Realtime
publishers

# The Definitive Guide™ To

# Cloud Computing

sponsored by

IBM®

Dan Sullivan

# Introduction to Realtime Publishers

**by Don Jones, Series Editor**

For several years now, Realtime has produced dozens and dozens of high-quality books that just happen to be delivered in electronic format—at no cost to you, the reader. We've made this unique publishing model work through the generous support and cooperation of our sponsors, who agree to bear each book's production expenses for the benefit of our readers.

Although we've always offered our publications to you for free, don't think for a moment that quality is anything less than our top priority. My job is to make sure that our books are as good as—and in most cases better than—any printed book that would cost you $40 or more. Our electronic publishing model offers several advantages over printed books: You receive chapters literally as fast as our authors produce them (hence the "realtime" aspect of our model), and we can update chapters to reflect the latest changes in technology.

I want to point out that our books are by no means paid advertisements or white papers. We're an independent publishing company, and an important aspect of my job is to make sure that our authors are free to voice their expertise and opinions without reservation or restriction. We maintain complete editorial control of our publications, and I'm proud that we've produced so many quality books over the past years.

I want to extend an invitation to visit us at http://nexus.realtimepublishers.com, especially if you've received this publication from a friend or colleague. We have a wide variety of additional books on a range of topics, and you're sure to find something that's of interest to you—and it won't cost you a thing. We hope you'll continue to come to Realtime for your educational needs far into the future.

Until then, enjoy.

Don Jones

## *Copyright Statement*

# Chapter 1: Changing the Way We Deliver Services with Cloud Computing

Computing is constantly changing, creating new hardware technologies, improving software, and optimizing business processes. The history of computing is almost a constant stream of advances. Mainframe computing was followed by mini-computers, which were followed by personal computers, and most recently mobile devices. Software development followed a similar trajectory with an evolution that started with batch-oriented mainframe applications and moved through client server models to highly distributed service-oriented architectures and Web applications. Business processes changed and computing expanded beyond the reach of large volume highly-focused back office systems supporting core operations to widely adopted collaboration and personal productivity applications. Sometimes the changes in hardware, software, and business processes converge in ways that create significant new opportunities for delivering business services. The advent of cloud computing is one of those events.

Cloud computing in its simplest form is a model for allocating compute and storage resources on demand. In practice, it is much more. Cloud computing offers new ways to provide services while significantly altering the cost structure underlying those services. These new technical and pricing opportunities drive changes in the way businesses operate. The *Definitive Guide to Cloud Computing* describes the technical, operational, and organizational aspects of cloud computing and provides a roadmap for navigating the emerging landscape of cloud computing.

## Overview

Cloud computing is a broad-ranging and still-developing set of technologies and business practices. This guide examines the essential technical and business aspects of cloud computing in order to provide a broad assessment of the benefits and challenges facing adopters of cloud computing. This book consists of 10 chapters; each deals with a significant aspect of cloud computing:

- Chapter 1, this chapter, introduces cloud computing and its impact on how we deliver services. In this chapter, we examine the business drivers behind cloud computing and the related issues of the changing economics of information technology (IT). The chapter concludes with a discussion on aligning business strategy with IT services, especially with regard to cloud computing.

- Chapter 2 identifies the essential elements of cloud computing, discusses different types of cloud computing services and different types of cloud delivery models, ranging from public to private cloud services.
- In Chapter 3 we examine the business advantages of cloud computing and the various sources of Return on Investment (ROI) in cloud computing.
- The business case for cloud computing continues in Chapter 4. Topics include identifying business priorities, assessing current capabilities, determining considerations for adopting a cloud model for service delivery and consumption, and measuring the value of a cloud.
- In Chapter 5 the topic shifts from the business case to understanding how to plan for a cloud and how to assess architecture options with regard to cloud computing. Use cases are included to highlight some of the practical considerations in developing a plan to move to cloud computing.
- Chapter 6 delves deeper into the technical issues introduced in Chapter 5. These include providing high-availability compute, storage, and network services. Cloud management and adapting IT procedures to the cloud are also discussed.
- In Chapter 7, we take a process-oriented approach and consider how to use the information developed in the previous chapters and apply it to specific business needs. Subject areas include performing workload analysis, managing cloud services, centralizing resources, and defining service level agreements (SLAs).
- The planning topics of Chapter 7 are followed by Chapter 8. The focus of this chapter is on establishing a private cloud, transitioning compute and storage services, and operational issues managing cloud services.
- Chapter 9 delves into long-term management issues ranging from controlling access to cloud services to capacity planning and risk mitigation.
- The *Definitive Guide to Cloud Computing* concludes with Chapter 10. This chapter consolidates and summarizes the essential aspects of planning, implementing, and managing cloud computing services.

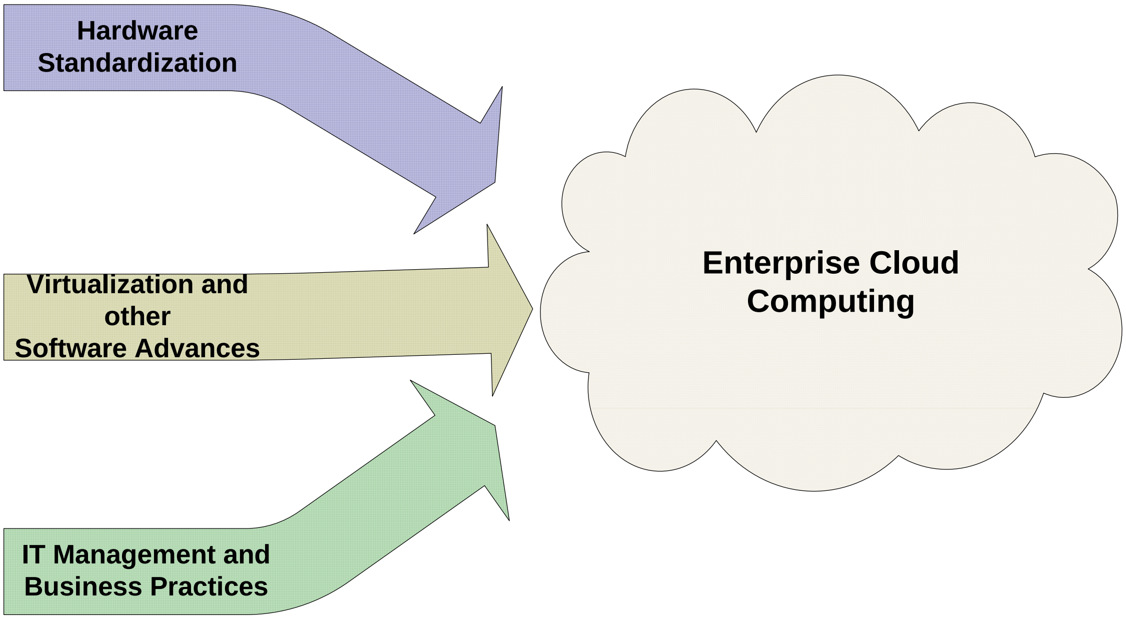

Hardware Standardization
Virtualization and other Software Advances
IT Management and Business Practices
Enterprise Cloud Computing

**Figure 1.1: Enterprise cloud computing is the product of the confluence of advances in three distinct areas: server hardware standardization, virtualization and other software advances, and IT management and practices. Without all three, enterprise cloud computing would not be possible.**

## The Moving Target that Is Cloud Computing

Given the speed at which IT changes, writing a definitive guide can be like designing and building a plane while flying in it. This is especially true of cloud computing. Public clouds are well established and private clouds are emerging as an alternative delivery model of cloud services. Identifying which existing applications are readily ported to the cloud while spotting others that are best run on existing platforms is an ongoing process. Applications are being built that take advantage of high-performance, distributed computing through the use of new programming paradigms and database designs. Vendors are revising their infrastructure management tools to support clouds. Cloud computing is a quickly moving target.

With the relentless pace of change in cloud computing technologies and practices, one might argue that it is too early and cloud computing too volatile to suggest a roadmap for understanding and adopting cloud computing. This argument has some merit, but its validity assumes we focus on low-level implementation details. Rather than try to define low-level best practices in this book (it is too early for that), we base this work on the principles and practices that IT professionals have long used to adapt and adjust to changing technologies and business conditions.

Change is nothing new to IT, and our past experience is a sound guide to understanding cloud computing. With that in mind, recognition of the following facts will guide the approach taken in this book:

- **Cloud technology will continue to evolve in intelligible ways.** We understand the current state of cloud technology and recognize that it is a product of earlier technologies.
- **Changes in cloud computing come from not just from changes in underlying technologies but also from the ways we combine and use these technologies.** Business processes, workflows, and cloud management will drive the way we combine cloud techniques.
- **The fundamentals of computing principles have not changed.** Basic building blocks of IT consist of computing, storage, and network resources. The underlying principles of serial and parallel computing have been known for generations. Design and management principles that have guided us in the past are still relevant.
- **Business services drive the adoption and continued use of cloud services.** Unless you are a computer scientist, cloud computing is a means to an end, not an end in itself.
- **In technology, as in the evolution of life, those that adapt what has worked well in the past to new conditions and find ways to build on those past successes to address novel challenges are rewarded.** There will be no single best model of cloud computing for all applications. The specific conditions and requirements of a service will shape the optimal use of cloud computing for that service.

Our goal in this book is not to prescribe precise regimens for implementing a specific cloud computing application. Instead, the objective is to provide the reader with a background in the underlying technologies and business practices of cloud computing along with a roadmap for moving from the theory to practice of cloud computing.

## A Brief Introduction to Cloud Computing

Cloud computing is a model for delivering information services that provides flexible use of virtual servers, massive scalability, and management services. With the dictionary definition out of the way, we can now proceed to describing cloud computing in terms of its essential features and how it functions alongside other information technologies. Cloud computing is a unique combination of capabilities which include:

- A massively scalable, dynamic infrastructure
- Universal access
- Fine-grained usage controls and pricing
- Standardized platforms
- Management support services

These capabilities enable a number of variations in cloud computing services. For example, one service might provide “raw iron” servers for running specialized applications, another offers on-demand relational database services, while yet another provides a fully-featured Customer Relationship Management (CRM) application.

**Cross Reference**

Chapter 2 will examine different types of cloud computing options in more detail; for now, we will restrict the discussion to features that are common to most cloud computing options.

### A Massively Scalable Infrastructure

If we had to choose one characteristic that most distinguishes cloud computing from other models, it is the massively scalable infrastructure. In theory, one has the potential for massive scalability without the cloud provided one has the financial resources to acquire and the skills to manage a massively distributed infrastructure. The cloud puts that kind of theory into practice.

Massive scalability from the service consumer perspective means the end user controls allocation of compute or storage services as needed. In the past, acquiring additional compute cycles required either procuring additional hardware, which could take weeks, or fitting jobs onto existing servers. Procuring new hardware has obvious time and cost drawbacks, but running jobs on other servers is far from a panacea. It is not uncommon to run into problems such as:

- Incompatibilities with the operating system (OS) or applications on the server
- Conflicts in the scheduling of workloads
- Difficulties allocating costs to owners of the jobs running on the server
- Irresolvable violations of security policies regarding access controls and data protection policies

These problems can occur when trying to share a single server across application or organizational boundaries let alone hundreds or thousands of servers that may be required for a compute-intensive job. The problems are avoided with cloud computing because of three characteristics of the technology:

- Rapid allocation of virtual servers
- Standardized hardware
- Persistent cloud storage

Together, these characteristics provide the benefits of sole use servers with the efficiencies of shared resources.

## Rapid Allocation of Virtual Servers

Cloud computing avoids these problems by decoupling physical servers from applications and single users. In the cloud, a user allocates the number and type of virtual machines needed to perform a task. The virtual machines run a task as long as required and then shut down when the task is complete. (Actually, the implementation details, such as whether a virtual machine is actually shut down or allocated to another job, are cloud-specific; logically, it appears to the cloud users that virtual machines are no longer allocated to them.) In a cloud, physical servers become shared resources without the drawbacks previously described. As Figure 1.2 shows, the distribution of jobs and number of virtual servers running on a set of physical servers can change quickly in a cloud.

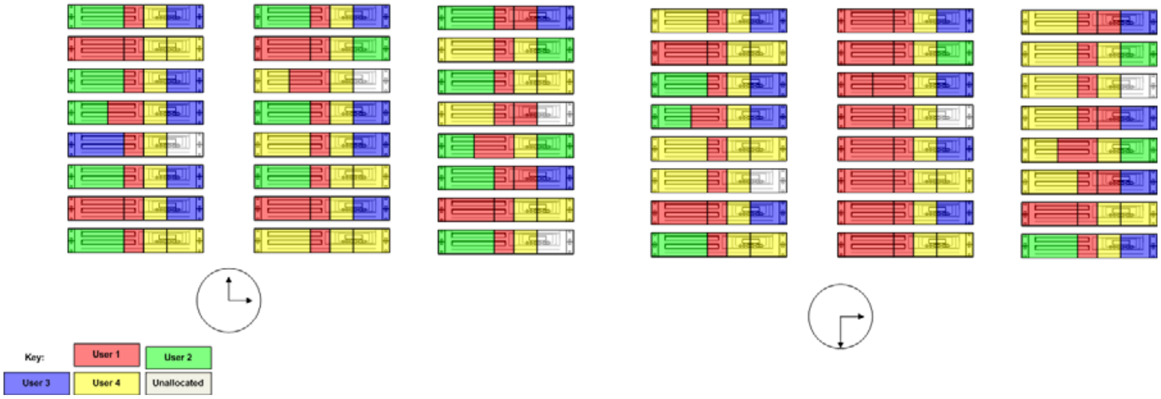


**Figure 1.2: Virtual machines are quickly allocated and deallocated to specific tasks in the cloud.**

Anyone who has waited hours or days to have an OS and application stack installed on a server may wonder how cloud computing servers can switch among uses so quickly. In a cloud, large numbers of physical servers are ready to respond to the specific requests for computing services. Often, these physical servers will support multiple virtual machines each dedicated to different tasks (see Figure 1.2).

Different cloud models require or support (depending on your perspective) different levels of configuration information from users. In a simple case, a user may only need to specify the number of servers she would like dedicated to her job. A slightly more complicated setup would require the user to specify a number of servers and the roles each server carriers out, such as a Web server role or application server role. Another model requires users to specify a specific virtual machine image to execute on each of the virtual machines requested. Regardless of which model is used, clouds can rapidly allocate virtual machines in response to the computing needs of users.

### Standard Hardware Platform

Another enabling characteristic of cloud computing is the use of standard hardware platforms, such as the x64 architecture. By standardizing on hardware, applications and OSs can run on many combinations of servers within the cloud without incurring additional overhead required to manage many different types of servers. Cloud providers may offer different levels of computing services by offering the functional equivalent of different physical configurations, such as:

- Basic server: 64-bit, 2 cores, 2GB of memory, and 320 GB of local storage
- Midsize server: 64-bit, 4 core, 8GB of memory, and 320 GB of local storage
- Advanced server: 64-bit, 8 core, 16GB of memory, and 1 TB of local storage

In practice the cloud provider may have all 64-bit, 8 core, 16GB of memory servers but will vary the number of virtual machines to accommodate the mix of services requested by users.

### Persistent Storage in the Cloud

Rapidly allocating and deallocating virtual machines allows for efficient allocation of computing resources, but many of the computations run on these servers will generate data that must be stored for extended periods of time. It is useful to have local storage on servers for temporary needs, but once the virtual server is deallocated, any locally stored data would be lost.

With persistent cloud storage, data is stored and made accessible to any server in the cloud, subject to access control restrictions. Decoupling persistent storage from servers is another way cloud computing provides for fine-grained control over resources. The combination of rapid provisioning of standard hardware and the use of persistent storage enable massive scalability.

**The Potential Network Bottleneck**

Three types of resources are fundamental to cloud computing: computation, storage, and networking. Technology is in place now to enable massive scalability of compute servers and storage capacity; the same cannot be said for network resources.

Within a cloud infrastructure, a cloud service provider has control over the network architecture and resources. If additional bandwidth is required to maintain service levels, cloud providers are in a position to make those changes. Problems potentially can arise when moving data into and out of the cloud. This is especially the case when there is an initial, large data upload from an existing non-cloud storage system. It can also occur if large volumes of data are generated rapidly and must be moved to the cloud.

In the case of private clouds, a single company would control the cloud infrastructure and the network resources between the source of the data and the cloud. Public clouds depend upon public network infrastructure, and that can vary widely. Figure 1.3 shows the wide variation in average national broadband speeds. Although businesses may have the resources to purchase additional bandwidth, these figures demonstrate the limits of large-scale public network infrastructure in different regions.

One way to mitigate the problem of the large initial data load is to physically ship storage media to the cloud provider. This may not be a viable option for repeated use; another option is to generate and store data in the cloud, avoiding the need to use public network infrastructure.

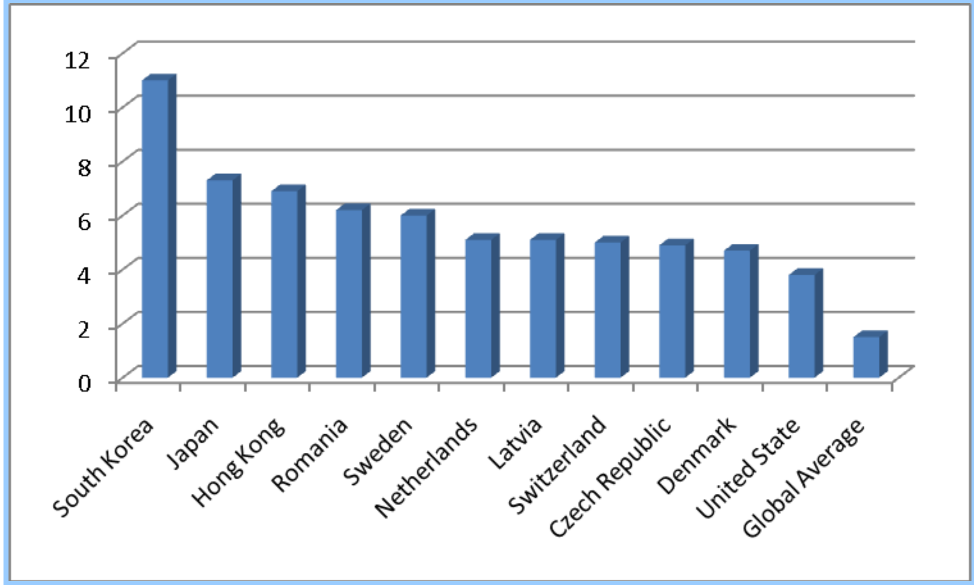


**Figure 1.3: Average national broadband speeds (Mbps) vary widely by region (Source: [The Akamai State of the Internet Report 2nd Quarter 2009. Volume 2 Number 2](#)).**

## Universal Access

Another defining characteristic of cloud computing is universal access from anywhere on the Internet. Today, we have universal email access over the Internet, although it was not too long ago that proprietary email systems required local network connections or virtual private network (VPN) access to use our email. Similarly, access to cloud computing resources can leverage Internet protocols to ensure widespread access.

Universal access should not be confused with open access, especially with regard to private clouds. Companies and governments deploying private clouds will have authentication and authorization systems in place to control access to private cloud resources. Even public clouds require some degree of identity management in support of management reporting and billing.

## Fine-Grained Usage Controls and Pricing

The economic benefits of cloud computing are one of the key drivers to adoption. One of the features that enable this benefit is fine-grained usage controls and pricing.

When we purchase servers, we pay up front for a substantial resource with approximately a 3-year useful lifespan and some residual value at the end of that period. Trying to optimize purchase decisions at this granularity is difficult because the ROI depends on many difficult-to-gauge factors, like the load on the system over the life of the server, which will vary with changing business conditions and requirements. If we undersize a server, we risk not meeting SLAs. If we opt for excess capacity, we incur unnecessary costs. Cloud computing can adjust the compute and storage services as application demand dictates.

Cloud computing models allow us to purchase compute resources based on the mixture of jobs that need to be done now. Similarly, we purchase and pay for storage based on what is actually needed now. We no longer have to make purchase decisions based on single server considerations, such as peak capacity requirements. During period of peak demand, we provision additional resources from the cloud and release them when the demand is met and pay only for what is used.

## Standardized Resources

Cloud computing provides standard hardware, virtualization, and application platforms. Standardization, however, is not homogenization. There is room for a range of options in cloud computing. For example, a cloud can provide a few different configured servers, a couple of different OSs, and several different application stacks to choose from, such as Linux or Microsoft OSs and LAMP (Linux, Apache HTTP Server, MySQL database and Perl/Python programming languages) or Microsoft .Net Framework application stacks.

By limiting the range of options, cloud providers avoid excessive management and maintenance expenses and keep the marginal costs of expanding the cloud to a minimum. This, however, has to be balanced with business requirements that may justify a greater range of customization.

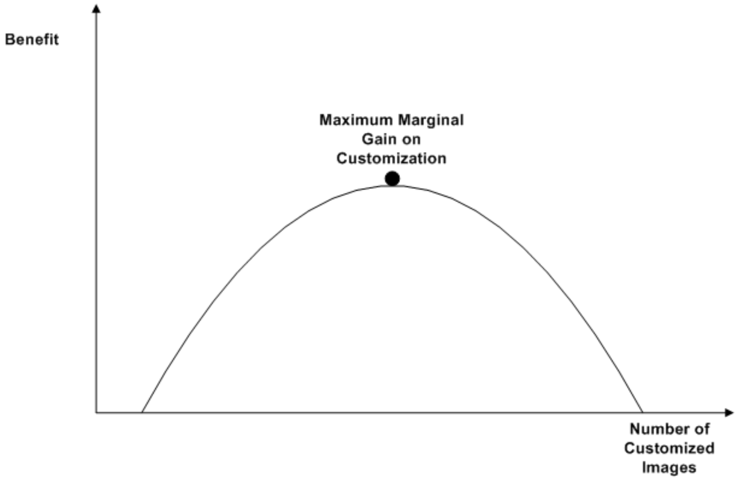


**Figure 1.4: At some point, increasing customization of images incurs additional management costs and an associated decrease in marginal benefit.**

### Management Support Services

Cloud computing is not a complete service without management support services. These services support both operational and management aspects of the use of cloud computing. Operational support services enable cloud users to provision the resource they need without additional support from IT staff. They include:

- Provision servers
- Search and select virtual images to run on server instances
- Allocate persistent storage
- Monitor jobs executing on allocated servers

Management reports are especially important for managing costs. These include reporting on:

- Time periods and number of servers allocated
- CPU utilization
- Storage use
- Network bandwidth consumed to upload and download data to and from the cloud

Management support services provide the information needed to refine the use of cloud services. For example, CPU utilization reports may indicate low utilization in jobs that have been spread over more servers than necessary. Storage reports and network bandwidth use reports might help identify jobs that involve transferring data into and out of the cloud at a cost greater than using persistent storage services to store that data in the cloud. Cloud computing services are not complete without this type of management support services.

This brief introduction has just scratched the surface of key aspects of cloud computing, such as massive scalability, universal access, fine-grained usage controls and pricing, standardized platforms, and the role of management support services. More details on these topics are provided throughout the rest of this book, but before we delve further into technical details, we will turn our attention to the drivers behind cloud computing adoption.

## Drivers Behind Cloud Computing

Cloud computing changes the way we consume and provide services and in the process improves the user experience. The combination of technologies described in the previous section enable these drivers but are not the drivers to adoption themselves.

## A Better Way to Consume Services

The early days of IT were dominated by monolithic applications that performed a series of related tasks in a fixed order. Applications processed accounting transactions to balance the books, calculated payroll for the company, and generated monthly statements for customers. This approach worked well, and still works well, for some business requirements, but it does have some drawbacks:

- Isolating specialized functions that might be useful in other applications
- Utilizing a fairly rigid flow of execution making it difficult to adapt to emerging requirements
- Offering few options to vary service levels according to varying needs

Cloud computing readily supports service-oriented architectures, which can provide a better way to consume services.

### Service-Oriented Architecture in the Cloud

Service-oriented architectures use loosely coupled services to deliver functionality. Each service is implemented in a way that does not require or depend upon knowledge of the way the service is used. For example, service to calculate the credit risk of a customer could be used by a customer sales portal as well as a back-office risk analysis application. Service-oriented architectures exchange data and invoke services standards such as Simple Object Access Protocol (SOAP) and frameworks such as Representational State Transfer (REST).

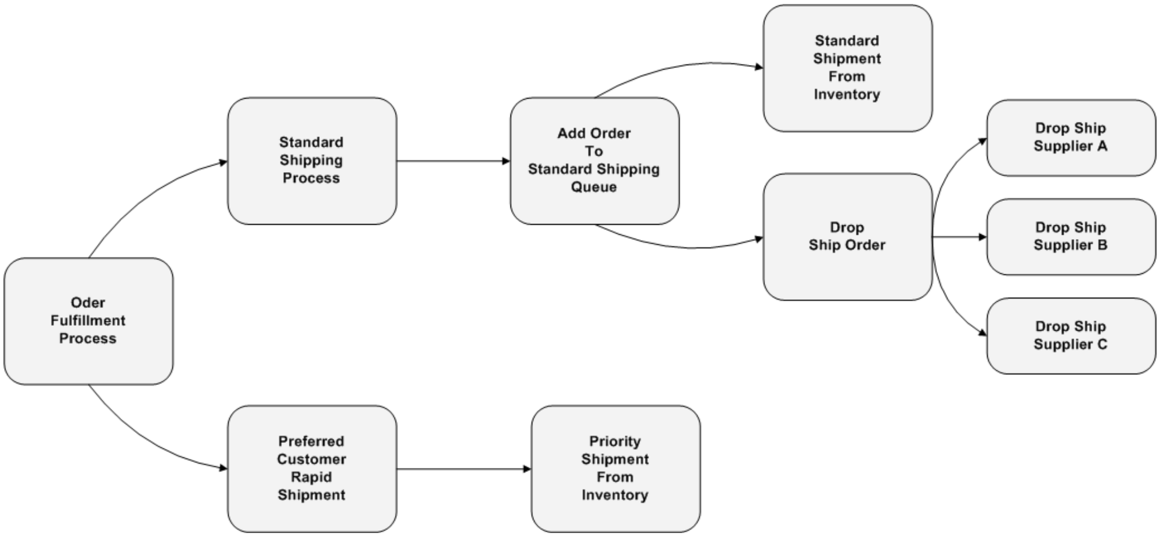

Oder Fulfillment Process
Standard Shipping Process
Add Order To Standard Shipping Queue
Standard Shipment From Inventory
Drop Ship Order
Drop Ship Supplier A
Drop Ship Supplier B
Drop Ship Supplier C
Preferred Customer Rapid Shipment
Priority Shipment From Inventory

**Figure 1.5: Services orchestration combines loosely coupled services in a flow of execution designed to complete a logical unit of work.**

By implementing a service-oriented architecture in the cloud, customers can consume only the services they need for as long as they need them and be billed only for that use. The same level of fine-grained control over resource use that the cloud provides at the level of servers and storage is available at the services level as well.

### Differentiated Levels of Service

The cloud model of computing also supports differentiated levels of service. Customers can choose the appropriate level for their needs. For example:

- A customer executing an online transaction processing application (OLTP) may need high throughput and rapid response times. This warrants a number of high-end servers with a single virtual machine instance running the customer's OLTP application.
- A marketing analyst data mining the results of several campaigns may be willing to have a longer turnaround time in return for running her application on a lower-cost low-end server.
- A team of developers performs continuous integration testing every night and needs guaranteed delivery of output at the start of the next business day. The jobs can run at any time during the night as long as the there are sufficient server resources to complete the job in time. The job could be allocated to low-end servers early in the night, or if demand for those is high, can run later in the night but on a number of higher-end servers.

Cloud computing enables customers to define the level of service they require, which in turn, allows the cloud provider to optimize workloads across customers and cloud infrastructure.

### More Efficient Delivery of Services

There are a number of ways to exploit the fine-grained controls over compute, storage, and higher-level services in cloud computing to make service delivery more efficient. Some of the most important are:

- Management infrastructure
- Optimization of workloads across shared infrastructure
- Self-service management
- Monitoring

These support services prove to be beneficial for both cloud consumers and providers.

### Management Infrastructure

Both public and private clouds support a large pool of potential customers with a wide range of diverse service requirements. Cloud computing supports these requirements with a well-defined set of basic service components, so a comprehensive management structure can be built on a small number of management services, such as:

- Tracking customer use of virtual servers in terms of number of servers and time used by server
- Tracking the amount of persistent storage used by customers for a given period of time
- Accounting for the data transfer into and out of the cloud
- Accounting for data transfer within the cloud
- Tracking the use of licensed software

This type of management reporting enables cloud providers to bill customers for resources used. Providers can help customers optimize their use of the cloud by providing near real-time updates on their resource utilization as well as aggregate billing and charge-back reports.

Cloud computing introduces new opportunities for software vendors to change how they price their software. Named user and number of user-based pricing schemes will fit well with cloud computing, but CPU or core-based pricing methods are problematic. A highly-parallelized application might run for 10 hours on a single server or in 1 hour on 10 servers. If the software were licensed to run only on a single server, the customer will lose a significant advantage of cloud computing. Expect vendors to experiment with new pricing models for enterprise software as businesses adopt cloud computing.

### Optimization of Workloads Across Shared Infrastructure

A large server farm is indistinguishable from a set of cloud servers when looking at the hardware. Servers, switches, routers, power supplies, and other components are the same. The difference lies in how these resources are used.

The servers in a typical corporate data center prior to the advent of cloud computing were assigned to a particular department or application use. The configuration was relatively fixed and changed only when the server was upgraded, reassigned, or decommissioned. These servers were configured to do one type of operation. This makes for a reliable compute resource, but not an efficient one.

Servers with fixed configurations are less likely to have high-utilization rates. Unless there is a steady stream of jobs that fits the machine's configuration, there will be idle periods. Without proper infrastructure for rapidly deploying virtual machines, the cost of reconfiguring a server is so high that it is done only for significant long-term changes. In the cloud, the cost of switching virtual machines is low enough that idle servers can be reconfigured with different virtual machine images allowing other applications to run on the same physical server that had just been running other types of jobs.

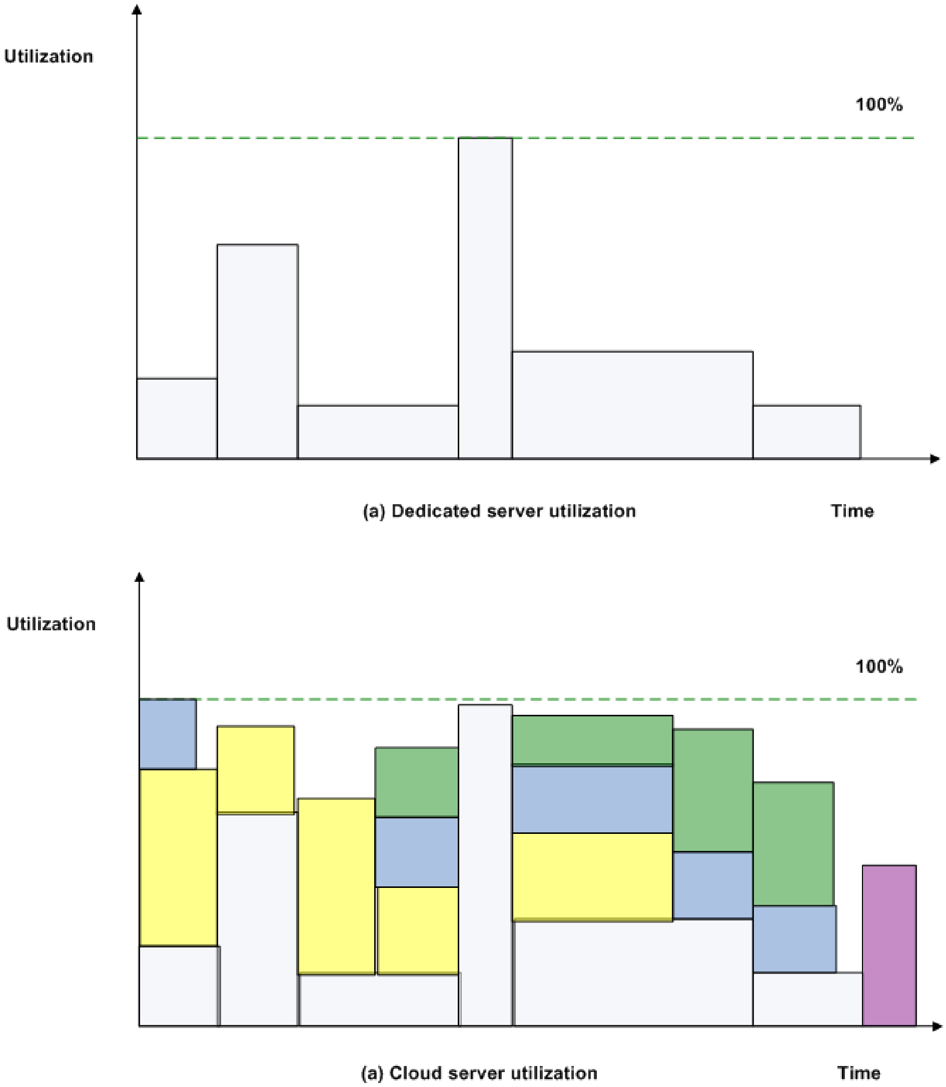


**Figure 1.6: In the cloud, server utilization can be significantly higher when workloads are distributed and optimized over available servers.**

### Self-Service Management

In cloud computing environments, provisioning and other management tools are made available to cloud service consumers. This lowers the cost of delivering cloud services by eliminating or significantly reducing the need for IT professionals to complete allocating and deallocating operations. Self-service management also eliminates IT staff availability as a potential bottleneck to using the cloud. Cloud consumers have the tools they need to acquire and use cloud resources themselves.

### Monitoring

The state of the cloud will frequently change. New images are loaded into some servers to execute jobs while other virtual server instances are shut down when jobs complete. It is important to monitor both the availability of servers and the workloads running in the cloud. After a server has been deprovisioned, it should be quickly allocated for a new job to maintain maximal utilization rates.

The combination of management infrastructure, optimization of workloads across shared infrastructure, self-service management, and monitoring of cloud resources creates a key driver behind cloud adoption—the more efficient delivery of services.

### Improving the User Experience through Cloud Computing

Another driver behind cloud computing is that it can improve the end user experience. As noted earlier, cloud service consumers have more direct control over the resources they use. Simplified, Web user interfaces makes this possible.

Users are also relieved of long-term management issues when using cloud services instead of dedicated servers. Concerns such as scheduling patches, ensuring security policies are enforced, performing backups, and developing a disaster recovery plan are addressed by the cloud service provider. Users are free to focus less on maintenance and more on core business issues.

Cloud computing also improves the user experience by lowering the barriers to experimenting with data or a new business process. For example, a marketing analyst might have an idea for increasing market share for a product in a particular region. Evaluating her idea requires a substantial amount of data and compute resources. The sales data warehouse makes use of cloud storage, so the data is readily available and provisioning servers is a simple matter with the cloud's Web interface. Without cloud computing resources immediately available, the cost of procuring or borrowing servers to run this job may have been so high that it was not done.

Cloud computing changes how we consume services, how we deliver services, and the way end users experience the use of these services. These three factors are fundamental drivers behind cloud computing. There are, however, other economic factors involved as well.

## Changing Economics of IT

The economics behind cloud computing make a compelling case for adopting this approach to delivering services. The economic benefits can be seen in at least three areas:

- Reducing capital expenditures
- Efficiently allocating resources
- Rapidly delivering IT services

A common thread among all three areas, as we will explore in a moment, is that cloud computing allows us to share computing infrastructure in a way not previously possible and, in the process, realize efficiencies unrealized until this point.

## Reducing Capital Expenditures

An obvious economic advantage of cloud computing from the consumer perspective is the reduced need for capital expenditures. Consumers of compute and storage services do not have to procure the underlying hardware that enables those services. Rather than follow a "pay up front" model, cloud service consumers follow a "pay as you go" model. The "pay as you go model" is especially advantageous when a consumer would have to purchase servers and storage to accommodate peak capacity but that peak capacity is needed for only relatively brief periods of time.

Consider the following example. An online analytic processing (OLAP) application generates weekly business intelligence reports that require a number of high-end servers to perform all calculations in the time allotted to the process. In this scenario, the servers are underutilized most of the time; nonetheless, in the dedicated server approach to consuming compute services, we have to plan for and purchase for peak demand. A better option is to use the elastic scalability of the cloud to provision the servers when they are needed and release them when the reports are complete.

## Efficiently Allocating Resources

Cloud computing more efficiently allocates compute and storage resources than dedicated server approaches. The source of the efficiency stems from several factors:

- Ability to manage workloads and allocate jobs to available servers through the use of rapidly deployed virtual machine images to servers with excess capacity
- Ability to share storage resources and realize the economies of scale with regards to centralized storage services
- More efficient support operations, such as backup and recovery; rather than manage many different types of backup jobs that vary according to the needs of dedicated servers, cloud providers can consolidate backup operations of centralized storage
- Clouds can be configured to use geographically distributed data centers and replication services between the data centers to provide disaster recovery for all cloud consumers; under the dedicated server model, we must plan for disaster recovery separately at the department or project level
- High availability of service without significant overhead—if a server were to fail in the cloud, it could simply be removed from the pool of available resources; jobs would continue to run on other servers; in the dedicated server model, a stand-by server would be needed to act as a backup for each primary server

- More efficient patch management—when servers have relatively fixed OSs, each system must be individually patched to keep up to date with security and performance patches; under the cloud model, virtual machine images stored in a centralized catalog can be patched and when new instances of virtual machines are started, the patched images are deployed
- Increased self-service with regards to procuring servers and storage reduce demand on IT personnel
- More efficient server utilization requires few servers which, in turn, leads to lower hardware costs and power consumption

As these examples show, efficiencies arise both from more efficient allocation of IT assets and of IT personnel. For consumers of cloud services, this translates into more direct control over how they use services and that can translate into more efficient business operations.

