Realtime publishers

The Definitive Guide™ To

Cloud Computing

sponsored by

IBM®

Dan Sullivan

Copyright Statement

Chapter 9: Maintaining a Cloud Environment: Governance, Growth, and Security

There is much discussion about how cloud computing is different from earlier models of service delivery. This book has followed a similar pattern for the first eight chapters by concentrating on what distinguishes cloud computing from mainframe, client-server, and other distributed approaches to delivering services. This chapter will be different. Now we will focus our attention on themes common to all forms of IT and delivery:

- The role of governance
- Capacity planning
- The need for security

Governance is the guiding framework that defines how we go about implementing service delivery in the cloud. It can be thought of as a set of constraints on possible solutions to a problem. Principles of governance are not technical principles, per se, but they do have implications on the technical solutions we implement. For example, a policy may dictate that especially sensitive private and confidential information may only be stored on devices under the complete control of the company. This limits the use of public clouds as an extension of a private cloud. The governing policy need not explicitly mention restrictions on public clouds but that is the practical implication. Other aspects of governance influence and constrain how we deliver other services, what types of services may be delivered, and to whom we may deliver them.

Capacity planning is often a challenging task in IT management. Throughout this book, we have discussed how cloud computing makes capacity planning easier, and it does—for the cloud consumer. The cloud services provider, however, still faces the typical challenges of forecasting demand for services, balancing peak load demand with average load demand, and formulating acceptable service level agreements (SLAs) with customers.

In addition to having enough capacity to meet the demands of SLAs, we have to ensure that infrastructure is reliable enough to be available as required by SLAs. Fortunately, cloud architectures are inherently distributed and therefore enable relatively straightforward failover approaches. Nonetheless, we still have to be careful to avoid single points of failure and ensure that supporting services, such as making redundant copies of data, happen fast enough and frequently enough to ensure sufficient recovery in the event of a data loss in one part of the storage system.

The need for security in information management is ubiquitous. Cloud computing has its array of information security requirements that are similar to those found in other service deliver models, including the need to:

- Maintain identity information about users
- Limit access to data and applications based on identity
- Ensure software is checked for vulnerabilities and patched as needed
- Prevent malicious applications from operating within the cloud
- Protect the privacy of confidential information

The fundamental security requirements are no different in the cloud than in other models, but the way we implement security controls can vary, sometimes for the better. For example, if an operating system (OS) vendor releases a security patch and a business determines that the patch must be applied to every server, that patch will have to be pushed to each server. Even with an asset management application that automatically distributes and installs software patches, there is likely to be some manual intervention required. Systems administrators will have to review patch reports to verify patches were applied correctly, determine where patches have failed, and apply corrective action to each instance of the failure.

In a cloud computing environment, images in the service catalog can be regenerated with the patch and deployed to the service catalog. The older, vulnerable version of the image could be removed from the catalog so that it is no longer instantiated within the cloud. There may be instances of the vulnerable image running in the cloud in which case cloud administrators would have to coordinate with the systems administrators responsible for those instances to shut down those instances and restart with the patched versions. This is similar to the kind of coordination that typically occurs when servers are dedicated to particular departments or applications.

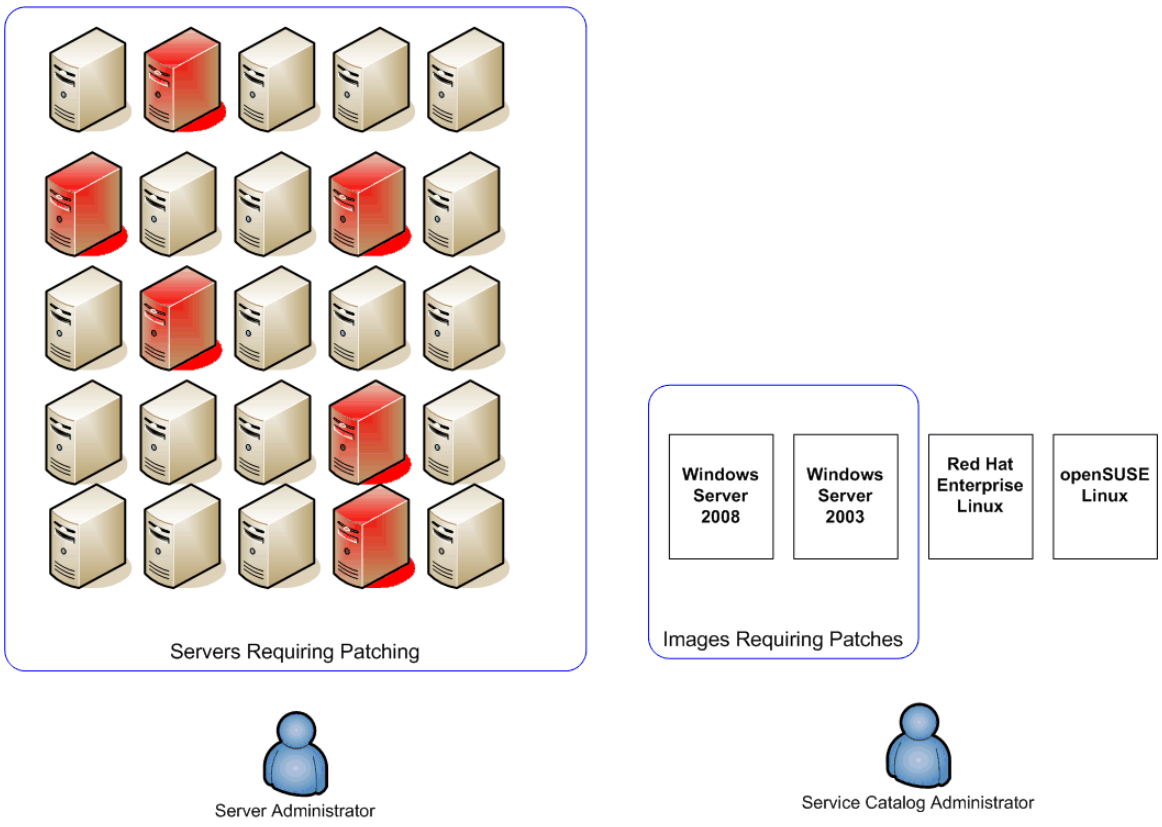

Figure 9.1: The need to apply security patches is the same with or without a cloud; however, the execution can be less problematic when working with a service catalog rather than individual servers where the patch may fail for different reasons.

