Cloud Computing Capacity Planning

Maximizing Cloud Value

Authors: Jose Vargas, Clint Sherwood

Organization: IBM Cloud Labs

Web address: ibm.com/websphere/developer/zones/hipods

Date: 3 November 2010

Status: Version 1.0

Abstract: This paper provides the reader a view to the elements involved in the careful planning for successful cloud computing environments and the tools available for administrators to manage those environments.

Executive summary

With its promise of elastic and easily accessible computing resources, cloud computing has understandably captured the attention of the information technology industry. Business executives see cloud as a way to take control of ever-escalating data center costs. Systems administrators see it as a way to automate and better manage available data center resources and requests. End users see cloud as a way to easily obtain the computing resources they need, and pay for the resources only while they're in use.

These are all great reasons to join the move to cloud computing, and highlight the powerful appeal of this new IT services delivery model. Cloud computing, however, as with most disruptive technologies, creates both opportunities and challenges. One key challenge is providing users with apparently unlimited computing resources while simultaneously reducing data center costs.

We must acknowledge first, as suggested by our use of "apparently," that there is no such thing as unlimited resources in cloud computing. All computing requests, even for virtualized resources, ultimately map to physical devices—finite resources. A virtual machine ("VM"), for example, has physical computing resources allocated to it: CPU, memory, disk storage, and network bandwidth. These physical resources must be planned for and delivered to the end user of cloud services in a way that creates the impression of unlimited resources. This impression is created through careful capacity planning—skillfully provisioning finite resources to meet highly variable demand. Too few computing resources create unhappy managers and users. Excessive computing resources pressure the cost side of the equation, diminishing the potential savings of the cloud delivery model.

The secret to cloud success is thus having just enough physical resources to meet anticipated demand.

This paper provides the reader a view to the elements involved in the careful planning for successful cloud computing environments and the tools available for administrators to manage those environments.

Note: Before using this information, read the information in "Notices" on the last page.

Contents

Introduction

The top three reasons cited by IT executives for moving to cloud computing are: (1) cost reduction, (2) improved speed to systems deployment, and (3) improved systems availability.[1] To achieve these goals, companies use planning processes and tools that 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 cost reductions for hardware, software licenses, energy, and maintenance. To fulfill this promise, and to successfully manage a cloud computing environment, it is critical to identify the optimal amount of cloud infrastructure required to meet the anticipated needs of end users. If there are too few computing resources, requests from users will either have to wait for resources to free up, or will be rejected until more hardware is added to the environment. A cloud that cannot fulfill requests effectively will not deliver on the promise of improved speed to systems deployment. If there are too many computing resources, on the other hand, hardware and other expenses negate the cost reduction promise of cloud computing.

Five questions should guide a systems administrator 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?

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.

The good news is that cloud computing does indeed deliver on its cost reduction and efficiency promises. A key to successful cloud planning is to understand that there is no magic involved. The recommendations provided in this paper can help organizations achieve the key objectives for moving to cloud computing.

Requests for IT services: traditional vs. cloud

According to Wikipedia, the primary goal of capacity planning is to ensure that IT capacity cost-effectively meets current and future business requirements.[2] A review of the way requests for IT services arrive at a data center can help us better understand capacity planning and the ways cloud can make it much easier to fulfill these requests.

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,

[1] Source: IBM Market Insights, *Cloud Computing Research*, July 2009. n=1,090

[2] http://en.wikipedia.org/wiki/Capacity_planning

traditional data centers can fulfill high priority computing 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, which is often more than they need. Once provisioned, these resources are jealously guarded, and even when the project ends are typically not given up unless the users are forced to do so. This attitude is understandable. After all, the success of the current project—and 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.

Cloud computing presents us with a very different scenario. Here, 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 they 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. And 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, and unless this date is extended, cloud resources are automatically returned to the available resources pool on that date. Thus, 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, cloud morphs a manual and 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.

Improving capacity through virtualization

A common problem for traditional data center administrators is low IT resource utilization, often as low as 10 - 20%. That is to say, on average 80 - 90% of a server's compute power is unused. And yet end users continue to request additional computing resources, each of which will almost certainly be likewise underutilized. 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 (rack space, power, network, or cooling) to add more systems.