### Rapidly Delivering IT Services

With a cloud, businesses can more rapidly deliver services to meet changing business requirements and market conditions. Once again, there is no single part of the cloud model that enables this; instead, it is a combination of factors.

Once again, the ability to rapidly provision and deprovision compute and storage resources is important. If demand for a service were to rapidly spike, for example, for a retailer during the holiday season, servers can be added to scale to meet demand.

Another consideration is the ability to expand the range of functions provided by IT applications. In this case, service-oriented architectures are well suited for rapid reconfiguration of applications through service orchestration (see the earlier discussion of service-oriented architecture in the cloud). Functionality developed for one application and delivered through the cloud using service-oriented architecture can be readily adapted to other applications as well.

The economic benefits of cloud computing emerge in different ways, including a reduction in the need for capital expenditures, more efficient allocation of resources, and the ability to rapidly deliver and adapt IT services. The efficiencies enabled by the reduced time and cost of cloud computing will be maximized only if business strategy is aligned with IT services.

## Aligning Business Strategy and IT

IT serves the strategy of the business, but keeping business objectives and IT operations in alignment is not always easy. We may have a clear business strategy mapped to detailed business processes that are ready to implement but still the execution stumbles. Why? One reason is that the information systems needed to execute the strategy are insufficient or poorly matched to the requirements. Cloud computing and service-oriented architectures can mitigate the risk of such misalignments, assuming they are used in ways supportive of business strategy.

Aligning business strategy and IT services is a several-step process, at least at the most coarse level:

- Identifying key business objectives
- Identifying IT services needed to support those objectives
- Assessing the current state of IT services and identifying gaps between the existing set and the needed set of IT services.
- Developing a plan for reducing the gap between the existing and needed set of information services

Key business objectives may include controlling and reducing costs, enabling more rapid response to changing market conditions, improving governance of the organization, or improving the resiliency of IT operations to adverse events, such as hardware failures, loss of power, or natural disaster. Many of the services needed to support business objectives can be readily identified once the business objectives are known. Cost controls and cost reduction come with more efficient server utilization, more self-service in systems management, and reduced overhead associated with infrastructure services such as backups, high availability, and disaster recovery.

The gap analysis process should take into account both technical and organizational considerations. For example, will existing hardware readily deploy in a cloud architecture or will new hardware be required? Are service management practices mature enough to implement in self-service delivery systems? Is a billing or chargeback mechanism in place if a private cloud is under consideration?

The first steps in creating a plan to move from the existing to the needed systems are to prioritize the gaps and identify dependencies in the process. This is certainly not a trivial process, but we will delve into a more detailed examination of the full alignment process in Chapters 5 through 7.

## Summary

Cloud computing is a model of service delivery that is enabled by a confluence of advances in hardware, software, and business processes. The availability of standardized servers capable of running multiple virtual machines, standardized virtual machine images for delivering complete application stacks to servers on demand, and mature service management practices that lend themselves to a significant level of self-service all contribute to enable cloud computing.

Cloud computing is different from other approaches to service delivery because of its unique combination of attributes, including:

- A massively scalable, dynamic infrastructure
- Universal access to services from any Internet-enabled device
- Fine-grained usage controls and pricing that allow for more efficient delivery of services
- Standardized platforms that lend themselves to lower procurement and operational costs
- Management support services for service consumers to control their use of cloud resources

Being able to build with these characteristics is not sufficient to warrant widespread adoption by business; there have to be additional drivers behind the technology. There are several business drivers behind cloud computing:

- Cloud computing offers an efficient way to deliver services
- Cloud computing coupled with service-oriented architectures improve on ways to consume services
- Cloud computing improves the end user experience by making it easier to work with services and apply them to new opportunities

In addition to these business drivers, there are compelling economic arguments for adopting a cloud model, such as reducing the need for capital expenditures and efficiently allocating compute and storage resources. Cloud computing is especially beneficial when aligned with business strategy to cost-effectively and rapidly deliver essential services.

In the next chapter, we will turn our attention to demystifying different types of clouds and their characteristics.

# Chapter 2: Demystifying Cloud Computing

The term “cloud computing” has become a shorthand way of describing a wide range of different computing services. When describing their cloud offering, a vendor might focus on the ability to rapidly provision instances of virtual machines to run applications of your choice. Another vendor might use the term “cloud” when promoting a new way to license and run the vendor’s applications on the vendor’s servers. Of course, there are any number of definitions in between.

The goal of this chapter is to demystify cloud computing by defining a set of common characteristics that should be included in any cloud service that could be considered ready for enterprise use. The common characteristics, as we shall see, still leave plenty of room for different types of cloud computing. We will examine several types of cloud services and the advantages and disadvantages of each. The chapter concludes with a discussion of different cloud delivery models that range from public to private clouds.

## A Note on Terminology

As noted in the first chapter, the types of computing services we are describing represent an evolution of information technology and service delivery. The elements of cloud computing are not radically new, but we are using and deploying them in new ways. This can sometimes lead to confusion in terminology.

Consider, for example, the term “provisioning.” In the past, provisioning a server almost always meant that a physical server was acquired, configured, and deployed to an organization’s network. The term still has that meaning, but it is not the only way the term is used when describing cloud computing. Provisioning can also mean creating an instance of a virtual machine, for example, to run a job in the cloud for some period of time after which the virtual machine is shut down.

The reason we use the same term for different processes is that both apply to making a computing resource available to a specific task. The key differences are tied to physical versus virtual servers, the duration for which the server is assigned to a specific task, and the time required to make the server available. (These difference underlie the efficiencies cloud computing introduces; however, before we can realize those efficiencies, we need to be clear about all the variables that are at work with services delivery. This chapter will make those variables clear.)

Throughout this chapter and the rest of this book, we will use explicit descriptions, distinguishing, for example, provisioning a physical server from provisioning an instance of a virtual machine. The text will also distinguish models of persistent storage when discussing databases. Relational databases are alive and well in clouds, but they are by no means the only database model available. "Systems management" is another term that is adapting to accommodate new tasks that application administrators are expected to handle when working with clouds.

Describing fundamental characteristics of cloud computing is a step to demystifying this new way of delivering services.

## Searching for a Common Definition: 3 Fundamental Elements of Cloud Computing

Reasonable people can disagree about precise definitions of new technologies. We will forgo well-constrained definitions of cloud computing and instead consider three characteristics that are required to deliver the types of services most of us have come to expect from cloud computing:

- Massive scalability
- Ability to easily allocate cloud resources
- A service management platform

There are other characteristics, such as security, that are entailed within these three and will be discussed shortly. Massive scalability, the ability to easily allocate cloud resources, and a service management platform are essential constituents of a cloud computing service.

### Massive Scalability

Massive scalability is the ability to rapidly allocate large amounts of computing resources on demand. This is not scalability in the sense of purchasing hundreds of servers, waiting for them to be delivered, configured, and deployed. Massive scalability in cloud computing is the ability to deliver significant resources in a matter of minutes, not days or weeks.

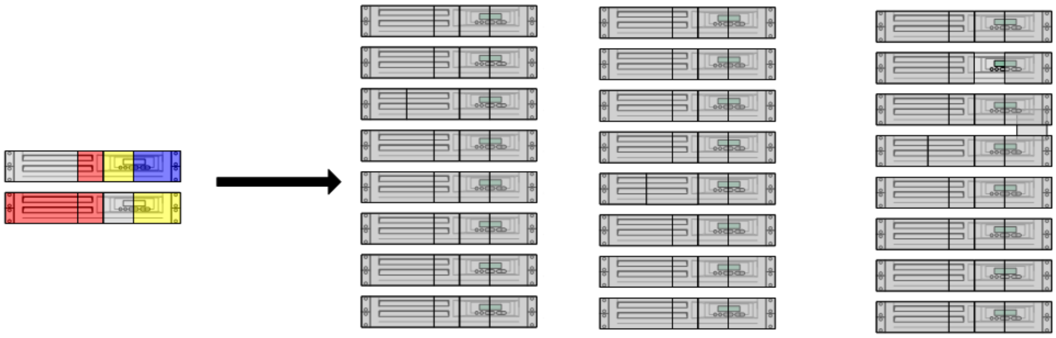


**Figure 2.1: Massive scalability provides the ability to rapidly increase the amount of allocated cloud resources as needed for a job.**

Three types of resources should be available:

- Computing resources
- Storage resources
- Network bandwidth

### Computing Resources

Computing resources are the means to process information. If there were a single workhorse in cloud computing, this would be it. Computing resources are provisioned for a cloud computing task in different ways, depending on the cloud model. At minimum, there is a smallest unit of computing resource that is allocated. This could be, for example, a virtual machine equivalent to an x64 architecture, 2GHz CPU dual-core processor with 32GB of memory and 300GB of local storage. Specifications such as this should be considered a logical specification. The virtual machine running jobs could be hosted on any of a number of physical implementations. This is one of the advantages of cloud computing: The details of the physical implementation are abstracted so that the consumer of cloud services does not have to concern themselves with such details.

Abstracting computing services can also lead to more efficient delivery. For example, a cloud provider can:

- Vary the amount of hardware running at any time according to demand—During periods of peak demand, many large servers may be running while during low demand periods, only the most energy-efficient servers are kept powered on.
- Run jobs in different data centers to better allocate work load—This functionality is constrained to some degree by business requirements. For example, businesses subject to European Union (EU) privacy directives may require that all personal information on EU customers be kept in countries that meet a minimum level of privacy protections.
- Execute workloads on physical servers that minimize the distance between the compute resources and the storage resources

Cloud service providers all abstract some level of implementation details, but that level can vary significantly. Consider a few different scenarios.

## The (Near) Raw Iron Approach

One cloud provider allows consumers to select a type of virtual machine (types vary by number of cores, amount of memory, and so on) and the virtual image to run on that machine. There may be several operating systems (OSs) to choose from as well as a variety of application stacks. This model has the advantage of giving cloud consumers a wide range of options but at the cost of additional configuration responsibilities. For example, a cloud consumer may have the option to configure and run a particular statistical analysis package on a preferred version of Linux with this provider, but she is also responsible for tuning and patching this image.

## The Server Role Approach

A second cloud provider may limit the range of options in return for a simplified deployment model. Rather than allow customers to build their own virtual machine images, the cloud provider may offer a small set of preconfigured images designed for specific roles, such as load balancing, running a Web server, or providing application services. Under this approach, cloud consumers could define the number Web servers they need and the number of application servers required without having to concern themselves with OS or application stack details.

## The API Approach

Another approach a cloud vendor may provide is a general computing platform that abstracts even basic distinctions such as Web servers and application servers. Under this model, cloud consumers develop applications that use a cloud provider's application programming interface (API), which might include, for example, functions for:

- Defining data structures
- Creating associative arrays (key-value pairs)
- Specifying queries
- Implementing transactions
- Utilizing task queues

When the application is run, the cloud consumer need only specify the number of servers to dedicate to the task. By limiting the range of options for implementing an application, the cloud consumer has fewer systems management issues to address. As Figure 2.2 shows, there is a tradeoff between flexibility and systems management responsibility.

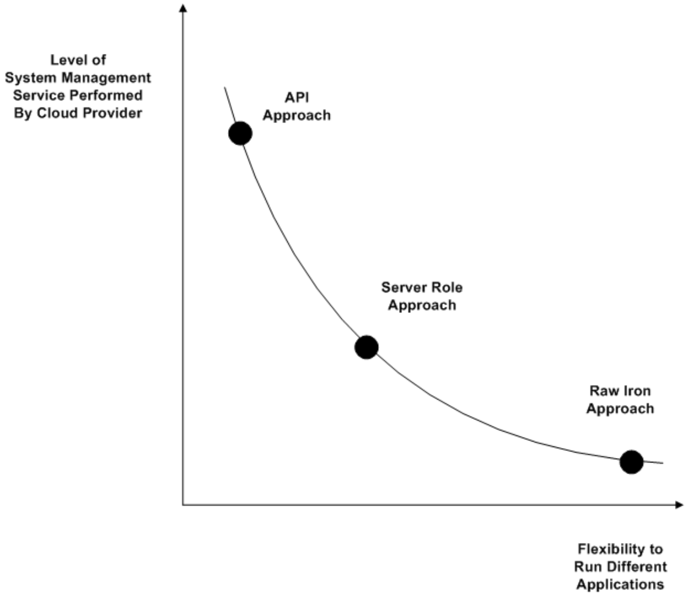


**Figure 2.2: Cloud consumers have a range of options that balance different levels of flexibility with the need for systems management tasks.**

### Storage Resources

Massive scalability implies the ability to persistently store raw data and computed results. For the cloud consumer, the amount of storage needed at any point in time should rapidly scale to meet demand. As with computing resources, storage resources should be allocated as needed and there should always be storage available.

Cloud storage is available in a few forms:

- As file-based storage
- As block-based storage in which arbitrary large objects are stored
- As relational storage in which data is maintained in relational database structures

Data stored in the cloud is manipulated in much the same way as it is in non-cloud architectures with some minor differences. Block-based storage may be accessed via URL. Relational data is queried the same in or out of the cloud, but database administrators will have less to manage with regards to the physical allocation of space and replication of data for high availability.

### Network Resources

The ability to move data from compute to storage resources must scale along with those resources. Within the cloud, the network capacity and infrastructure is defined and managed by the cloud provider. Providers can reasonably plan for moving data from servers to storage arrays or replicating data between storage devices. The situation changes when data has to move into or outside of the cloud.

Cloud service providers are more constrained in their ability to deliver network scalability because of dependence on the outside networks. Cloud consumers transfer data into and out of the cloud using whatever network services they have acquired. This may or may not be sufficient for the volumes of data that need to be transferred. In response, some cloud providers offer “sneakernet to the cloud” services: physical storage devices are shipped to the cloud provider where they are uploaded to the cloud.

Part of optimizing cloud-based services is determining the best way to move data into and out of the cloud and minimizing transfers outside the cloud. The network bottleneck is one reason to generate, process, and store data in the cloud as much as possible.

Massive scalability is a fundamental characteristic of cloud computing. Cloud providers offer different approaches to providing computing resources that tradeoff between flexibility in applications that can run in the cloud with demands on cloud consumers to manage system resources. Similarly, massive storage scalability is fundamental to cloud computing. At this point in time, networking resources outside the cloud are a potential bottleneck to moving data to and from the cloud.

### Ability to Easily Allocate Cloud Resources

Cloud computing can significantly reduce the need for systems administration support by providing easy-to-use tools for allocating cloud resources. One of the advantages of abstracting many implementation details is that it allows for greater automation of the cloud resource provisioning. As noted earlier, cloud providers offer different levels of abstraction of services, but in all cases, the provider should offer tools that enable application administrators the ability to adjust the usage as demand dictates.

Consider a simple example. A marketing analyst has just acquired several large data sets on product sales over the past several months. This is a onetime task and the analyst needs to aggregate the data for business reporting as well as run some statistical analysis programs over each data set. Outside the cloud, the analyst would need to perform several time-consuming steps:

- Find a department server with availability and convince the owner to allow the jobs to run on that server.
- Next, assuming a server is found, the analyst would then submit a ticket to systems administrators to install the necessary analysis software.
- When that is done, which could be a few hours to a few days depending on the IT support backlog, the analyst would need to upload the data. If the data is compressed, additional storage will be required to store both the compressed and decompressed files until the decompress operation is complete.
- Run the analysis jobs. This is a compute-intensive job, so the time to complete it will depend on the number of CPU resources available. If the analyst was provided with a virtual server running on a host with several other virtual machines, the workloads on the other virtual machines can adversely impact the data analysis job.

The same process in the cloud is significantly less arduous.

- Select a virtual machine image to run on cloud servers from a catalog of images. These can range from OS-only images to complete development or analysis environments.
- Specify the number of the virtual instances to run. In some cases, cloud vendors may offer options on the size of servers (for example, small, midsize, high-end), in which case, the size would need to be specified as well. As multiple servers are available, the analysis job can be subdivided into smaller jobs and run in parallel.
- Load the data into cloud storage and decompress if necessary.
- Run the analysis jobs.

These steps would be performed in a Web browser using a resource management interface by the analyst. There is no need for specialized IT support, no need to search for a server with available capacity, and no need to allocate disk space to a file system. The combination of massive scalability and easy-to-use interface to allocate resources provides two of the three core elements of cloud computing. The ability to manage services is the other.

## Service Management Platform

Once we move beyond simple scenarios like the one previously described and start to consider enterprise-scale management issues, the need for a services management platform becomes clear. A cost-effective cloud service will offer a management platform that supports four aspects of service management:

- Support for automated provisioning and deprovisioning of resources
- Self-service interface
- A service catalog of standardized services
- Policy definition and enforcement

Support for automated provisioning and deprovisioning and the self-service interface were covered in the previous section, so we will focus our attention on the other topics here.

### Service Catalog of Standardized Services

A service catalog introduces consistency and reusability to the cloud. A catalog includes virtual machine images that can run within the cloud with minimal setup on the part of the cloud consumer. These images capture design patterns that have worked well in other use cases.

For example, a basic Web server service might include the latest version of the Apache Web server, a fully patched and hardened Linux OS, and a properly configured firewall. Another image in the service catalog could provide an extraction, transformation, and load (ETL) application for use with data warehousing applications. With the ability to instantiate a fully functional ETL system in a matter of minutes using a self-service interface, the barriers to entry to business intelligence and data analytics is significantly reduced.

### Policy Definition and Enforcement

A service management platform can ensure that operations in the cloud comply with organization policies. Technical policies can address issues such as:

- Authentication and authorization required to use resources
- Resource limits, such as the number of concurrent virtual servers a user can have instantiated during peak load periods
- Pre-instantiation checks, such as ensuring images are properly patched before executing or virtual machines use currently approved versions of supported OSs

Organizational policies can be enforced as well. These include:

- Adjusting the cost of using resources according to demand—This could be implemented with a policy of peak load pricing or bidding based spot pricing.
- Prioritizing workloads in the event sufficient resources are not available during peak demand periods
- Controlling the number of instances of a particular application that is running at any one time—This is would be used to ensure compliance with software licensing agreements

A service management platform is essential to reducing labor costs associated with delivering information services. It enables non-IT professionals to allocate resources they need when they need them while still ensuring organization policies are followed.

## A Cloud by Any Other Name

Cloud computing has the potential to significantly reduce costs and improve the delivery of business services. It is no wonder vendors would want to offer something in this area. Simply calling a service offering a "cloud" is not enough, at least for *The Definitive Guide to Cloud Computing*. This guide has and will continue to argue that cloud computing entails massive scalability, easy to allocate resources, and a service management platform that includes a service catalog. These three elements are essential to offering a viable cloud computing service in an enterprise.

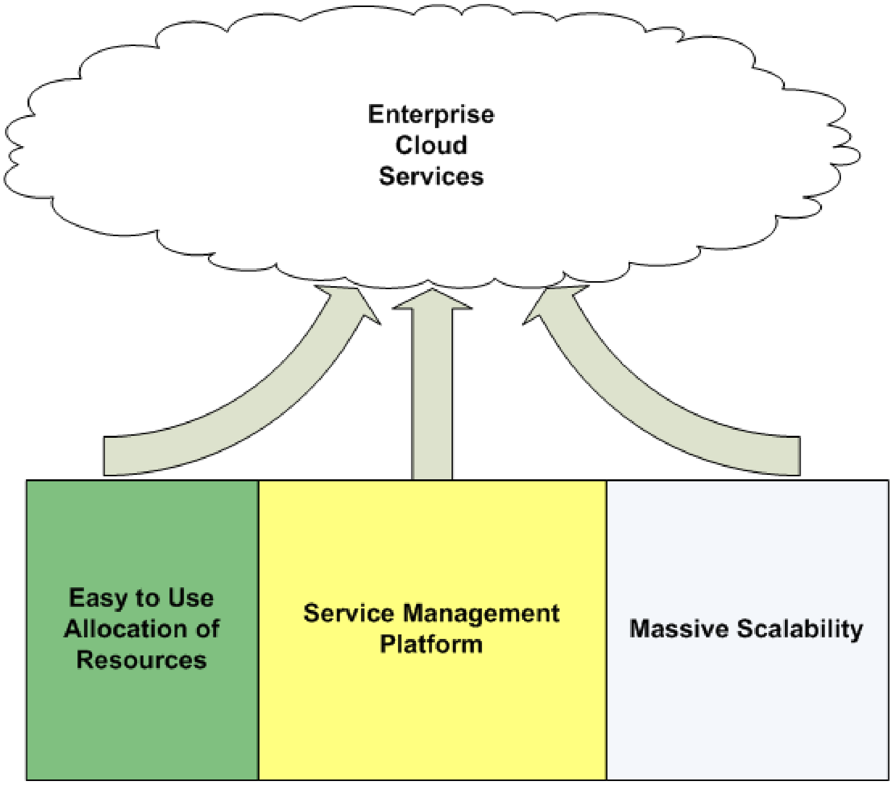

Enterprise Cloud Services
Easy to Use Allocation of Resources
Service Management Platform
Massive Scalability

**Figure 2.3: Cloud computing requires three fundamental elements to be effectively used in enterprise computing.**

If any doubts remain, consider if any one of these three characteristics were missing. Without massive scalability, there would not be the resources required to meet fluctuating demand. Cloud consumers would have to have backup resources in place in case cloud resources were not available. Traditional service delivery models would continue to exist and undermine the cost benefits of cloud computing. Without easy provisioning, cloud consumers would still have to depend on IT support, creating the potential for backlogs and driving up labor costs. Without a service management platform, cloud consumers would not have a well-managed service catalog, the lack of which would drive up costs of creating and maintaining virtual machine images. IT support would not have a mechanism to enforce policies, leaving the potential to violate governance and compliance regulations. Billing and resource management would require more manual processes, driving up costs in turn.

Cloud computing lends itself to a wide array of services and service delivery models. As we will see in the next section, there are many ways to provide cloud services.

## Different Types of Cloud Computing Services

Cloud computing can encompass a broad range of services, so it is not surprising to see a number of broad options emerging. These services range, in increasing order of specific type of service, to include:

- Infrastructure
- Platform services
- Application services

Each level of service meets a distinct set of needs.

### Infrastructure Services

Infrastructure services deliver computing and storage services. This type of service is the one used as a model in the previous section describing the three defining characteristics of cloud computing. Here we will turn our attention to describing how this type of service can be used along with an example use case to show how cloud computing can significantly improve some types of service delivery.

## Computing on Demand

The ability to provision computing resources for just about any computing requirement is valuable enough to drive the adoption of cloud computing even if none of the other types of cloud computing services were available. With computing on-demand services, organizations have the ability to allocate virtual machine resources for a variety of tasks:

- Executing proprietary workflows
- Meeting peak demand for computing
- Performing disaster recovery
- Running highly distributed applications

By allocating just basic computing services, cloud consumers can run proprietary workflows that do not depend on preconfigured services. A broad set of service images in a service catalog can provide a starting point for building proprietary workflows. For example, the service catalog would have virtual machine images with OSs and application servers, which users could instantiate and then add custom applications to complete the set of components needed for the workflow.

This type of cloud service also works well for accommodating peak demand periods for either standardized applications or proprietary workflows. Existing infrastructure may be sufficient for average loads, but during peak periods, such as the holiday shopping times in the retail industry, additional computing services may be needed for relatively short periods of time.

Maintaining a disaster recovery site can add significantly to the cost of providing a service. Even if a disaster recovery site is never used, businesses pay for the housing equipment, power to keep a minimal infrastructure running, and maintaining servers and other equipment. There may be marginal labor costs as well to maintain the site. An alternative, and one enabled by the computing on-demand model, is to use a cloud provider as a disaster recovery service. To do this, a business could:

- Maintain a set of virtual machine images that would run the business applications in the event of a disaster
- Maintain copies of data in cloud storage using an appropriate combination of backups and near-real-time replication
- Establish a plan for provisioning cloud services to meet disaster recovery requirements; for example, some services may be run on smaller, and therefore lower-cost, servers while in disaster recovery mode

Of course, as these requirements demonstrate, computing on demand can be closely coupled to storage on demand.

### Storage on Demand

Storage on demand can provide file, block, or relational storage to meet a variety of requirements. In some cases, such as the need for offsite backup, the need for storage is fairly consistent. Cloud storage offers the ability to protect backups from site-specific damage but without the need to maintain another physical site. When dealing with multiple remote sites, copying backups to the cloud can be an appealing option rather than physically transporting tapes from those sites or maintaining additional disk storage at a data center to accommodate those backups.

Demand for storage can vary widely. For example, an accounting firm may have peak demand for 2 to 3 months prior to tax-filing deadlines when large amounts of data are coming into the firm. After the deadline, data can be archived and moved off disk, but without an option such as cloud-based storage on demand, the firm would have to maintain peak storage capacity all year. The wide potential for on-demand computing and storage can be demonstrated with a more generally applicable example as well.

### Business Intelligence Use Case

Business intelligence reporting is driven by large volumes of up-to-date information. Collecting and processing this data can impose significant demands on computing and storage resources, especially when the ETL phase has to occur in a limited window of time. With on-demand computing and storage, data can be uploaded from multiple local sources simultaneously. That data is then aggregated at low and mid levels in parallel before being aggregated at a global level and finally stored in a cloud database for later report generation.

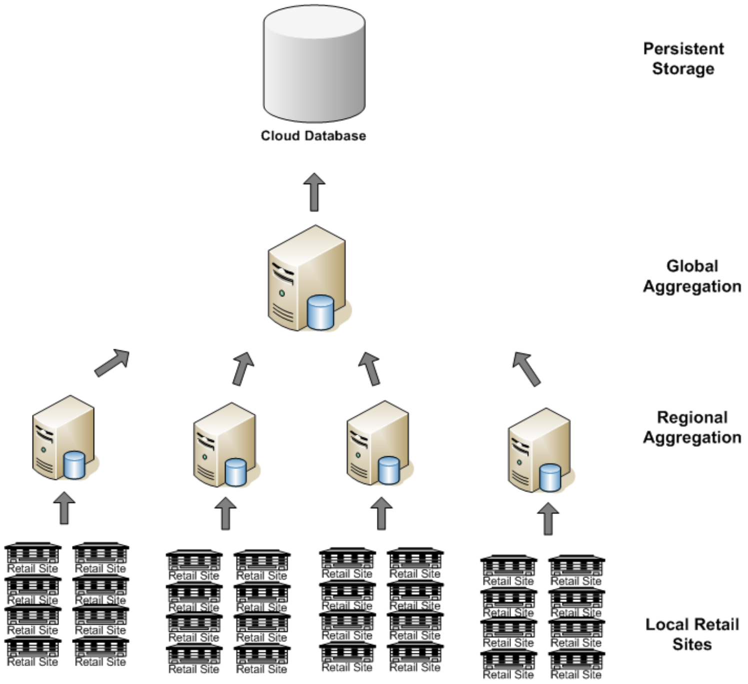


**Figure 2.4: With on-demand computing and storage, time-critical operations like aggregating data for business intelligence reporting, can be done in parallel. This ensures the job completes within the allotted time.**

Running the aggregation operations in parallel allows the process to complete faster than if done sequentially. Running the operations with cloud resources eliminates the need for maintaining dedicated servers that would otherwise be underutilized.

Infrastructure services are an appropriate delivery model when organizations require basic computing and storage resources. When those needs include components commonly found in application stacks, the platform services delivery model may be a better fit.

### Platform Services

Platform-based cloud services deliver higher-level services than the infrastructure-based model offers. Platform-based services include tools for designing, developing, and deploying applications using a set of supported application components, such as relational databases and application security services that span multiple layers of the application stack.

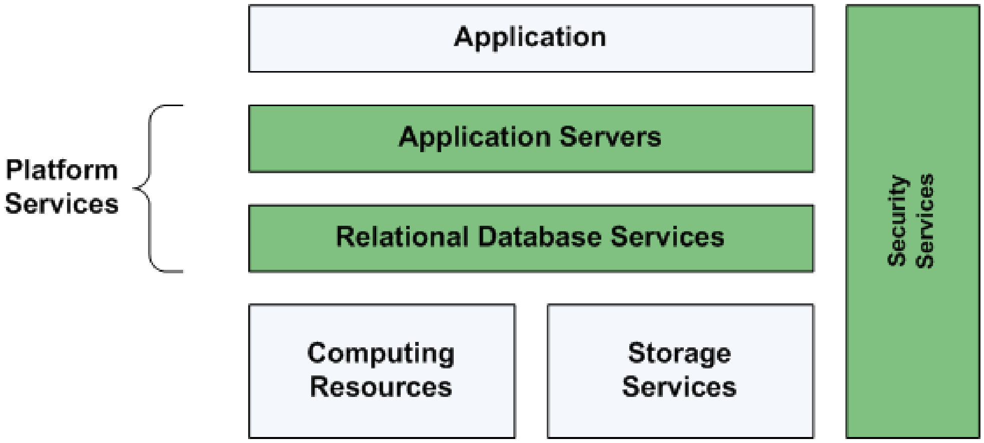


**Figure 2.5: Platform services (in green) provide application development components built on lower-level cloud services.**

### Relational Database Services

Relational databases are the data backbone of most enterprise applications. Since the later 1970s, relational data models have offered significant advantages over other database frameworks. Continuous improvement in relational database management systems have allowed relational databases to keep up with growing and changing demands for managing persistent data. One of the latest advances is the ability to host relational databases in a cloud.

To avoid any confusion, it is worth noting that there are two ways one could host a database in the cloud. One method is suitable for small projects with short life spans, the other takes advantage of cloud infrastructure for a more scalable solution.

## A Simple Relational Database System in the Cloud

The first method basically transfers the same approach to database management we typically use outside the cloud and applies it in the cloud. Under this method, a database administrator provisions a virtual server and installs the database management system on that server using local disk storage for database files. This approach may be suitable for limited needs but is not a general solution for persistent relational storage in the cloud.

One drawback is that local storage is allocated to a user's virtual machine instance only as long as the instance is running. One of the advantages of the cloud is that virtual machine instances are started and stopped as needed. Unless the instance hosting the database is kept running, the database will be lost. Another drawback is that the versions of relational database management systems running on typical enterprise servers are not designed to take advantage of cloud storage services based on allocating blocks or buckets of storage for arbitrary data. Although this is one way to use relational databases in the cloud, it is not what is generally considered a relational database service.

## Relational Database Services Optimized for the Cloud

Relational database services for the cloud take advantage of the scalability of compute and storage resources of the cloud. As one might expect, relational database services attend to a number of low-level implementation details that are typically the responsibility of a database administrator. For example, within the cloud, database administrators do not have to concern themselves with:

- Managing disk space
- Specifying how to distribute low-level data structures, such as tablespaces, across multiple disks to optimize performance
- Monitoring I/O patterns to detect bottlenecks in disk operations
- Replicating data to ensure high availability since persistent data is typically written to multiple locations within cloud storage

Of course, this does not mean the end to database administrators as we know them any more than cloud computing is putting an end to systems administration. Database administrators working with relational database services can focus more attention on the logical aspects of database design:

- Defining schemas
- Optimizing indexes
- Tuning stored procedures and triggers
- Creating views and other abstractions to better support application development

Also, expect cloud providers to support the three fundamental characteristics of cloud computing with respect to relational databases: massive scalability, easy to allocate resources, and a service management platform.

### Application Servers

Application component services provide middleware services in the cloud. Like relational databases, middleware applications, such as application servers and portal servers, can be optimized for the cloud. This ensures the components can take advantage of scalability, high-availability, and service management platforms provided in the cloud.

### Security Services

Security is not a component one can isolate like a database or a messaging queue. Security is a product of specialized components, such as authentication and authorization services, as well as systems design. The fundamental principles of security are no different in the cloud than outside the cloud. We cannot, however, simply use the same security procedures in the cloud that we use outside the cloud anymore than we can simply run a database management system built for a single server in the cloud and expect cloud-like benefits.

Security services need to be embedded into cloud platform services and, at a minimum, include support for:

- Authentication
- Authorization
- Auditing and reporting
- Key management
- Security token management

Authentication and authorization are necessary to determine who is using a system and limiting what they are allowed to do. Auditing and reporting are required to ensure policies and procedures are enforced and to detect unauthorized activity as soon as possible. Key management and security token management are especially important in distributed systems where multiple systems depend on trusted identity management systems to perform authentication, authorization, and other security services on their behalf.

Above infrastructure services and platform services in the hierarchy of cloud services, we find applications.

### Application Services

Today's complex enterprise applications are often built on application frameworks and design patterns, so it is not surprising to see support for these in the cloud. The frameworks vary but include components such as runtime libraries, development frameworks, and higher-level application components. The level of support for different frameworks will vary by cloud provider, especially if providers specialize in supporting one type of framework. In some cases, a cloud provider may offer a framework specifically designed for the cloud and not available in other architectures.

Even with variation in frameworks and programming languages, a number of application services may be available that allow programmers to take further advantage of what a cloud infrastructure has to offer. Two such services are messaging queues and support for highly distributed, parallel processing.

### Messaging Queues

Messaging queues provide for asynchronous communication between processes running in the cloud. Messaging is useful for constructing workflows, implementing distributed transactions, and accommodating the failure of a component within a distributed system. Consider as an example a Web interface running on one server accepts requests from users. In a tightly coupled application, the interface may pass the request to one instance of a backend service and wait for a response. If the backend service is down, the application fails. In a loosely coupled design, the interface would submit the request to a queue. Any one of a number of instances of the backend service could read the request from the queue, respond to it, then delete the request. If a single instance of the backend server is down, the request can still be serviced. If one of the backend instances crashes while processing a request, another instance can still read the request because it is not deleted from the queue until the response is generated.

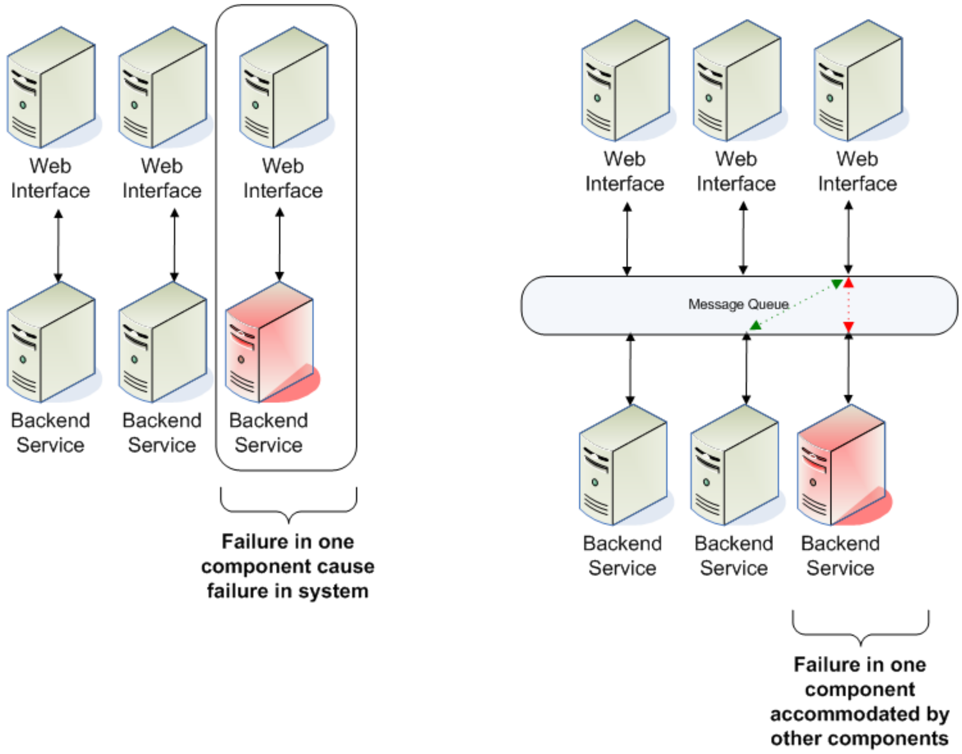


**Figure 2.6: Tightly coupled systems are more likely to have single points of failure; messaging queues enable more robust application design.**

Application services within the cloud also include higher-level components that enable enterprise application functionality.

### Distributed, Parallel Processing

One of the advantages of cloud architectures is access to a large number of servers. This introduces opportunities for performing operations in parallel that would normally have to be done sequentially when only a small, fixed number of servers are available. A programming paradigm known as map-reduce is one suitable-for-clouds method to implement parallel applications.

The basic idea behind map-reduce is that some problems are inherently parallel: Some steps in the computation can be done independently of other steps and the results of individual computations can be combined to produce the final result. The ETL example cited earlier highlights a problem with course-grained parallelism. That problem can be broken down into a small number (for example, on the order of 10) steps followed by an aggregation process to combine results. Other problems, especially those with large amounts of data, can be divided into even larger numbers of sub-problems.

Take for example, analyzing click-stream data. A business is analyzing patterns of activity on their e-commerce site to determine whether there are common characteristics shared across customer interactions in which the customer abandons his or her cart. The click stream data from the Web site contains information about what products the customer viewed, reviews that were read, and navigation paths taken to the point where a product was added to the cart. As one customer's activity is independent of others, this is a good candidate for highly parallel analysis.