The long-term maintenance of a cloud computing environment requires attention to governance, capacity planning, and security. In this chapter, we will consider each in turn and outline key considerations in each area. Not surprisingly, the same types of issues we see in governance, capacity planning, and security in other architectures occur within the cloud. This presents a significant advantage for cloud computing administrators: We can adapt the best practices that have evolved over the past decades of IT management to cloud computing.

Governance Issues in the Cloud Computing

Governance is about establishing a framework for directing, monitoring, and reporting on the implementation activities of an organization. Businesses have boards of directors for governing the company at large. Cloud computing governance is a subset of corporate governance. The directions and principles established at the corporate level define the environment in which cloud computing governance occurs.

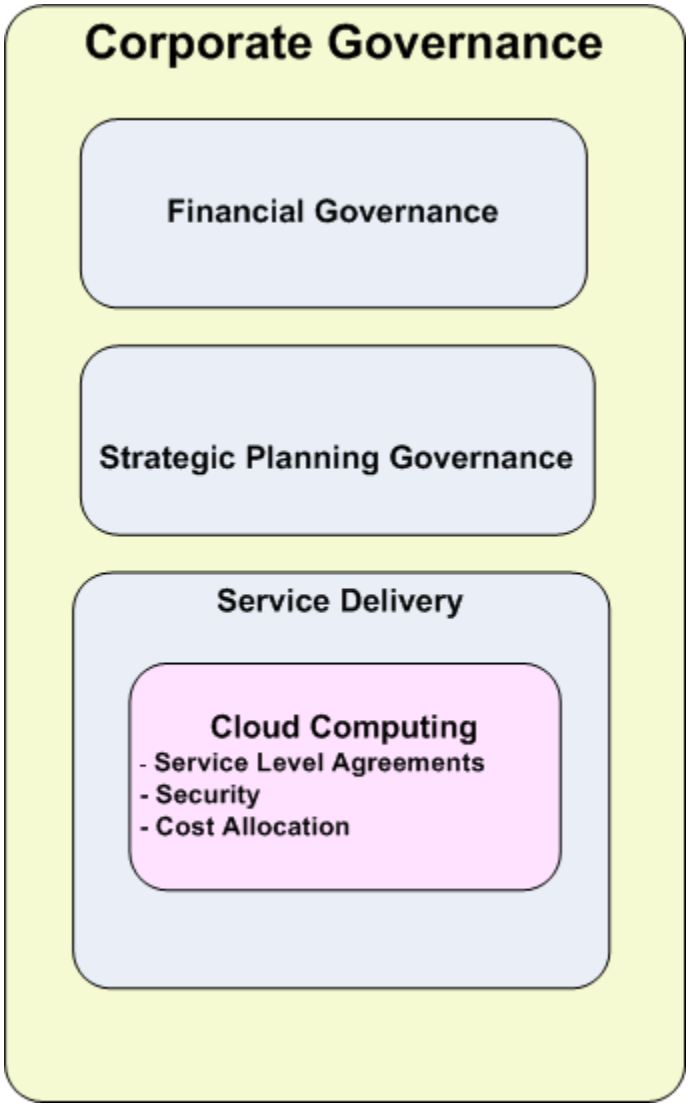

Figure 9.2: The hierarchy of corporate governance subsumes cloud computing governance.

Corporate governance establishes direction and management principles for the entire company with some specialization, as required, for areas such as finance, strategic planning, and service delivery. Within service delivery, we can place cloud computing governance. Some of the most important aspects of cloud governance include:

- Protecting the integrity of business services
- Controlling access to cloud services
- Allocating costs for cloud services

These areas all have implications for how we implement cloud services, but they are primarily business issues, not technical issues. The technical aspects of these issues come into play when we start to implement the policies defined by governing bodies. Cloud governance defines what is to be implemented; cloud implementation defines how it is implemented.

Protecting the Integrity of Business Services

The integrity of business services entails two parts:

- Ensuring individual transactions and operations in the cloud function as expected without compromising the confidentiality of those transactions and operations
- Ensuring cloud services are available as expected and as agreed to in SLAs

Confidentiality in the Cloud

What level of confidentiality should a cloud consumer expect when using cloud resources? For example,

- Who will have access to the data transmitted between the cloud and outside data stores?
- Who will determine who will have access to data stored in cloud storage?
- What efforts are made to reduce the risk of inadvertent disclosure of data?
- Under what circumstances will normal confidentiality protections be suspended in order to prevent or investigate malicious activities in the cloud?

It is the responsibility of the governing body to specify policies that answer these and similar questions that will arise. (Again, governance addresses what should be done not how to do it. Implementation details are delegated to others, so we will not delve into the technical details of how to meet these requirements right now.)

Policies on confidentially should specify a combination of protections that should be in place as well as a description of the limits to those protections. For example, policy may dictate that cloud administrators make available encrypted communications between client devices and the cloud resources. Cloud consumers can make use of encrypted communications if they want, but they may not be required to. At the same time, policy may require cloud administrators to avoid deploying software with known vulnerabilities that could compromise the security of the cloud. This may lead cloud administrators to not offer basic ftp services and instead require a secure form of ftp. This may seem contradictory but it is not.

In one part of a policy, we state that cloud consumers, not administrators, can decide on the level of security they desire for communications. In another part, the policy states that vulnerable software should not be deployed, and this limits cloud consumer choices. It is not unusual for complex policies to lead to seemingly contradictory indications. In these situations, one part of the policy has to take precedence over the other. In this example, protecting the cloud resources and its users is worth constraining the options of users.

Governance and Balancing Acts

This kind of balancing act is commonly seen in law. The freedom of speech is a well-known right to many but that does not permit us to yell "Fire!" in a crowded theater when there is no fire.

