Cloud success secret: Flexible capacity planning

Uncover the elements and tools to plan for successful cloud computing environments

Jose Vargas
Clint Sherwood

One promise of cloud computing is that virtualization will reduce the number of servers needed. It is therefore critical to identify the balanced amount of cloud infrastructure required to meet the anticipated needs of users. The authors introduce basic concepts to help you understand cloud capacity and how to calculate for it. They also introduce a tool that can help you plan for the optimal resources necessary to make your cloud environment a success.

As a part of moving to a cloud computing environment, companies use planning processes and tools to achieve cost reductions, improve speed to systems deployment, and improve systems availability. These processes and tools provide systems administrators with the information they need to manage their environment and plan for future computing needs.

One promise of cloud computing is that virtualization will reduce the number of servers needed, leading to reductions in hardware, software licenses, energy, and maintenance. Therefore, it is critical to identify the optimal amount of cloud infrastructure required to meet the anticipated needs of customers and users. With too few computing resources, requests from users must wait for resources to free up or those requests will be rejected until more hardware is added to the environment. With too many computing resources, the hardware costs and other expenses negate the cost-reduction promises of cloud computing.

Is too much virtualization a bad thing?

A mistaken notion is that the virtualization, automation, and volume of cloud computing can make up for a bad financial model. Unfortunately, if a traditional computing environment is losing money on each transaction, automation may only exacerbate the problem. Proper capacity planning is crucial to understanding the benefits, savings, and costs associated with cloud computing. Remember, a key to successful cloud planning is to understand that there is no magic involved.

A systems administrator should be able to answer five questions in order to successfully plan a cloud environment:
1. How much capacity is available in the data center?
2. How much of available capacity is currently being consumed?
3. When will capacity free up?
4. What is the forecast for new requests?
5. What is the return on investment?

This article introduces some concepts to help you understand cloud capacity and how to calculate for it. It also introduces a tool, the IBM® Infrastructure Planner for Cloud Computing, that can help you achieve the key objectives for moving to cloud computing.

**Understanding cloud capacity**

A cloud computing environment is composed of physical servers that contain resources that can be shared by many users and applications. Each server has disk storage and one or more central processing units with memory. Because cloud environments are virtualized, a fraction of the total CPU, memory, and disk storage is allocated to each user request. This fractional allocation of resources ensures maximum flexibility.

For example, some applications require a lot of disk storage but not a lot of CPU power. Others have the opposite requirement — lots of CPU use and small amount of storage. Cloud computing allows users to specify the amount of each system resource needed for their application.

**The central calculation: Defining virtual CPUs**

When planning for a cloud environment, keep in mind that a **system CPU** is not the same as a **virtualized CePU**. It is often difficult to compare the processing power of modern systems. For example, systems manufactured last year will most likely have processors that are slower than systems manufactured this year. Newer systems also have CPUs with multiple cores.

To ease the challenge of accurate systems resource allocation and capacity planning, some cloud environments have standardized on a cloud CPU unit equal to the processing power on a 1GHz CPU. When a user requests two CPUs, for example, they will get the processing power of two 1GHz CPUs. This means that a system with two CPUs, each with four cores, running at 3GHz will have the equivalent of 24 CPU units:

\[
2 \text{ CPUs} \times 4 \text{ cores} \times 3\text{GHz} = 24 \text{ CPU units}
\]

This calculation is helpful because users can plan for the number of CPUs they need and have a reasonable expectation about performance; and administrators can more easily share the resources provided by one system across multiple requests. Total CPU capacity can be calculated by adding the CPU units available in the environment.

**One note of caution:** When comparing cloud CPU units on different platforms, the processing power of a 1GHz CPU on an IBM PowerVM™ processor system is not the same as 1GHz on an Intel®-based processor. For accurate results, only compare processors within the same platform.
What impact do physical CPUs have?

The number of physical CPUs available within systems is another consideration for capacity planning. A cloud may have 100 CPU units available, but if the most powerful system in the cloud has only 20 physical CPU units, this becomes a limit for a virtual machine request.