A map-reduce approach to this problem could be defined as follows:

- Split the set of all click stream data by customer session
- Partition the customer sessions across 100 instances of the analysis program
- For each customer session, scan the click stream for the number of times each possible 3-page sequence pattern occurs; to simplify the pattern, look for types of pages, such as product details, reviews, search results—this is the map phase
- Combine the results of each map phase to produce the aggregate number of times each pattern occurred—this is the reduce phase

A key advantage of this approach is that large volumes of click stream data can be analyzed much faster in parallel than sequentially, thereby creating the possibility for greater amounts and more in-depth analysis of customer interaction behavior.

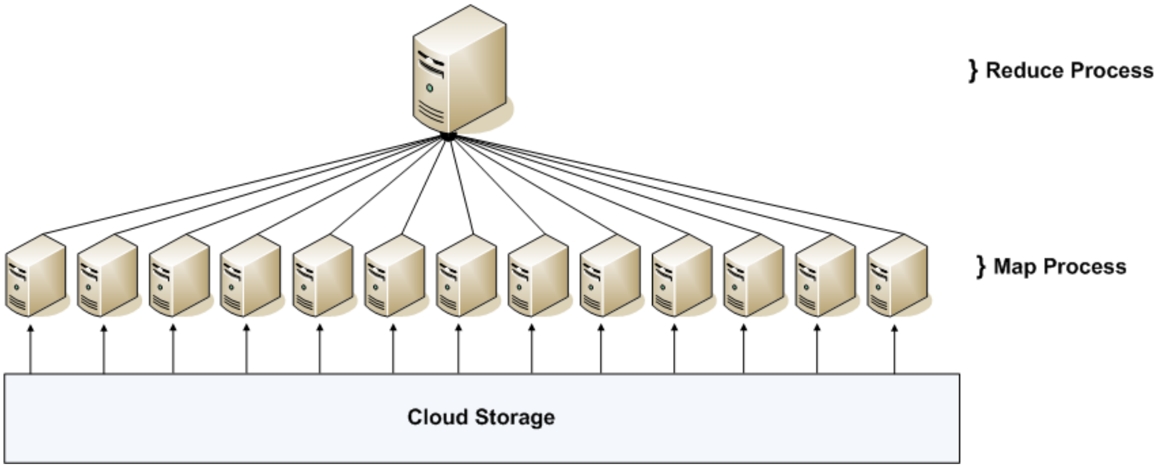


**Figure 2.7: Map-reduce is a parallel programming framework that works well with cloud computing and storage.**

Application middleware, such as application servers; design patterns, such as the use of messaging queues for asynchronous communication across multiple processes within a cloud; and programming frameworks, like map-reduce that exploit the parallel capabilities of a cloud, are all enabling components for delivering enterprise applications in the cloud. As cloud technology adoption grows, we can expect to see more enterprise applications being offered directly to cloud service consumers.

## Applications and Business Services

Providing application and business services from the cloud presents an opportunity to consolidate those services. The beneficial features of cloud computing, such as flexible scalability and a service management framework, can enable organizations to reduce the number of separate instances of applications running throughout the enterprise.

### Consolidating Enterprise Applications

Consider a few common types of enterprise applications:

- Customer relationship management (CRM)
- Enterprise resource planning (ERP)
- Business intelligence

Each of these types of applications can have broad reach throughout a business. With the commonly used “one server/one application” approach that has been used for years, businesses may find themselves limited to how many users they can support with these applications.

For example, consider a company that runs a CRM application on a server sufficient for current needs as well as some moderate growth. The company then merges with another business that also needs CRM support. The IT staff of the new company will have to determine whether a single server can support the newly merged enterprise or multiple instances of the system will have to be run. The latter option can lead to fragmentation and arbitrary divisions that in turn can lead to organizational problems down the road.

Let's assume the business decides that running two instances of the CRM application is the more cost-effective alternative. The customers are divided geographically with North America, South America, and Southeast Asia customers in one instance, and Europe, Middle East, Africa, and other Asia customers in the second instance. A host of questions arise:

- How should customers in global, transnational companies be divided?
- Will regional subdivisions of customers be separated?
- How costly and time consuming will it be if the allocation of customers has to be re-arranged to align with new business strategy?
- What is required to support a federated identity management system so that users in one system can access the other system as needed?

Similar questions can be asked about ERP systems; instead of customers though, the questions would focus on budgets, inventories, financial projections, and accounting issues.

In the case of business intelligence, fragmentation can occur around tools and procedures. Enterprise-scale data warehouses may have dedicated database administrators who are able to tune and manage complex database management systems. Departments with more limited requirements may build locally managed data marts employing easier-to-use databases and reporting tools. This may be the most expeditious approach in the short run but over time it can lead to duplicated data, increased software licensing costs, and redundant administration costs.

Moving enterprise applications such as CRM, ERP, and business intelligence systems to the cloud can help reduce costs and improve the delivery of business services. With standardized virtual machine images and centralized cloud storage, additional compute resources can be brought online as demand for services grows. As data is consolidated in the cloud, we can avoid data fragmentation problems. Standardized virtual machine images deployed through a services management platform reduce the demand for specialized database and systems administration expertise in departments running local applications, such as data marts.

### Managing Business Services and Workloads

As applications move to the cloud, there will be a need to manage according to service level agreements (SLAs) and other expectations for performance and availability. This will require both technical and management approaches to the problem.

On the technical side, application administrators will need to utilize performance reporting provided by the service management platform to ensure SLAs are met in cost-effective ways. Running multiple instances of an application and load balancing across those instances can help maintain performance and provide a level of reliability to the system.

On the management side, we need to be cognizant of utilization. There is no point running six instances of an application with an average server utilization of 25% when running three instances still leaves plenty of margin for spikes in demand without the need to instantiate another virtual machine image.

It is clear as we consider the different types of services, from infrastructure to platform to application services, there are many ways to leverage cloud services and the benefits generally arise from a set of common attributes.

### Common Attributes of Cloud Service Models

The three defining characteristics of clouds—massive scalability, easy to allocate resources, and a service management platform—describe key architectural elements of computing and storage clouds. A consumer of cloud services may see a different set of attributes from their perspective:

- On demand self service—The ability to allocate, use, and manage computing, storage, application, and other business services at will without depending on IT support staff
- Ubiquitous network access—The ability to work with cloud resources from any point with Internet access; cloud service consumers are not dependent on being in corporate headquarters or in a data center to have access to an enterprise cloud
- Location independent resource pools—Compute and storage resources may be located anywhere that is network accessible; resource pools enable redundancy and reduce the risks of single points of failure
- Elastic scalability—Cloud consumers decide how much of any resource they utilize at any time; allocation is driven by immediate demand not the need to maintain capacity for peak demand
- Flexible pricing—Cloud providers typically charge with a "pay as you go" model; as cloud computing matures, we will likely see a variety of pricing models, including prices that vary by level of demand

We have described cloud services from an architectural view, in terms of services delivered, and from the perspective of a cloud consumer. One remaining dimension we should consider is the public/private cloud distinction.

## Cloud Delivery Models

When cloud computing first emerged as a viable platform, the term generally applied to what we would now call a public cloud. As cloud computing expanded, so did the delivery models to the point where we have at least three distinct delivery models:

- Public clouds
- Private clouds
- Hybrid clouds

Public and private clouds have advantages and disadvantages; hybrid clouds attempt to capture the best of both worlds.

### Public Clouds

Public clouds are computing and storage services that are open to any consumer. An immediate advantage of using a public cloud is that there is no upfront capital expenditure required of business users. Cloud consumers purchase computing and storage services as needed and pay as they go. There are likely costs associated with transferring data to and from the cloud, and these costs can easily grow beyond the cost of computing and storage for high-transfer rates. Another disadvantage is that businesses are dependent on the viability and reliability of the cloud provider. If there is a significant service outage, data and services will be inaccessible. Risk assessments and mitigation strategies are called for when working with any cloud, but they are especially necessary when critical business services are dependent on third parties.

### Private Clouds

Private clouds are owned and operated by businesses for their internal use. This delivery model can be especially appealing when compliance, security, and other risks factor significantly when developing a cloud strategy. A key advantage of a private cloud is that the business is in control of the service: it can set pricing and policies, control access, and define its own service catalog of virtual machine images for use in the cloud. A private cloud does require capital expenditure to procure hardware and software for the cloud. A staff of IT professionals must also be available to administer and manage services. To realize the greatest benefit of the cloud architecture, multiple data centers will implement distributed storage and compute infrastructure. Capacity planning is also an issue. A business could find a successful private cloud creates demands that exceed current capacity. Expanding a private cloud can require substantial capital expenditure; a hybrid model could be a better alternative.

### Hybrid Clouds

A hybrid cloud combines public and private clouds. A business that has implemented a private cloud can use public cloud resources as an extension of their own cloud. There are a few different ways to do so.

The two clouds could be separately managed service platforms. Policies are established to govern what kinds of jobs can run in the public cloud, and cloud consumers have the option to run and manage their jobs in the public cloud. This approach gives cloud consumers freedom to choose between two services. There may be cases where the public cloud is less expensive or can provide capacity unavailable on the private cloud.

Another way to manage the hybrid private-public cloud is to enable access to the public cloud from within the service management platform. The two services are still independent, but cloud consumers would have a single point of management.

Finally, the public cloud could be treated as an extension of the private cloud by implementing a virtual private network (VPN) in the public cloud. Under this model, a portion of the public cloud is treated as an extension of the private cloud.

As is so often the case in information technology, there is more than one way to deliver a service, and the best option in any situation is highly dependent on specific requirements.

## Summary

Cloud computing is relatively young, but in the short time since its inception, it has managed to create a host of competing definitions, architectures, service models, and delivery methods. Across all of these varying ways of looking at cloud computing, we find common characteristics, including massive scalability, ease of allocating resources, and a service management platform. Building on this foundation, cloud providers can deliver a range of services, from infrastructure to platforms to applications and business services. No single delivery model meets all needs, but the combination of public, private, and hybrid clouds offer a range of options suitable for many business requirements.

# Chapter 3: Enabling Business Innovation by Using Cloud Computing

Many discussions of cloud computing focus on its technological advantages—and there are many—but there are business advantages as well. This chapter shifts focus from questions of architecture and operations to issues of service delivery and return on investment (ROI). After all, cloud computing is not an end in itself (unless you are a computer scientist or systems architect) but a means of delivering existing services more efficiently and enabling the delivery of new services that may not be practical under other models.

The chapter is divided into three main sections:

- Launching a new business service—The first section compares service delivery under traditional IT service models and under cloud computing. Example scenarios will illustrate some of the key differences.
- Advantages of doing business with cloud computing—The advantages of doing business with cloud computing include the reduced time required to deliver new services, new means to control costs, the ability to scale to demand, and the adaptability of cloud computing.
- Sources of ROI in the cloud—ROI in cloud computing comes from both reduced capital costs and lower operational costs. As with other technologies, the ROI in the cloud is highly dependent on more than just the technology; how you implement and manage cloud services contributes to how much of the potential ROI is actually realized. As a first step to understanding the source of ROI in cloud computing, let's consider a couple of hypothetical examples of how service delivery in the cloud differs from traditional IT service delivery.

## Launching a New Business Service

There is nothing like launching a business service to combine the exhilaration of creating something new with the apprehension associated with choreographing all the elements required for a smooth launch. And there is no shortage of pieces that must be in place:

- The computing, storage, and network services required to support the service
- Software that captures the functional requirements of the new service while providing a usable interface
- A well-developed plan for deploying elements in the proper order so that dependencies are in place as new components are put in place
- Policies and procedures to govern how the service infrastructure is managed and maintained
- A recovery strategy and corresponding systems to mitigate the risk of data loss or service delivery failure

It is easy to see how essential each of these technical and business elements is to the ultimate success of the project.

Take away sufficient computing, storage, or networking, and the service can degrade to the point of failure. Skimp on usability engineering or otherwise shortchange the user interface, and you lose customers at the proverbial front door. Those of us who have worked on projects with inadequate planning know the frustration and futility that come with ad hoc, reactive management. The worst part is that the delays, rework, and missed steps could have been avoided. As we consider the advantages of cloud computing for service delivery, you will see how some of these potential problems can be reduced. Needless to say, cloud computing is no panacea and no amount of technology can compensate for poor management practices. Cloud computing can, however, reduce some of the burdens and challenges that typically come with planning and implementing new projects.

Once a service is deployed, it is time to move into an operation maintenance mode. Planning is just as important here as it was during design and deployment. The difference is that now you shift from a project planning framework of deliverables, milestones, and resource balancing to operations guided by policies and procedures that define what is to be done and how to do it. Policies governing everything from service level agreement (SLA) monitoring to backups to security should be in place at launch. Procedures, which turn those polices into executable tasks, must also be in place to ensure proper operations. Of course, even with the best planning and policies in place, hardware fails, software errors manifest themselves, and natural disasters strike. A recovery management strategy, commensurate with the value of the new services, can help you respond effectively and efficiently when adverse events occur.

As Figure 3.1 depicts, successful service delivery is dependent on these and other technical and business factors. One of the questions facing business strategist and systems architects is, What is the best service delivery model for realizing project objectives?

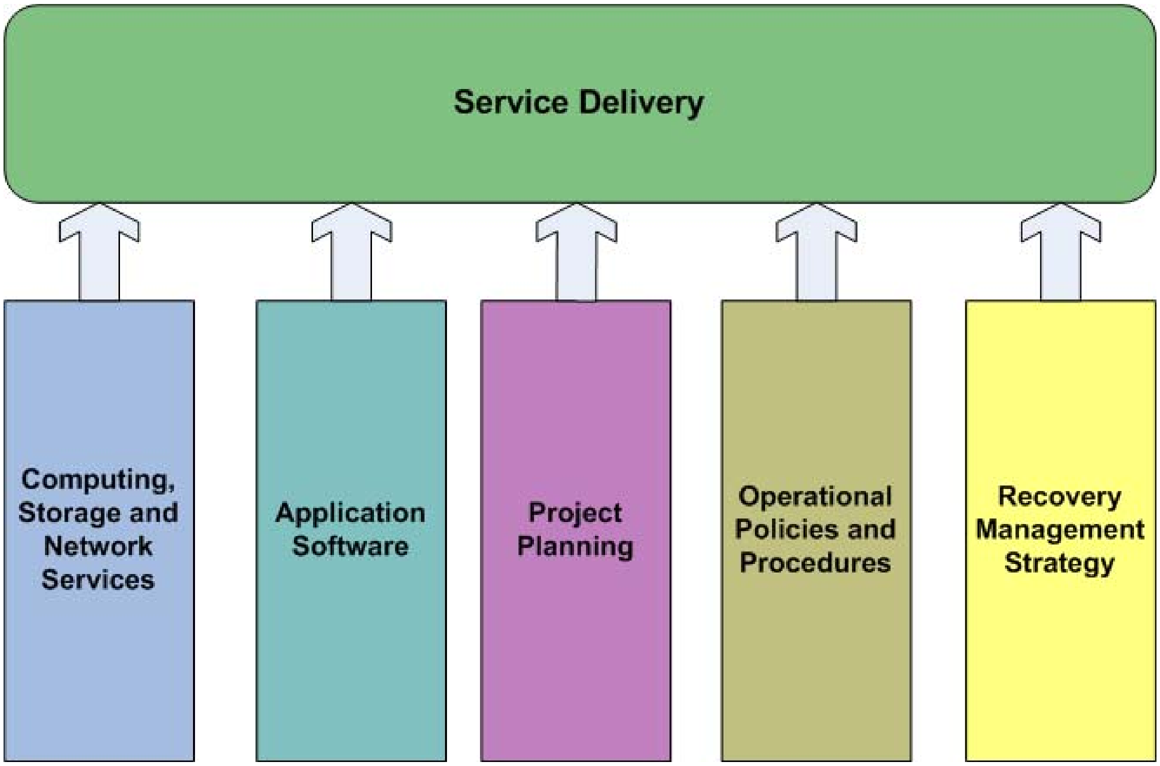

Service Delivery
Computing, Storage and Network Services
Application Software
Project Planning
Operational Policies and Procedures
Recovery Management Strategy

**Figure 3.1: Service delivery is built on a foundation of technology and business services and practices. Remove, disrupt, or undermine any of these, and services delivery is adversely affected.**

To better understand how service models influence service delivery, let's assess delivering a couple of different types of services under different models. In the first example, we will consider a home improvement retailer with a plan to offer tutorial videos on home improvement projects for the do-it-yourself (DIY) customer. In our second example, we will see how business analysts deal with the problem of "big data" and the need for advanced business intelligence and analytics services. These examples are chose for several reasons:

- They are significantly different types of services—one is a customer-facing Web application and the other is a more batch-oriented back office service
- They require a different combination of computing resources
- They have different usage patterns over time
- Cloud computing can reduce the cost of delivery of both services regardless of the differences in the type of application an demand profile

First, let's explore the steps involved in deploying these two services under a traditional IT service model. Next, we'll look at how the same service could be deployed in the cloud.

## New Services Under a Traditional IT Service Model

Project management, software development, testing, and deployment practices are well developed under traditional IT service models. They all come into play in our two hypothetical scenarios.

### Scenario 1: Tutorial Videos for the DIY Customer

Not all of us are gifted carpenters or skilled plumbers, but some of us think we could do a fairly decent job around the house if we just had the right tools and a few tips to get us started. A home improvement retailer that has traditionally done well serving the small contractor segment of the market has decided to target the potential DIY customer in an effort to improve sales and expand their share of that market segment. The following list highlights key features and non-functional requirements:

- The service will provide short tutorial videos on a range of home improvement topics. Videos will range from 1 to 10 minutes in duration with an average of 5 minutes.
- Videos will be streamed over the Web and delivered through the company's Web site.
- The service will be launched in beta to customers in the Northeast United States for 4 weeks followed by an extended 4-week beta to the Northeast, Mid-Atlantic, and Southeast United States. After that, it will be made available throughout the company's North American market.
- The initial launch will support up to 500 videos; at the end of the beta testing phase, 1000 videos will be available. Content will grow at an average rate of 200 videos per month after that.
- Metadata will be assigned to each video to improve search and browsing. Tags will include structured data, such as repair type, tools required, and time to complete the task. Unstructured data describing the video content is also included.
- Videos will be accessible through a centralized "How-to Video Library" in the Web site as well as through product pages that link to relevant videos.
- Customers will be encouraged to review and rate videos. The results will be analyzed to improve the overall quality of instruction, expand the scope of topics, and eliminate the least-useful content.

Using current Web site statistics, business planners anticipate peak demands Wednesday and Thursday evenings between 6:00pm and 10:00pm and Saturday mornings between 7:00am and 11:00am. The anticipated demand pattern is depicted in Figure 3.2.

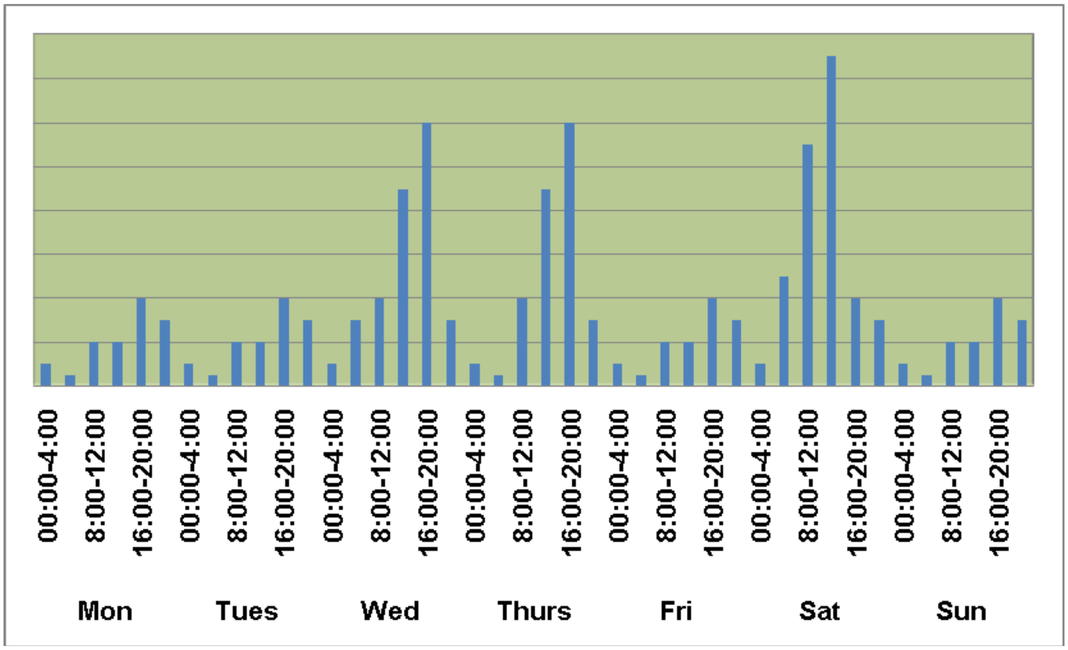


**Figure 3.2: Service demand will vary widely by day of week and time of day. (Times are relative to the time zone of the data center hosting the service).**

As the systems architects and application designers plan the infrastructure for this service, they have to take into account a number of considerations. The service will require servers to meet peak demand, although those periods are relatively few and fairly short. The irony of running a "how to fix" tutorial service on a poorly functioning platform could undermine the brand image and is not worth risking.

On the business side, this project will require a capital expenditure and C-level approval. The IT professionals on the team know that they will have one chance to get the resources they need within the next 12 months. They do not have sufficient data to confidently predict demand for the service, so they resort to the next best thing: making a best guess estimate and then add another 20% for contingency. The combined concern for not performing to customer expectation with the inability to get a second round of resources rapidly enough push the applications designers and systems architects to choose a more costly solution than may ultimately be required.

The major components they decide on include:

- Several servers to stream the video tutorials
- A load balancer to distribute user sessions across several servers
- A storage array with sufficient redundancy (for example, RAID 6)
- Application licenses to support the service

Figure 3.3 shows the configuration.

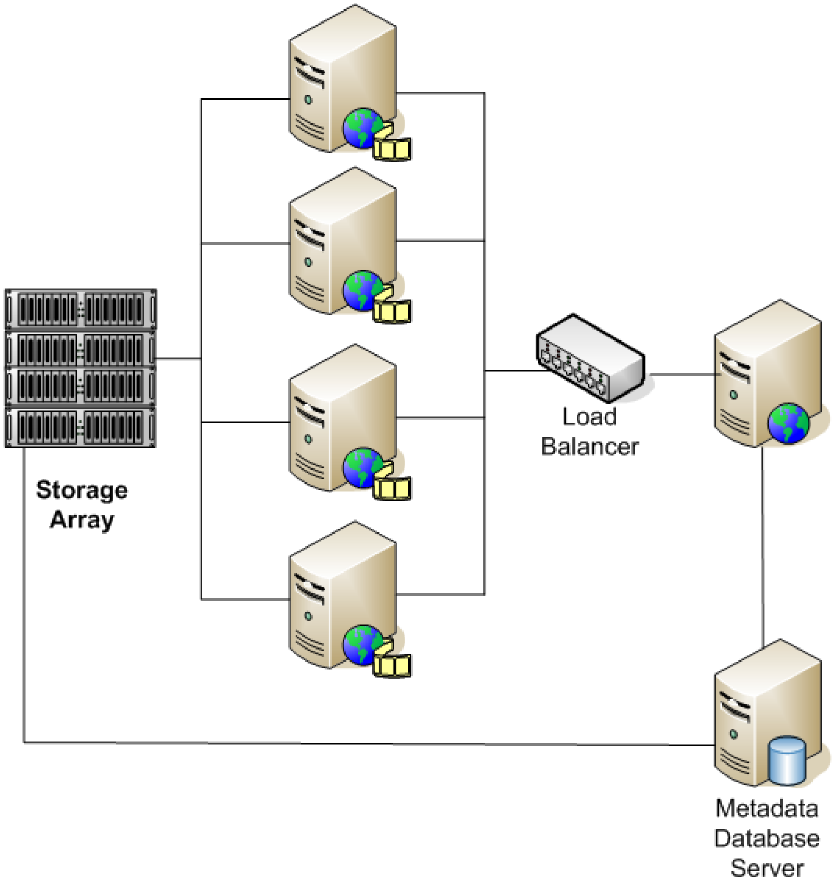


**Figure 3.3: The video tutorial service requires hardware to meet peak demand even though the average demand is significantly less.**

It is clear from this example that building out this service following a traditional strategy requires that you build for peak demand before you even have sufficient information to determine the actual level of need. Not only can you not adjust to changing needs, you have to make a fairly long-term commitment to the architecture early in the process.

### Scenario 2: Advanced Analytics for Auto Insurance Premium Calculations

The auto insurance industry is a competitive business. As with any type of insurance, premiums have to correlate with risks. For auto insurers, there are many factors to consider, including the age and sex of the driver, past accidents, number of moving violations, primary garaging location of the vehicle, and so on. From a competitive perspective, using just these factors is insufficient to gain any competitive advantage; after all, competitors use the same data. Using the same data can lead insurers to cluster drivers into similar groups making it difficult to compete on price within those groups.

In this scenario, several auto insurance analysts propose expanding the base of data used to categorize customers and then applying data mining techniques to create finer-grained clusters of customers. Premiums can be adjusted to these finer groups of customers so that customers posing greater risks can be charged higher premiums allowing for lower premiums for safer drivers. Ultimately, this could reshape the risk pool by attracting better drivers with lower rates than competitors offer while giving incentive to higher risk drivers to look elsewhere for insurance.

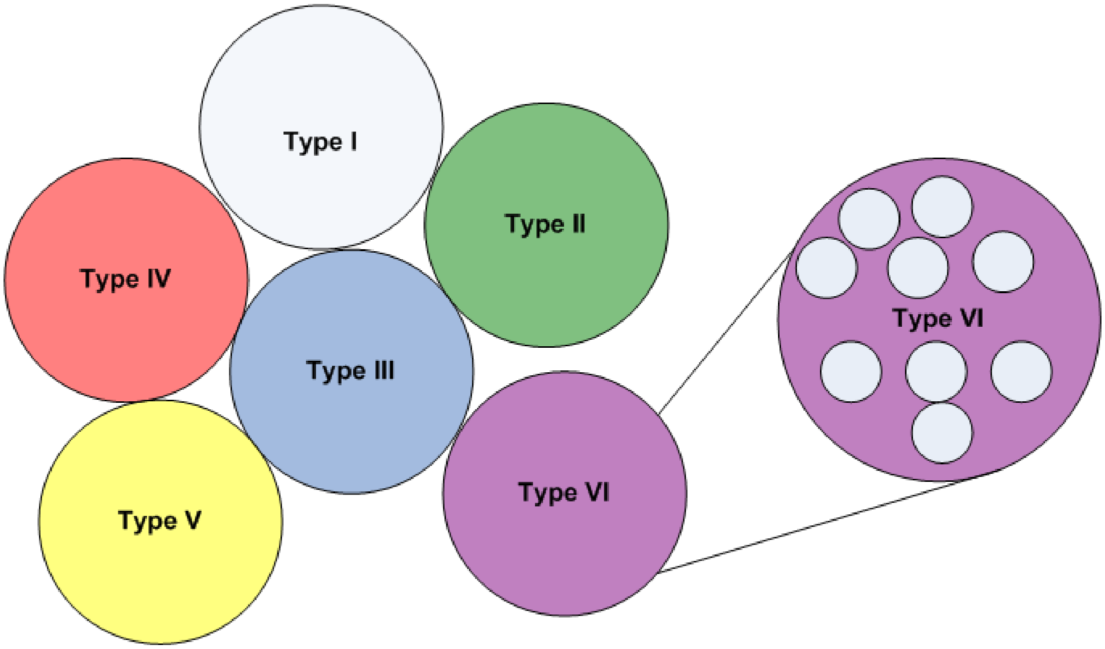


**Figure 3.4: Finer-grained clustering of customers can create a competitive advantage by allowing more precise and accurate premium pricing.**

The following list highlights key features and non-functional requirements:

- Existing data sets on age and sex of the driver, past accidents, number of moving violations, primary garaging location of the vehicle, and so on must be available for data mining
- Additional data on household income, including income by age, disposable income, household net worth, disposable income, and so on; consumer spending data by category, such as financial services, automotive, medical, recreation, and so on; business activity data by location; and publically available data, including census data and crime statistics
- On a monthly basis, internal and external data will be collected and analyzed to build a predictive model that categorizes each customer by fine-grained risk estimate

- New extraction, transformation, and load (ETL) procedures will be developed to collect data from multiple sources and copy it to project storage; data will not be stored once the model is constructed
- To improve the quality of predictions, multiple prediction models will be constructed and results will be combined to make final classifications.

This application is compute intensive during the times when the data mining systems are running and predictive models are being created. After the models have been created, the models will be executed on to categorize new customers and reassess the premiums on existing customers during policy renewal. Running models are significantly less compute intensive than generating them.

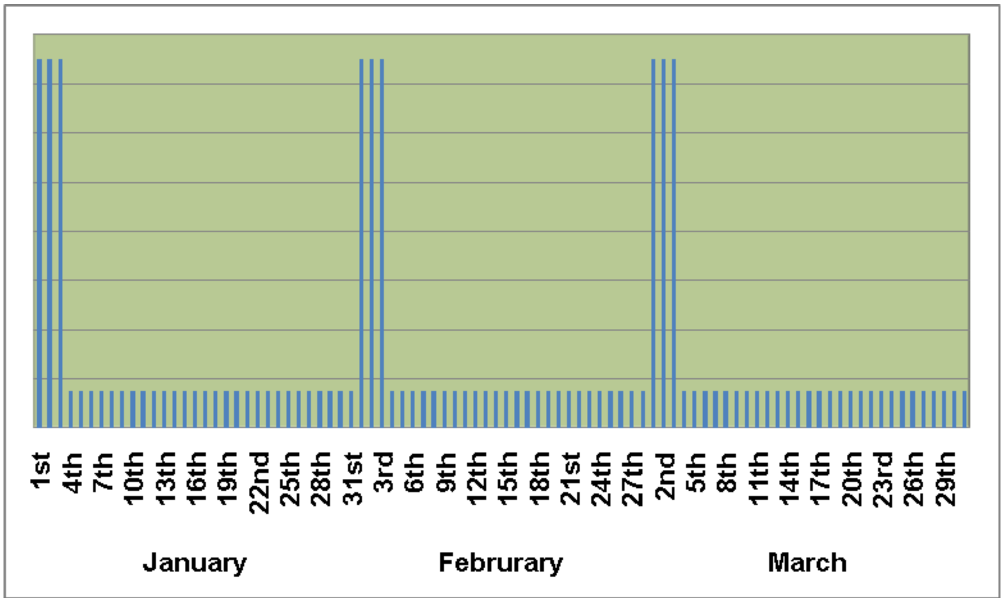


**Figure 3.5: Analytic operations have fairly predictable demand patterns that include significant periods of peak demand followed by analysis operations.**

Once again, this service requires that you build an infrastructure for peak capacity. A cluster of high-end servers each with multiple multi-core CPUs and significant amounts of memory are required to build the individual predictive models combined into an ensemble prediction service. Although data will only need to be stored during the time the models are built, architects will have to purchase storage sufficient to support copies of all the various data required.

Both of these scenarios manifest common difficulties with the traditional IT model of service delivery. Dedicated resources are not used efficiently. Capital spending decisions may have to be made with insufficient usage data. It is difficult if not impossible to scale the infrastructure up or down according to demand. The cloud computing model offers an alternative method for deploying services.

## New Services Under the Cloud Computing Model

The cloud computing model provides a flexible infrastructure that allows service providers to acquire the compute and storage resources they need, when they need them, for as long as they need them, and to pay for only what is used. Both of the example scenarios would benefit from deployment on the cloud.

### Scenario 1: Tutorial Videos in the Cloud

The tutorial video service is a new customer-facing service that could have wide-ranging demand patterns. Initially, the systems architects decide to allocate two virtual servers for the beta-test period; however, if demand warrants additional or fewer servers, systems administrators will adjust as needed. Planning for long-term storage is not a significant issue because additional storage will be allocated as needed. There is no need to purchase peak-load storage. As the project moves from the beta testing stage to full production, the systems administrators will add virtual servers as needed. Rather than focus on predicting what the peak demand will be over the next 12 months, systems administrators can focus on immediate demand and server allocation to efficiently and cost effectively meet that demands.

### Scenario 2: Advanced Analytics in the Cloud

The cloud is a much more cost-effective method for delivering the kind of advanced analytics described earlier. In this case, there is a recurring demand for a significant amount of storage and computing resources. The demand is for only a few days every month, so purchasing dedicated hardware is not cost effective. Deploying to the cloud is relatively straight forward and includes:

- Creating virtual images with the required software, such as ETL systems, and pre-processing scripts and statistical and data mining packages
- Instantiating servers to run parts of the workflow as needed; for example, based on the type of source data and it's configuration, it might make sense to instantiate 10 virtual servers for ETL operations that run in parallel—as the ETL operations execute, they write data to cloud storage, which is taken as input to pre-processing scripts that output data into the proper format for the data mining application
- Allocate storage to store the raw and processed data; once the data has gone through the pre-processing stage, the raw data is deleted; once the predictive models are built, the output of the preprocessing stage is deleted as well

This method improves upon traditional implementation models in at least two ways. First, you can run the workflow as a sequence of steps allocating servers for each step as needed and then shutting them down and starting servers with software for the next step. With virtualization and service catalogs, this is a simple matter. In theory, you could do this with a set of dedicated physical servers by running different virtual machines at each step of the workflow; however, the virtual machine image management would be more difficult without a service catalog and it would still not address the problem of having to purchase hardware for peak demand.

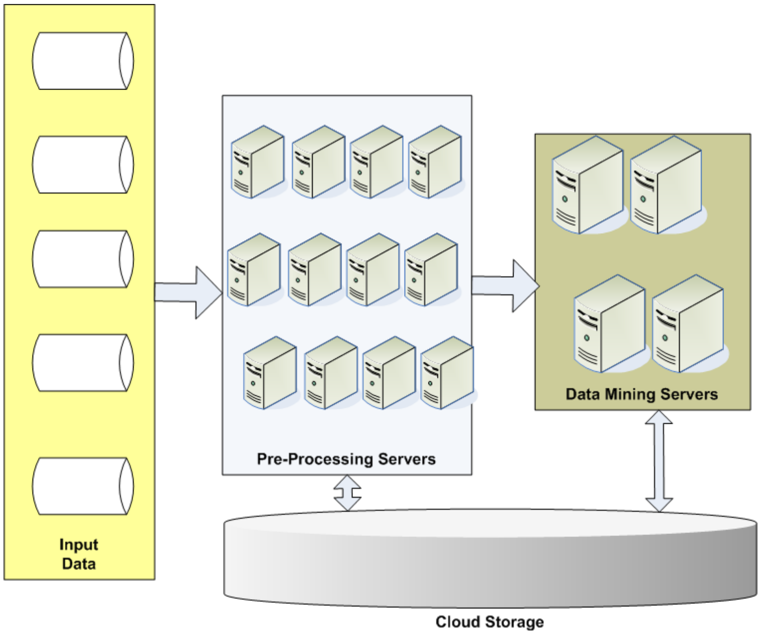


**Figure 3.6: In the cloud, servers can be allocated to do task as long as needed and released at which point other servers are instantiated for the next step in the workflows. Service providers only pay for when they are using compute and storage resources.**

The traditional model of service allocation has worked well for us. The many critical business services are running today on dedicated infrastructure. Cloud computing models improve on the traditional deployment model by allowing you to easily share compute and storage resources and allocate only what is needed when it is needed. This approach reduces the need for *ad hoc* solutions to mitigating risk, like adding an arbitrary percentage to a project budget in case additional hardware is needed. As these two scenarios show, even with diverse types of projects targeted to different users with different compute and storage requirements, cloud computing can offer significant advantages. Next, we will identify the advantages alluded to in the scenarios just described.

## Advantages of Doing Business with Cloud Computing

The advantages of deploying services with cloud computing infrastructure fall into four categories:

- Time to deploy new services
- Cost control
- Ability to scale to demand
- Adaptability of resources

Each of these advantages is closely tied to the architecture of cloud computing combined with management practices for allocating the costs of compute and storage services.

## Time to Deploy Services

When hardware is dedicated to specific functions, it can be difficult to find compute and storage resources for a new initiative. In the early stages for development, would-be service providers may be able to squeeze in some applications on underutilized servers. The likely success of this approach depends on the availability of server or storage capacity and the ability to find that excess capacity. If one has to cross organizational boundaries to find these resources, the chances of securing them can drop significantly. If successful, these stop-gap measures will eventually have to be replaced with a more permanent solution.

Procuring hardware can be time consuming. Capital expenditures for multiple servers, storage arrays, and other equipment can require multiple levels of approval. Plans may have to be reviewed and approved from both a budget and technical perspective. Delivery of hardware can take weeks, and in some cases, months. Once the hardware arrives, the next stage of deployment begins.

Installing hardware is a multifaceted process. It needs to be configured according to organizational standards and incorporated into support systems, like backup schedules and patch management systems. Some of the most frustrating delays come when a single piece of hardware, such as a storage controller, has to be ordered separately and installed when the server arrives. In terms of frustration, order glitches are second only to having to wait for a simple task, like running a fibre to the new server, to get to the front of the service queue. Many of these configuration tasks are unavoidable. The integrity of infrastructure depends on keeping hardware and software in accordance with policies. Fortunately, cloud computing provides a framework that preserves the integrity of infrastructure without many of the time delays (and frustrations) encountered in traditional IT deployment models.