It is conceivable that governing regulations will impose constraints on what business units might want to do. One department might want to negotiate an SLA that allows them to rapidly upload large volumes of data from external resources. Internal regulations, however, require that any files uploaded from external resources be scanned for malware. The scanning will cause the loading process to exceed the time window the customer wants. The governing principles exist for a reason and in spite of how it might limit what business units conceive, they are in place to protect the cloud infrastructure, data within the cloud, and the business operations that depend on it.

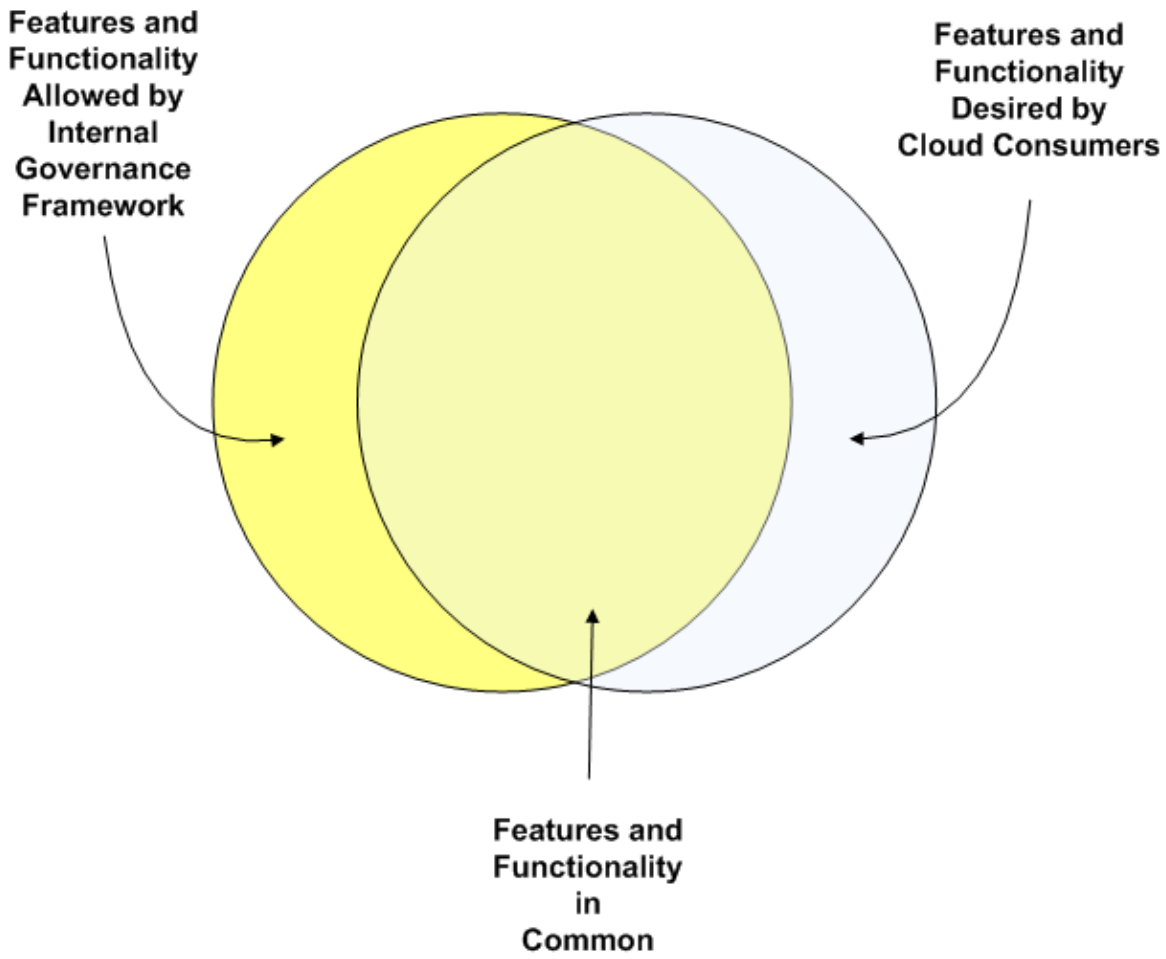

Figure 9.3: Governance policies define how cloud resources may be used. Business units might want additional features or functionality that are not allowed in the cloud; instead, they are constrained to the features they would like that overlap with those allowed by governance regulations.

Availability and SLAs

Another topic for governance is availability and the role of SLAs. A governance framework does not dictate specific rules about availability, but it does set guidelines. For example, the governing body may specify that SLAs will contain specifications for:

- The number and types of servers that will be available to the cloud consumer on a regular basis
- The percentage of time that the agreed upon number and types of servers will be available
- Compensation for violations of SLAs

These are SLA-specific issues that would be negotiated between the cloud administrators and users of cloud services. The governing body may also specify global guidelines, such as requiring that not more than X% of servers, storage capacity, or other resource be down for routine maintenance at the same time. This type of global constraint further defines the boundaries of actions that cloud administrators can take.

The integrity of cloud services is protected in part by policies protecting confidentiality of data and preserving the availability of services. It is also highly dependent on security controls, including access to the cloud.

Controlling Access to Cloud Services

One of the most fundamental considerations in the governance of cloud resources is determining who has access to those resources. If a company invests in a private cloud, will the company make the cloud available to

- Any employee or contractor with an interest in using the resource
- Members of research and development, engineering, or other product development efforts that require significant computational resources
- Employees in any department with the funds to cover the costs of the resources

Once it is determined who will have access to the cloud, security controls, such as identity management, authentication, and authorization systems, can be used to enforce those policies.

Within the group of users eligible to use cloud resources, there may be a further division by priority. Some departments, such as finance, may be given top priority under the assumption that their needs are immediate and critical. Research and development and engineering groups may be in a second tier of users because their work is essential to the long-term viability of the company and they have demonstrated the need for large amounts of CPU time. A third tier may be everyone else in the company who will have access to resources not consumed by the other two groups.