Balancing CPUs with memory and storage

Keep in mind that CPU power is not the only factor in achieving successful capacity planning. While capacity planning involves making sound decisions about the number of CPUs, it also involves balancing the CPU information with the amount of memory and disk storage purchased for each system.

For example, purchasing a system with 24 CPU units of processing power and only 2GB of memory makes little sense in a cloud environment. In this case, when a user asks for a virtual machine with two CPUs and 2GB of memory, the server will be fully allocated to fill this single request. The 22 unallocated CPU units would remain unavailable to other users and therefore idle for the life of this request.

It makes sense to correctly balance systems resources when making hardware purchases for the cloud environment.

How you request your IT service has an impact

A primary goal of capacity planning is to ensure that IT capacity is the right size at the right moment, whether that moment is now, tomorrow, or 20 years from now. So one important tool to enable effective capacity planning is to look at the way requests for IT services arrive at a data center.

The traditional view

In traditional data centers, system administrators receive requests for IT resources from software engineers for prospective development projects. Administrators typically review IT requests on a weekly basis to determine what resources are available and which projects have the highest priority. Higher priority projects usually get their requests answered first.

In many cases, traditional data centers can fulfill high priority requests in as few as three weeks from the time a decision is made to allocate the resources. If the IT resources need to be purchased, however, the process can take months.

Projects that are low on the priority list may need to wait a long time, depending on budget and resource availability. In some cases, these low-priority projects may not get their requests fulfilled at all!

Given this lengthy, uncertain process, users become conditioned to request as much computing resources as they can get. Unfortunately, these requests are often more than the users actually need. Once provisioned, these resources are jealously guarded and even when the project ends, the resources are typically not relinquished unless the users are forced to do so. This attitude is
understandable within the limitations of the traditional IT paradigm. After all, the success of the current project, as well as the next, depends on having sufficient IT resources.

But the sad lesson of this traditional model is clear: Excessive resources often arrive late in the development cycle, impacting productivity and competitiveness. When the project ends, those same resources, now hoarded by the users, become underutilized, wasted capacity.

**The view from the clouds**

Cloud computing presents us with a very different scenario:

- Developers access a web site where they can enter their request for IT resources — servers, software, storage, etc.
- Users know immediately if the resources are available.
- If resources are available, the request can be immediately submitted and automatically routed to the cloud administrator for approval.
- Because the process is automated, requests are often fulfilled within an hour of the request.
- When the project ends or winds down, developers using the cloud no longer hoard the computing resources, knowing they can easily and quickly access the same resources in the future as the need arises.

For future projects, developers using the cloud will likewise only request the resources they need rather than over-provisioning as they are conditioned to do with traditional IT resource delivery. In addition, cloud users must typically specify a project end date; unless this date is extended, cloud resources are automatically returned to the available resources pool on that date. So even if resources are not intentionally released by the user, they still become available for use by others.

From an administrator's point of view, a cloud environment morphs a manual, time-consuming process into a one-click, automated approval process. Information about the availability of data center cloud infrastructure and resources is provided in near real time, giving the administrator an immediate window into the total capacity and remaining resources of the environment.

**Determining resource needs**

Let's examine resource needs using a common development organization scenario:

*A company is implementing a new cloud environment for their development and test organization consisting of 150 software engineers. One hundred of the software engineers develop software, 40 perform software quality assurance, and 10 are responsible for running and maintaining their production environment.*

How large should the cloud be to meet this organization's computing demands? There are two major pieces of information we need to answer that question (each breaks down into sub-units of information, of course): users' requirements and systems resources. It looks something like this:

1. Users' requirements:
   a. Average resource requirements for software developers
      i. Two VMs per developer on average
ii. CPU: 6 CPU units, memory 2GB, disk storage=100GB
iii. Environment needed for 90 days on average

b. Average resource requirements for software assurance engineers
   i. Three VMs per developer on average
   ii. CPU=4 CPU units, memory=2GB, disk storage=50GB
   iii. Environment needed for 30 days on average

c. Average resource requirements for production environment
   i. One VM per application environment
   ii. CPU=12 CPU units, memory=16GB, disk storage=500GB
   iii. Environment needed for one year on average