In the cloud model, provisioning becomes a matter of instantiating a virtual machine instance. There are no hardware orders, delivery delays, or waiting for IT support to get around to installing your hardware. With the ability to rapidly adjust the number of instances, there is less need to analyze projected demand. Inefficient and time-consuming efforts to find existing servers with spare cycles are also eliminated. Hardware resources are centrally managed and allocated on demand. The new bottlenecks to deployment are establishing a charge account for the cost of cloud services, selecting a virtual image to run, and deciding how many instances to start.

## Cost Control and Ability to Scale to Demand

Another advantage of using cloud as a delivery platform is greater cost control, and that is tightly linked to the ability to scale to demand. This comes from the ability to make fairly fine-grained decisions about resources. Whereas you might have to decide between purchasing a $10,000 and $15,000 server under a more traditional deployment scheme, in the cloud realm, you have to decide whether you want to run the $0.50/hr server or the $0.90/hr server. You are not committed to using these servers for 2 to 3 years either; in the cloud, you could be charged by the hour. If you make a mistake and underestimate your need, you add more servers. When utilization reports show that the virtual servers you have allocated are underutilized, you scale back the number of servers you are running.

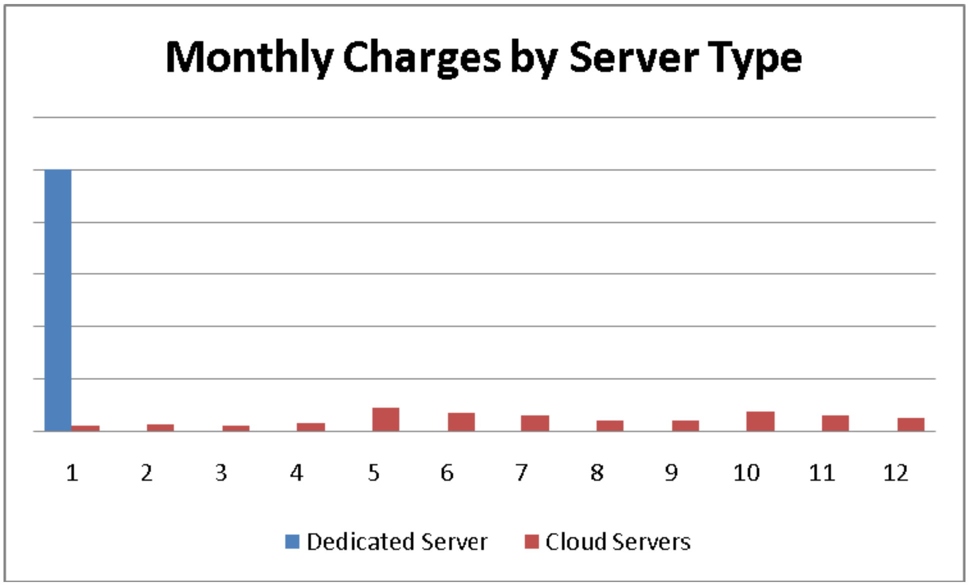


**Figure 3.7: Dedicated servers incur high initial cost inline with anticipated peak demand. Cloud servers incur costs for actual use over time.**

Systems administrators and service managers have greater control over the allocation of resources in the cloud and therefore can provision as needed for current demand. With cloud computing, they have effectively escaped the challenge of needing to constantly dedicate resource for peak demands.

There is also a potential for cost savings with software licensing. Traditionally, software is often licensed to named users or for a specific number of concurrent users. The cloud opens the opportunity for new software pricing models, such as charging by the hour. Ultimately, any cost savings on software licensing will depend on vendors adapting their pricing models to the cloud.

### Adaptability of Resources

Through the course of IT's history, there has been a trend toward making computing resources more adaptable. For example, in the 1960s and 1970s, if you purchased a mainframe or mini-computer from IBM, Digital Equipment, or one of the few other hardware vendors of the day, you would get "the" operating system (OS) for that machine, such as OS/360 for the mainframe or RSTS for the mini-computer. Each machine was used for different purposes, such as batch processing business applications or interactive scientific programs. By the 1980s, hardware and operating vendors started to separate, with Microsoft providing the dominant OS for the IBM PC while Apple introduced its OS to run on Motorola hardware. In the 1990s, it was not uncommon to run different OSs on the same type of hardware. Cloud computing has moved this trend to the next stage with the ability to rapidly switch virtual machine images running on a hardware platform.

In the cloud, hardware resources are not tightly coupled to any single platform. The same resource that runs an instance of Windows Server 2008 an hour ago may be running Ubuntu Linux now. A set of servers that were tasked with generating reports for a data warehouse might be used to generate customer invoices after that. Removing restrictions on the type of software and radically reducing the time and expertise required to change OS platforms significantly improves the adaptability of hardware.

The advantages of cloud computing stem from the ability to deploy new services faster than possible under more traditional models; the ability to control costs at a much fine-grained level of detail than possible before, including the ability to rapidly scale to needs and the adaptability of resources to different tasks. The movement away from dedicated servers for single tasks to using cloud resources brings with it several sources of ROI.

## Source of ROI in the Cloud

The ROI of cloud computing is realized in two forms: reduced capital expenditures and improved operational costs.

### Lowering Capital Costs with Cloud Computing

With cloud computing, business services can be launched without the same type of capital outlays required in traditional IT deployment models. The shifts in capital expenditures occur for three reasons:

- Reduced need for initial capital outlay
- Reduced need for building for peak capacity
- More efficient utilization through virtualization

As we saw in earlier, just getting a new business service started requires access to hardware and software. Traditionally, this means procuring dedicated servers right from the start even if the full capacity of the server is not needed for some time. Tying up working capital in hardware brings with it opportunity costs. The capital that went into purchasing a server could have been invested in a resource that begins producing an ROI right from the start instead of having to wait months before the service requires the extra initial capacity.

Another advantage from a capital cost perspective is that you do not have to invest for peak capacity. With the cloud model, your costs over time are more closely aligned with the average cost of delivering a service, not the peak capacity costs. The savings can be significant, especially when peak demand is highly skewed relative to other demand periods. For example, in the case of the advanced analytics application, there was relatively modest average demand for computing resources but substantial peak demand, providing for substantial savings in capital costs.

Another source of ROI is due to virtualization. The utilization of a physical server is no longer tied to a single application's usage pattern. A server dedicated to the advanced analytics application would sit idle most of the month; however, the same server in a cloud configuration could have multiple virtual machines running on the physical server constantly if there is sufficient demand. Of course, one of the objectives of managing a cloud service is to have enough physical servers to meet demand but not so many that overall utilization rates drop.

Part of the ROI realized with cloud computing can be traced to the reduced cost of capital expenditures, but even more substantial benefit can be accrued by lowering operational costs.

### Lowering Operational Costs with Cloud Computing

The most important drivers in ROI relative to operational costs can be grouped into four areas:

- On-demand provisioning
- Reduced marginal cost of systems administration
- Standardization and automation
- Service management reporting

The ROI in operational costs are subject to the economies of scale. These savings are particularly important in larger cloud installations.

### On-Demand Provisioning

IT support services are necessary in any deployment model, traditional or cloud. The amount of support that is needed for provisioning servers can vary significantly, though. Consider the steps involved in provisioning a virtual server in a traditional IT environment (the "to do" list is even longer when dealing with physical servers), which includes:

- Submitting a service desk ticket requesting a virtual machine instance
- Identifying which physical server will host the virtual machine
- Determining the configuration parameters for the new instance
- Specifying required support services, such as backups
- Coordinating with other users on the shared hosts to avoid common peak demand periods—for example, running a full backup on one virtual machine instance while an I/O intensive job is running on another instance.

The process can be time consuming because there is a division of labor that separates those who know what has to be implemented from those who know how to implement what is needed. This is a typical scenario in IT. The complexity of IT systems demands a pool of specialized IT knowledge. Service developers and business users require their talents to deploy new services and that creates a potential bottleneck. Cloud computing avoids this problem with support for self-provisioning.

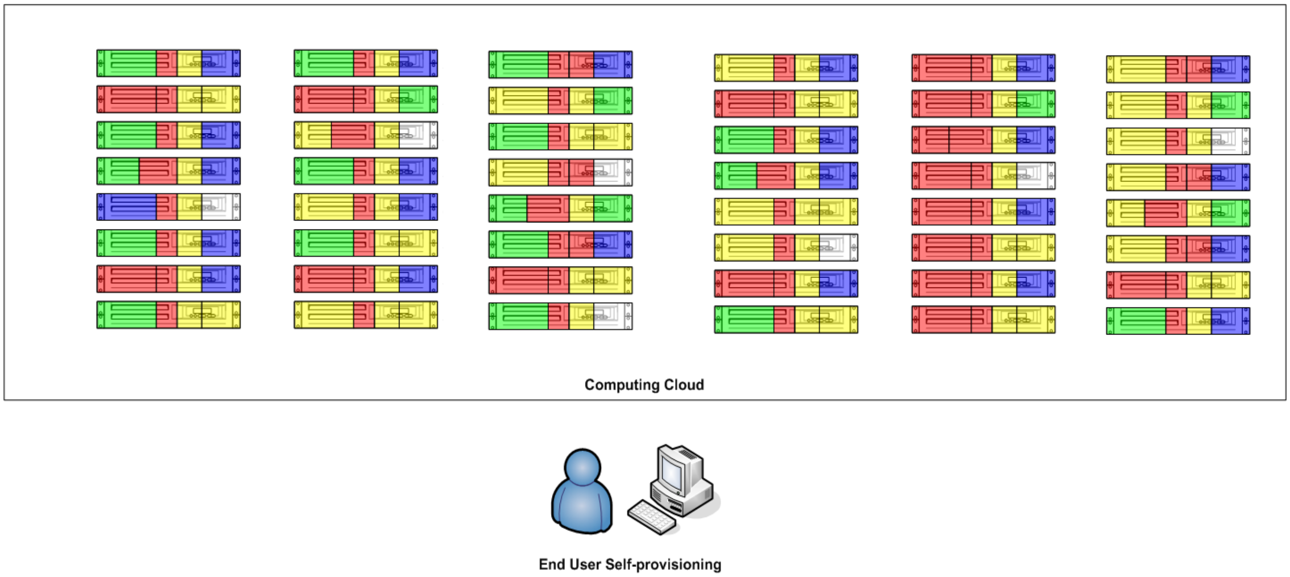


**Figure 3.8: Self-provisioning allows cloud consumers to allocate and manage their own resources.**

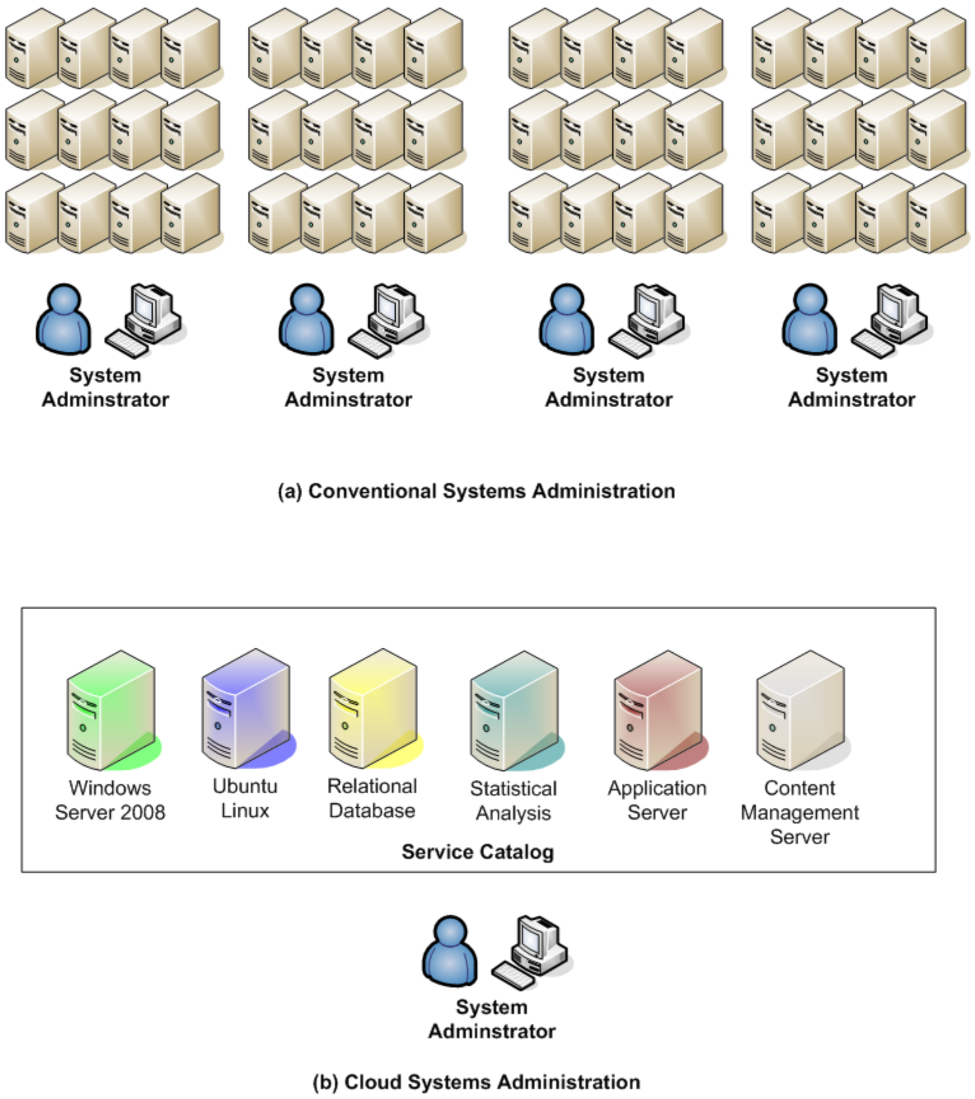
With a self-provisioning system, cloud consumers have access to management systems that allow them to specify the type and number of virtual instances to create. All the hardware in the cloud is managed centrally and virtual machine images are maintained in a service catalog, so cloud consumers do not have to deal with low-level details. For instance, details about what device drivers have to be installed or which libraries are needed to run an application have already been addressed when the virtual images were created. Also, cloud infrastructure abstracts implementation details such as allocating memory or CPUs to particular virtual machine instances.

### Reducing Marginal Costs of Systems Administration

To understand how a cloud infrastructure can result in significant ROI, you only need to look at how systems administration changes with the cloud. A typical list of systems administration tasks include:

- Installing new applications and packages on servers
- Patching OSs and applications on each server
- Backing up local storage on each server
- Allocating space to file systems as needed
- Reviewing and purging log files
- Performing security checks, such as running vulnerability scanners and reviewing results for each server

In conventional environments, systems managers have to repeat these tasks for each server. Fortunately, service management tools support these efforts, but they can still be time consuming. Consistency across servers is important to reduce the amount of time required to maintain systems; however, as the number of servers grows, so does the chance of human error during systems management operations.



(a) Conventional Systems Administration

(b) Cloud Systems Administration

**Figure 3.9: Cloud systems administration entails maintaining images in the service catalog, unlike traditional systems administration, which is linked to each physical server.**

In the cloud, maintaining individual servers is swapped for maintaining virtual machine images in the service catalog. The service catalog is the set of images available for running in the cloud. For example, there may be several Windows server and Linux images that have been configured for general use. There may also be more specialized images for relational databases or content management systems. Still other images may be designed for developers who need to routinely instantiate application servers for development and testing as well as for production use. Having a centralized repository of virtual machine images can significantly reduce the time required to perform routine tasks. Consider a simple example.

A midsize business could easily run 200 servers with a mix of OSs and applications. If a critical security patch is released and has to be applied to 50 servers, the patch has to be applied 50 times. Even with patch management applications to help, systems administrators will have to verify the success of the patch in each case. In cases where automated tools are not available, systems administrators will have to apply each patch manually. Now compare that with patching a service catalog. The existing image is removed from the catalog; a new patched version is generated and uploaded into the catalog. What could have taken 50 distinct tasks is done in one step.

This example does raise another difference from a systems management perspective. The service catalog image is patched, but there may be instances of the unpatched image running in the cloud. Where are those images? How long will they continue to run? At what point should the instances be shut down and restarted using the patched version? The first two questions can be addressed using cloud management software. The last issue is a question of policy analogous to deciding when to schedule a critical patch for a server. Systems administration in the cloud may be less labor intensive but sometimes difficult decisions about balancing security or stability with business expectations remain.

### Standardization and Automation

Another reason for operations-related ROI is that by standardizing on a set of general purpose virtual machine images, you reduce the overhead in maintaining them. Images are deployed and virtual machine instances are started using a management console, so a cloud user who knows how to deploy a Windows server knows how to deploy a Linux server or a relational database as well. Standardization also enables behind-the-scenes automation that further reduces the demand for systems administrator expertise.

For example, when you install Linux on a server, you have to decide what type of file system to use and how to partition the disk. These are not particularly difficult tasks, but you do need to know something about how partitions are used, how much space to allocate to each, and the tradeoffs between the different kinds of file systems. When you instantiate servers in the cloud, you do not have to worry about storage services, they are provided for you. The images in the service catalog are configured to work with cloud storage services. Much of the tedium of setting up monitoring processes to collect performance and usage data is also automated with service management systems.

### Service Management Reporting

ROI is not just about technology but about how you manage it. With service management reporting, service providers can better understand the resources they use and adjust their allocations accordingly. Some of the measurements service providers might use include:

- Number of server hours allocated
- Overall average server utilization
- Average server utilization by hour
- Average server utilization by instance type
- Total storage space used
- Amount of network I/O

Data on these measurements can help determine how many servers to allocate and how long to run them. Data on storage use and the amount of network I/O can help guide optimization of application performance, especially if there are charges based on network traffic.

Many aspects of cloud computing contribute to the ROI in the technology. Capital expenditures are significantly lower, if not eliminated, for new service deployment when using the cloud. The big savings, however, comes from reduced operational labor costs enabled by self-service management, automation, and standardization.

## Assessing the Business Value of Cloud Services

The ROI in cloud technologies will vary from one business to another. Much will depend on factors out of your control, such as economies of scale that will benefit larger businesses than smaller ones, as well as factors you can manage, such as server utilization rates. To assess the value of cloud services to a business, consider several cloud metrics as well as the source of ROI for your particular business.

The reason to track particular metrics in cloud computing is no different than that of any other business operation: to quantify the costs and benefits of the service. This is especially important when using a private or hybrid cloud model. Key metrics for these clouds are:

- **Utilization of all cloud resources**. If resources are underutilized, servers can be powered down to save on energy costs. IT may also want to promote the use of the cloud and publicize availability of resources.
- **Systems management hours**. Labor can account for significant portions of IT operating budgets but should be significantly less for cloud services.

- **Virtual machine image use**. All images in a service catalog have to be maintained. If some images are not used, or used infrequently, they may be incurring more costs than they recoup through usage charges. Infrequent use or use by only one user can also indicate specialized or "one off" images. These are sometimes necessary to meet business requirements, but if the number of specialized images grows, the cost of maintaining them will increase. Charges may need to be adjusted to recoup the full costs of maintaining specialized images.
- **Time to provision**. This metric can indicate insufficient resources in the cloud. If a sufficient number of servers are not available, users will have to wait for other jobs to finish in the cloud before there virtual machine instances will be provisioned.

In addition to these more global metrics, looking at ROI based on specific elements of cloud infrastructure is useful as well. These include the ROI realized from:

- Improved hardware utilization, especially when fewer servers are required to meet a workload leading to reduced capital costs, lower maintenance costs, and reduced energy costs
- Lower software costs because software licensed per server can have improved utilization that parallels hardware utilization
- Self-service management, which reduces systems administration
- Increased productivity due to reduced wait time to deploy servers and applications

Cloud computing is an evolution in information technology and so it is not surprising that many of the same metrics and ROI factors we have used in IT for years have analogs in cloud computing as well.

## Summary

Cloud computing offers new ways to deliver business services. As the two example scenarios highlighted, different types of business applications can benefit from deploying in the cloud. The ability to scale compute and storage resources as needed reduces the need to build for peak demand. This, in turn, reduces the cost of delivering services while avoiding costly risk mitigation strategies, such as adding contingency funds to a project budget to purchase additional hardware to meet unexpected demand.

Further benefits of cloud computing accrue with regards to reducing the time to deploy new services, more ways to control costs, and the adaptability of resources. Servers in the cloud can be repurposed rapidly and with minimal technical expertise, reducing the need for dedicated servers and their typical low utilization rates.

Perhaps the primary driver for the adoption of cloud computing is the ROI. Capital costs are reduced largely due to higher utilization rates of servers. Even more substantial savings can realized with self-service management and savings in systems management. With standardized images, automation, and service management reporting, cloud users can not only deploy services in the cloud but also manage them effectively.

The first three chapters have introduced cloud computing, examined some of the technical aspects, and described in general how cloud computing can improve service delivery. In the next chapter, we will turn our attention to the question of how to begin planning for cloud services in your business.

# Chapter 4: How Cloud Computing Will Help Your Business

Cloud computing changes the way we do business. Much of the coverage of cloud computing has focused on the technical aspects of this computing model: the consolidation of servers, virtualization, security, and so on. This is understandable, as you must have a clear idea of what cloud computing offers from a technical perspective before you can appreciate what it can do for you from a business perspective.

This chapter turns attention to the business side of cloud computing. In particular, this chapter considers the following:

- How cloud computing can help your business
- Assessing current capabilities
- Introducing cloud computing as a new model for consumption and delivery
- Measuring the value of a cloud

The discussion will start by identifying key business priorities; move on to looking at the current state of IT infrastructure, services, and procedures; describe how to transition those capabilities to a cloud-enabled enterprise; and finally assess the financial value of a cloud to the organization.

Cloud computing has and always will be a rich technical area, but its application to real-world business problems requires the examination of organizational issues that range from bordering on the technical to being the provenance of business executives responsible for overall strategy.

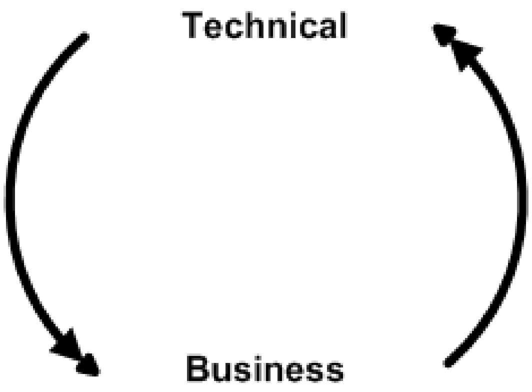


**Figure 4.1: The technical aspects of cloud computing shape what can be done from a business perspective; the business drivers determine how cloud computing is applied.**

## How Cloud Computing Can Help Your Business

Adopting cloud computing is a major change from the traditional distributed systems models many of us use today. No rational business person would make such a fundamental change to core infrastructure without understanding the consequences for the businesses. After all, if your current computing and storage systems are meeting your needs, why change? Why bring on the risks associated with a new technology. Certainly, there is some allure to being on the cutting edge and having the latest technology, but chasing technology trends for their own sake is not a sound strategy for long-term business success. Instead, technology is adopted in service to a business strategy.

Moving your business to the cloud should begin with considerations that have nothing to do with clouds, at least not yet. Cloud computing is a solution; the first question to ask: What is the problem? To get to the answer to that question, you need to:

- Identify business priorities
- Identify operation inefficiency
- Identify barriers to innovation

These steps help you identify what the business is striving to achieve and what is hindering those efforts. Only after you understand that can you turn your attention to addressing the problems that thwart, delay, and increase the cost of your business operations.

Before proceeding to consider priorities, inefficiencies, and barriers to innovation, let's consider a hypothetical scenario that highlights the issues relevant to aligning business goals and technology infrastructure.

### Business and Technology Alignments: The Ideal vs. Reality

Consider a healthcare provider with several hospitals, tens of clinics, and hundreds of doctors serving thousands of patients. As part of a strategic plan to improve the quality of services while controlling costs, executives at the healthcare provider decide to disseminate information on patient conditions, treatments, and outcomes. The executives believe, with sufficient feedback on the results of treatment choices along with details on the cases where particular treatments work and do not work, physicians will be able to reduce uncertainty associated with selecting treatment options.

To implement this plan, the healthcare provider will have to:

- Create a consolidated reporting system such as a data warehouse
- Develop procedures for extracting and loading data from multiple sites into the data warehouse
- Create a reporting infrastructure to deliver information to physicians in a way that is easy to use and fits with their work patterns
- Establish governance over the data warehouse and reporting procedures to ensure compliance with HIPAA and any other relevant regulations
- Define a mechanism to collect feedback from users to improve the system

With the high-level requirements in place, the next step is to determine how the IT department will proceed to implement the plan. Some of the issues that would likely arise include:

- Acquiring servers and storage to house the data warehouse
- Purchasing licenses for database, reporting, and extraction, transformation, and load (ETL) tools
- Assembling a team to install and configure the infrastructure once it is acquired
- Designing logical and physical data models for the data warehouse
- Developing reports
- Establishing access controls over reports and data
- Creating a support team to monitor data warehouse processes and provide end user support

There are other items you could add to the list, but the list is sufficient to demonstrate the potential drag that IT operations can have on business initiatives. First, though, let's depict an ideal scenario.

IT has sufficient server and storage capacity for the data warehouse. Development work can begin immediately. Fortunately, IT has standardized on a relational database, a data warehousing methodology, and reporting tools. These applications already work with the identity management system in place at the organization, so access controls can be readily established and managed. The support services group within IT is already familiar with these standardized tools, so there is minimal marginal cost to support another set of users.

This ideal scenario is one in which infrastructure, standardized applications, and support services are in place and readily available for new initiatives. To many, this is a fantasy; the reality that many of us have experienced in projects like this is far different from this scenario. Here is a version of the scenario that might ring true for more readers.

The requirements for the reporting project outstrip the available budget. Requirements will have to be prioritized and some features will have to be delayed until later phases. There is insufficient storage available to the business department that owns this project. (There is plenty of storage on another department's disk array, but organizational boundaries rule out using it.) Hardware will have to be procured and installed. Rack space and cabling are a problem that can be worked out with the infrastructure management group, which, given their backlog, will be in a few weeks. The company has a site license for the database software but this project requires several additional packages that will have to be purchased. Several reporting tools are used in other business intelligence projects, so an internal evaluation will be done to determine the best tool for this effort.

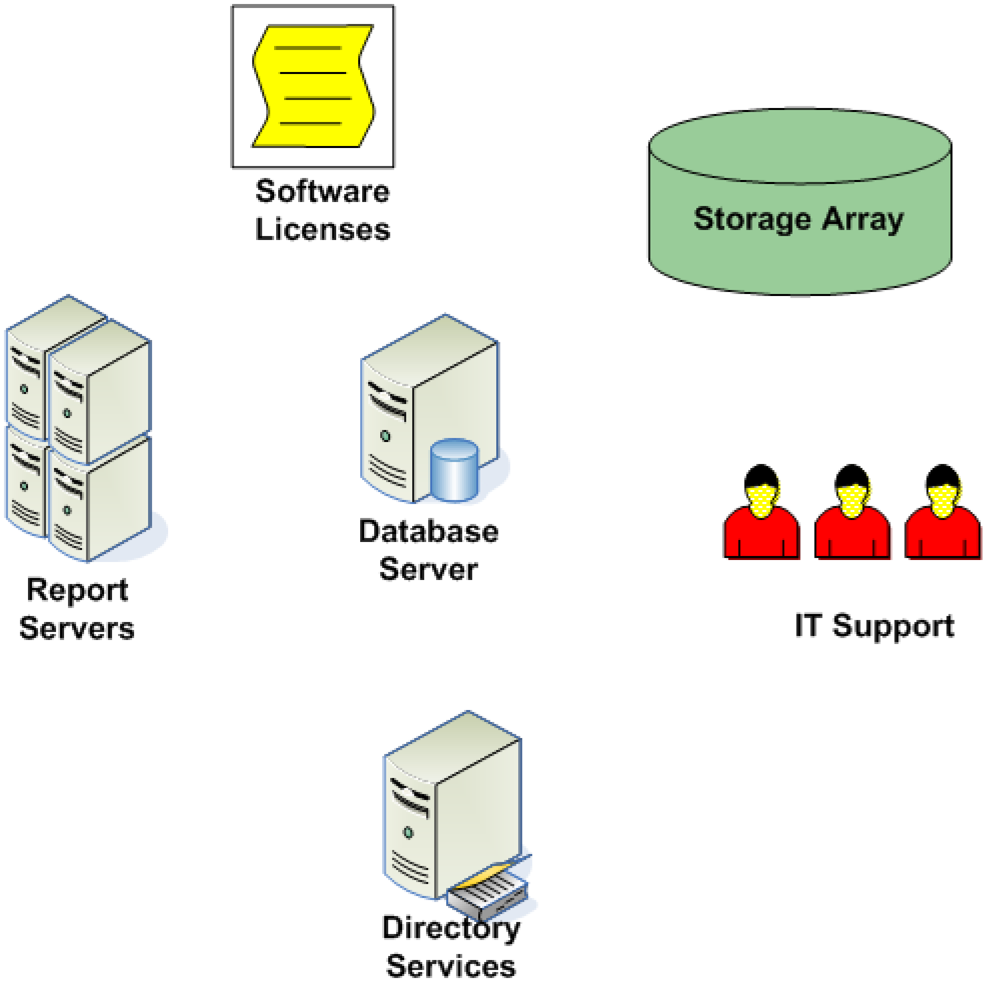


**Figure 4.2: Innovative IT projects require an array of technology and services that can be difficult to coordinate and integrate.**

The collective effect of seemingly small and, in many cases, expected problems is to slow down and prolong the implementation. In this scenario, the needs of the business for rapidly deploying an important medical decision support tool is hampered by the way budget for projects, allocated resource across organizational lines, failure to standardized as much as possible on software, and tax support staff with inefficient deployments.

## Identify Business Priorities

One of the most important aspects of successful IT services is that they align with business goals. That is a short way of saying IT services support business objectives in a straightforward manner and do not introduce unnecessary cost, delays, or other burdens on a business strategy.

Although there is no set of priorities that would apply to all or even most companies, common priorities include:

- Controlling costs
- Expanding market share in mature industries
- Expanding into new markets in growth industries
- Improving customer service
- Improving customer retention
- Increasing cross selling

Whatever the priorities and their relative importance, it is critical to identify these for a business. Knowing these will help determine whether and how cloud computing can help your business. For example, if cost control is a top priority, increasing server utilization through virtualization and increasing storage utilization through consolidation with cloud computing architectures can help. If improving customer retention is important, you may need to invest in advanced analytics, such as data mining and statistical analysis, to detect early warning signs of churn. Advanced analytics can be compute-intensive and is a good application for cloud computing. Of course, knowing your business priorities may lead to the conclusion that cloud computing is not something you need at the moment. Whatever your conclusion, if you start with business priorities, you will at least justify why or why not to pursue cloud computing or any other technology.

> **Caution**
>
> To be clear, cloud computing is not a panacea that will solve all your problems. There are times when cloud computing is not the right solution. It may be an appealing option at a later time, but your business may not be in a position to move to the cloud until it improves its IT governance practices, for example.

### Identify Operational Inefficiencies

Operational inefficiencies are a drain on the bottom line. When an employee has to perform ten steps to complete a task that could be done in six steps, the business loses productivity. When servers are powered on and functioning but not running productive jobs, the business is realizing an opportunity cost as well as incurring unnecessary energy costs. Operational inefficiencies, ironically, are often found in IT departments that have traditionally been a source of increased productivity. Operational inefficiencies in the IT realm come from the way we deploy and utilize hardware and the way we manage software.

Low server utilization is a common inefficiency. Prior to the widespread adoption of virtualization, many organizations used a “one application, one server” approach to deployment. This approach minimized problems with conflicting requirements and allowed administrators to manage servers and applications as a tightly coupled unit. The price we paid for this was wasted CPU cycles.

Less obvious was the management inefficiencies that came with the one application, one server model. Configurations could be tailored to individual applications, so an IT group could soon find that many of its servers had different operating system (OS) components installed, with different applications and services and all requiring slightly different support. As a result, the number of servers a single administrator could manage was limited. Had a standard configuration been developed for a limited set of roles, systems administrators would have less variation to contend with, and this would result in more productive systems management.

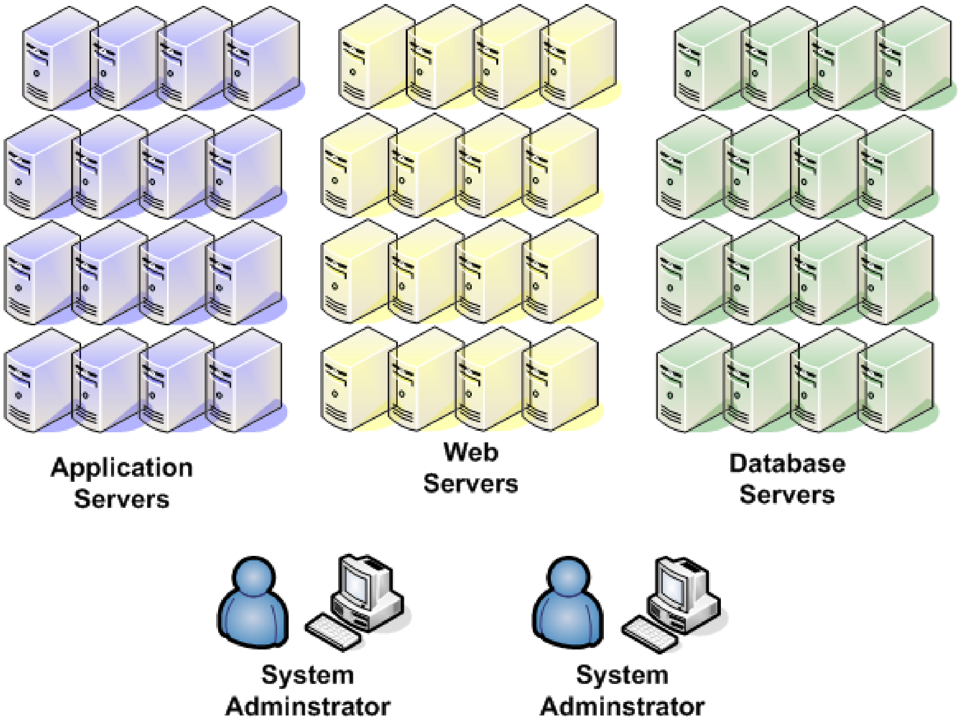


**Figure 4.3: Standardizing server configurations is one way to reduce systems management inefficiencies.**

## Identify Barriers to Innovation

The scenario described earlier shows how innovation can stagnate because of technology barriers. It is important to note that the barriers in that scenario were not caused by poor management or unskilled IT professionals; the problem arose from the constraints on procuring new hardware, configuring software, and ordering a sequence of deployment events that account for a range of dependencies between steps.

The scenario can help you see the different types of barriers to innovation that can creep into IT operations:

- Delays in procuring and deploying hardware
- Initial capital costs
- Ongoing operational costs
- Insufficient support staff
- The need to evaluate, select, and coordinate multiple software components

Any or all of these can be significant barriers to innovation. In the time of a globalized economy, customers have more options than ever before, businesses have access to a wider pool of suppliers and business partners, and the list of potential competitors is more likely to grow than not. Add to this list the demands on companies to consistently meet performance expectations quarter after quarter, and you see that barriers to innovation can be a potential long-term drag on the company overall.

The first steps to understanding how cloud computing can help your business is to formulate a clear picture of business priorities, pinpoint operational inefficiencies, and identify barriers to innovation. These three elements comprise the key business drivers that can guide the successful use of cloud computing in your organization. As noted earlier in this chapter, business requirements drive technology-based solutions, but before adapting new technologies, it helps to have a clear understanding of current technical capabilities.

## Assessing Current Capabilities

Technology capabilities are a combination of hardware and software infrastructure within an organization as well as the management practices that govern the use of that technology. For the purposes of understanding the role of cloud computing in improving business services, let's consider several types of capabilities:

- Infrastructure
- Platforms
- Applications
- Governance
- Management and reporting

It is important to have a clear and comprehensive understanding of these capabilities because they are all relevant to adopting the cloud computing model. Cloud computing is an evolutionary advance in computer architecture and service management; it improves on what came before it but does not represent a wholesale replacement. Sound management practices, software development life cycle methodologies, and systems administration practices are as relevant to delivering services through a cloud as they are to other delivery methods.

### Infrastructure Capabilities

IT infrastructure for the purpose of this discussion includes server and storage hardware as well as networking components. When assessing current capabilities with regards to infrastructure, consider:

- The inventory of servers currently in place
- The geographic location of servers
- The costs of maintaining each server
- Utilization metrics for servers
- Network infrastructure between sites

At the end of the infrastructure assessment, you should have a clear idea of overall server utilization. If your operations are similar to most, you will have many servers running single applications, and those servers were configured to handle peak, not average, capacity. If this is the case, a move to cloud computing is an opportunity to consolidate servers and decommission those with high maintenance costs, no standard configurations, and relatively low performance. Such consolidation can have an immediate impact on the power and cooling costs of a data center.

There may also be an opportunity to consolidate data centers or at least servers currently located in remote offices. Reducing the number of sites can help ease management overhead and streamline IT operations such as backups.