Within each group, there may be limitations on the resources they can acquire. For example, the top-tier Finance group may have access to as many servers as they like but can run them continuously for only 48 hours if other jobs are waiting to run in the cloud. Engineering may need to run large calculations for extended periods of time, so they may run their virtual server instances for as long as they like but are limited in the number of virtual servers they can instantiate at any one time. Regardless of who can access cloud services, someone has to pay for them.

Pricing Cloud Services

There are two broad approaches to determining the costs for cloud services: cost allocation and competitive pricing. In practice, the actual prices cloud consumers pay be me a mix of both approaches, but we will discuss them separately and then see how they can be merged.

Cost Allocation

Cost allocation is a pricing model that is driven by the costs incurred by the provider of the service. At its most basic level, the cost of a service is equal to the cost of purchasing and maintaining equipment and providing labor to support the service divided by the units of the service provided. An example can help clarify some of the details.

Let's assume a basic server can run four virtual servers. The server runs 24 hours a day, 7 days week for 3 years for a total of 26,280 hours. Let's also assume the server was purchased for $5000, requires $1000 in labor to maintain over the course of 3 years, and incurs $300 in power, cooling, rack space, and other miscellaneous charges for a total of $6300 in costs over 3 years. (For simplicity, we'll assume that this server only runs open source software so that there are no software licensing costs). The hourly cost of providing this server is 26,280 hours divided by $6300 or $0.24 per hour.

In practice, this simple cost allocation model will need some modification. For example, the assumption that a single server will run 24×7 for 3 years straight is unrealistic. Also, clouds are designed to accommodate varying peak demand periods, so there will be time when some servers are not utilized and therefore not charged to any customer. Finally, servers in the cloud may have been acquired at different times for different prices. Trying to assign each server its own individual total cost of ownership (TCO) would generate more accounting work than it is worth. A better approach is to use an average cost and an average utilization rate for each server.

In the cost allocation model, we have to make some assumptions about utilization rates and availability of servers. When we set prices, we have to hope we have made good estimates. If we are overly optimistic about utilization and availability, we may find that in fact we do not recover all the expenses we had planned for and are left with a revenue or cost recovery shortfall.

This kind of cost allocation model is found in government institutions where pricing is driven by the need to recover costs rather than to earn a profit. The same model may work well within a business where IT units are treated as cost recovery centers and not profit earning centers.

Competitive Pricing

Another approach to pricing, which is common in business, is competitive pricing or pricing according to what the market will bear. Presumably public clouds use a competitive pricing model where their price for a unit of service includes the costs we described earlier plus an additional amount for profit. This certainly makes sense for a public cloud, but does this pricing model have a place with private clouds used only by internal customers? Yes, in some cases.

By charging more than the actual costs, a cloud provider can generate a reserve of earnings that are not allocated to cover the costs of providing the cloud services. (This is similar to profits or retained earnings, but those have specific accounting definitions, so we will try to avoid using those terms.) This reserve can be used in several ways:

- As a resource for funding future expansion of cloud infrastructure
- To mitigate the risk of unanticipated problems, such covering the costs associated with replacing failed devices that may or may not be under warranty
- To fund experimental cloud services that are provided for free in return for feedback on the services

The cost recovery model does not provide a mechanism for this kind of retained reserves funding. One could imagine incorporating the cost of future expansion, risk management, and service development into the cost of providing services, but that is a bit counter to the intention of the cost recovery approach.

Neither cost recovery nor competitive pricing is inherently better or worse than the other. It is up to the governing body to determine which approach better serves the long-term goals of the enterprise.

Cloud computing governance is a subset of corporate governance. Regulations put in place at the enterprise level constrain what can be done with cloud services. Such high-level constraints are insufficient guidance for providing a governing framework for a private cloud. Further regulations around protecting the integrity of services, limiting access to cloud services, and allocating the costs of the cloud are all required. Another facet of long-term maintenance is capacity planning.

Planning for Growth

One of the key benefits of using cloud computing is that users of the cloud can rapidly scale their resource use up and down. As workloads increase, the number of servers dedicated to the task can increase. As data volumes grow, so can the storage utilized. Users no longer need to worry about maintaining peak capacity infrastructure—it is available in the cloud when it is needed. Cloud computing does not eliminate the need for capacity planning; it centralizes the burden on the cloud provider.

