in that time interval. Figure 36 illustrates a spiky workload depicted in green. It’s possible that this virtual machine may be marked as both undersized and oversized in the displayed interval based on respective threshold settings.

Because the oversized and undersize virtual machine’s analysis is based on two thresholds, it is recommended that you configure these values appropriately. A workload may show steady resource utilization with few variations, while another may show significant variations in resource utilization. Figure 36 tries to illustrate this with two virtual machines, one shown in blue and other in green. Based on degree undersized, the spiky workload in green may not be marked as undersized. In general, we recommend the following:

• For virtual machines with wide variations, use a lower more than or less than threshold value, and a higher % oversized or undersize threshold.

• For virtual machines with relatively steady workloads, use a higher more than or less than threshold value, and a smaller % oversized or undersize threshold.

**Figure 35.**
Summary

Capacity management is a balancing act and tends to become more of an art than a science. Capacity management in a virtual environment is more challenging due to shared resources and vSphere hypervisor optimizations. CapacityIQ is the purpose-built tool from VMware that takes into account these nuances and helps you address capacity-management challenges in virtual deployments. This guide provides background into the key CapacityIQ concepts and methodology. It also describes the different configurable settings that affect the way CapacityIQ analyzes the data. This information should help you understand and interpret the capacity data specific to your environment.