2. Systems resources:
   a. Systems used: IBM BladeCenter® HS22 8-way 2.8GHz blade servers
   b. Memory per server: 48GB
   c. Disk storage per server: 1,200GB

Figure 1 shows that capacity planning estimates that on average, 113 systems are needed. To ensure the environment has available resources to fulfill all requests 100 percent of the time, planning would recommend 124 servers. So capacity planning can determine the number of systems needed to support this organization. (Later in the article, we'll introduce a capacity planning tool, the IBM Infrastructure Planner for Cloud Computing, to make the planning task simpler. The image in Figure 1 is from the tool's results in planning for this scenario.)

**Figure 1. Estimate for this scenario**
Improving capacity through virtualization

Before we introduce the IBM Infrastructure Planning tool, let's look at how virtualization can improve capacity.

A common problem for traditional data center administrators is low IT resource utilization, often as low as 10 to 20 percent. That is to say, on average, 80 to 90 percent of a server's compute power is unused. So requesting additional resources is a waste in a traditional situation.

In addition, data centers often have limited raised floor space for their systems, so even if a business has the financial resources to buy more equipment, it may not have the physical resources to add more systems.

By contrast, the virtualization that is a key component of cloud computing makes it appear that one system is many individual servers. With this technology, a hypervisor running on top of the host computer's operating system allows multiple operating systems to run concurrently. Rather than wasting 80 percent of valuable compute resources as happens in traditional compute environments, the hypervisor ensures that every server operates at the most efficient, productive level. These efficiencies are even more pronounced in today's high-performance, multi-core processors systems with large amounts of memory and disk storage.

Virtualization gives cloud administrators the ability to handle more requests with fewer systems.

Real world, real trends

You are always better able to anticipate the future by understanding the past. In the case of capacity planning, it is easier to forecast an organization's computing needs if you have a clear picture of IT resource consumption over the previous six months. Historic usage patterns and trends allow an IT manager to estimate when resources should be added and how many resources will be needed.

For example:

- Online shopping sites know that during the holiday season there is a spike in website visitors.
- They also know which items are most popular during the holiday rush.
- There's a corresponding increase as well during this time in the number of follow-up visits to check on the status of orders.

The increased traffic translates to requests for more computing resources during the last two months of the year; however, user traffic tends to go back to normal after the beginning of the year.

Knowing the patterns helps an administrator better plan for future seasonal spikes. Plotting traffic over time helps separate true spikes from possible overall increase in web traffic.

Demand that changes over time is considered a trend rather than a spike as (shown in Figure 2).
Consider this scenario: A successful company needs more resources to facilitate growth. The administrator uses IT growth rate information to anticipate the need for additional resources, allowing those resources to be requested in a timely way. In a well-managed cloud computing environment, this capability is provided in an automated way. The environment is able to meet current needs because by nature it is an elastic IT supply model.

Knowing the rate at which demand has been increasing is important. Using cloud tools, it is possible to estimate, based on growth trends, when more resources will be needed. With this information, the manager is better able to estimate the additional capacity needed and when it will be needed.

For accurate forecasting administrators need to monitor the following information:

- Number of user requests
- Number of virtual machines requested
- Allocated CPU, memory, and disk capacity
- Actual consumption of CPU, memory, and disk capacity
- Total cloud capacity

It's important to understand the relationship of resources allocated versus resources consumed. If experience is any judge, users will most likely request many more resources than they actually need. This makes it reasonable for the administrator to consider lowering the amount of CPU allocated, for example, if CPU utilization for a particular virtual machine is consistently at or below 10 percent.
Figure 3 illustrates the way trending data can be used for capacity planning decisions.