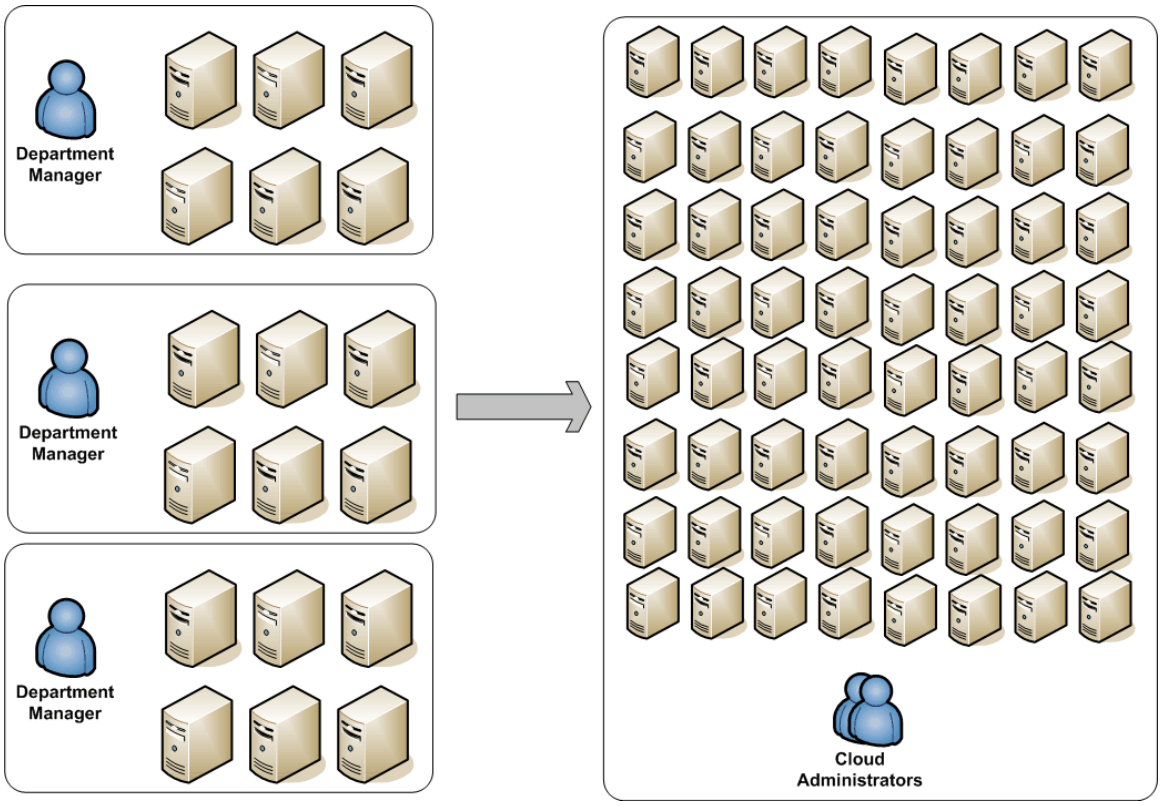

Figure 9.4: With the adoption of cloud computing, the scope of capacity planning shifts from individual applications and departments to a centralized cloud service provider.

Centralized cloud providers will have to address capacity planning issues common throughout IT:

- Researching customer expectations for current and future resources
- Estimating costs of future services
- Planning how to deliver needed capacity in the most efficient manner
- Identifying dependencies that can influence how new capacity is added

Capacity planning begins by identifying key resources that affect the ability of service providers to meet SLAs. Then we turn our attention to understanding how demands for capacity of various resources are expected to grow.

Key Resources in Cloud Computing

The key resources in cloud computing are those that limit the ability to deliver services:

- Physical servers
- Storage
- Network bandwidth

Each is a limiting factor because in spite of adequate capacity in two of these, a shortage in the other will inhibit the ability to deliver services. If there are ample servers and sufficient network capacity but we run out of storage, storage-dependent workflows will be blocked. Similarly, if network bandwidth is saturated, the ability to move data into and out of the cloud is constrained.

How are we to accurately predict the future needs of cloud users? Especially when their workloads and peak demands can vary so much? The answer is SLAs. These contracts between cloud providers and cloud consumers specify what levels of resources are expected by cloud consumers and what the cloud provider commits to. Cloud consumers are responsible for estimating their current and future requirements in terms of computing, storage, and network demands. Cloud providers are responsible for ensuring that the cloud can meet the aggregate demand for resources specified in SLAs.

Another factor that is easy to overlook is the physical environment in which the cloud infrastructure resides. Servers, storage devices, and network equipment require space, power, and cooling. There are limits to how many racks can fit in a data center, how much power can be reliably and consistently delivered, and how much heat generated by equipment can be adequately cooled or vented. SLAs probably will not explicitly state requirements related to environment; instead they have to be derived from the details about servers, storage, and network services. With these key components and details of SLAs, we can begin to formulate baseline and future growth projections.

Baseline and Initial Growth Projections

SLAs and historical data provide a starting point for establishing baselines for the amount of resources required to meet service delivery needs. One of the advantages of starting with SLAs and historical data is that it is reasonably reliable and accurate data. Assuming historical data is collected properly, we have a detailed record of what happened in the past. SLAs provide guidance on what will occur in the near future, and possibly longer if customers use long-term contracts to lock in favorable pricing.

Baseline Measures

We can think of a baseline measure as the average load on the cloud for computing, storage, and network services at some point in time. The purpose of taking a baseline is to understand what level of service can be delivered by a particular amount of cloud infrastructure. A baseline measure of cloud service delivery might include:

- Number of servers with all servers normalized to a standard, such as a single quad-core processor with 16GB RAM
- Total amount of storage available
- Network throughput
- Average server utilization
- Number of virtual machine instances available in the service catalog
- Percentage of time SLAs are met

The first three metrics capture the basic capacity of the cloud. They measure, in some ways, the overall throughput of the cloud infrastructure. These metrics are not precise enough for all performance-related tasks. For example, these metrics are not adequate for comparing the performance of different implementations of the same algorithm. For that, the implementations should be run on the same hardware under the same network load running the same OS and application stack. The purpose of collecting these measures is to be able to compare cloud infrastructure capacities in order to estimate what is required to meet a set of SLAs.

Average utilization is important because it influences the total throughput of the cloud. If utilization is low, there will be excess capacity that is not utilized. One way to improve the throughput of a cloud is to increase utilization. For example, to double the throughput of a cloud with 40% utilization, we double the number of servers and other infrastructure while maintaining a 40% utilization rate, or we could maintain the same level of infrastructure and increase the utilization to 80%.

Growth Projections

After establishing baseline measures, we can plan for growth projections. There are two types of growth we need to account for: growth in capacity and growth in usage or throughput. It is worth noting that increasing utilization and throughput can happen in a fairly incremental manner while the addition of infrastructure tends to happen in a more step-wise manner, as Figure 9.5 shows.