### Platform Capabilities

Platforms are the OSs and application stacks that run on a company's IT infrastructure. Enterprises typically have a number of platforms:

- Windows
- Linux
- Unix
- Mainframe OSs

Windows and Linux often run on servers that provide specialized functions, such as email servers, content management servers, databases, and directory services. Unix and mainframe OSs are typically found on enterprise-scale computers running high-volume, mission-critical applications.

Cloud infrastructure can be built using low-cost commodity hardware, so such hardware are ideal candidates for hosting Windows and Linux platforms; of course, Unix OSs run on these servers as well.

For the purposes of assessment, you should collect information about:

- The number and version of Windows OSs
- The number, version, and distribution of Linux OSs
- The application stacks that run on these OSs

The goal here is once again to consolidate as much as possible.

### OS Consolidation

Standardizing on a reduced number of platforms will reduce systems management tasks and provide a step toward the type of self-service management that is such an important factor in cloud computing's ROI. Standardizing in this case does not mean committing to using only Windows or only Linux but to reducing the amount of variation in the platforms. For example, if a department is still running an instance of Windows Server 2000, this is a good time to move those applications to Windows Server 2008. Similarly, if several distributions of Linux are currently supported, consider reducing that number. It may not be possible to find a Linux distribution that is optimal for all needs, but you might find you can use fewer different distributions than you currently have.

### Application Stacks

Application stacks are middleware that reduces dependencies between applications and OSs. When applications are written directly to an OS, they can be difficult to port. Even similar OSs, like different versions of Unix, can harbor enough differences to make porting software difficult. Application stacks and middleware abstract low-level OS details and provide a consistent programmatic interface and set of services. They will be just as important in a cloud environment as they are in today's distributed system environments.

Common application stacks are:

- Microsoft .Net
- LAMP (Linux/Apache/MySQL/Perl or Python)
- J2EE (Java 2 Enterprise Edition)

Application stacks are chosen for their fit with system requirements and the skills of developers working on the applications. Moving applications from one stack to another can be a considerable undertaking, so there is probably less opportunity to consolidate at this platform level. Just as important, though, you will want to ensure that all application stacks currently in use and needed in the future are supported in the cloud.

## Microsoft .NET Framework

Microsoft .NET Framework is a development framework for building Web applications for Microsoft platforms. The framework includes several components:

- A common language runtime that acts as an abstraction layer above OS functions
- Base class libraries
- Support for both compiled languages, such as Visual Basic and Visual C#, as well as dynamic languages such as IronRuby and IronPython
- Windows Presentation Foundation, a user interface (UI) framework
- Silverlight, a set of .NET tools for building rich Internet applications (RIAs)
- Windows Communication Foundation (WCF) for service-oriented architectures
- ADO.NET, a set of data access services
- Windows Workflow Foundation

Not surprisingly, the .NET Framework is designed to leverage SQL Server database and other OLE/ODBC data sources.

## LAMP (Linux/Apache/MySQL/Perl or Python)

LAMP is a set of commonly used open source systems for building Web applications. Unlike the Microsoft .NET Framework, the individual components of this set of platform tools had long and well-developed histories prior to the advent of LAMP. Each of the four components provides a basic service commonly needed in Web applications:

- Linux is the OS underlying the LAMP stack
- Apache is the Web server and related modules that may be installed as needed for particular applications
- MySQL is a popular open source database suitable for a range of applications sizes and needs
- PHP, Perl, and Python are scripting languages used to implement custom application functions

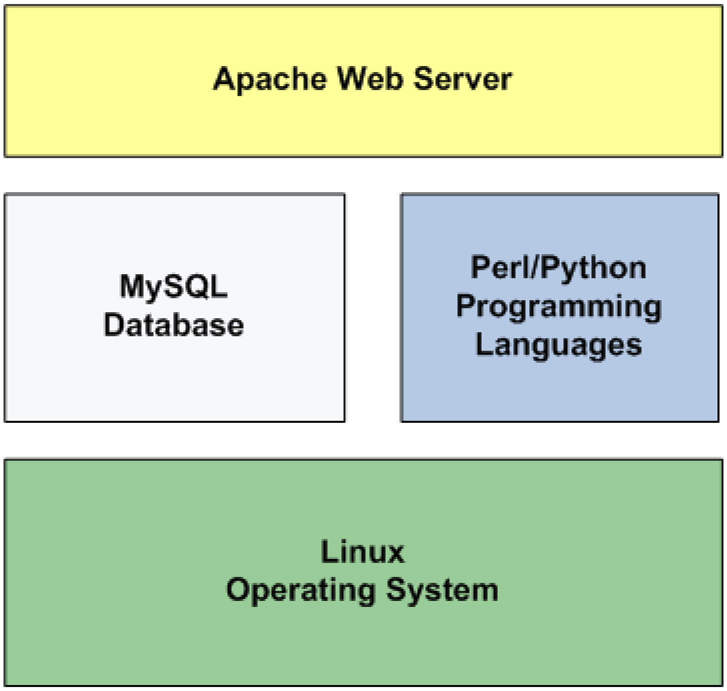


**Figure 4.4: The LAMP stack consists of a small number of independently developed open source components that are commonly used together for Web application development.**

## Java Platform Enterprise Edition

The Java Platform Enterprise Edition, sometimes referred to as J2EE, is a middleware framework designed for deploying distributed Java applications. Like the Microsoft .NET Framework, there are multiple components providing a range of services for application developers. These include:

- Enterprise JavaBeans, a distributed object container
- Java Transactions API
- Java Messaging Service API
- Java EE Connector Architecture
- Java XML Streams
- Java Persistence API
- Java Server Faces, a UI framework

These components are bundled together and incorporated into Java application servers. The application servers sometimes run higher-level additional application development components, such as portals and content management systems.

As you can see from the descriptions of three common application platforms, they provide many of the same functions but do so with fundamentally different tool sets. When assessing existing capabilities, it is important to catalog all the platforms and main platform components used so that they can be supported in the cloud as well.

### Application Capabilities

Enterprises run a wide range of applications and many of these are suitable for running in the cloud. The goal of assessing application capabilities is to determine:

- The relative priority of moving an application to the cloud
- The difficulty to move the application to the cloud
- Changes to application management practices that may be needed after the move to the cloud
- Potential risks and mitigating strategies

When it comes to prioritizing moving applications to the cloud, you should look for those systems that are (a) under-utilizing the servers they run on, (b) are running below needed performance levels because the hardware does not adequately serve current loads, or (c) have peak demands that could take advantage of elastic allocation of CPU and storage capacity of the cloud.

At the same time, you want to avoid immediately moving applications to the cloud that may have special requirements. For example, high-security applications that would require any deleted data be not only deleted but overwritten multiple times to reduce the risk of unauthorized reconstruction of that data. (Deleting data can be done by marking a data block as available for use, so old data can continue to reside on the disk even after files have been logically deleted.) It is not the case that these applications can never be moved to the cloud; they can once security procedures are in place to meet the application requirements.

Running applications in the cloud and on a dedicated server will require different management routines. For example, a cloud storage provider may provide sufficient redundancy in data duplication that you may reduce the number of backups performed. Also, as departments will likely be billed for the time virtual machine instances are running, they will want to optimize their workflows to keep the virtual servers utilized as much as possible when they are running.

Billing rules should also be considered when scheduling jobs. For example, if a department is charged for a full hour of virtual server time regardless of how much of that hour is utilized, it would be best to schedule jobs continuously rather than shutting down and restarting. Of course, this assumes that jobs can be scheduled together that require the same platform. As even this simple scenario shows, the way you manage cloud applications will have to account for new billing structures and server use patterns.

Another area you need to consider during the application assessment is the risk of moving applications to the cloud. These risks include:

- If using a public cloud provider, ensuring security and compliance standards can be met
- Being able to run applications on a standardized platform that may require specialized runtime libraries or utilities
- Having access to virtual machine images that support the application; this can be an issue if a cloud provider only offers the latest patch version of an OS and the application does not run correctly under that particular version

Moving applications to the cloud entails sharing control with the cloud provider. This may be less of an issue when using a private cloud, but it still must be considered at the level of application requirements. These considerations with regard to applications represent just some of the broader issues that must be addressed as part of governance processes.

### Governance Capabilities

A capabilities assessment should include an assessment of governance practices as well. Although cloud architectures are fault tolerant and resilient, the governance practices for clouds are a potential single source of failure. Poor governance affects all users of the cloud.

Governance of cloud operations is required for all types of clouds: private, public, and hybrid. The policies that are implemented will vary by type of cloud, but in general, they will include:

- Complying with government and industry regulations
- Defining and enforcing audit controls and security procedures
- Establishing cost allocation and cost recovery policies
- Setting policies on the management of the service catalog
- Adjusting existing policies to accommodate cloud services

These governance requirements should not be new with cloud computing. The need for governance is independent of IT architecture choices. As noted earlier, though, the cloud changes the way you deliver services and provides new opportunities to change management or governance policies. For example, change control policies may become more flexible with regards to platform-level changes because multiple versions can coexist in the service catalog.

The goal of the governance capability assessment is to understand the mechanisms that are already in place to guide IT operations, identify weaknesses, and make necessary changes. Cloud computing will not improve governance practices, but poor governance can eventually undermine the value of the investment in cloud computing.

## Management and Reporting Capabilities

With cloud computing, service consumers have greater control over how they use computing and storage resources. To optimize their use of these resources, they need information about their workloads, levels of utilization, costs, and other metrics. Management reports are the key to delivering that information.

Reports and data on cloud usage should be available for both front-line managers responsible for scheduling jobs and budgeting for services and for back-office billing operations. Front-line managers should have access to near real-time billing information on CPU utilization and storage allocations so that they can tune workflows. They should also have comparative historical data so that they can detect trends and properly plan for future needs. When a private cloud is used, back-office billing systems will need to accommodate billing or charge backs for cloud services. Existing financial reports would then provide an additional set of reports for front-line managers.

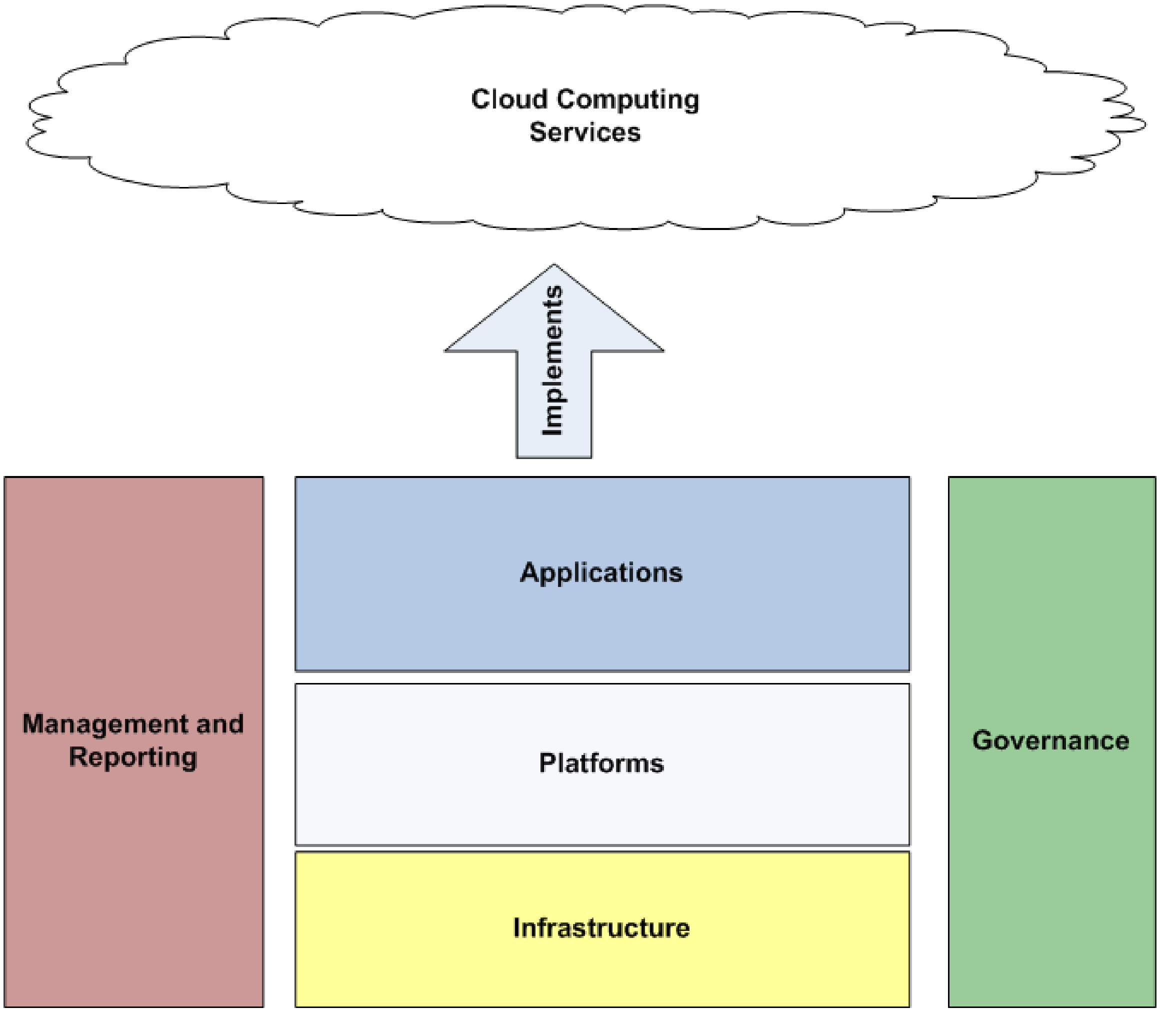


**Figure 4.5: Organizational capabilities in the form of infrastructure, platforms, applications, governance and management, and reporting enable the deployment and use of cloud computing services.**

An assessment of organizational capabilities, spanning existing infrastructure, platform, applications, governance and management, and reporting procedures will provide an organization with a starting point for introducing cloud computing services.

## Introducing a New Model for Consumption and Delivery

Introducing cloud computing can be done in two ways: by using a public cloud or by using a private cloud. We will focus most of our attention on the latter, but we will briefly address the use of public clouds.

### Introducing Public Cloud Consumption Model

Public clouds can be introduced quickly for small, experimental evaluations that do not involve confidential data, specialized workflows, or complex security requirements. In a very short time, a department-level manager could:

- Establish an account with a public cloud provider
- Upload data for analysis into the public cloud provider's storage system
- Select from the public provider's catalog of OSs and other platform software
- Allocate the necessary number and types of servers
- Run the job
- Shut down the servers, collect the results, and complete the task

This type of isolated, tactical use can also be done in cases where confidential data, specialized workflows, or complex security requirements exist, but it would take significantly more planning, along the lines of what we will be describing shortly in the discussion of a private cloud deployment.

Public clouds allow consumers to experiment with the cloud delivery model without fully committing hardware, software, and management to a full-scale deployment. It is also a viable option for meeting peak demands of jobs that are readily moved to a public cloud. Running significant portions of your business services in the cloud for extended periods can certainly be done but will require the type of attention and planning that one finds with the use of private clouds.

### Introducing Private Cloud Consumption Model

There is nothing inherent in cloud computing that requires the cloud be owned and operated by another business. Cloud computing is an architecture and a set of services that enable access resources on demand. The infrastructure and services are managed by the provider and used by service consumers. The provider can be a third party offering a service to the public or an IT division within a company offering computing and storage services to other departments within the company.

A private cloud may appear to lack some of the economic advantages of cloud computing, such as lower management costs and no need for capital expenditures. This may or may not be the case with private clouds; the economic benefit will depend on circumstances within the business providing a private cloud. If the business has a large existing infrastructure with low utilization and high systems management overhead, the company could benefit from redeploying their infrastructure to a private cloud. The number of servers could be reduced because fewer will be needed to meet existing demands. Management overhead could be simplified with cloud management software. In cases where capital expenditures are required, businesses can still benefit from spending less on infrastructure than they would if they did not use a cloud-based approach.

Introducing a private cloud will entail changing procedures and practices; these changes fall into three areas:

- Deploying existing infrastructure in a private cloud
- Enabling application services in a cloud
- Managing a cloud

### Deploying Existing Infrastructure in a Private Cloud

The first step is to establish the hardware infrastructure for running the cloud. Existing hardware may be used for this, but of course, it will require planning to ensure existing services are not disrupted during the transition.

The first step is to identify the servers to use for cloud services. One of the goals of cloud computing is to increase the server utilization, so you would expect to use fewer servers for the same level of demand. If this is the case, older servers with lower overall performance and higher maintenance costs are obvious targets for elimination. Some of the factors to consider when selecting hardware:

- Number of CPUs and cores in the server
- Amount and speed of RAM
- Network interface card throughput
- Cost of maintenance contracts, if any
- Cost of leasing contracts, if any
- Power consumption
- Cooling requirements
- Standardization

Most of these are common sense considerations. The last, standardization, addresses the fact that you should expect hardware failures in the cloud. Actually, you should expect hardware failure in architectural configuration. By standardizing hardware, you reduce the need to maintain multiple types of backup components and streamline troubleshooting procedures.

Storage hardware should be selected for comparable reasons: speed, capacity, throughput, power consumption, cooling, and so on.

Network capacity and throughput should also be considered at the early deployment state. If data centers are being consolidated, additional network capacity may be required. Also, consider the levels of redundancy on the network to ensure services can continue at needed levels if, for example, one Internet access provider is down.

### Enabling Application Services in the a Cloud

Application services begin with a service catalog. This is the set of all OSs, middleware, and applications that will run in the cloud. As with deploying hardware, this is an opportunity to standardize on software components. The advantage of standardizing is that there are fewer pieces of software to manage, patch, and configure and that ultimately leads to reduced support costs.

Software services in the catalog should be based on business requirements. There will be needs for different OSs and application stacks, possibly in multiple configurations. For example, to support existing business services, the service catalog may need to include:

- Windows Server 2008 with .NET Framework
- Windows Server 2008 with Java Enterprise Edition framework
- Linux with LAMP framework

In addition to the applications needed in the existing configuration, there will be additional software needed to manage the cloud.

### Managing a Private Cloud

Managing a private cloud requires software and procedures. Operation management software is needed to track the use of compute and storage resources in the cloud. As noted earlier, cloud consumers should have the ability to track their use and costs as they make use of services. They should also have the ability to:

- Monitor their jobs
- Schedule their jobs
- Establish complex workflows
- Track storage use
- Create specialized virtual machine images with custom configurations

Management tools are also needed by IT support staff to maintain the services catalog. For example, systems administrators should be able to track metadata about every item in a service catalog, such as:

- Description of a virtual machine image
- Date it was created
- Applications, libraries, and utilities included
- Patch levels
- Number of times instantiated
- Location of sources used to construct the image

Many of the same service management procedures used in non-cloud environments are still relevant to the cloud. Images will need to be patched, access controls will need to be applied, identities will need to be managed, and charges will have to be made to departments.

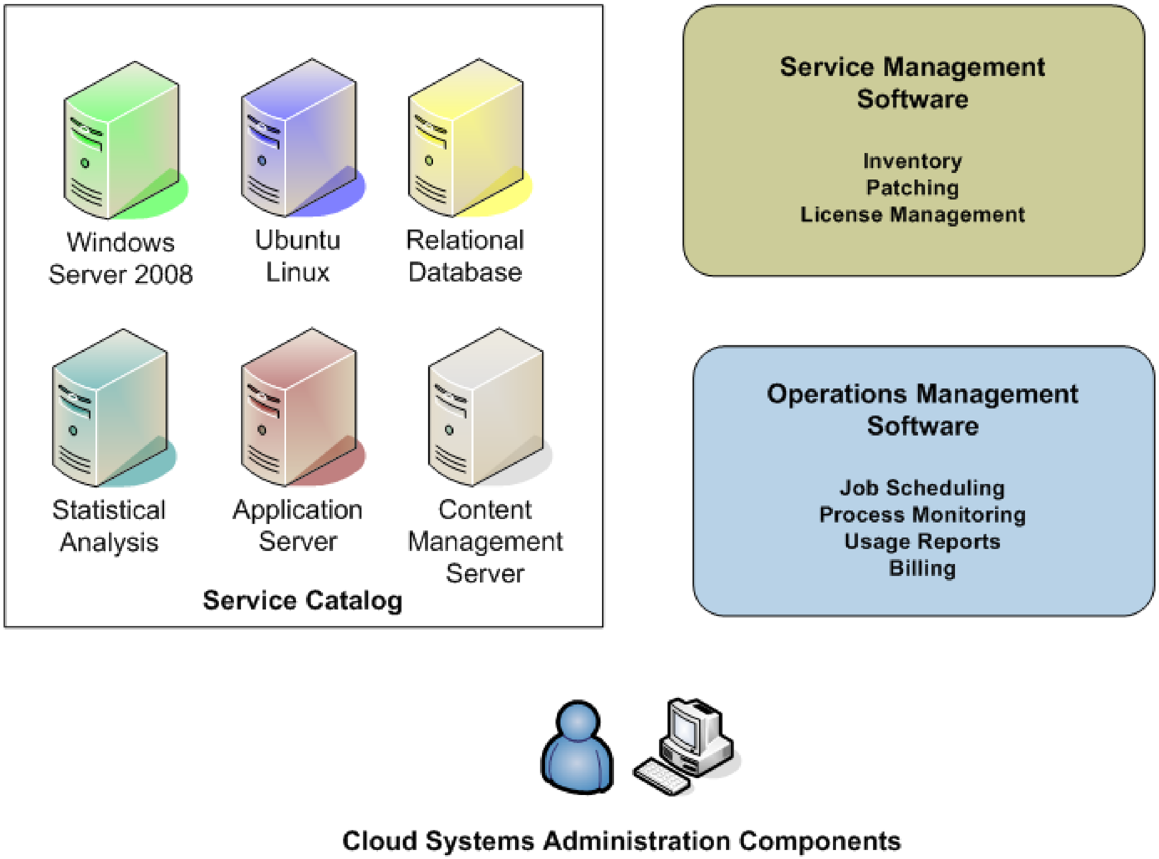


**Figure 4.6: Management components include the service catalog of a platforms and applications available in the cloud as well as management support software.**

## Measuring the Value of a Cloud

Moving to a cloud computing environment will change the IT cost structure and impact both capital cost structures and operational costs.

### Changes to Capital Cost

In a conventional IT model in which departments or service managers use dedicated servers, they often have to plan for capital costs. These are infrequent but significant costs that are budgeted outside the normal operations budget. Although the cost of a server or two may be accommodated in an operational budget, that is not the case for a fully functional application environment.

Consider the costs of developing, testing, deploying, and maintaining applications. For hardware, you would need development and test servers. For small projects, a single high-end server may serve for both as long as each ran in its own virtual environment. The production server may actually be a cluster of servers and a load balancer in order to scale to peak demand. The load balancer will provide some degree of high availability, but disaster recovery procedures dictate a backup set of servers in an offsite location. Storage will be required as well, adding to the capital expenditure. In addition to these hardware costs, there will be the cost of application and OS licenses.

In the cloud, these costs do not go away, but they are reduced. The key is to efficiently share resources rather than dedicate servers and storage arrays to single services or departments. Rather than having multiple service managers develop their own capital budgets, complete with wide ranging contingency funds either explicitly or implicitly added to the budget, central IT can plan for capital costs across a wide base of users. The end result is less capital expenditure because of more efficient use of infrastructure, platforms, and applications.

### Changes to Operational Cost

Cloud computing can prove advantageous for operational costs in four areas:

- Labor
- Infrastructure maintenance
- Facilities operations
- Simplified accounting

#### Labor Costs

Labor costs are reduced with the use of self-service management enabled by cloud management software. Consumers of cloud services have the ability to choose the virtual servers they want to run, determine what applications to run, and schedule them when needed. The standardized service catalog reduces the need for costly software configurations, which in turn depend on access to a skilled IT professional to perform the configuration.

The cloud can reduce IT support labor costs in other ways. With a centralized service catalog, updating and patching becomes less labor intensive. For example, if an OS vendor releases a critical patch that has to be pushed to servers, then hundreds of servers may be involved. This requires identifying which servers need the patch, deploying the patch through an automated delivery system, reviewing the results of the patching operation, and manually applying the patch to those servers that failed to be patched correctly using the automated method. This can be a time-consuming burden on IT support staff with other regularly scheduled tasks to complete. The same patch could be applied to images in the service catalog in fewer numbers because only one copy of each configuration is needed. Also, the patch would be available to users of those images the next time they instantiate their virtual machines.

### Infrastructure Maintenance

Standardization is a well-established method to reduce costs. Standardizing on infrastructure is no exception. Adding new components, such as servers, to a cloud will have low marginal costs if they are configured similarly to servers already in the cloud. If there are failures (and there will be), the new units are readily swapped in without requiring configuration changes. Inventories of spare components are kept to a minimum as well. Clouds run in a centralized data center, so there is less need for remote office visits to deal with failed hardware.

### Facilities Operations

Another contributor to savings in operational costs comes in facilities management. IT infrastructure can consume significant amounts of power leading to high energy costs. Of course, all that power that comes into the data center gets converted to useful computation, but the conversion from electricity to computation is not perfect. The inefficiencies in conversion are realized in the form of heat; heat that has to be removed with costly cooling systems. By driving up the average server utilization, a business can reduce the number of servers needed, which in turn reduces power and cooling costs.

### Simplified Accounting

One of the advantages of cloud computing is that it provides a way to standardize computing and storage units of service. For example, a virtual machine running on a dual core processor (or its functional equivalent) for 1 hour can be defined as a unit of computing resource with a standard price attached to it. Similarly, a gigabyte of storage stored for one day could be a unit of storage for accounting purposes. From these fundamental units, you could build pricing schedules that could account for additional costs for OS or application licenses.

With this type of model, cost recovery is simplified. Cloud consumers can readily plan their expenditures. Reporting and integration with financial systems is less complex than if a large number of specialized cases and accompanying business rules have to be accommodated. Cloud computing presents clear cost benefits in both capital and operational costs—as long as proper planning and assessment are done.

## Summary

Cloud computing is an efficient framework for utilizing computing resources. To get the most of your investment, begin by assessing the current state of business and technical operations. This includes identifying business priorities, operational inefficiencies, and barriers to innovation. It also entails assessing the current capabilities in terms of infrastructure, platforms, applications, governance and management, and reporting. Deploying a cloud is a multi-stage process that includes deploying existing infrastructure, enabling application services, and managing the cloud. The value of the cloud will be measured in both capital and operational cost savings.

# Chapter 5: Strategies for Moving to the Cloud

Cloud computing is a framework for delivering services that, as we have seen in previous chapters, offers a number of compelling benefits. Now it is time to turn our attention to strategies for moving an organization from thinking about cloud computing to using cloud computing. Many of the same rational methods and management techniques we use in IT planning and deliver today are relevant to cloud computing. This is not surprising. As I have noted in this book, cloud computing is a phase in the evolution of IT services delivery; it builds on previous practices to deliver new levels of efficiency, control, and manageability.

This chapter focuses on how to plan for the organizational and technical issues around the move to cloud computing. It is specifically structured around three broad topics:

- Planning principles
- Architectural principles
- Use case scenarios

The first section on planning principles will describe a process for understanding the current state of IT services and framing them in such a way that we can properly start delivering these services in a cloud-based environment. In the second section on architectural principles, we examine issues such as scalability, manageability, and service delivery in terms of design and implementation issues. High-level discussions about planning and architecture in the first two sections of this chapter are complemented by a set of use case scenarios in the third section of this chapter. The goal of the use cases is to provide concrete examples of applying the planning and architectural principles to typical scenarios facing cloud computing adopters.

## Planning Principles for Moving to Cloud Computing

Planning a move to cloud computing starts pretty much the same as any other planning process: understanding where you are and where you are trying to go. In the realm of IT, this generally means understanding the business drivers that dictate the services to be delivered, the expectations for those services, and the constraints on actually delivering them. From there, we can move to a detailed definition of requirements. With a clear and well-defined set of requirements, we can document workloads that we expect to utilize the cloud. Each of these steps will be considered in turn.

## Prioritizing According to Business Drivers

Business drivers are the strategic objectives of an organization that frame the need for IT services. These can include:

- Increasing productivity
- Reducing time to market in new product development
- Reducing production costs
- Optimizing product distribution and delivery
- Increasing market share
- Increasing customer retention

Business drivers are so high level that they can apply to many different businesses. This is expected because businesses all have the same high-level goals of maximizing returns for owners.

What distinguishes businesses in terms of strategies is how they prioritize these objectives and how they define and implement strategies to realize their goals. For example, one company may decide to focus on increasing productivity in order to remain competitive in an increasingly global market. Another company may realize that it is their intellectual property (IP) that drives their growth, and they need to invest more in computational resources to develop new IP. Still another company operating in a mature market may decide to grow by acquiring new customers by targeting perceived weaknesses in their competitor's product line.

The first step in planning a move to the cloud, then, is understanding what business objective is served by that move. Certainly, moving to cloud computing because it is a more efficient vehicle for delivering computing services is a sound reason. We do not need to settle for just that, though. If we press for an even more detailed set of drivers, we can more precisely plan our cloud services. This will help us to plan for short-term capacity demands, plan for long-term needs, as well as deploy needed applications and other software to support those objectives.

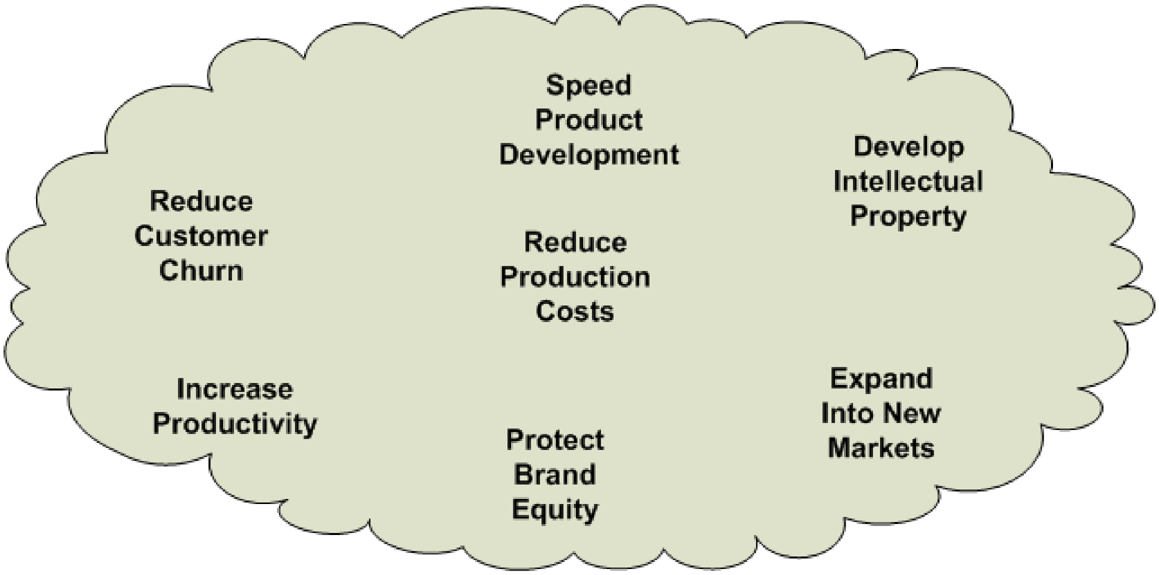


**Figure 5.1: High-level, non-prioritized business objectives are less helpful in shaping cloud computing planning than more precise, prioritized objectives.**

## Defining Requirements

Defining requirements that will drive a cloud computing adoption can be a daunting task. It is difficult enough to elucidate and define requirements for one application let alone gathering requirements for multiple applications serving different business needs and managed by a range of departments. Fortunately, we are not starting from scratch. Applications, documentation, policies, and operational procedures are probably already in place. Our job is then one of understanding the details of existing systems because these reflect, at least to some degree, the current application requirements. We can then build on this by assessing additional requirements going forward.

### Existing Applications Infrastructure: The Current State of Affairs

An inventory of existing applications and workloads is a valuable asset for planning a move to the cloud. An inventory should include all applications that might migrate to the cloud. Implementation details will vary from one application to another (even among the same software used by different departments or for different business purposes), so it is important to include in the inventory key information in three areas:

- Business requirements and related details
- Technical and implementation requirements
- Operational details and requirements

Business requirements specify who within the organization is responsible for a service, how critical that service is to the business, and what strategic objective is served by the application. These requirements are not necessarily long, detailed documents; a simple half-page summary is probably enough. Our goal is not to create an encyclopedic resource on every application in the organization but to create a planning tool that highlights the services that will run in the cloud and identify the core requirements for those services.

The technical details catalog some of the implementation details about existing services. This includes details such as:

- Server configuration
- Workloads on servers
- Dependencies and interoperability considerations
- Use of shared resources, such as disk arrays

All the necessary details about existing services may be documented in a form like that shown in Table 5.1.

| Type | Requirement Area | Description |
|---|---|---|
| **Business** | Service Description | A high-level description of the service |
| | Business Owner | Person or department that funds and governs the IT service |
| | Service Level Agreements | Key requirements on service delivery |
| | Business Objective | Describes the strategic business objective that is served by this IT service |
| | Criticality | Ranking of relative importance of this service. |
| **Technical** | Servers | List of servers and description of configuration; role of each server |
| | Shared resources used | Shared IT resources, such as disk arrays, network, backup services |
| | Platform Services | Operating system required, libraries, utilities and other packages required to run the applications |
| | Applications | Commercial, open source and custom applications |
| | Physical distribution of servers | Location of primary servers, backup servers and disaster recovery sites |
| | Utilization | Description of server, disk array, network utilization. |
| | Peak Periods | Times and duration of peak loads, frequency of peak periods, periodicity of peak demands |
| | Dependency on other Services | Other IT services that are required to deliver this service |
| **Operations** | Backup requirements | Recovery point objectives, recovery time objectives, etc. |
| | Disaster recovery | Time to recover services, level of services to be restored, critical dependencies |
| | Compliance issues | Summary of key compliance and governance issues with this service |

**Table 5.1: Requirement categories for summarizing existing applications, software stacks, servers, and related hardware**

### Additional Requirements for New Applications

If there is one thing we can count on with IT services, it is that requirements will change. A move to the cloud will open new opportunities to deploy additional services, change the way services are consumed, and consolidate resources. These should also be captured during the requirements-gathering stage. We certainly want to capture applications and workloads that fall into the "more of the same category" (for example, more departments will stand up small databases because the overhead with managing them is reduced) but the most interesting, and perhaps the most influential in the long term, are those that change the way we do business. Consider examples such as:

- Using cloud storage to store single copies of data that are accessed by multiple applications rather than duplicating data sets
- Reducing the number of ad hoc reporting tools as users standardized on the "best of the breed" tools offered in the cloud's service catalog
- New applications, such as statistical analysis and data mining of large customer transaction data sets enabled by on-demand access to compute and storage resources

In the best cases, we will be able to devise reasonable estimates on compute and storage impact of some of these new requirements. For example, in the case of reducing duplicate data for business intelligence applications, we can develop fairly accurate estimates. The more innovative applications, such as advanced analytics, are more difficult to pin down. The CPU demands of such applications are highly dependent on the type of analysis, the algorithms used, the implementation of the algorithms, and the amount of data we are analyzing. Even with these limitations, we can at least provide best estimates (sometimes guesses) for these new types of applications. The next step in the planning process after prioritizing business drivers and defining known and estimated requirements is to analyze the potential workload for the cloud.

## Assessing Workloads

Workloads are as varied as business requirements. Some workloads place a heavy load on CPUs while others are more I/O intensive. Sometimes workloads are fairly consistent over time and others have well-defined peak demand periods. It is important to understand workload profiles for a few reasons.

### Capacity Planning

First, it helps to estimate the overall capacity of cloud services the business will consume. This is especially important if you are implementing a private cloud and want to ensure adequate capacity for peak demand periods. Public cloud customers will also find this data useful for budgeting and long-term planning although there is no need to be concerned about the hardware capacity of your provider (at least in theory). For hybrid cloud configurations, this type of detail can help you understand when internal capacity will be exceeded and public cloud resources will be required.

### Scheduling

Another reason to assess workloads is for scheduling purposes. Some jobs have fairly predictable workloads. For example, services provided to the customers through Web applications will have generated historical data that can be used to determine demand patterns. These applications may have minor periodic variations, for example, Mondays have heavier workloads than Fridays, or longer, seasonal variations such as those retailers experience just before the Christmas holiday.

Cloud providers can use knowledge of workloads to optimize scheduling. Ideally, at any time, we would have a mix of jobs that have different levels of demand on CPU, I/O, and networking. We would not want, for example, to have all the I/O and CPU intensive extraction, transformation, and load (ETL) processes running at one time. Depending on the level of control one has over the workload scheduling, a cloud provider can schedule jobs in an optimal manner or use variations in pricing schedules to provide incentives for users to schedule their jobs in ways that coincide with the scheduling goals of the provider.

One way to globally optimize scheduling is with a bid/accept model for pricing. Cloud consumers can bid a price for a server or CPU time based on the value of having a particular job run. If it is a high-priority job, the customer will bid a higher price; if the job can wait, the customer will bid less. This approach will optimize the allocation of resources in the way a free market optimally allocates resources. This model, however, is subject to the same limitations as free markets; the model breaks down when there is, for example, insufficient information or time to fully evaluate options.