By contrast, with cloud computing's virtualization technology, one system can be made to appear as 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% or more of valuable compute resources, as happens in traditional compute environments, the hypervisor ensures that every server operates most efficiently and productively. These savings are even more impressive in today's high-performance, multi-core processors

systems with large amounts of memory and disk storage. Cloud administrators can thus use virtualization to handle more requests with fewer systems.

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 one or more central processing units (CPUs) with memory and disk storage. 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. Some applications, for example, require a lot of disk storage but not a lot of CPU. Others have the opposite requirement—a lot of CPU and small amount of storage. Cloud computing allows users to specify the amount of each system resource needed for their application, ensuring that only the amount needed for that particular application is allocated.

When planning for a cloud environment, keep in mind that a system CPU is not the same as a virtualized CPU. 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 one gigahertz CPU. When a user requests two CPUs, for example, they will get the processing power of two 1 GHz CPUs. This means that a system with two CPUs, each with four cores, running at 3 GHz 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 in capacity planning. Users can plan for the number of CPUs they need and have a reasonable expectation about performance. 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 1.0 GHz CPU on an IBM® PowerVM™ processor system is not the same as 1.0 GHz on an Intel-based processor. For accurate results only compare processors within the same platform.

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

Successful capacity planning thus involves making sound decisions about the number of CPUs, as well as 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 2 GB of memory makes little sense in a cloud environment. In this case, when a user asks for a VM with two CPUs and 2 GB 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 thus idle—for the life of this request. It therefore makes sense to correctly balance systems resources when making hardware purchases for the cloud 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? Here's the information we need to answer that question:

1. Users' requirements:
    a. Average resource requirements for software developers
        i. Two VMs per developer on average
        ii. CPU: 6 CPU units, memory 2 GB, disk storage=100 GB
        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=2 GB, disk storage=50 GB
        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=16 GB, disk storage=500 GB
        iii. Environment needed for one year on average
2. Systems resources:
    a. Systems used: IBM® BladeCenter® HS22 8-way 2.8 GHz blade servers
    b. Memory per server: 48 GB
    c. Disk storage per server: 1,200 GB

The number of systems needed to support this organization can be determined using a capacity planning tool such as IBM Infrastructure Planner for Cloud Computing.[3] Figure 1 shows that the capacity planning tool estimates that, on average, 113 systems are needed. To ensure the environment has available resources to fulfill all requests 100% of the time, the tool recommends 124 servers. Other information provided can be seen in Figure 1.

[3] See Appendix A: IBM Infrastructure Planner for Cloud Computing.

Required Systems

	xSeries-Blade HS22 8-way 2800
Average	113.12
90th percentile	122.00
100th percentile	124.00

VM Requirements

	Supported Users	Required VMs	Supported VMs per Node	Supported Users per Node
Developers	100.00	210.00	2.39	1.14
Testers	40.00	126.00	1.43	0.45
Production	15.00	16.00	0.18	0.17
Total	**155.00**	**352.00**	**4.00**	**1.76**

End-Point Server Resources Used by VMs (Percentage)

Cloud CPU units used	70.42 percent
Total memory used	16.87 percent
Total disk storage used	23.68 percent

Figure 1: Infrastructure Planner estimate

Trends and forecasting

We are 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 we 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 Web site 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. This 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 such 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. (Figure 2)

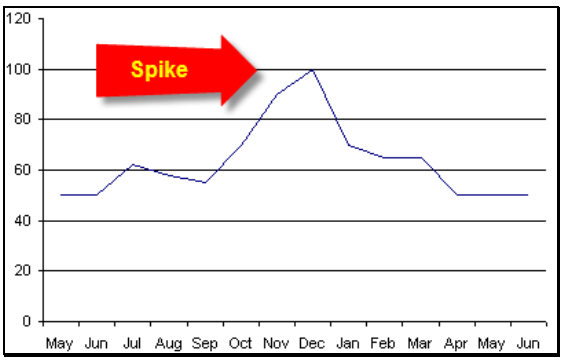

Figure 2: Example holiday season spike

Demand that increases or decreases over time is considered a trend rather than a spike (Figure 3). For example, 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. A well-managed cloud computing environment provides this capability in an automated way. The environment is able to meet current needs, because by nature it is an elastic IT supply model. Using cloud tools, it is also possible to estimate, based on growth trends, when more resources will be needed. Knowing the rate by which demand has been increasing is important. With this information, the manager is better able to estimate the additional capacity needed, and when it will be needed.