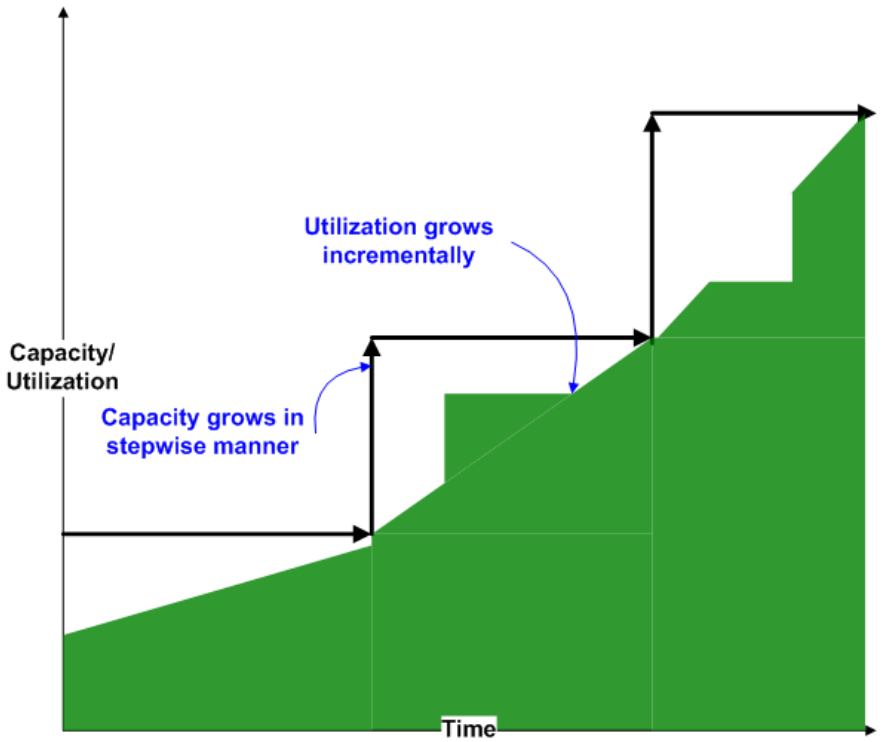

Figure 9.5: Capacity is often acquired in bulk, giving a stepwise growth in capacity. Utilization tends to grow incrementally, although there may be spikes or temporary drops in utilization.

Growth in Utilization

Utilization grows at a rate determined by a number of factors, such as an increase in the

- Volume of work performed by existing cloud consumers executing existing workflows
- Number of distinct workflows executed by existing cloud consumers
- Number of cloud consumers

For each of these types of increase, there can be corresponding decreases. For example, a department may re-engineer its processes and stop using an application that had run in the cloud.

Some of these growth factors are likely to lead to incremental growth. As a line of business expands into new markets or launches new product lines, there can be a progressive growth in the volume of transactions that need to be processed. In some cases, there may be sharp and sudden rises in the number of transactions (think of the Apple iPad launch).

Sudden and dramatic growth in demand can arise from changes in the organization. A merger or acquisition can add a large pool of potential cloud service customers to a company and drive demand for services sharply higher. Similarly, divesting in a line of business can cause sudden drops in demand and therefore overall utilization.

Growth in Capacity

Although demand for cloud services can change in fairly incremental ways, capacity changes tend to be more bulk, stepwise changes. This reality is driven by economics. Conceivably, a company could follow a steady incremental growth plan. For example, a company could add 100 high-end servers to the cloud every week for the foreseeable future. If the company is a rapidly growing Web infrastructure provider, this might make sense. In many cases, a stepwise growth in capacity makes more sense.

Consider a typical budget cycle. An IT manager creates an infrastructure budget based on projected demand. The CFO takes into account revenue growth, cash flow projections, borrowing costs, and other factors and determines that 25% of the budget will be available in the first quarter, 50% in the third quarter, and if revenue projections are on target, another 25% in the fourth quarter. The IT manager will likely purchase the equipment in three periods as the funds become available. The hardware will be brought online as soon as possible. The funds are not available any sooner, so there is no way to accelerate the purchases. It makes no sense to leave equipment in the shipping containers, unless demand is low, in which case the purchases were unnecessary.

Another factor that leads to the stepwise growth in capacity is the economics of hardware installation. If one goes to the trouble to install a single rack in a data center, the marginal cost of installing a second, third, fourth, and so on is so low that it often makes sense to perform these operations in bulk. As the practice of cloud computing has matured, another option has become available for providers of private clouds: expanding by using public cloud compute and storage resources.

Expanding Using a Public Cloud

The reasons that a private cloud provider would want to make use of a public cloud parallel the reasons that end users are drawn to public clouds: elasticity and cost effectiveness. The combination of private and public clouds, known as a hybrid cloud, has several advantages as well as some disadvantages.

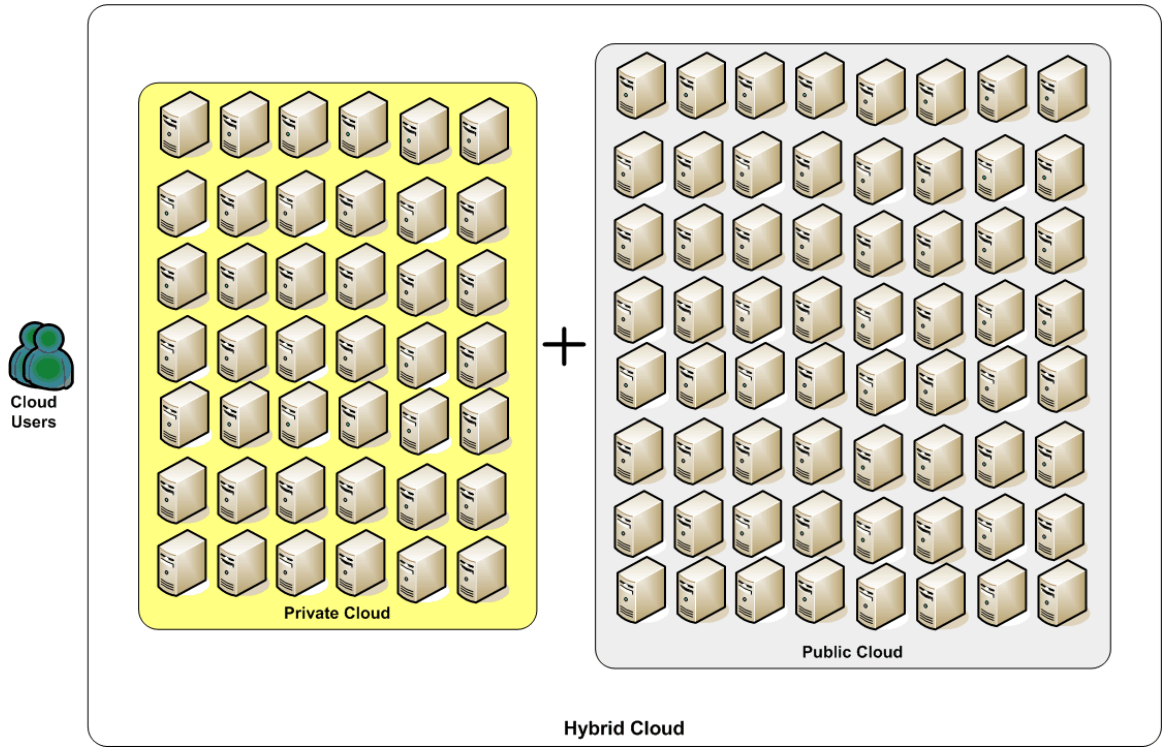

Figure 9.6: Hybrid clouds appear to users to be functionally equivalent to private clouds. Private cloud administrators hide the implementation details from end users.