### Cost Recovery

Public cloud providers set their rates to cover costs and earn a profit. The IT department, or other organization structure charged with providing private cloud services, will likely charge for services provided as well. Internal service providers generally are more concerned with recovery costs than making a profit, and a shared cost model is a common means for charging for these services. Charges are based on a simple formula:

(Total Cost of Providing Service / Number of Units Consumed) = Cost Per Unit

Units of service can be CPU hours, server hours, or gigabytes of storage per month. Basically the idea is that the service providers recover whatever the cost of providing a service.

> **Note**
>
> This is different from a simple market model in which price is determined by supply and demand. In the case of a cost recovery model, when demand goes down, price per unit could actually go up because the number of units consumed goes down. Conventional free market economics predicts the price will drop in such situations.

The mix of workloads and their distribution over time are important factors when aligning requirements to the cloud model.

### Aligning Requirements to Cloud Services

At the end of the planning phase, we should have:

- A set of high-level requirements for existing applications that will move to the cloud described in terms of business, technical, and operational requirements
- Rough estimates for new applications enabled by the cloud
- Workload information that can provide the basis for capacity planning, scheduling, and cost recovery

To ensure a cloud service meets the expected needs, we want to have sufficient capacity. How we do so will depend on whether we are using public, private, or hybrid cloud services. When a private or hybrid cloud model is used, we are both the provider (for some of the services in the hybrid case) and the consumer. As the provider of cloud services, we have to redeploy existing hardware and/or procure additional hardware and deploy it in a cloud infrastructure along with management applications and a service catalog of machine images and related software. When a public cloud provider is used, we have to demonstrate the provider can offer the levels of service needed at the times they are required. As we get into these issues, we move away from the planning aspects and start to focus on more architecture-oriented issues related to moving to the cloud.

## Architectural Principles for Cloud Services

The architectural principles underlying the cloud model are designed to maximize the utility of computing infrastructure by making it available to a broad range of users for a variety of applications without unnecessarily coupling hardware and software to single uses. To do so, we design around a number or architectural principles focused on:

- Designing for scalability
- Designing for manageability
- Deploying layered technical services
- Delivering business services

Before discussing each of these in detail, it is worth noting the importance of virtualization to cloud architectures. Virtualization is a fundamental aspect of cloud computing and is used at numerous levels of service delivery. We virtualize computing and storage, which hides the implementation details of these low-level services. Higher-level services, such as database management, content management, and identity management, are provided as services abstracted away from implementation details.

An immediate benefit of virtualization is flexibility. Hardware can run different operating systems (OSs) at different times. Different software stacks can be deployed to run for some period of time and then shut down. Legions of IT professionals are not needed to do this; virtualization enables greater levels of self service than have been possible in the past.

The degree of flexibility, and the benefits derived from it, varies with the amount and method of virtualization. For example, at one end of the spectrum, we can deploy single servers dedicated to single tasks. If additional resources are needed to accommodate growing workloads, either the server needs to be upgraded or additional servers need to be dedicated to that purpose. This is especially costly if the additional resource requirements are only for short peak demand periods.

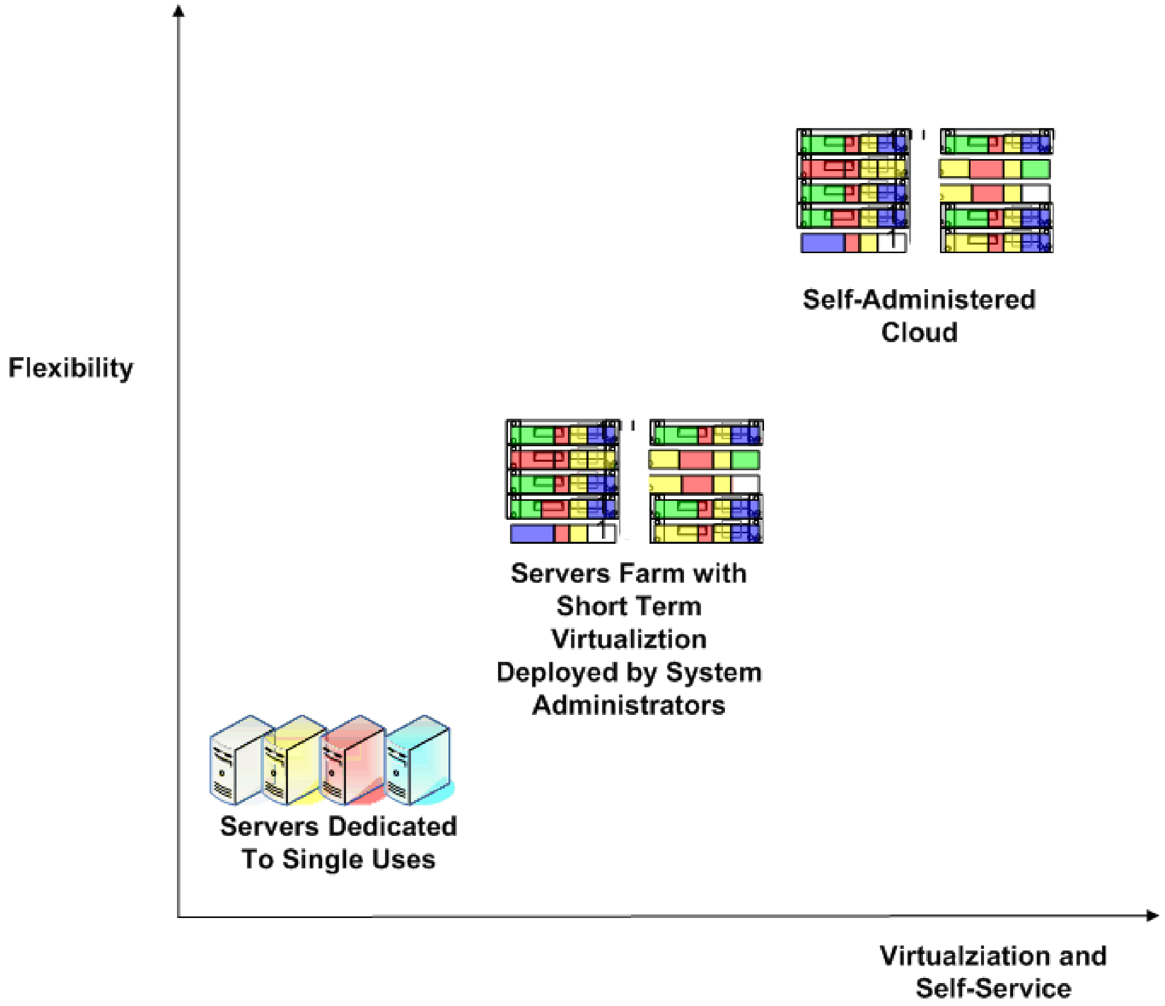
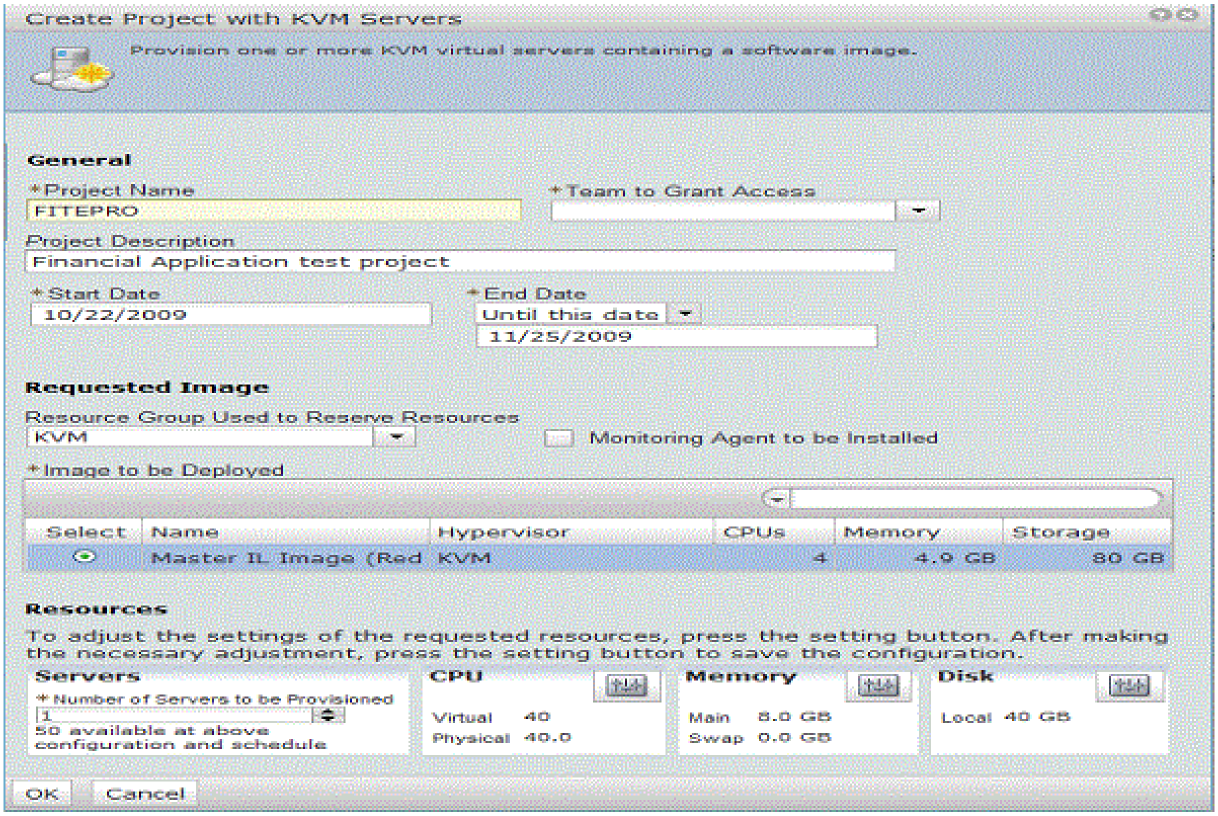

Flexibility
Self-Administered Cloud
Servers Farm with Short Term Virtualiztion Deployed by System Administrators
Servers Dedicated To Single Uses
Virtualziation and Self-Service

**Figure 5.2: The greater the virtualization and support for self-administration, the greater the flexibility in adapting computing resources to changing service needs.**

A step away from the dedicated server model toward a highly virtualized environment like the cloud is a server farm in which servers are reallocated according to changing needs. There are a number of advantages of this approach over the dedicated server model. First, policies and procedures are in place to change the roles of servers fairly rapidly. Systems administrators shut down applications and supporting software, install machine images with other applications needed at the time, and redeploy the servers in their new roles. A second advantage is that hardware is fairly easily reallocated; there is no need to procure new hardware for small, incremental increases in workloads.

Although the virtualized server farm is a step in the right direction, it is still hampered by the need for IT support to reallocate resources. This creates a certain amount of overhead cost associated with the switch. Granted, it is smaller than the cost associated with switching dedicated servers, but it is still greater than the cost associated with the self-service switching costs found in cloud environments.



Create Project with KVM Servers

Provision one or more KVM virtual servers containing a software image.

**General**

*Project Name: FITEPRO

*Team to Grant Access:

Project Description: Financial Application test project

*Start Date: 10/22/2009

*End Date: Until this date 11/25/2009

**Requested Image**

Resource Group Used to Reserve Resources: KVM

□ Monitoring Agent to be Installed

*Image to be Deployed

| Select | Name | Hypervisor | CPUs | Memory | Storage |
|---|---|---|---|---|---|
| ● | Master IL Image (Red | KVM | 4 | 4.9 GB | 80 GB |

**Resources**

To adjust the settings of the requested resources, press the setting button. After making the necessary adjustment, press the setting button to save the configuration.

**Servers**

*Number of Servers to be Provisioned: 1

50 available at above configuration and schedule

**CPU**

Virtual 40
Physical 40.0

**Memory**

Main 8.0 GB
Swap 0.0 GB

**Disk**

Local 40 GB

**Figure 5.3: Virtualization combined with self-service administration lowers virtual machine deployment costs. Non-technical cloud consumers can manage their own workloads.**

In a cloud environment, the process of deploying virtual machines is highly automated with the use of self-service software. In addition, resource tracking modules in the cloud administration software can track the images used, the time servers are up and running, and the amount of storage used by job. This can further reduce administration costs and facilitate charge backs and cost recovery. In addition to flexibility, virtualization enables critical qualities such as scalability and manageability.

## Designing for Scalability

Concerns about scalability affect both cloud providers and cloud service consumers. In the case of cloud providers, the designing for scalability entails addressing several requirements for meeting varying workload demands. For cloud consumers, the issues tend to be around the question of how to most effectively utilize the computational resources available in the cloud.

### Providing Scalable Computing Resources

At first glance, cloud scalability may look like just a matter of hardware. With enough physical servers, disks in storage arrays, and network bandwidth, we can meet scalability demands, right? Not exactly, or at least that is not the entire story. Cloud service providers also have to provide services and features in addition to raw hardware to enable a functional, scalable cloud. Some of these services and features include:

- Security services
- Standardized catalog of applications
- A service oriented architecture (SOA)

These requirements are comparable to those we find outside the cloud.

## Security Services in the Cloud

Security in the cloud looks much like security outside a cloud environment. When we deploy applications to the cloud, we have to concern ourselves with several security requirements:

- Identity management
- Access controls
- Auditing and logging
- Vulnerability management and threat assessment

Identities are independent of workloads running in the cloud. Identities persist over time and should be maintained with authentication and authorization information as well as encryption keys. This type of information is needed, for example, to control limits on resource allocation in the cloud and to store keys used to encrypt data stored in the cloud.

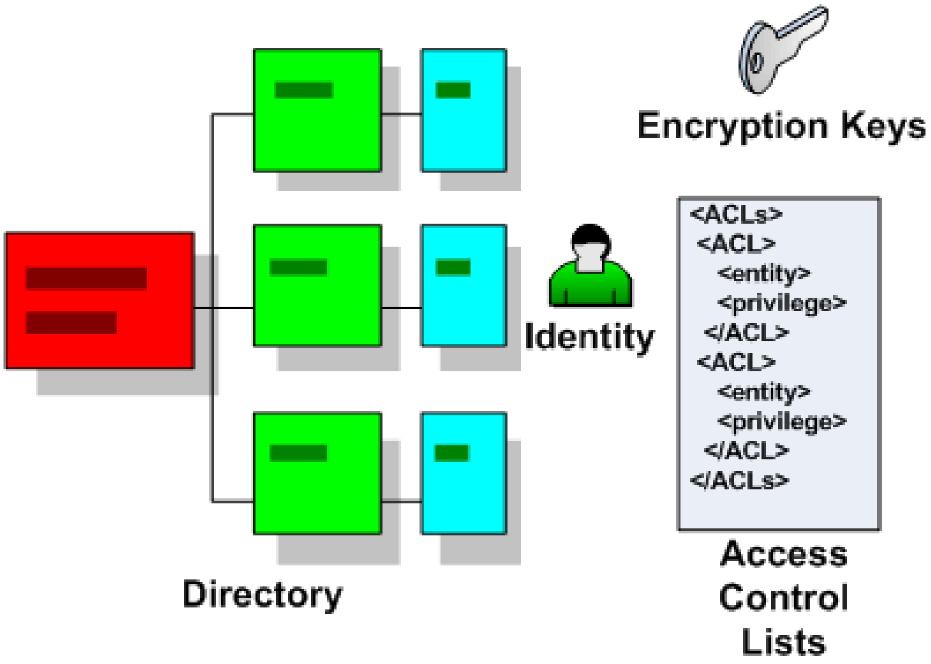


**Figure 5.4: Security controls in the cloud depend on identity, access control lists, and encryption keys.**

In addition to user-centric security information, cloud providers need to support process-oriented auditing and logging. As in any deployment architecture, audit and other logs must be tamper-proof and sufficiently detailed to meet security and compliance requirements.

The images that comprise the service catalog will support a wide range of OSs, utilities, libraries, and applications. These are all sufficiently complex to require regular vulnerability scanning, patching, and upgrading. Cloud providers will also need to have procedures in place to perform vulnerability scans on images, track patch levels, and update images as needed. One of the advantages of cloud architectures is that once an image is scanned or patched, every cloud user that deploys that image will have access to the latest version. There is no need to push patches to servers or desktops, verify installation, and then manually correct failed patches.

### Standardized Catalog of Services

Scalability often implies repeated use of a small set of constructs. Take, for example, a cluster of computers comprising identically configured servers, distributed database running the same database management system in different sites, or even the ubiquitous desktop OS. These examples show that benefits of standardization can often outweigh the disadvantages of not having customized solutions to a particular problem.

In the cloud, standardization at the platform and application level comes with a standardized catalog of services. Cloud users can instantiate virtual machines running images from the catalog. The data we collect in the planning stages about application requirements can form the basis for building the service catalog. Cloud users are still free to bring or develop their own custom applications, but the service catalog provides a supported foundation for all cloud users. Cloud providers have to weigh the benefits of adding specialized images to the catalog against the additional overhead of managing more images.

### SOA

Services in any architecture have to be sufficiently accessible to be of use; when we are working with highly-scalable architectures such as the cloud, it is even more important. In the cloud, we have the possibility of running a large number of services under varying workload conditions which are subject to different constraints. In environments such as this, there should be as few dependencies as possible between applications.

SOAs decouple services through agreed upon interfaces and message passing. This model scales to different types of services, a wide range of inputs and outputs, and can scale to a large number of services.

Scalability requires design and implementation considerations beyond those of just hardware and infrastructure. Scalability in the cloud requires providers to plan for and support security services, a standardized catalog of services, and an SOA.

## Using Cloud Services in Scalable Ways

A cloud architecture is, by definition, scalable; however, to realize the full benefit of the cloud, we as cloud consumers need to use application architectures that take advantage of the cloud's underlying scalability. This requires that our applications avoid processing bottlenecks, such as a service that is provided only on a single server. As other parts of the application scale up to meet demands, that service would be bound by the constraints of the single server. Two common ways of avoiding this type of bottleneck are to distribute workloads in either a round robin manner or by partitioning workloads.

### Scaling with Round Robin Load Balancing

Consider an online retailer that experiences peak demands during the holiday shopping period. The holiday season lasts several weeks, so scaling their Web site with cloud-based applications makes sense. There will be many users all accessing the Web site and most of the demands on the server will be to deliver Web pages, so the retailer will deploy multiple Web servers each hosting the same content. A load balancer receives all HTTP requests from shoppers and distributes them evenly across all the Web servers. In this way, no single server becomes a bottleneck and additional Web servers can be deployed from the cloud if needed. Furthermore, this approach provides high availability as well because the failure of any one Web server will be compensated for immediately by other servers in the cluster.

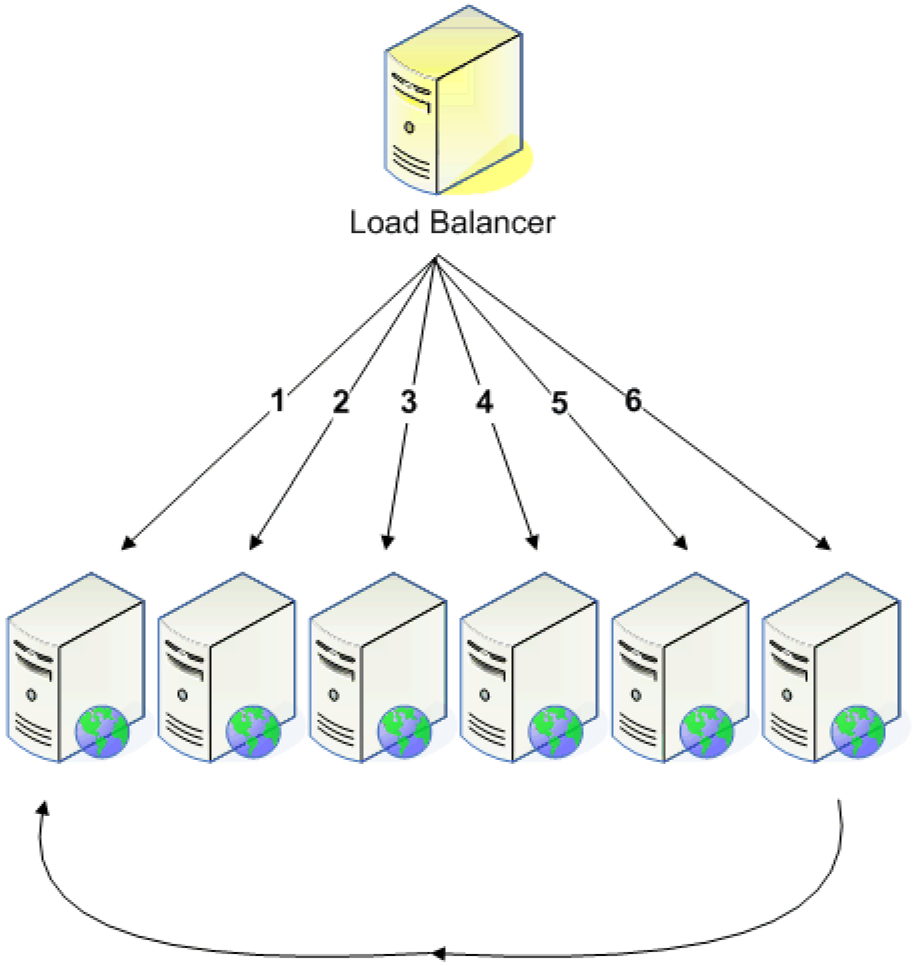


**Figure 5.5: Round robin load balancing assigns each new connection or transaction to the next server in an ordered list of servers; when the last server is reached, the next connection or transaction is assigned to the first server in the list.**

## Partitioning by Data Characteristics

Another way to scale applications is to divide workloads by some characteristic of the data:

- Geographic location of customer
- Distribution center fulfilling an order
- Product category
- Customer name

Ideally, the criteria for dividing a workload will lead to roughly equal size partitions of the data. This helps to ensure scalability because no one server supporting a partition would become overloaded faster than the others. Also, it can help long-term maintenance if the partitioning scheme allows for changes to the partition criteria without significant overhead. For example, if a geographic partitioning scheme is used and one area grows faster than the others, one could subdivide the fast-growing geographic area into two subdivisions.

Partitioning data and storing it in different databases is sometimes used when a single database server cannot keep pace with workloads. Geographic distribution is especially helpful in localizing network traffic and improving the responsiveness of applications that run on the same local network as the database server. In the cloud, this is less of a concern at least for the cloud service consumer. Nonetheless, this type of partitioning is still useful for performance.

Databases use a combination of in-memory caches and persistent disk storage. Queries that can be answered using cached data are significantly faster than those that require disk operations. In the cloud, multiple instances of a database can run on multiple servers. Each server will maintain a cache of partitioned data and, presumably, use cloud storage for persistence. The total amount of memory available for caching is the sum of cache memory across all database servers. This can result in a much higher ratio of queries being answered from the cache rather than from disk.

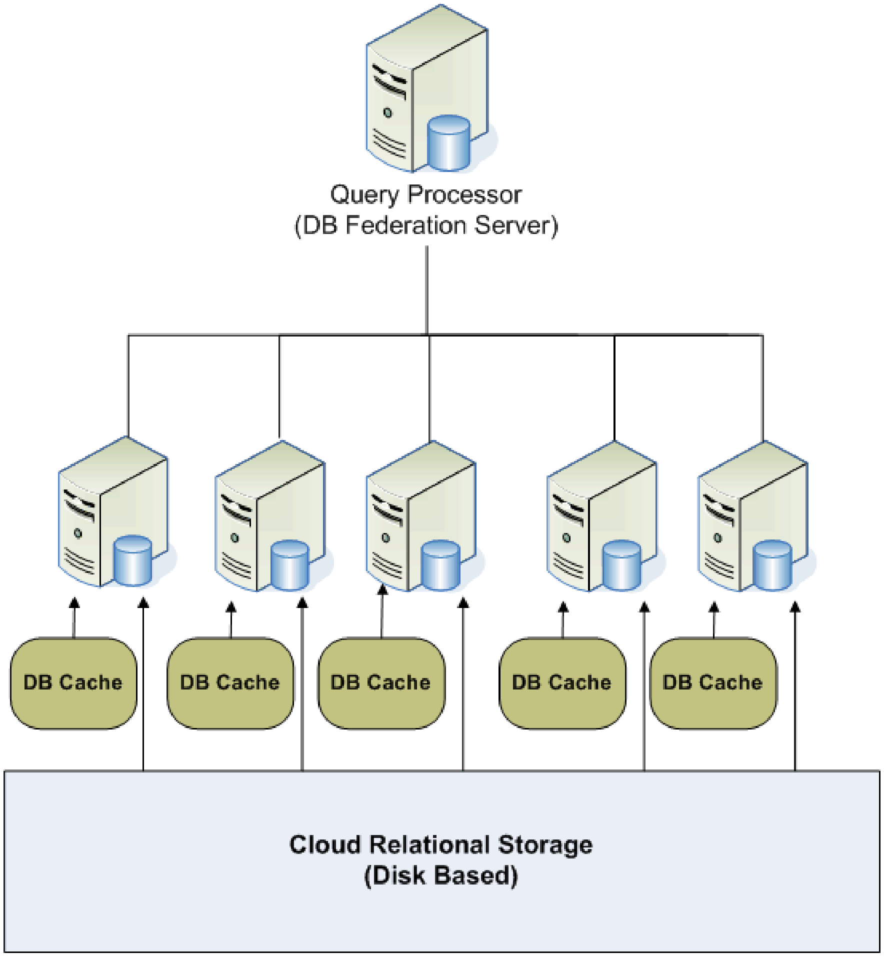


**Figure 5.6: Partitioning data across multiple database servers can improve the scalability of data-intensive applications.**

Designing for scalability is concern for both cloud providers and cloud consumers. Providers need to address obvious hardware and networking infrastructure issues with scalability, but those are not the only scalability issues they face. Security, a standardized catalog of applications, and an SOA are also essential for ensuring scalability. Cloud consumers also have a role in ensuring scalability by designing their applications appropriately using techniques such as round robin load balancing and data partitioning.

## Designing for Manageability

Manageability is another architectural principle that strongly influences how we implement and consume cloud services. This is an important principle for both cloud providers and consumers. Three key points in this area are:

- Provisioning
- Monitoring
- Usage and accounting

The more these services can be automated, the more efficiently a cloud can deliver services to its users.

### Managing Cloud Provisioning

Provisioning in the cloud is the process of instantiating one or more virtual servers running a particular machine image. In the simplest case, a user needs to start a single server, and after running a process, the user shuts down the server. This is a fairly straightforward task but still requires management software to allow non-IT personnel to manage the process. Even in a simple case, there are issues:

- Selecting a machine image to run on the virtual machine
- Determining the time to start the virtual instance
- Deploying additional applications needed to process the particular workload
- Starting services on the virtual machine
- Executing a workflow
- Shutting down the virtual server

Provisioning operations can be more complex if they involve multiple instances running different applications. For example, a workflow may require six virtual servers running a Java application server and a load balancer for distributing transactions across the six other servers. The servers may be shut down at different times as the workload varies or other application servers may be added to the set of servers to meet peak demand. Easy-to-use software is essential to low-cost provisioning.

### Monitoring Jobs in the Cloud

Once servers are provisioned and jobs are running, we will need to monitor them. This includes tracking:

- CPU and memory utilization to determine whether additional resources are required or some should be shut down
- Disk I/O to ensure sufficient throughput on I/O operations to meet requirements and service level agreements (SLAs)
- Application logs to look for adverse events or warnings of potential problems
- Jobs and workflows running in the cloud, including running time, resources allocated, and costs for those resources

This type of monitoring is primarily for managing running jobs. It is also important to have management reports that summarize jobs, resources used, and costs over longer periods of time.

## Usage and Accounting Reports

Usage and accounting reports are especially important for verify billing and analyzing trends in cloud usage. For providers, these reports show aggregate information about:

- Who is using cloud services
- Number of virtual servers run per job and the duration of jobs
- Machine images instantiated in the cloud
- The amount of storage in use
- The amount and type of I/O operations

Cloud users may find these reports especially useful for optimizing how they schedule jobs. Unlike running a dedicated server, there are easily controlled marginal costs associated with running jobs in the cloud. There may be cost advantages to running jobs on larger servers but running fewer instances when the pricing scheme provides such an advantage. There may be advantages to aggregating jobs and running them less frequently. This can be the case when cloud providers charge in minimum units of one hour and jobs are consistently finishing in well under one hour.

Designing for manageability means planning for end user provisioning, process monitoring, and usage and accounting reports from the start. Cloud service consumers should make use of these reports to run their jobs in the most efficient manner possible.

## Deploying Layered Technical Services

Layering services is a long-standing approach to dealing with software complexity. OSs have long used layering to isolate the need to deal with hardware-specific issues or manage low-level operations, like virtual memory. Layering services is a sound approach in cloud environments as well. At the most course description, cloud services are layered as:

- Infrastructure services
- Software platforms
- Applications and information services

Infrastructure services are the lowest-level service and include virtual machines, virtualized storage, and network services. On top of this layer, we run middleware software such as relational database management systems, Java application servers, content management systems, portals, and so on. This middle tier provides the building blocks for business-specific applications such as customer relationship management (CRM) systems, business intelligence reporting systems, and customer-facing Web applications.

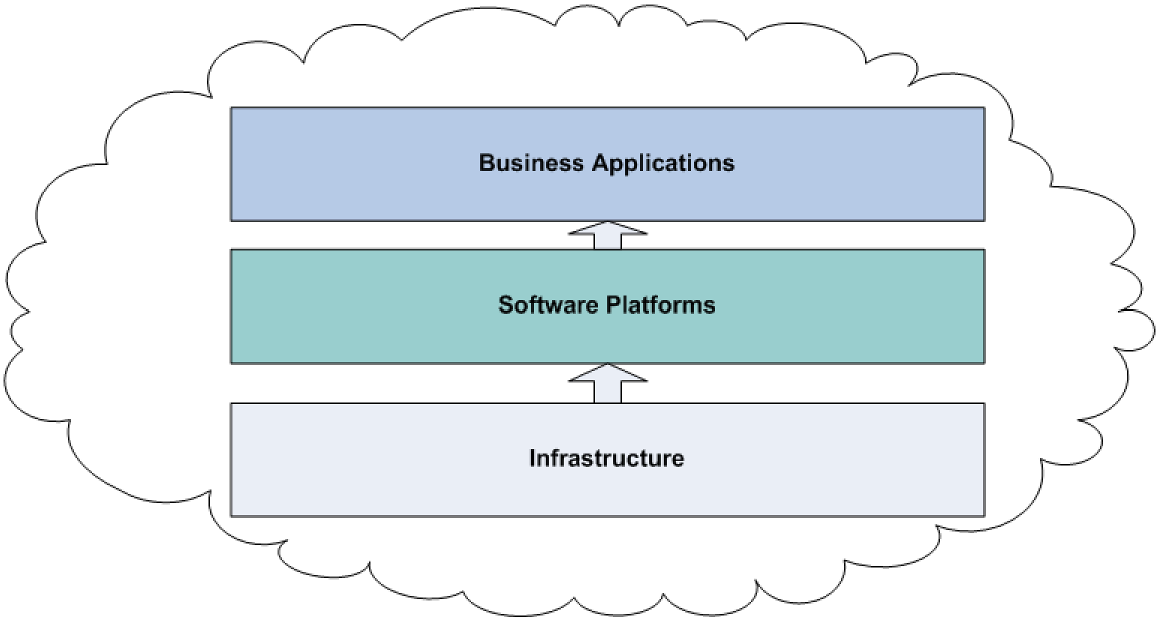


**Figure 5.7: Cloud services are delivered in layers, each providing service to the layer above with the top-most layer providing end use business applications.**

## Delivering Business Services

Usually we would stop discussing architectural principles once we reach the top of the application stack where business services are delivered. We'll veer from the normal case here to address one other essential part of delivering and consuming cloud services: the need for managing service delivery.

The service catalog discussed earlier is part of this process. As noted, the contents of the service catalog are driven by existing and anticipated business requirements. The service catalog has its own long-term maintenance issues, just as software distributed throughout the organization. One of the advantages of the cloud is that service management is less complex. Servers are generally concentrated in the data center and there is less need for maintaining desktop clients.

Policies are needed to govern cloud operations and services to ensure their long-term stability. Basic policies, such as the following, should be in place:

- Pricing and cost recovery
- Patch management
- Security policies
- Acceptable usage
- Auditing
- Data retention

Policies define how cloud services will be governed and managed and provide the final piece of the planning processes for deploying business services in the cloud. In the next section, we will turn our attention to two use cases to provide examples of applying the planning process and architectural principles to typical business requirements.

## Business Services in the Cloud: Use Case Scenarios

We will consider two use cases: a new customer service initiative and a business intelligence application. We will also examine some of the workload considerations that factor into managing cloud-based services.

### New Customer Initiative Use Case

The first use case scenario is motivated by the business driver to improve customer retention. A company has been experiencing moderate but increasing turnover in the customer base; this is commonly known as *churn*. In an effort to reduce churn, the company has determined that it can gain a competitive advantage over others in the market by improving customer experience. In particular, the company has decided on a two-pronged approach. First, it will allow customers to access their entire account history rather than just the past 4 months, as currently implemented. Second, it will provide more targeted offers based on a customer's account history.

As part of the planning process, the company reviews the business, technical, and operational requirements for these services (see Table 5.1 for a list of requirement categories). The business area requirements focus on this imitative as mid-level criticality (that is, not essential for core day-to-day operations but a long-term priority).

The technical requirements include platform services such as relational database management services, customer identity management services, and access to a portal to provide presentation-level services. Estimates are compiled on the amount of data that will be stored, the number of customers querying their account histories each day, and the processing load required to update account histories on a daily basis.

Operational requirements include backup recovery and, because this is a customer-facing application, disaster recovery. Compliance requirements are minimal, but company policies protecting private customer information must be followed.

The requirements are well met by cloud architecture. Accessing entire account histories for all active customers requires the ability to rapidly scale both computing and storage resources. The incremental growth in storage required to accommodate new customer activity is also readily met by the cloud. Analyzing customer account history to generate custom offers is a compute-intensive process but will not require significant additional storage. This type of analysis will be done periodically but not more frequently than once a month. The peak CPU demands generated by this process will last for 1 to 2 days. The need for additional compute resources can be met by the cloud as well.

The service catalog already supports the middleware required, including the database, portal, and statistical analysis software. Each of these platform services is available in different images, so each will be running on one or more virtual machines. This is a customer-facing Web application, so the portal servers will be configured in a load-balanced cluster and the data will be partitioned to evenly distribute the customer database over multiple database servers.

### Business Intelligence Use Case

A company has decided to consolidate its business intelligence reporting services to improve the efficiency of business intelligence operations and lower overall costs. One of the defining characteristics of business intelligence and advanced analytic operations is that they entail large amounts of data and they are computing intensive.

Traditional data warehouses and similar business intelligence architectures inefficiently allocate resources. They can be deployed around dedicated department-level servers and storage. This tends to lead to low CPU utilization between data loads and report generation. Unless there is high demand for ad hoc queries outside of data loads and report generation operations, the server runs well below capacity.

Another potential area of significant inefficiency is in storage. It can be difficult to estimate storage requirements, especially when various performance techniques, such as excessive indexing, denormalization, and materialized views, may be used to improve performance. The best combination of optimization techniques may not be discovered until the business intelligence system has been in use for some time. In a traditional deployment, that storage hardware would have been purchased already. That inconvenient fact often leads to purchasing more storage than is needed for fear of not having adequate storage.

Business intelligence as a cloud service can be implemented more efficiently. Let's assume the business drivers behind this project include improving sales by providing detailed and timely reports to sales managers while reducing the total cost of business intelligence services in the company. Technical requirements include large volumes of storage and a large number of servers to perform ETL operations to populate and update the data warehouse on a daily basis. Once the ETL process is complete, reports will be generated. Once the reports are complete, the peak demand period is over but an estimated 25% of peak computing resources will be needed during the rest of the data for ad hoc reporting.

The cloud allows this initiative to start servers as needed for the ETL and reporting operations, then scale back to a smaller number of servers. An additional benefit is that a single copy of data can be shared among different departments. For example, the marketing department and the quality control group may both want to use sales data but in different ways. In cases where each department maintains its own data mart, the sales data would be duplicated. The same data marts can run in the cloud but share a single copy of the source data.

### Mixing Workloads

Jobs that do not need to run on strict time schedules can be arranged to optimize utilization. For example, loading schedules can be optimize to increase utilization by performing extraction and copy operations during times when there is a low demand on cloud resources. Similarly, workloads can be mixed so that some I/O intensive jobs are run at the same time as other CPU intensive jobs that can run at the same time as jobs with more constant and predictable workloads, such as development and test environments or collaboration services.

Both of these use cases demonstrate common characteristics of business services that fit well with the cloud model:

- Minimal or moderate security requirements
- Minimal dependencies between services
- Moderate audit requirements
- Minimal customization

As a result, these applications can meet the requirements of the business drivers that motivate their development; they can be deployed using the infrastructure, platform, and application services provided by the cloud; and they can be managed using the self-service provisioning, monitoring, and usage accounting services provided by the cloud management software.

## Summary

When formulating a strategy for moving a business to adopt cloud services, we should bear in mind both business planning and architectural considerations. On the planning front, start with the business drivers and ensure that services deployed in a cloud support those drivers. To do so, be sure to analyze requirements in terms of business, technical, and operational needs. Also understand workloads and related issues, such as capacity planning, scheduling, and cost recovery.

Key architecture and design consideration also have to be taken into account by cloud service providers and cloud service consumers. Scalability is essential. Cloud service providers ensure scalability by providing sufficient hardware, software, and networking services but also by supporting security services and a standardized catalog of applications in an SOA. Manageability is also a factor in realizing scalable services, especially related to provisioning, monitoring, and usage reporting.