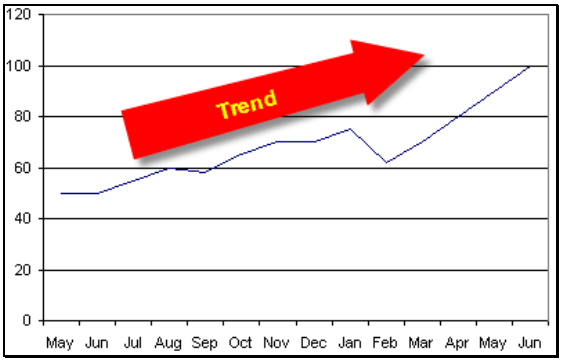

Figure 3: Example of increasing trend

For accurate forecasting administrators need to monitor:

1. Number of user requests
2. Number of virtual machines requested
3. Allocated CPU, memory, and disk
4. Actual consumption of CPU, memory, and disk
5. Total cloud capacity

It is important to understand resources allocated versus resources consumed. Users may request much more than they actually need. It is 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%.

Figure 4 illustrates the way trending data can be used for capacity planning decisions.

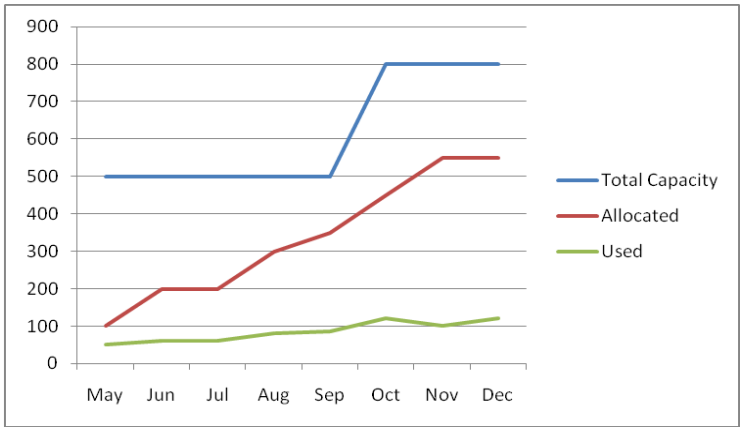

Figure 4: CPU allocation vs usage trend

The chart shows total CPU capacity (blue), allocated CPU (red) and CPU actually used (green). 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). "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. For this example, a substantial amount of CPU resources could be overcommitted and still meet user demand.

In this example, we also see that the "Allocated" trend line would provide justification for the resources added in October. However, the "Used" trend line shows that even at the 500 total capacity limit, there was enough capacity to meet the user's demands.

Tools that monitor cloud resources and provide trending information are available from IBM. We highlight one of these tools in Appendix A—the IBM Infrastructure Planner for Cloud Computing—but a complete description of these tools is beyond the scope of this paper. More information about cloud management and automated trending tools is available by contacting your IBM representative.

Public vs. private clouds

Cloud computing environments are often classified as public or private. A combination of the two is called a hybrid cloud. Public clouds provide computing resources to users in a utility-style model from third party providers. Users don't need to be concerned about acquiring physical systems, configuring them on the network, or managing the environment. They simply request the needed computing resources—for example, CPUs, disk, storage.—for a specific period of time. A service level agreement (SLA) guarantees the availability of resources, including network bandwidth. A key selling point for public clouds is the flexibility to pay only for what you use.

Private cloud environments provide computing resources for users within a single company's internal network. Private cloud computing resources are managed and dedicated for the specific use of the company. This type of cloud allows for better control of security and availability of cloud resources. In addition, legal or regulatory requirements for data safekeeping and storage (within the company or within a particular country) can sometimes only be guaranteed with a private cloud.

As the name suggests, a hybrid cloud environment includes elements of both private and public clouds. In this case, the public cloud may be leveraged for certain projects, for example, but storage and specific applications remain within the firewalled and managed environment of a private cloud.