Elastic Scaling and Hybrid Clouds: The Benefits

Combining resources with a public cloud allows private clouds to rapidly expand capacity without the capital investment of expanding a private cloud. Also, resources in a public cloud can be commissioned and decommissioned faster than adding or removing comparable physical resources in a private cloud.

The cost of a private cloud may be less than that of a public cloud. This is not criticism of private clouds. The two are designed for different purposes and serve different needs. Private clouds are designed according to the particular needs of a single business and governed by policies needed to protect that business. Public clouds are generic computing and storage resources with policies designed to accommodate a wide range of users. Public clouds may be able to offer lower prices because they benefit from economies of scale that are not available to private cloud providers. Also, public clouds may have less in the way of security, auditing, and control over the service catalog than a private cloud does. As is often the case in IT, choosing between the two is a matter of choosing a solution that best fits a particular set of requirements.

Elastic Scaling and Hybrid Clouds: The Disadvantages

The primary disadvantage of a hybrid cloud is that some data is moved outside the corporate firewall. Public cloud providers can make significant efforts to protect their customers' data (they certainly have no incentive to risk a data breach of one of their customers) but that may not be enough for security-conscious executives and managers.

Moving large volumes of data can also be a hindrance. In a cloud computing version of the old "sneaker net" (that is, running data back and forth between data centers on portable disks), public cloud providers offer customers the option of shipping disks to a data center for bulk loading rather than copying data over the Internet.

Hybrid clouds are a viable option in many cases when expanding a private cloud is not a practical option. When the public cloud can be used to run applications that do not instantiate protected intellectual property, the volumes of data to transfer are low, and the security requirements are minimal, then public cloud services make sense. Public clouds can supplement private cloud capacity for conventional workloads; public clouds can also contribute to mitigating the risk of hardware failures.

Mitigating Risks Through Architecture

Capacity planning should take into account the need for excess capacity in case of failures in some parts of critical infrastructure. When a small number of servers fail, the jobs running on those servers can be restarted on other servers. This situation can often be accommodated by the excess capacity that can exist because of the difference in the capacity deployed and the capacity required to meet utilization requirements (see Figure 9.5).

Catastrophic failures require advance planning. For example, if an entire data center becomes inaccessible or a large number of servers is down because power distribution is disrupted to a large number of racks within a data center, the excess capacity in the cloud may not be enough to accommodate for the loss. In such cases, we need to plan to maintain additional capacity. Two factors should be considered when planning such excess capacity: the physical distribution of data centers and the need for redundant infrastructure.

Physical Distribution of Data Centers

Data centers in different geographical locations reduce the risk that two or more data centers will be struck by the same catastrophic event (for example, regional power loss, earthquake, and flood). In addition to building data centers in different areas, we need to keep replicas of data in different data centers, maintain redundant copies of the service catalog in different data centers, and ensure that policies and procedures are defined and implemented in the same way across data centers.

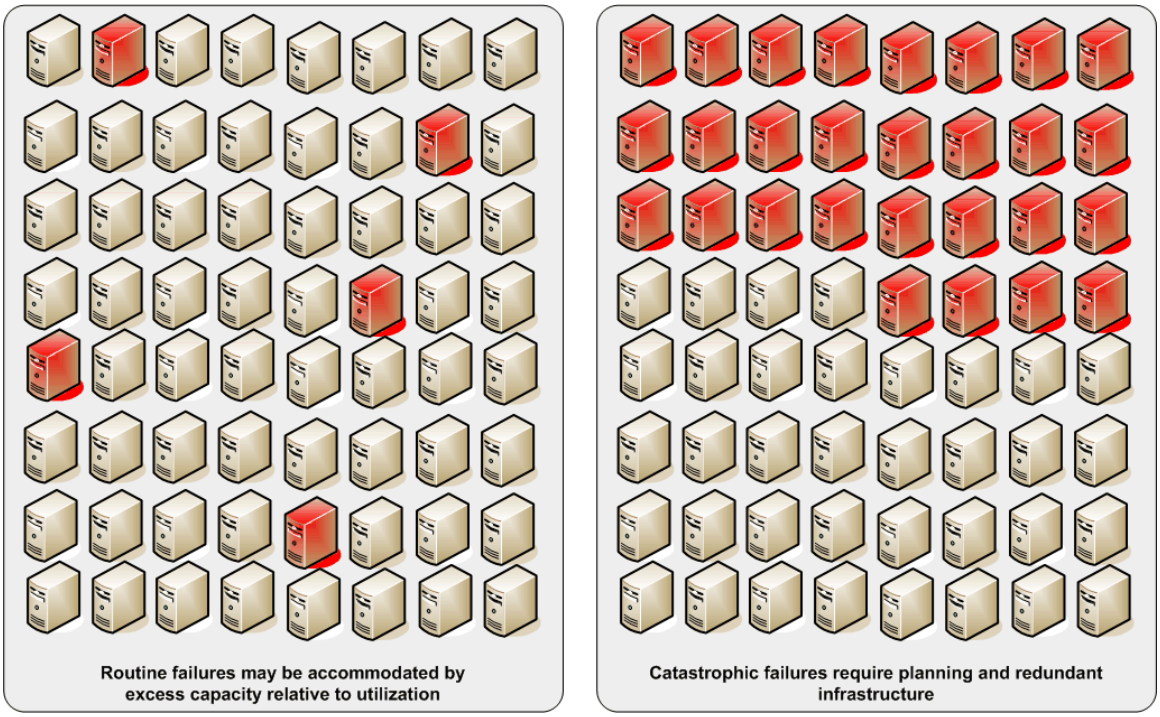

Figure 9.7: Routine failures are readily accommodated in clouds but catastrophic failures require failover planning and additional infrastructure.