In the next chapter, we will delve deeper into technical and architectural issues with a look at identifying further details of cloud architectures and their impact on your business.

# Chapter 6: Identifying the Right Cloud Architecture for Your Business

Cloud computing is a general model for delivering computing and storage services. The model lends itself to a range of implementations with no single architecture constituting a "true cloud" architecture. This is hardly surprising. The defining characteristics of cloud computing (the ability to allocate and release compute and storage resources on demand, a pay-as-you-go funding mechanism, and high levels of self-service) allow cloud providers to deliver a wide range of services using a number of implementation models.

This range of variability means businesses can choose the right cloud architecture for their environments. In this chapter, we will examine several aspects of selecting a cloud architecture:

- Levels of cloud architecture
- Issues in providing compute services
- Issues in providing storage services
- Considerations for network services
- Cloud operations management
- Service layers and adapting IT operations to infrastructures
- Topics in service management

We will start with a brief review of architectural elements common to all cloud architectures.

## Levels of Cloud Architecture

Cloud architectures can be thought of in terms of layers of services in which each layer depends on services provided by the next lower layer. As with other layered models of abstraction in software engineering, layers in a cloud control the potential complexity of cloud design by following a few basic principles:

- Services are provided as logical abstractions that hide implementation details. When a program needs to allocate additional storage, for example, it makes a call to a storage service requesting a particular amount of space. There is no need to delve into details about directory structures, files systems, or disk configurations.
- Services are isolated to appropriate layers in the architecture. An application programming interface (API) for storage allocation may make calls to additional services that are not available outside of the storage system. For example, when allocating new storage, an API procedure might call an isolated procedure to add the allocated disk blocks to a list of blocks that are replicated to storage devices for backup and performance reasons.
- Services are provided at a functional level appropriate to the users or services that consume the services. The higher up the stack of services we go, the broader and more business oriented the services. Although lower-level services might operate on storage blocks, upper level services might initiate business process workflows.

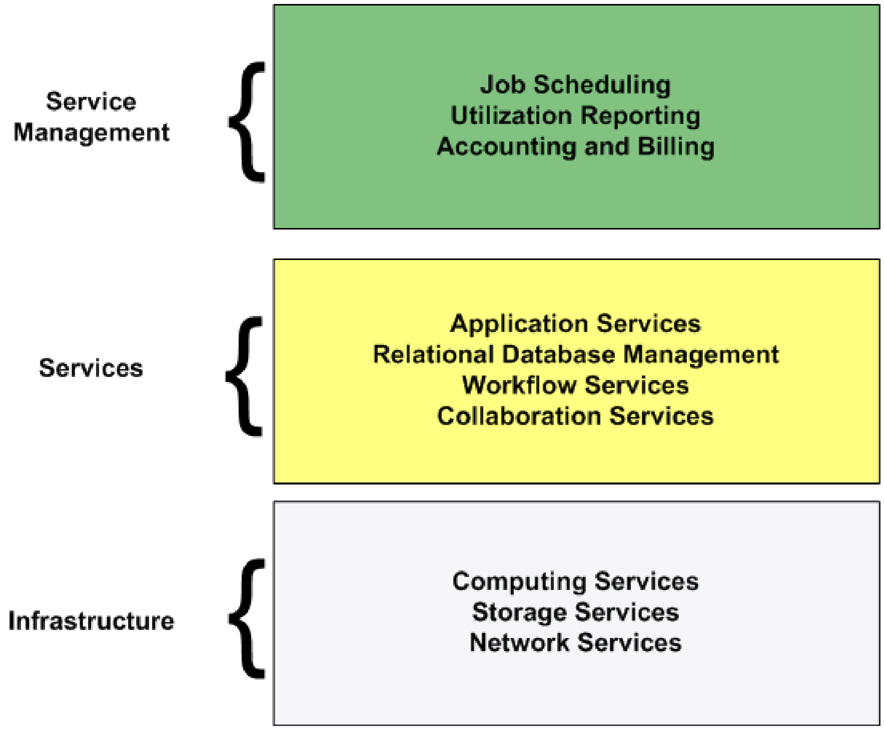


**Figure 6.1: Cloud architectures can vary in detail and levels of services provided but most include some combination of infrastructure, platforms, and services management.**

Broadly speaking, we can think of three coarse-grained levels of services in a cloud architecture:

- Virtualization of resources
- Services layer
- Server management processes

Each of these levels can be further subdivided.

## Virtualization of Resources

One of the hallmark characteristics of a cloud is the virtualization of resources. Virtualization can be thought of as a way of abstracting computing and storage services away from implementation details and toward a more logical and less physical view of resources.

Many of us use virtual servers routinely although we might not know it. We connect to servers across the Internet that run Web sites, email servers, databases, and other business applications. Most of the time, we do not think of the implementation details about these services. Is the email server running on a single physical server? A cluster of load-balanced servers? Or perhaps the application is hosted on a virtual server that shares a physical server with several other virtual machines running an entirely different set of applications. These details are often unimportant, at least from our perspective.

The ability to hide implementation details without adversely affecting services is essential to providing cloud computing. Virtualization is especially important for efficiently using computing and storage infrastructure. (We will focus primarily on server virtualization here and address storage virtualization later in the section entitled "Providing Storage Services.")

### Logical Units of Computing Resources

Server virtualization allows us to manage compute resources in finer-grained units than just a physical server allows. One of the first advantages of this approach is that we can work with a standardized set of features, such as the number of CPU cores and amount of RAM. For example, a standard virtual server might be equivalent to a physical server with one Intel Xeon 5600 series or AMD Opteron 6000 series processor and 8GB. One could also define virtual servers in terms of performance relative to standard benchmarks, such as the Transaction Processing Performance Coucil's (http://www.tpc.org/tpcc/default.asp) online transaction processing (OLTP) benchmarks (TPC-C and TPC-E) and the ad hoc, decision support benchmark (TCP-H). How the logic units are defined is less important than the fact that we have a standard for allocating computing resources that is not tied to a particular physical implementation.

By decoupling how we allocate computing resources from the underlying hardware that provides those resources, we gain flexibility in managing how we consume compute services and manage them.

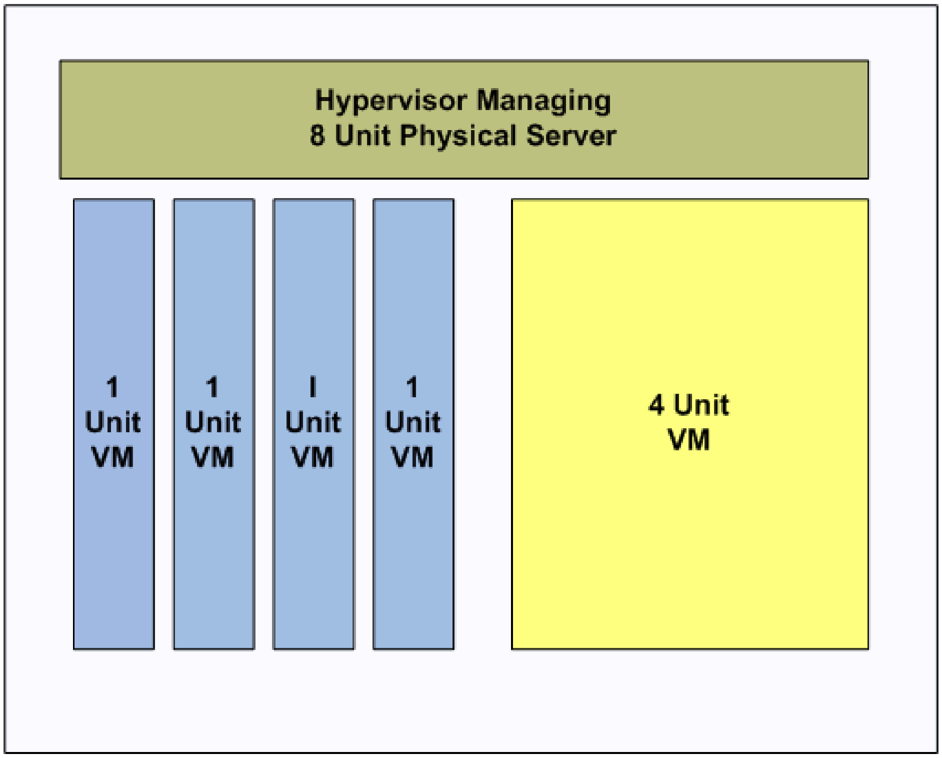


**Figure 6.2: Server virtualization allows cloud service consumers to use standardized units of computing services without concern for the physical implementation details.**

### Hardware Independence

Another advantage of virtualization for cloud service providers is hardware independence. Cloud consumers can allocate the level of computing resources they need without having to worry about whether a particular physical server is a 2, 4, or 8 core server. Cloud providers can deliver those logical units using the most economical way possible. For example, a cloud might have several types of physical servers running in the cloud. The less energy efficient servers are only used when the more efficient servers are running at peak capacity. The first time a cloud consumer runs a job, the job might run on one of the more energy efficient servers; the next time the same job runs on the other type of server.

### Standardized Service Pricing

Along with logical units of computing resources and hardware independence, virtualization allows for standardized service pricing. Although this is not a technical issue, it has a direct impact on how cloud service consumers plan and manage their use of the cloud.

Virtualization of services is an essential element of a cloud architecture. It directly enables the most efficient allocation of resources, reduces the need for cloud service consumers to understand the nuanced differences in physical servers, and provides for a straightforward pricing model that consumers can use for planning and budgeting.

## Services Layer

The services layer is another common characteristic of cloud architectures. At this level, we work with not just virtualized hardware but also operating system (OS) and application services. It is certainly possible to provide a cloud that offers only infrastructure services (that is, the virtualized equivalent of bare metal machines), but for business users of cloud services, the services layer can provide additional benefits.

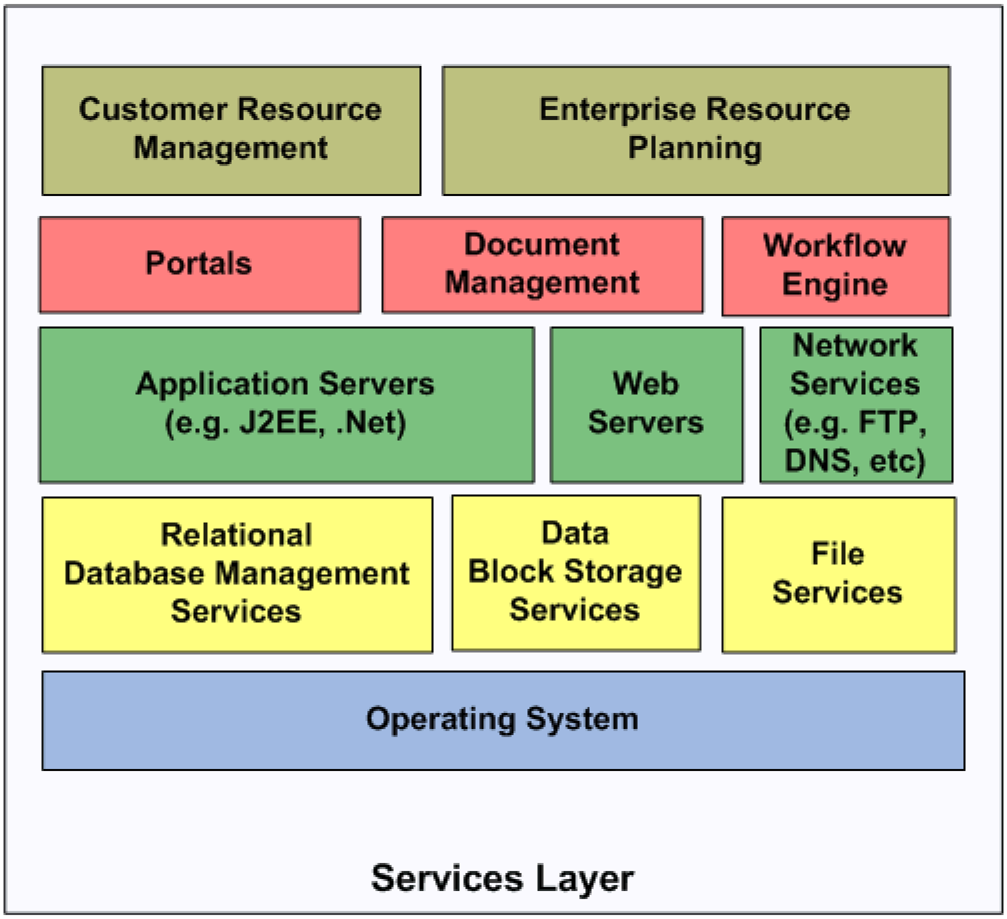

Customer Resource Management
Enterprise Resource Planning
Portals
Document Management
Workflow Engine
Application Servers (e.g. J2EE, .Net)
Web Servers
Network Services (e.g. FTP, DNS, etc)
Relational Database Management Services
Data Block Storage Services
File Services
Operating System
Services Layer

**Figure 6.3: The services layer consists of a wide range of service types, some of which build on others within the same layer.**

Services such as those shown in Figure 6.3 might be delivered in different ways to customers. OSs of course are included in the virtual machine images, but other services might be independent of virtual machine instances. Persistent storage services, such as block storage and relational database services, might be available as services available to all virtual machine instances running in the cloud. Higher-level services, such as application servers, portals, and workflow engines, might be embedded within virtual machine instances along with other software stack components. At the highest levels, business applications such as CRMs and ERPs may be provided as Web applications that run in the cloud. At this level, service consumers are completely divorced from implementation details and are solely concerned with business-related functionality.

## Service Management Processes

A third major aspect of cloud architectures are the service management processes that support the delivery of services. These include:

- Virtual machine image management
- Image deployment
- Job scheduling
- Usage accounting
- Management reporting

The first two of these services supports a catalog of images preconfigured to particular applications, software stacks, or OSs that can be deployed by cloud service consumers.

Job scheduling applications help with routine processes that run repeatedly on a schedule as well as large, one time jobs that can be submitted to run in the cloud as services are available. Job scheduling services are especially useful when services pricing varies by point in time demand or time of day.

Usage accounting and management reporting are necessary for billing or charge-backs on the part of cloud service providers and for cloud service consumers who must plan and manage their budgets for IT services.

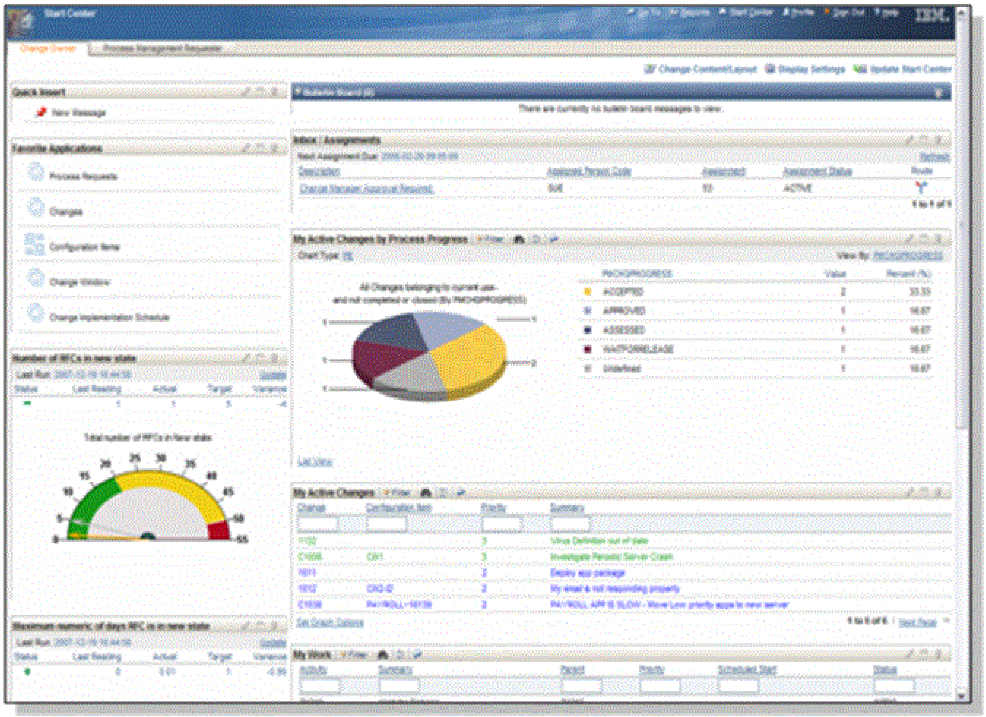


**Figure 6.4: Management reporting serves the needs of both cloud service providers and consumers.**

To summarize, cloud architectures can be described in terms of infrastructure, services, and services management. Variations in these layers allow for different types of cloud architectures. Multiple factors will determine the best choice of architecture for a particular set of business requirements. The remaining sections of this chapter will delve into those factors.

## Providing Compute Services

There are compelling reasons to adopt a cloud architecture that include an internal or private cloud. Businesses maintain total control over computing resources with a private cloud. This can significantly reduce compliance issues with cloud computing. Private and confidential data is not moved outside the company, data destruction policies and procedures are defined by the business, and systems are not shared with outsiders, including potential competitors. With the advantages come additional functional responsibilities.

Businesses that choose to provide private clouds or hybrid private/public clouds must be in a position to provide the physical infrastructure and basic management services needed in a cloud. (Businesses can provide higher-level services, such as enterprise applications, as cloud applications while using a public or other third-party physical infrastructure.) Those that will deliver computing services directly though a private cloud should consider:

- Hardware selection
- Implementing virtualization
- Failover and redundancy
- Management reporting

A business' ability to address each of these issues can strongly influence their success in delivering computing services in a cloud.

### Hardware Selection

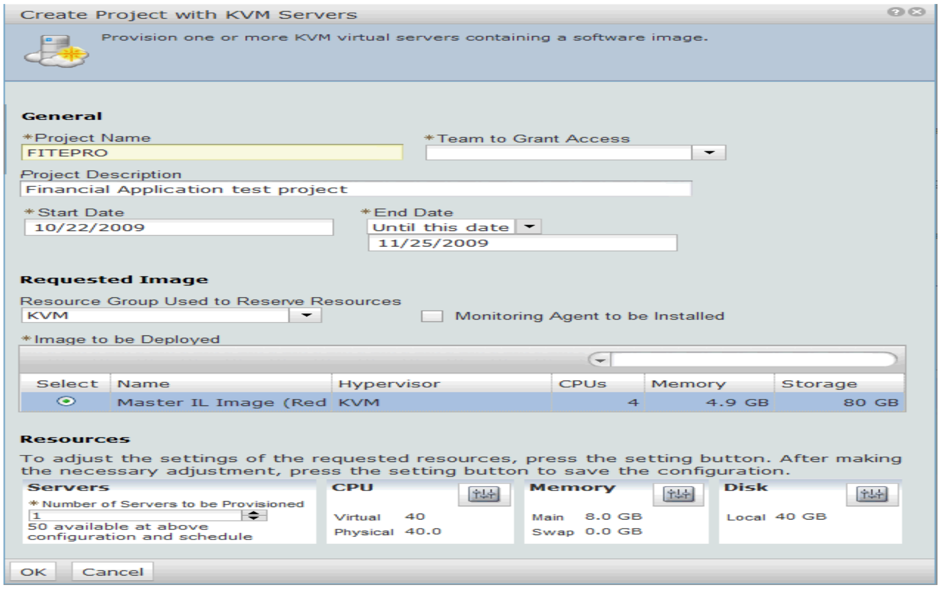
Hardware selection for clouds depends upon two competing interests: controlling costs by redeploying existing hardware versus acquiring a standardized server platform that is configured specifically for cloud computing. Using existing hardware can lower initial capital expenditures but might lead to higher costs over the long term. Older machines that require more maintenance, need parts that are difficult to procure, or consume more electricity can have a larger total cost of ownership than new servers. One option is to use existing hardware initially and replace it over time as the cost of new servers becomes competitive with the cost of continuing to operate the older devices.

An advantage of new hardware is that the cloud can be configured with standard servers optimized for cloud computing: large numbers of CPU cores, significant amounts of memory, high speed I/O for connections to disk arrays, and so on. Standardization also helps reduce maintenance costs.

## Implementing Virtualization

Many organizations use virtualized servers outside of clouds; however, virtualization in the cloud requires more management services than typical in IT environments. Conventionally, managed virtual servers are installed by systems administrators and run for extended periods carrying out a fixed set of functions. Additional controls are available in some environments that support virtual machine migration from one physical server to another. This is especially useful in situations in which a single server is running at or near capacity and one or more of the virtual machines needs to be moved to a less utilized physical server. Even this, though, does not meet the level of virtualization management needed in a cloud.

Clouds should support end user management of computing resources. A knowledgeable user should be able, for example, to select a virtual machine image from the catalog and instantiate a specified number of virtual servers.



Create Project with KVM Servers

Provision one or more KVM virtual servers containing a software image.

**General**

*Project Name: FITEPRO

*Team to Grant Access:

Project Description: Financial Application test project

*Start Date: 10/22/2009

*End Date: Until this date 11/25/2009

**Requested Image**

Resource Group Used to Reserve Resources: KVM

□ Monitoring Agent to be Installed

*Image to be Deployed

| Select | Name | Hypervisor | CPUs | Memory | Storage |
|---|---|---|---|---|---|
| ● | Master IL Image (Red | KVM | 4 | 4.9 GB | 80 GB |

**Resources**

To adjust the settings of the requested resources, press the setting button. After making the necessary adjustment, press the setting button to save the configuration.

Servers: *Number of Servers to be Provisioned: 1. 50 available at above configuration and schedule

CPU: Virtual 40, Physical 40.0

Memory: Main 8.0 GB, Swap 0.0 GB

Disk: Local 40 GB

**Figure 6.5: Providing computing services in a cloud requires significant support software, such as services for selecting and instantiating virtual machine instances.**

## Failover and Redundancy

An advantage of cloud architectures is that we move away from tightly coupling applications and services to particular physical or virtual servers. Applications are run on virtual servers that meet a set of configuration requirements defined by the cloud service user. Applications that are well suited for the cloud do not need specialized hardware or a particular server. This reduces the challenge of providing failover services.

Clouds are inherently redundant. If a physical server fails for any reason, it can be removed from the pool of available resources. Virtual machine images are deployed to other physical servers until the failure is corrected. This type of failover and redundancy is at the server level, not the application level.

If a physical server were to fail while an application were running on it, recovery would depend on the application. For example, if the application provided stateless Web services, it could be restarted on another virtual machine instance on another physical server and start responding to service requests again. In cases where the application writes state information to persistent storage and checks for prior execution information each time the application starts, the application could also recover fairly robustly on another virtual machine.

### Management Reporting

Different types of management reports are required when providing computing services in a cloud. In a traditional "one server-one application" approach, the business owner of a process is responsible for identifying the servers needed to support a business process and covering the cost of the servers, either virtual or physical. In this model, there is fairly little to report outside of utilization rates. The business process owner is paying for sole use of servers, so there is not much incentive to monitor server use as long as it does not adversely affect performance.

Cloud service consumers can use and should expect detailed usage reporting. With a pay-as-you-go pricing model, there is an incentive to allocate the fewest number of virtual servers and run them for the shortest time possible while still meeting business requirements. Cloud service consumers can use reports detailing:

- Number of virtual servers allocated to a job and the time the servers ran
- Peak and average utilization rates of servers
- Amount of data stored persistently
- Amount of data transferred across the network
- Charges for compute, storage, and network services

Detailed utilization information will help business process owners optimize their applications. For example, if analytic servers are running at 40% utilization because they are dependent on data preprocessing operations that are not processing data fast enough, additional servers could be instantiated for preprocessing. Presumably the cost of running the additional preprocessing servers would be offset by reducing the length of time the servers have to run. The analytic servers would run at higher utilization and for shorter periods of time reducing the overall cost of the process.

Providing computing services in a private or hybrid cloud requires a combination of hardware, virtualization management and deployment systems, a server configuration that supports failover and redundancy, as well as robust management reporting.

## Providing Storage Services

If a business moves forward with providing private cloud computing services, it will have to provide storage services as well. This would require additional support services:

- Storage virtualization
- Backup or other redundant storage
- Disaster recovery

### Storage Virtualization

Storage virtualization, like server virtualization, abstracts the services provided by hardware. Consumers of these services can allocate resources without concern for implementation details. For example, details like the logical unit number (LUN) mappings to storage volumes and storage devices are managed by storage virtualization software. When persistent storage is needed, the cloud services consumer simply makes a call to a programming interface specifying the amount of storage required.

**Local vs. Cloud Storage**

Virtual machine instances can provide local storage for temporary storage during the life of the virtual machine instance. The data in this storage is lost when the virtual machine is shut down. The persistent cloud storage described here is provided by devices that are independent of virtual machines. Multiple virtual machines can access the same storage blocks and the data continues to exist regardless of how virtual machines are started and stopped.

The advantages of virtualized storage are similar to those of virtualized servers:

- More efficient use of storage—rather than dedicating large units of storage to a single use for extended periods of time, storage is allocated in smaller increments and for only as long as it is needed
- Lower capital expenditures for individual projects and business units that do not have to acquire storage hardware
- Lower operating costs associated with the pay-as-you-go model typical in cloud computing storage
- More efficient delivery of backup and recovery services

This last benefit is especially important.

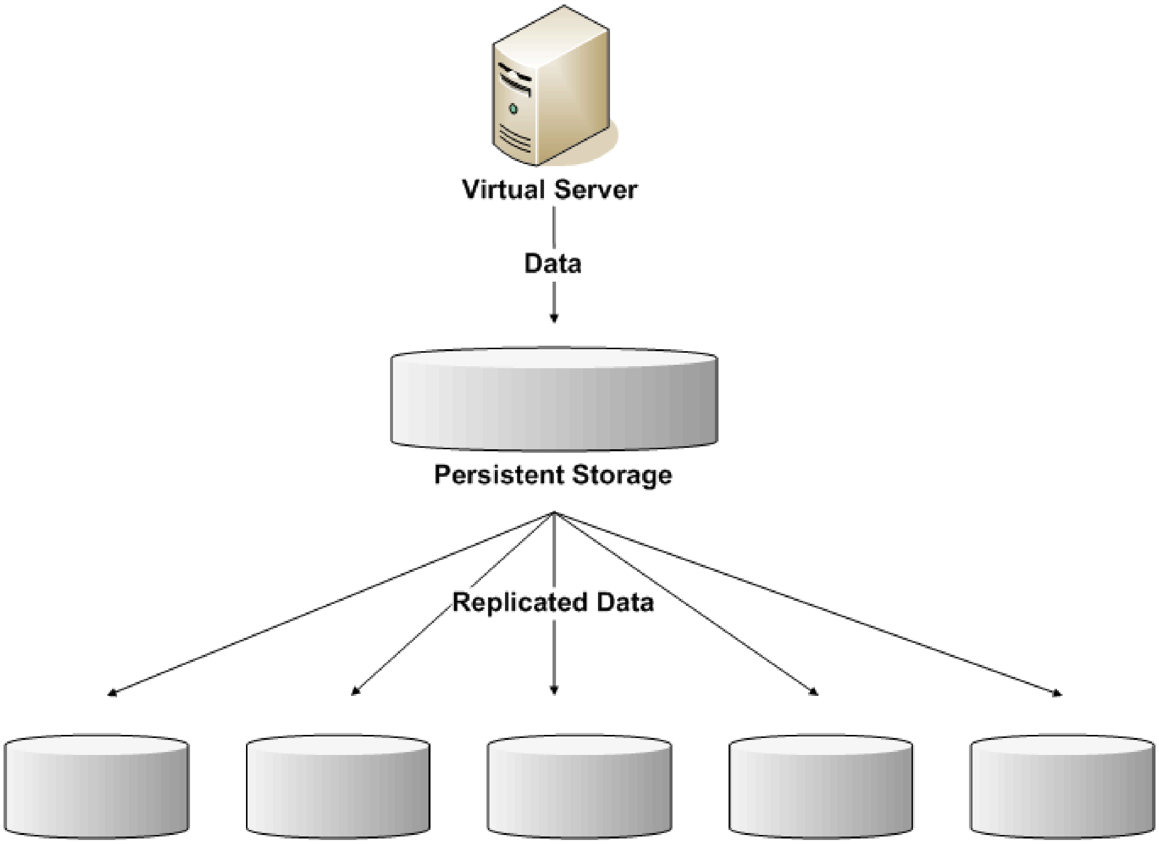


**Figure 6.6: Cloud storage systems can use data redundancy to improve data management performance and reliability of data services.**

One of the advantages of virtualized storage is the ability to provide large amounts of storage through a single logical device—the storage cloud. Behind the scenes, of course, we have multiple disk arrays possibly located in different facilities. This setup creates an opportunity to replicate data across multiple storage arrays to improve reliability and performance.

Reliability is preserved because multiple copies of data are available. If a storage device should fail, there is no need to restore from backup tape; the data is immediately available from another device. The particular device that returns the data is irrelevant to the user. Replication can be done asynchronously so that I/O operations return as soon as data is written to the primary storage device. A background replication process can add new or changed blocks to a queue of blocks that will be copied to devices.

Users can also benefit from improved performance with multiple copies. Data warehousing and business intelligence applications often query large amounts of data. Users contending for access to a single copy might experience bottlenecks and associated drops in performance. In the cloud, different queries can be served by different copies of the database, relieving contention for the same resource.

This type of replication also supports disaster recovery. In the event of a catastrophic failure in one data center, users could be re-routed to another data center that maintains replicated copies of the lost data.

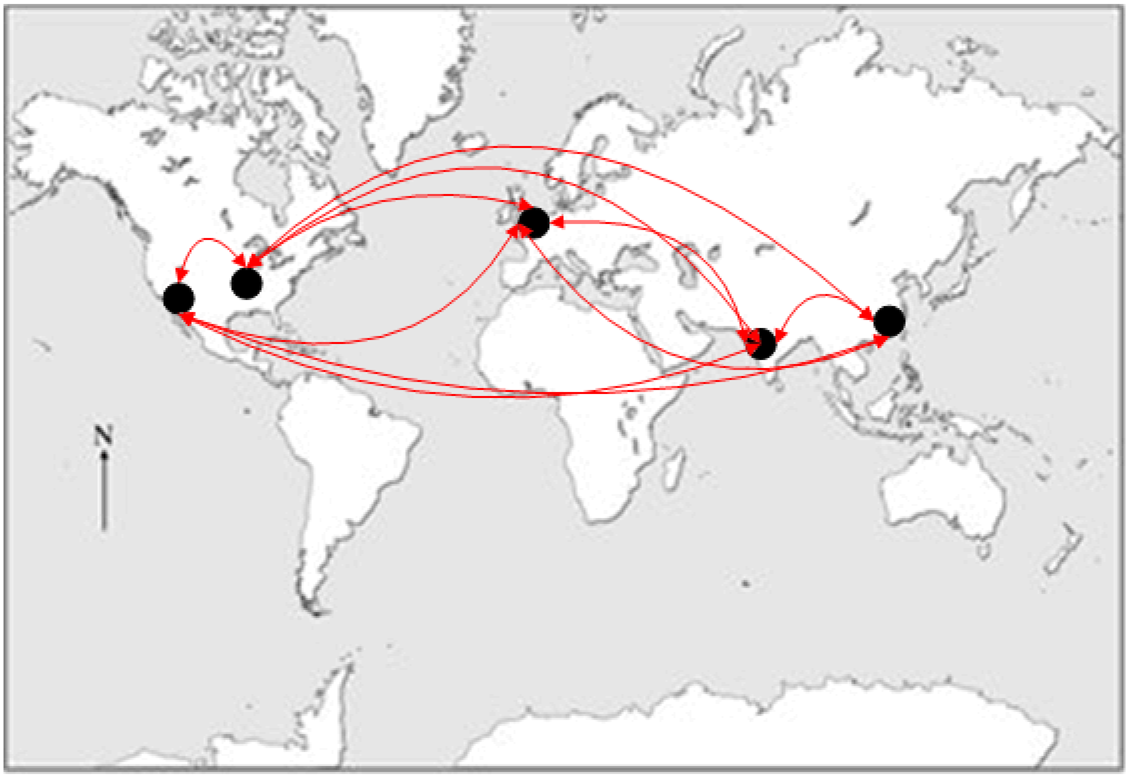


**Figure 6.7: Storage virtualization supports data replication across data centers, which improves reliability and performance.**

This type of replication does not eliminate the need for backup, however.

## Backups and Cloud Storage

Data replication as just described is a valuable asset in cases of disaster recovery, but it cannot meet all recovery requirements. The ideal replication solution maintains multiple copies of data in near real time, so any errors generated in the source system will be replicated to other storage devices as well. Without a separate backup copy of data, there would be no way to restore the database back to a point in time before the error was introduced.

Backup services are generally specified in terms of recovery point objectives (RPOs) and recovery time objectives (RTOs). An RPO defines points of time in history that can be restored; examples include previous day at midnight, previous end of week, or in the case of highly volatile databases, a previous time in the same day. RTOs define the maximum period of time between request of a restore operation and the time the restore operation completes.

Traditional backups are easily accommodated in the cloud. Source data is backed up from the cloud and written to cloud storage. The process could be as simple as copying and compressing data files or block storage from one storage area to another. If backup software supports direct reads and writes to cloud storage, backup processes can take advantage of incremental and differential backups reducing the total amount of space needed to store backup files.

### Management Reporting for Storage Virtualization

A reporting framework, similar to one needed for server virtualization, is required for storage virtualization as well. Businesses that deploy shared disk arrays will probably have a storage reporting system in place that provides much of the needed functionality:

- Reporting on storage used by project, department, or other billable unit
- Cost of storage by type, such as primary storage versus archival storage
- Trending reports on growth in storage use

Infrastructure managers should have additional detailed reports on such things as replication performance.

Storage virtualization and server virtualization share many of the same benefits and management requirements. Together with networks services, they constitute the core infrastructure for cloud services.

## Network Services for Cloud Computing

Networking can be the most resource constrained part of cloud infrastructure. Public cloud providers are necessarily dependent on public Internet providers for connectivity between their data centers and their customers. Private cloud providers might also depend on public Internet providers, especially for access from remote offices or smaller corporate facilities. Dedicated network connections can be employed between sites, but cost is a limiting factor. The key issues we must consider when evaluating different cloud architecture options are:

- Capacity
- Redundancy
- Management reporting

### Capacity

Network capacity limits the amount of data that can move between data centers and between cloud service consumers and the cloud. This directly affects a number of services within the cloud.

### Intra-Cloud Replication

From an infrastructure management perspective, network capacity and speed directly affect replication. As noted earlier, replication is an essential element of creating and maintaining a reliable, high-performance cloud. Heavy demands for loading data into the cloud not only create demand to get data into the cloud but also lead to additional network I/O due to replication. Cloud administrators might determine, for instance, that given the mean time between failures (MTBF) on disk drives, cloud-stored data should be replicated four times to reduce the probability of data loss to whatever threshold they have defined. This means that all data loaded into the cloud plus data generated or updated by cloud-based operations will need to be copied over the network four times.

### Loading Data into the Cloud

Cloud computing is an ideal approach to analyzing large amounts of data. In fact, the phrase "Big Data" has become a moniker for use cases where traditional data management methods break down. The need to deal with multi-terabyte and even petabytes of data used to be a problem limited to specialized niches, such as national intelligence and astrophysics; today, the problem spans industries such as financial services, retail, pharmaceuticals, government, and life sciences.

Businesses with large data sets can leverage large numbers of servers to process and analyze "Big Data" in parallel using platforms such as Apache Hadoop (http://hadoop.apache.org/). It is not always practical to move large amounts of data over networks to load it into the cloud. In such cases, it is best to bypass the network and employ a cloud version of "sneaker net" (that is, ship hard drives to data centers).

**Hadoop and Related Tools**

Hadoop is an open source implementation of the map reduce model made famous by Google. In addition to supporting massively parallel processing over clusters of computers, it includes a scalable database (HBase), a data warehouse infrastructure (Hive), a high-level data flow language (Pig), and a coordination service for distributed applications (ZooKeeper).

Network capacity can be a limiting factor in cloud architectures if a large amount of data (relative to network capacity) has to be moved into the cloud. In some use cases, this is only a problem during the transition to cloud computing when initial data is loaded; after that, data is generated in the cloud using cloud-based servers. In other cases, data may be generated outside the cloud by sensors and other instrumentation; in such cases, we would need to design network capacity to meet large-scale data transfers over the long term.

### Redundancy in the Network

Both computing and storage services in the cloud use redundancy to mitigate the risk of failures. When servers fail, they are removed from the pool of available resources. When storage devices fail, data is retrieved from another device with a redundant copy of the data. Network services require similar redundancy to avoid a single point of failure.