**Figure 3. CPU allocation vs. usage trend**

![Graph showing CPU allocation vs. usage trend](image)

The tool path: IBM Infrastructure Planner for Cloud Computing

The IBM Infrastructure Planner for Cloud Computing tool makes it easier for IT administrators to ensure that sufficient actual hardware, software, and infrastructure are in place so that cloud users enjoy a sense of endless compute resources. The tool — currently targeted to estimate capacity in the IBM Smart Business Development and Test Cloud, IBM WebSphere® CloudBurst, and IBM Tivoli® Service Automation Manager (TSAM) environments — enables administrators to

- Model cloud capacity planning using sets of unique user class templates associated with a planned production roll-out.
- Model the performance of generic and custom business applications targeted for a variety of traditional and cloud computing environments.

Further information on this product can be obtained by sending an email to planner@us.ibm.com.

Figure 3 shows:

- Total CPU capacity (the blue line)
- Allocated CPU (the red line)
- CPU resources actually used (the green line)

Total CPU capacity through September was 500 CPU units. In October, 300 additional CPU units became available when more systems were added to the environment. The allocated line shows that CPU capacity is consistently being added based on user requests (a typical scenario for new cloud computing centers).

The used capacity line shows how much of the resources are actually being used. Although requests for CPU resources are on a steep curve, the actual usage is staying at around 100 CPU units. Using this information, the administrator can make a decision about how much to over-commit CPU resources. That is, the administrator may decide to promise to provide more than
100 percent of available resources when resource requests are for more than 100 percent of available resources, but user resource demand at any given time is for less than 100 percent of available resources. In this example, a substantial amount of CPU resources could be over-committed and still meet user demand.

You can also see that by going by the allocated trend line, it appeared justified to add the resources that were added in October. However, the used trend line tells a different story — it shows that even at the 500 total capacity limit of the system before the October additions, there was enough capacity to meet the users’ demands.

Automated monitoring and reporting on cloud resources makes trend spotting and capacity planning easier and more accurate. Tools that perform these tasks (like the IBM Infrastructure Planner for Cloud Computing) are often worth their weight in gold since they streamline the process and allow the administrator to quickly provision needed resources.

**In conclusion**

**The next step**

Besides using the resources provided at the end of this article, take a look at this developerWorks search result ... computing resource capacity planning has been around a long time and developerWorks has always included it in its technology coverage.

Capacity management is a vital activity in the context of cloud computing. Done properly, capacity planning provides users with the needed computing resources to create innovative solutions and meet the performance goals of a business application while at the same time contributing to an organization’s financial goals.

Today's high performing multi-core servers have large amounts of memory and huge disk storage capacities that can best be fully utilized by using the virtualization technologies that are a key component of cloud computing. This resource-rich IT environment leads us to new and better ways to plan for optimal resource allocation.

Cloud computing environments enable easy access to computing resources. With careful planning, a cloud environment can create the appearance of an endless supply of computing resources. Organizations that employ the right set of processes to monitor and plan for the use IT resources can position themselves to reap the promised benefits of cloud computing.
Related topics

- **Cloud Computing Capacity Planning: Maximizing Cloud Value** is the original paper that forms the foundation of this article.
- **IBM WebSphere CloudBurst Appliance information center** also, check out developerWorks WebSphere resources.
- **IBM Tivoli Service Automation Manager information center**; also, check out developerWorks Tivoli resources.
- This **VMware Resource Management Guide** provides information about VM overhead and architecture.
- The **HiPODS whitepaper library** delivers many papers on capacity planning and high performance on demand solutions.
- Take a deeper look at the **WebSphere CloudBurst Appliance**.
- Take a deeper look at the **Tivoli Service Automation Manager**.
- In the developerWorks **cloud developer resources**, discover and share knowledge and experience of application and services developers building their projects for cloud deployment.
- Access the **IBM Smart Business Development and Test on the IBM Cloud**.

© Copyright IBM Corporation 2010

**Trademarks**
(www.ibm.com/developerworks/ibm/trademarks/)