Redundant Infrastructure

Data centers will of course need servers, storage, and network infrastructure. They will also require comparable backup power systems, multiple Internet service providers (ISPs), and backup cooling and venting systems to reduce the risk of a single point of failure in the infrastructure.

Capacity planning has traditionally been challenging in IT. When working within the constraints of department or line of business budgets, it might be difficult to realize a highly redundant, rapid failover architecture without significant cost. Centralizing the management of infrastructure within the cloud allows for pooled utilization and capacity. It also provides for more efficient deployment of redundant infrastructure, which can mitigate the risk of failures in the cloud.

The third and final topic we will consider with regard to long-term maintenance of a cloud is the need for security.

Security in the Cloud

Key considerations for long-term planning for security in the cloud are similar to those for other aspects of enterprise security:

- Identity management
- Entitlements and access controls
- Vulnerability assessments
- Patching and image management

These are not fundamentally different from what needs to be done in other IT environments but, as is so often the case, different implementations of similar services and functionality bring with them varying dependencies and maintenance requirements.

Identity Management in the Cloud

Identity management is the practice of maintaining information about users of IT resources and services. A primary concern in the cloud is how to maintain an accurate and up-to-date database of identities. Common questions that arise with identities in the cloud are:

- Who should be added as a user in the cloud? All employees? Full-time employees only? Should contractors be added, and if so, according to what criteria?
- How should identities be removed to ensure the least risk of failing to remove someone's identity that should be removed?
- What type of monitoring on the activity of identities is required?
- How frequently should identities be audited?

The concern here is with long-term management and maintenance, so implementation issues are not considered, although they are certainly important. They are just outside the scope of this discussion.

Before we can address whose identities should be added to the cloud, we have to have a clear understanding of the purpose of the cloud. The looser the purpose (for example, to provide general computing and storage services to all business units for all purposes), the more broadly defined is the set of potential users. More restricted clouds, such as those for research and development and engineering purposes, will have correspondingly restricted groups of users.

Removing identities is also an issue. Ideally, changes to a centralized HR system would trigger the removal of identities in the cloud when an employee leaves the company. This may not account for contractors and consultants who are granted access to resources. It may not be sufficient for employees changing roles and losing privileges to the cloud.

Routine monitoring of activities associated with identities can help detect anomalous events. For example, if one or two individuals are using cloud resources at rates significantly higher than others in the same role, it may be an indication of unauthorized use. Less frequent but routine auditing of the identity management database can help detect cases where identities that should have been removed or disabled remain active.

Identities provide a means to associate privileges with users. These privileges, or entitlements as they are sometimes called, also require oversight.

Entitlements and Access Controls

Entitlements should be associated with well-defined roles in a business. For example, financial analysts should have access to historical financial transactions and various data marts and business intelligence applications; however, access to product designs, marketing strategies, and sales forecasts may be restricted to a small group of executives. Under ideal conditions, no one would ever be granted entitlements to data or applications that are not required for them to do their jobs. Employees change roles, controls on data change, and new applications are brought online sometimes with overly broad execution privileges.

Policies and procedures should be in place in the cloud to protect a number of entitlement related issues:

- Granting access to data according to a data classification scheme. These often are based on four categories: public data, sensitive data, private data, and confidential data. Public data can be shared without harm; sensitive data should not be shared broadly but would not cause serious harm if it did; private data is about a customer or other person and is not to be shared outside a restricted group; and confidential data is company-related data that would cause significant harm if disclosed.
- Applications should be controlled along similar lines as data. Some applications contain proprietary knowledge, such as a risk scoring program, and should be restricted to individuals who have a legitimate need for the application.
- Software licensing may restrict the number of users that can simultaneously run an application or restrict an application's use to a set of named users. Software licensing models tend to evolve along with server technology, so it is reasonable to expect software vendors will quickly adapt their pricing models to the cloud.

Entitlements and access controls protect how data and applications are used. Next, we will turn our attention to ensuring those applications are functioning as expected.

Vulnerability Assessment and Patching

It is widely assumed that complex software has flaws. Sometimes bugs are the result of programmers making mistakes in their coding. Other times, designers create applications that although coded according to specification, function in unanticipated ways. At other times, software developers create better ways of performing the same task and release new versions of applications with better performance. In all of these cases, there are reasons to update the software with vendor-provided patches.

Patching is a common practice and can significantly improve the security and quality of the software we run. It is not without risk, though. A patch may correct one flaw while introducing another. A patch could render an application that worked well in one configuration non-functional. Policies should be defined for the cloud service catalog that specify when and how patches should be applied to virtual machine images in the cloud. These policies should consider:

- What would trigger the decision to apply a patch? Reasons include a regular patch release from a vendor, a notice in the trade press about a newly discovered vulnerability in a popular software application, or through the use of vulnerability scanning software with the company.
- What testing should be done prior to releasing a patched image? In some cases, it may be sufficient to release a new version while maintaining the older version in the service catalog. Users would then be free to choose which to run. This may work for non-security patches, but images with known, high-impact vulnerabilities should not be left for general use.

As with other security aspects, patching and vulnerability management practices outside the cloud can be readily adapted to the cloud.

Summary

Long-term management and maintenance of a cloud environment requires attention to governance, capacity planning, and security issues. Governance issues include framing policies for the cloud that fit with overall corporate governance, defining the scope and structure of SLAs, and formulating a cost recovery mechanism for cloud services. Capacity planning is based on SLAs and strategic direction of the company. SLAs provide a baseline for determining the capacity needed to meet SLAs while maintaining reasonable utilization rates with some tolerance for the inevitable hardware failure. Long-term security concerns include the need to address identity management, entitlements, vulnerability assessment, and patching. These are not new management considerations for IT professionals and many best practices that have been created over the past decades can continue to serve us well if we adapt them to the particular requirements of a cloud environment.