A variety of factors must be considered when deciding which cloud environment is best—public, private, or hybrid. Among these factors are company history, business model, security concerns, speed to deployment, and the probability that resources will be needed even if projects are canceled. While this is not an exhaustive list, several questions should typically be asked:

1. Are data center operations one of your business's core strengths? In other words, do you want to be in the business of managing data center resources? A grocery chain operating on tight margins may decide that the cost of data center operations—equipment, staff, energy consumption, etc.—provides only minimal return on investment. As a result, the company may decide that renting computing resources from a public cloud provides them with most of the IT resources they need to run their business.
2. Can your company's data reside outside your company's walls? Or, in the case of some financial, medical, and other regulated data, can that data reside outside the country or state where your company is located? Perhaps your company's intellectual property needs high levels of protection, which can be better offered with internal resources.
3. Public cloud companies often offer service level agreements. Can you live with the terms of such an agreement?
4. In a downturn, do you have the ability to quickly dispose of IT capital equipment purchased during the last growth cycle?
5. How long does it take to acquire—or do you already have in place—the computing resources needed to start a major project?
6. How much does it cost on a CPU/hour basis to run your environment?
7. Are some of your computing capacity needs tied to specific goals during a specific time frame, or are they seasonally based?

Cloud providers are well positioned to address specific client needs when they are able to offer the right mix of public, private, and hybrid clouds.

Conclusion

Capacity management has been, and continues to be, an important activity that ensures users have the needed computing resources to create innovative solutions, and meet the performance goals of a business application, while at the same time contributing to the 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 using virtualization technologies. This resource rich IT environment has led to new and better ways to plan for optimal allocation of needed computing requirements for today's business applications.

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.

References

- Learn about IBM Smart Business Development and Test Cloud at www.ibm.com/cloud/enterprise/beta/dashboard
- Learn about IBM CloudBurst at www.ibm.com/software/webservers/cloudburst/
- Learn about IBM Tivoli Service Automation Manager at www.ibm.com/software/tivoli/products/tsam-facts.html
- Refer to the *VMware Resource Management Guide* at vmware.com/pdf/vi3_35/esx_3/r35u2/vi3_35_25_u2_resource_mgmt.pdf for information about VM overhead and architecture.
- Many of the workloads are described in further detail in separate white papers. See all the IBM High Performance On Demand Solutions white papers at: www.ibm.com/Websphere/developer/zones/hipods
- The WebSphere Application Server Performance Web site provides a centralized access to many helpful performance reports, tools and downloads. See: www.ibm.com/software/Webservers/appserv/performance.html

Appendix A: IBM Infrastructure Planner for Cloud Computing

IBM offers the Infrastructure Planner for Cloud Computing tool to meet the needs of IT professionals facing the challenge of planning for and providing scalable and efficient performance for their user communities. While users in cloud communities enjoy a sense of endless compute resources, IT administrators must do as they've always done—ensure that sufficient actual hardware, software, and infrastructure are in place to meet computing needs.

The IBM Infrastructure Planner for Cloud Computing enables IT 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.

Cloud computing environments provide unique challenges for capacity planning because workloads tend to be more dynamic, and the range of users and user classes can be more expansive than traditional compute environments. The IBM Infrastructure Planner for Cloud Computing allows administrators to estimate the numbers of users that a cloud compute environment can support. It also estimates the computing resources required for multiple sets of users of different classes.

The current version of the planner is targeted for estimating capacity in three cloud environments: IBM Smart Business Development and Test Cloud, IBM CloudBurst, and IBM Tivoli Service Automation Manager (TSAM).

The IBM Infrastructure Planner provides performance results that allow users to assess the adequacy of a given configuration for their requirements, as well as provide insight of likely bottlenecks. The Planner can therefore be useful for capacity planning, evaluation of infrastructure and workload changes, and projecting cloud and noncloud site scalability.

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

Notices

Trademarks

IBM, the IBM logo, BladeCenter, CloudBurst, PowerVM, Tivoli, and WebSphere are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Special Notice

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While IBM may have reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Anyone attempting to adapt these techniques to their own environments do so at their own risk.

While IBM may have reviewed each item for accuracy in a specific situation, IBM offers no guarantee or warranty to any user that the same or similar results will be obtained elsewhere. Any person attempting to adapt the techniques contained in this document to their own environment(s) does so at their own risk. Any performance data contained in this document were determined in various controlled laboratory environments and are for reference purposes only. Customers should not adapt these performance numbers to their own environments as system performance standards. The results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.