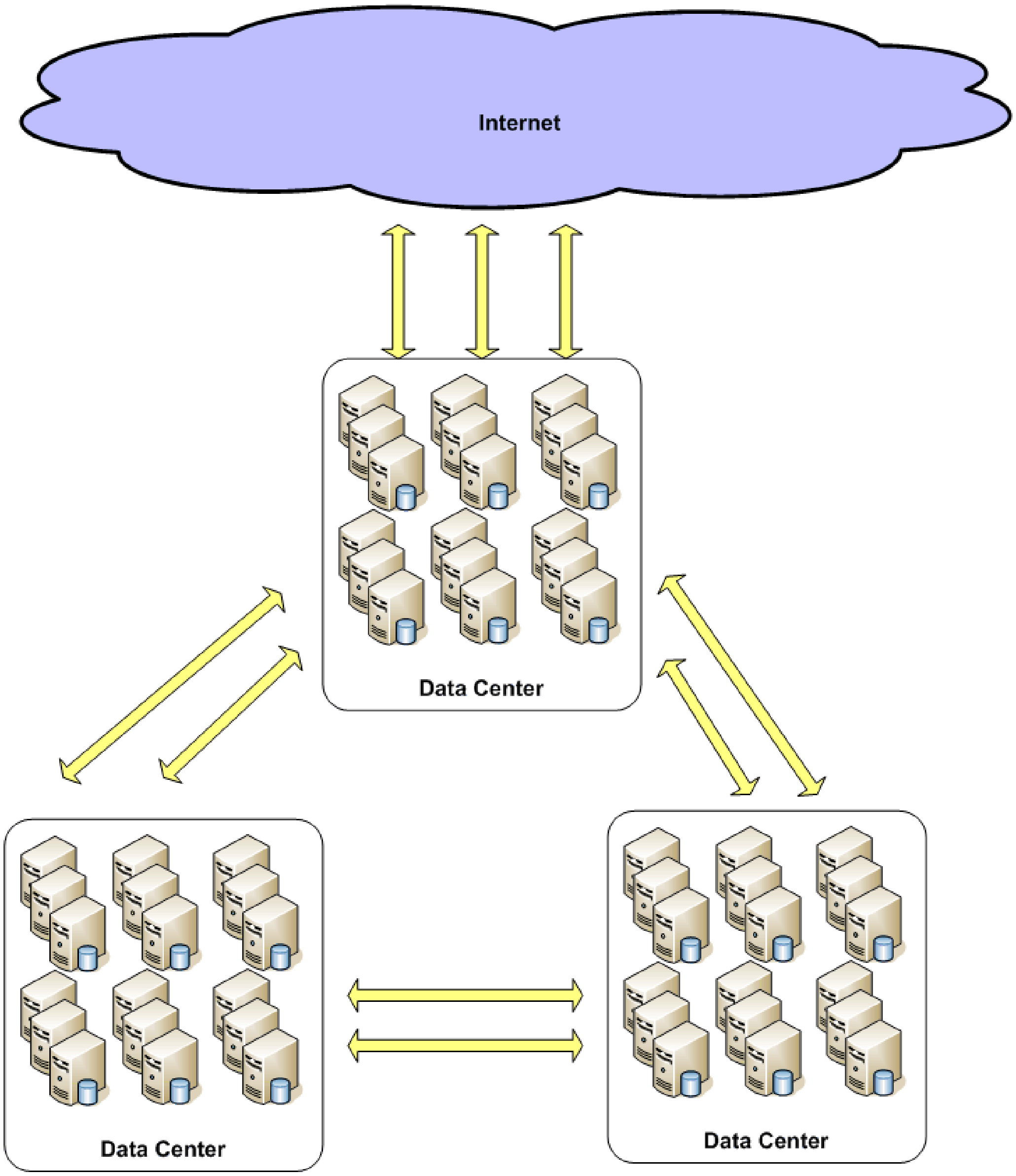


**Figure 6.8: Redundant network connections are necessary between data centers as well as to the Internet.**

## Management Reporting

Cloud service users will be interested in network use reporting as a means to control costs and to monitor trends in network usage. We should not underestimate the cost of network services. For example, when dealing with "Big Data," the cost of network I/O can exceed the cost of computing and storage resources. Management reports can be especially useful if they provide a detailed breakdown of network use by time period and by job. Aggregate reporting over extended periods of time are also needed to determine baseline usage rates, cyclical patterns of variation in network utilization, and long-term growth trends.

Network services, computing services, and storage services are the foundation of cloud computing. Each of these components are provided in redundant manners supporting reliability and increased performance. Management reporting is required in all three areas. In addition to the requirements mentioned, there are further demands for operations support.

## Cloud Operations

Maintaining an efficient cloud operation requires management support mechanisms in addition to those previously described; in particular, image management and workload management. These are tasks associated more with overall cloud management than with individual uses of cloud services.

### Image Management

A cloud can only instantiate the virtual machine images available in the cloud's catalog. The catalog constitutes the baseline set of services provided in the cloud. Users can install additional services, of course, but once a virtual machine is shut down, those changes are lost. The next time that system is required, the additional software must be installed again. For many situations, the cloud catalog constitutes the set of applications and platforms that can run in the cloud.

Machine images can include a fairly wide range of software in addition to the base OS:

- Application servers
- Software libraries
- Analytic software
- Business-specific applications

The base OS as well as the optional software will need to be maintained over time. Each image in the catalog will have to be routinely patched, scanned for vulnerabilities, and rebuilt as new versions of core components become available.

### Workload Management

Workload management functions can vary from basic job scheduling to job optimization. Job scheduling software is useful for queuing large jobs or for repeated jobs in the cloud. The information managed in the job scheduler is useful for tracking future use of cloud services. If metadata about previously run jobs—such as number of servers used, duration of jobs, amount of network I/O, and so on—is collected, it can provide data for estimating future demands on various cloud resources.

Clouds, like any other IT resource, can be optimized. All things being equal, users might prefer to run large jobs overnight and shorter jobs during the workday. This may lead to peak demands that are significantly higher than low demand periods. For example, users may run most data loading jobs at night, leading to periods where demand exceeds capacity while network capacity is underutilized during the day. This type of skewed demand schedule may be smoothed by adjusting price of services. If network resources are in high demand at night, the price is higher than in the day. If demand for computing servers is low in the early hours of the business day, the hourly price for servers is reduced.

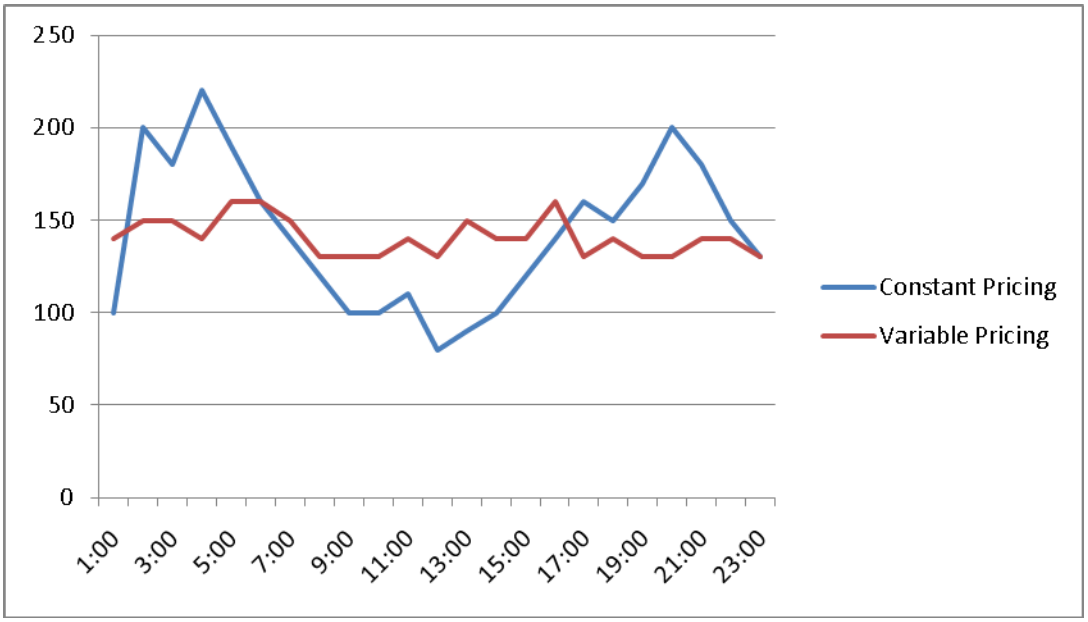


**Figure 6.9: Demand for cloud resources can be smoothed by varying the price of services to shift demand away from peak periods to low-demand periods.**

Software for cloud operations such as image management and workload management are necessary to ensure clouds operate in an efficient manner. Up to this point in the discussion, we have focused on lower-level services, such as virtual servers, storage, and network I/O, and management of those low-level services. Information technology services also provide high-level functions.

## Services Layer: Adapting IT Operations to Cloud Infrastructure

The cloud is an ideal platform to run many, although certainly not all, business services. Applications written to take advantage of mainframe capabilities and tuned to run on mainframe OSs are probably best run on that platform. Many business applications are already running on distributed platforms, taking advantage of clusters of servers, shared storage devices, and high-speed network interconnections. These applications are ideal candidates for deploying to a cloud, but there are still additional factors that should be considered when moving systems to the cloud:

- Designing for recoverability
- Managing workload
- Performing maintenance and upgrades
- Maintaining security

These are all considerations in service delivery, but cloud architectures influence how we address them.

### Designing for Recoverability

Recoverability is an issue at the application level as well as at the infrastructure level. An application that depends on a large pool of servers to analyze data should address questions such as:

- What happens if a single server fails?
- Will the job have to be restarted from the beginning?
- Is there a way to detect what data was being analyzed when the server failed?
- Is there a way to roll back to a prior state without starting from the beginning?

There are many design choices for addressing these types of questions. For example, each server can receive a subset of data from a distribution node. The distribution node maintains a queue of data sets to distribute to analysis servers. When the distribution node receives a message that a data set has been analyzed, it is removed from the queue. In this way, if a server fails while analyzing data, the data will simply be sent to another server for processing. To avoid a single point of failure, this solution would also require a failover mechanism to start another distribution node should the primary one fail. Alternatively, multiple distribution nodes could run simultaneously and use persistent cloud storage to maintain the queue of data sets that could be read by any of the distribution nodes. This is just one example of a resilient application design for distributed computing; there are many others.

### Managing Workload

Providing services through the cloud will require us to think of jobs and workloads in ways that we do not necessarily need to when we have full control of dedicated servers. In particular, we will want to maximize server utilization when we run our jobs while ensuring jobs finish in whatever time window required. If, for example, our cloud charges a minimum of 1 hour of server time for each instance, and we have several small workloads, we should run those in tandem on a single virtual server rather than run them on different servers each incurring the minimum charge.

### Performing Maintenance and Upgrades

Maintenance and upgrades of applications will have to be coordinated with the cloud service provider. When departments or projects manage their own servers, they can determine their own upgrade schedule (within broader company policies, anyway). In the cloud, applications are delivered through virtual machine images maintained in the centrally managed image catalog. Similarly, patching and other maintenance decisions will have to be coordinated with the cloud provider.

### Maintaining Security

Fundamental security considerations continue to persist in the cloud. Of particular importance is the need to manage identities and entitlements in the cloud. If private information is stored in the cloud, appropriate application-level controls will have to be in place to prevent unauthorized access. Direct access to the private data via the persistent storage API will also have to be blocked through authentication mechanisms and access control lists (ACLs) or other authorization control.

In addition to access controls, we must consider application-level security issues such as vulnerability scanning. Ideally, security concern is addressed by the cloud service provider, but customizations might be the responsibility of the application owner.

## Service Management Layer

A final piece of the software and infrastructure architecture that makes up a cloud is the service management layer. Throughout this chapter, we have considered core computing, storage, and network services from both the service provider and the service consumers' perspective. We have seen the overlap in concerns between both parties for issues such as image management, workload management, and optimization of resources. This overlap and shared need for support service continues as we consider the service management layer.

Service management includes additional services necessary for managing the business of providing and using a cloud. These include:

- Provisioning, which are services that allow non-IT professionals to deploy cloud services as needed
- Performance management, which provides additional management reporting and monitoring services that allow cloud providers to understand detailed operations in the cloud as well as plan for longer-term management issues
- Usage accounting, which is necessary for tracking who uses which services and for how long; this is essential for proper cost allocations or billing for cloud services
- License management services are important for compliance; running a cloud does not necessarily entitle one to run as many instances of a commercial off-the-shelf product as one would like—cloud service consumers cannot not be expected to monitor the number of copies of licensed software running in the cloud or to know licensing details, thus license management systems are needed to ensure compliance

Support services such as these, and others related to service monitoring and availability, provide the higher-level management services necessary, especially when running a private cloud.

## Summary

Cloud services can be provided with a number of architectures, and a wide range of factors need to be considered when choosing to deploy a cloud. Issues related to providing computing services, storage services, and network services all come into consideration at the most fundamental levels. Reliability, performance, and management reporting are recurring themes when considering those three core services. In addition, cloud operations management, adapting IT operations to cloud architectures and topics, and service management must be examined as businesses choose the right cloud architecture for their situations.

# Chapter 7: Roadmap to Cloud Computing: The Planning Phase

The benefits of cloud computing are well established: This model of service delivery is efficient, scales well, and meets a wide range of business needs. These benefits are maximized when business drivers, infrastructure, and policies are properly aligned to take advantage of the clouds method of delivery services. Cloud computing is not a universal panacea and some business processes are better delivered by other approaches. Not all businesses will benefit equally from cloud computing; much depends on how well they prepare for the adoption of cloud computing. The purpose of this chapter is to outline a planning process that will help maximize the benefits of cloud computing. The planning process consists of several steps:

- Assessing readiness for cloud computing
- Aligning business strategy with cloud computing services
- Preparing to manage cloud services
- Planning for centralized resources
- Committing to service level agreements (SLAs)
- Meeting compliance requirements

The chapter concludes with a pre-implementation checklist to help manage your own planning phase.

## Assessing Readiness for Cloud Computing

The ancient Greek aphorism “know thy self” is surprisingly relevant to planning for cloud computing. The first step in the planning process is to assess where the organization stands with respect to

- Web application architecture
- Self-management of compute and storage services
- Standard platforms and application stacks

Each of these three areas is relevant to the delivery of cloud services. At this stage of the planning process, it is not necessary to have all three in place at ideal levels; in fact, most organizations not already supporting a cloud infrastructure will likely not have fully deployed and standardized around these three areas. This is not a problem. This is the planning process and the point of the assessment stage is to understand what resources are in place when we begin the move to cloud computing. The information gathered in this process will help to guide later planning and design efforts.

## Web Application Architecture

Applications are designed using a variety of design principles that are roughly grouped into what we call application architectures. These architectures vary in terms of a number of characteristics, such as:

- Level of centralization
- Coupling of components
- Accessibility of components
- Ability to execute multiple instances
- Platform independence

We need to consider how existing applications are designed with respect to each of these to understand how well those applications are adapted to cloud infrastructure. As we will see, those with characteristics most closely aligned to Web application architectures are best suited for the cloud; but first, we will briefly describe each of these characteristics.

### Levels of Centralization

An application may be centralized with all application code executing on a single machine, in a single process, and under the control of a single component. Centralized applications range from small utilities to large enterprise-scale applications. For example, a simple text editor can be realized with a single executable that runs a simple accept input-process input-generate output loop. Also in the most centralized application category, we have large, complex batch-oriented mainframe applications that have developed over years to incorporate many functions. A billing system for a telecommunications company, for example, may have millions of lines of code that, although divided into sub-modules, is largely controlled by a single control module and executes on a single machine. These applications are at one extreme of the centralization spectrum.

The middle ground of centralization is typified by client/server applications. In this application architecture, the work performed by an application is divided between servers, which perform the bulk of computing and storage operations, and client devices that are responsible for user interactions. A simple example of an application employing this approach is an order entry system consisting of a .Net user interface running on a Windows desktop and a SQL Server database. The client and the server components are fairly tightly coupled but they execute on separate devices and the components, with some effort, could be exchanged for a different form of the component. For example, the SQL Server database could be replaced with an Oracle database with little impact on the client.

Decentralized applications execute multiple processes over multiple devices. Web application architectures take advantage of decentralized applications to combine services. A typical Web application may require persistent data storage provided by a relational database, user management provided by an LDAP server, compute services provided by a Java application server, and user interaction services provided by a Web server. Decentralized applications are especially well suited for cloud architectures because services can be run on virtual servers as needed and new services can be easily added without disrupting the loose coupling between services or requiring one to provision additional dedicated hardware.

### Coupling of Components

The components of an application, such as a service, module, or procedure, may be tightly coupled with other components. For example, a procedure for calculating the shipping costs of an order may be part of a larger order entry program that calls that procedure at specific points in the execution of the order entry process with a data structure specific to that program. This is an example of a tightly coupled set of components.

Loosely coupled components can execute in more autonomous ways. They may run on different servers, they may be executed on the behalf of multiple calling programs, and they exchange input and output in ways that support a broad array of calling applications. Applications built on loosely coupled components work well in cloud architectures because the number of instances can be adjusted to meet demand and the services they provide are available to other applications running in the cloud.

### Accessibility of Components

Accessible components are those that are available to different services. To be accessible, a component must:

- Be programmatically discoverable so that other components can find it
- Exchange input in well-generalized formats, such as XML
- Respect authentication and authorization requirements
- Maintain reasonable response rates under varying loads

Web application architectures are built on accessible components using standards such as SOAP and WS-Security to meet some of these requirements. Others, such as the ability to maintain reasonable response rates, are met by using scalable architectures such as compute clouds.

### Ability to Execute Multiple Instances

The ability to execute multiple instances might seem an odd requirement at first. After all, why couldn't one run multiple instances of an application? The answer: You couldn't run multiple instances when components are tightly coupled and exclusive use of a resource is required. A monolithic application, for example, may assume that it can lock a file of customer data for exclusive use preventing other processes from operating on that resource. If the application cannot finish processing in the time window allotted to it, the application manager could not simply start another instance of the program on a different server and finish in half the time.

Applications that are well suited to the cloud do not require that only a single instance of the program execute at any one time. Older applications may not have been designed with this characteristic in mind, but Web application architectures, built on decentralized, loosely coupled components, generally do not have these problems.

### Platform Independence

Another characteristic of Web applications is that services are not required to run on a single type of platform. Services are decoupled so that requirements define how data is exchanged between those services but not how the services execute. A service that needs to retrieve information about a user could just as easily do so by calling an LDAP service running on a Linux platform as by calling Active Directory (AD) running on a Windows server.

Web application architectures are characterized by decentralized, loosely coupled components that are accessible to other service components and can scale to meet loads placed on them. This combination of characteristics is seen in the service bus model that uses message passing and service abstraction. Applications that use this approach are well suited to the cloud. Applications that do not use this model can still benefit from the management and cost benefits of using cloud services. The more decentralized and loosely coupled the application, the greater the potential benefits.

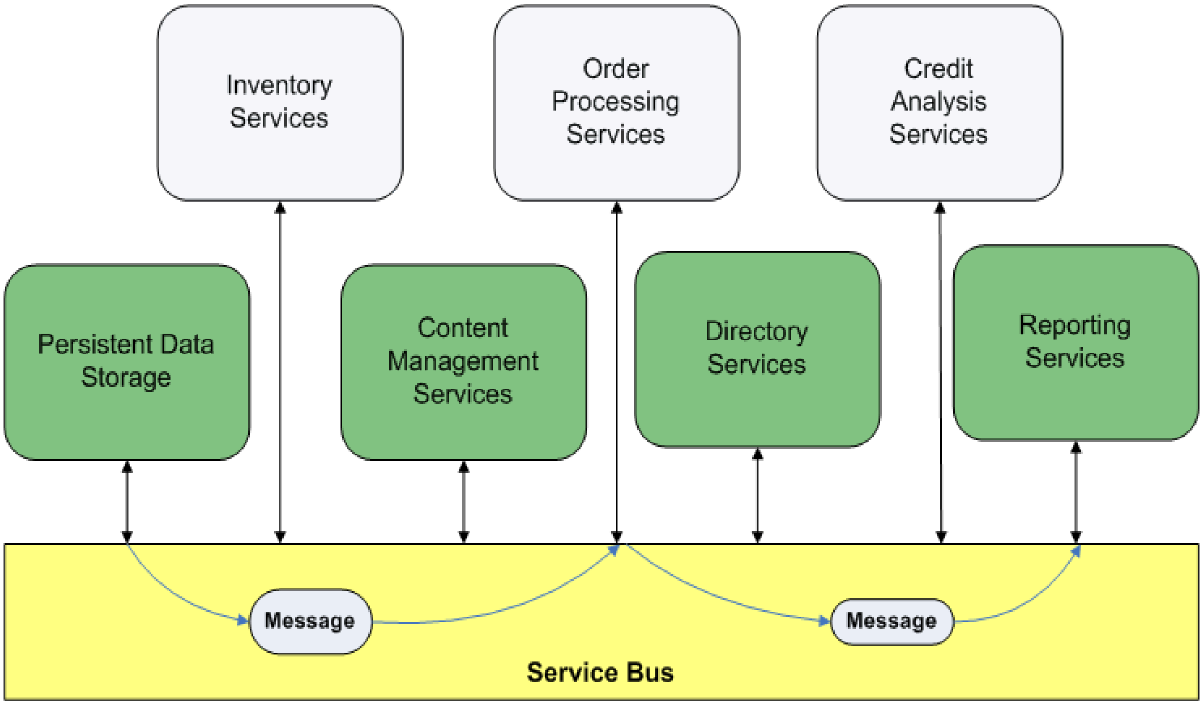


**Figure 7.1: Web applications that utilize a service bus model are well suited to executing in the cloud.**

From an assessment perspective, a business should try to determine how closely existing applications use a Web application architecture. Even without a formal service bus, other application architectures can exhibit the characteristics that fit well with cloud computing. For example, the common 3-tier architecture that Figure 7.2 shows has many of the characteristics previously described.

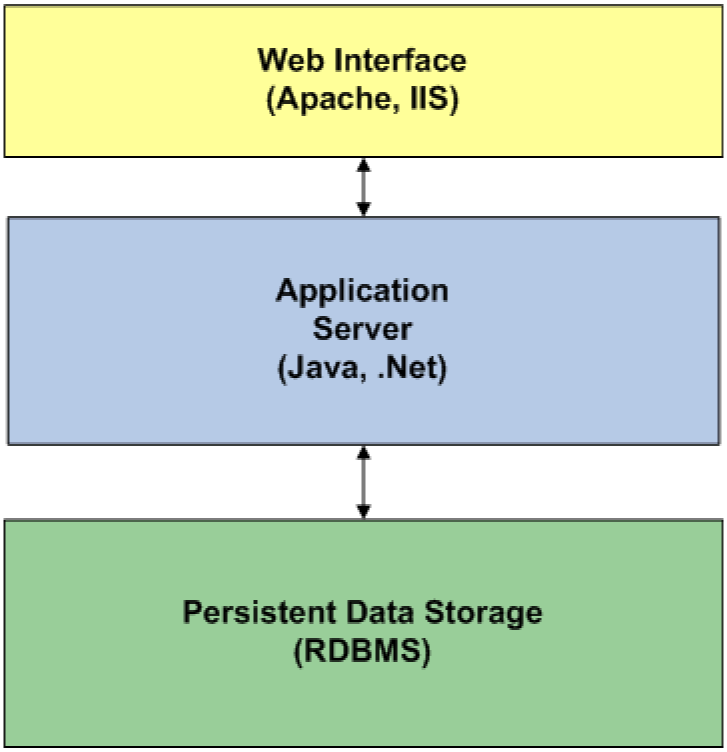


**Figure 7.2: The 3-tier architecture also exhibits decentralization, loose coupling, and platform independence.**

Another factor to consider when assessing readiness for cloud computing is support for self-management of resources.

### Self-Management of Compute and Storage Resources

The efficient allocation of compute and storage resources requires the ability to start and stop services on demand in response to changing conditions. As we have discussed in previous chapters, one of the inefficiencies in dedicating servers to a single application is that such servers have to be configured for peak capacity and this often leads to underutilization during non-peak periods. The same problem could occur in the cloud if cloud consumers were not able to rapidly respond to changes in demand. This is true for both computing and storage resources. It is not uncommon for users of storage arrays to have to submit a ticket to IT support to have additional disk space allocated to their dedicated servers. This could take minutes to days depending on the backlog in IT support. The potential delays can lead to application managers allocating more storage than needed simply to avoid any possible risk of running out of space and not getting additional storage in time.

Ideally, application managers would be able to allocate compute and storage resources as needed. In many cases, self-management software is not in place prior to adopting cloud computing. This certainly will not prevent a business from moving to cloud computing but it will require that such a system be put in place. When evaluating compute and storage service self-management software, consider the following factors:

- Ease of use
- Management reporting for cloud consumers
- Integration with accounting and billing systems for chargeback purposes
- Adequate authentication and authorization
- Job scheduling features or integration with existing job scheduling systems
- Ability to scale to the number of cloud consumers

Another factor that will influence ease of management is the ability to standardize on platforms and application stacks.

## Standard Platforms and Application Stacks

Standardizing on a limited number of operating system (OS) platforms and application stacks can ease the management of a compute/storage cloud. Many organizations may find something akin to an 80/20 rule applies to them: 80% of application needs can be met with a relatively small number of platforms and application stacks, possibly 20% of all the platforms and stacks that are currently in use in a business.

### Determining Required Platforms and Application Stacks

For planning purposes, compile an inventory of applications including OSs, application servers, directory servers, Web servers, and other core components. With that inventory, one can derive a list of distinct combinations of platforms and application stacks. It is likely that many of the applications run on similar sets of OS and application stack. Those compose the “80%” side of the 80/20 rule.

For the remaining applications, assess the level of difficulty in transitioning from the existing combination of OS and application stack. For example, if many applications are running on a Red Hat version of Linux while a handful are running on SUSE versions, the effort required to migrate between those should be fairly low in most cases. An application that depends on a Windows server platform or on components that only run on Windows platforms would be significantly more difficult to port to a Red Hat platform. The goal in moving to a cloud architecture, however, is not to redesign existing applications but to leverage the benefits of the cloud.

This calls for something of a balancing act. First, we want to minimize the number of distinct application stacks we support in the cloud but we also want to maximize the number of applications that can be supported in the cloud. Adding application stacks should increase the ability to support either a significant number of general applications or targeted mission-critical applications that would benefit from running in the cloud.

Organizations that already have large portfolios of Web applications will likely find that they can address many of their requirements with a small number of different application stacks, such as:

- LAMP stack, with Linux, Apache, MySQL, and Perl/Python/PHP
- Windows stack, with .Net applications and servers
- Commonly used application servers, such as Java application servers and Java portals

Regardless of the combination of application components and OSs, there are services and policies that should be standardized across platforms in the cloud.

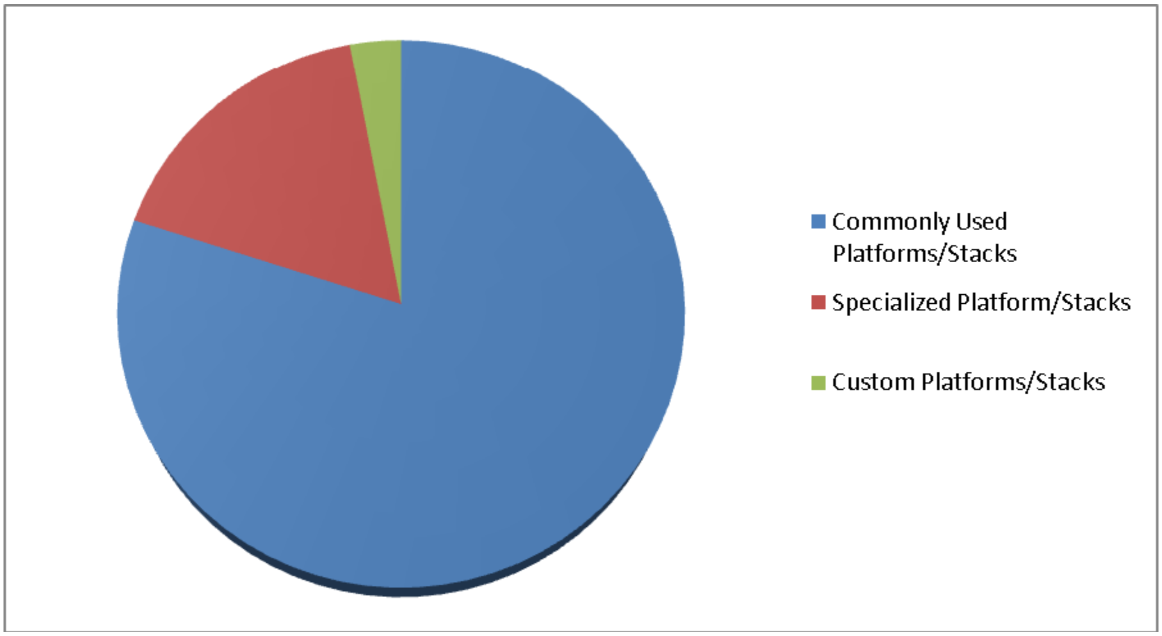


**Figure 7.3: Relative distribution of platform/stack needs that can be met by a small set of commonly used stacks, specialized stacks for less common requirements, and custom platform/stacks for single, custom needs.**

### Required Support Services

The cloud should provide identity management services such as authentication and authorization services. These are necessary to properly administer a cloud. For example, these systems would be used to:

- Determine how users or agents are authenticated to self-service applications used to manage cloud services
- Determine limits on cloud consumers, such as the maximum number of instances a user can start at one time or the length of time a single instance can be running a single virtual machine
- Allocate charges for cloud services to the proper department or billing code

The same authentication and authorization services could be made available to applications running in the cloud, reducing the need for application-specific identity management systems.

### Customization and Specialized Requirements

Another issue to consider around standardizing platforms and application stacks is the need for specialized versions of cloud-provided standards. The company may have standardized on Java or .Net for all application development but a department needs to host a third-party application developed in Ruby. Ruby is an interpreted programming language akin to Perl and Python. Ruby must be available on a server to execute a Ruby application. If this language is not part of the standard cloud offerings, the department may want to create a specialized virtual machine image to meet their needs.

There are advantages to allowing customized combinations of OSs and applications stacks. The most compelling is that cloud consumers have access to exactly what they need. There is no need to port applications to other platforms or find alternative solutions that run on standard platforms.

The disadvantage of allowing customized virtual machine instances is that they are more difficult to manage. For example, who is responsible for patching and maintaining customized virtual machine images? The creators know the components and applications best, but IT support staff may be most familiar with lower-level details, such as OS vulnerabilities. Also, if a patch were to break the application, how would it be dealt with? Will users have the knowledge and time to test patches before deploying in production? Will metadata about the contents of custom images be kept up to date? Will this task duplicate efforts already carried out by cloud providers? We are starting to see the potential for the kind of inefficiency that drives up IT costs in non-cloud environments.

Assessing readiness for moving to a cloud architecture is a critical first step in the planning process. This stage of planning requires an assessment of which applications will fit well with the cloud; those using Web application architectures, such as a service bus architecture or a multi-tiered application stack are well suited for the cloud. Once those applications are in place in the cloud, cloud consumers will want precise control over how they execute and the storage they use. Self-management services are essential to realizing the efficiencies of the cloud. Finally, during the assessment stage, one should identify the standard platforms and application stacks that are needed in the cloud. The benefits of the cloud should not be undermined by unnecessary management overhead.

The first stage of planning considered primarily technical aspects of delivering services in from a cloud. In the next stage, we consider more business-oriented aspects.

## Aligning Business Strategy with Cloud Computing Services

Clouds are deployed to deliver services and services are established to meet business requirements. To ensure cloud services are deployed in a way that is aligned with business strategy, we should consider existing workloads and their corresponding value metrics.

### Workload Analysis

Right now in your business there are hundreds, thousands, or even more applications executing business processes. Some of these are transaction-processing systems that provide high-volume, rapid processing of orders, inquiries, reservations, or a broad array of other narrowly focused business activities. Other applications are performing batch operations, such as generating invoices, reviewing inventory levels, or performing data quality control checks on databases. Still others are extracting data from one application, transforming the data into a format suitable for analysis, and moving it into a data warehouse. There is a wide array of different types of applications that are needed to keep an enterprise functioning.

These different types of applications have different requirements and constraints that must be considered when moving them to the cloud. For example, they might need:

- To start and finish executing within a particular time period
- To wait for another job to complete before it can begin
- To limit the functionality of some services, for example, write-locking a file to perform a backup
- To provision a significant number of servers for a short period of time for a compute-intensive operation

Any cloud will have finite resources. As part of the planning process, we need to understand what types of jobs can run in the cloud (that was addressed in the previous section) and how to run them efficiently.

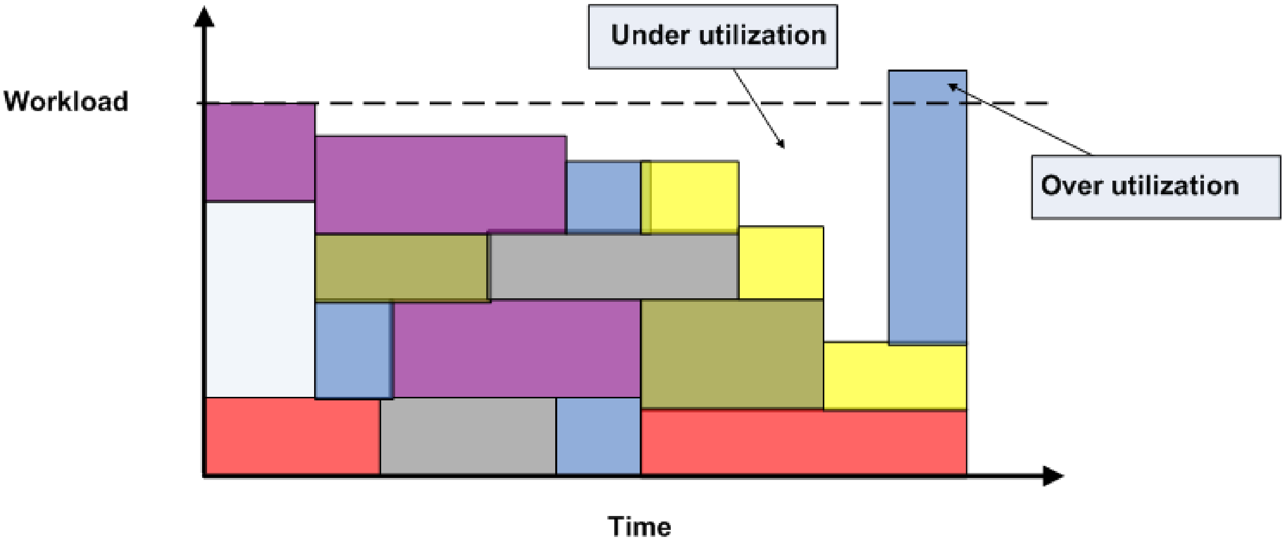


**Figure 7.4: The combination of workloads running in the cloud determines overall utilization at any point in time; ideally, periods of under-utilization and over-utilization are minimized.**

Cloud consumers are the ones who will decide when to start and stop jobs and how many virtual servers to provision for particular tasks, but in the planning stages, we will want to ensure there is sufficient capacity. To do so, we can look at existing workloads and take into account:

- How often jobs execute on dedicated servers
- The level of utilization of those servers
- Time constraints on when those jobs execute

Once again we are faced with a balancing act. We want to deploy sufficient cloud infrastructure to avoid periods when cloud consumers want to run more jobs than there is capacity for (over-utilization) at the same time we do not want extended periods of time when servers are idle (under-utilization). This brings us to the second aspect of business alignment: value metrics.

### Value Metrics

Developing a precise and generally accepted ROI function for any IT investment is difficult at best. To assess the value of cloud computing, we can consider more targeted measures such as the value relative to hardware and software investments and relative to labor costs.

### Hardware and Software Values

We will begin with hardware and software value measures by considering the constituent costs of running an application on a set of dedicated servers. They include:

- OS costs
- Application software licensing and maintenance costs
- Database management licensing and maintenance costs
- Hardware procurement and maintenance costs

The costs are relatively fixed, so it does not matter whether you run your application 24 hours a day or 1 hour a day; the hardware and software costs are the same when running that application on a dedicated server. The cost model of a cloud is different.

In a cloud model, the cost of licensing and hardware can be divided among multiple users. For example, one department might run an application for 2 hours a day, another for 6 hours a day, and a third user runs the application for 10 hours a day. Prorating the cost of licensing and maintenance over 18 hours of daily utilization lowers the cost for all three users, especially the user who only needs 2 hours of application services per day.

### Labor Value

The cost of labor in the cloud model is lower than dedicated server models for a couple of reasons. First, in the cloud, there is an opportunity to standardize hardware. Large numbers of servers all built using the same, or very similar, components are easier to maintain. If a hard drive fails in a server, replace it with a spare that would work just as well in any other server. There is less overhead to manage inventory and fewer chances for errors in configuration if all servers use the same type of components.

> **Standardizing vs. Repurposing**
>
> When first deploying a cloud, you might want to repurpose hardware that had been dedicated to applications that will now run in the cloud. Some of this hardware may not match the cloud's hardware standard. Once again, we have to balance the benefits of standardizing on hardware with the cost savings of repurposing hardware. One option is to repurpose non-standard hardware but replace it with standard equipment as it fails or no longer meets functional requirements.

Second, with self-service management, cloud consumers can manage their own applications and workloads. IT support staff that had been dedicated to responding to basic server support (for example, installing software, allocating disk storage, and running backups) can now be dedicated to higher-value tasks. The cloud infrastructure will require IT support services that can be provided more efficiently in the cloud than with servers dedicated to particular applications. For example, if a vulnerability is discovered in an OS, a single administrator can patch the OS, regenerate virtual machine images, and deploy those images to the service catalog. Compare that task with the patching of hundreds of servers across the organization. By analyzing workloads and calculating initial value measure in the planning process, we are better able to align business requirements in a cost-effective way with cloud services.

## Preparing to Manage Cloud Services

Up to this point in the planning process, we have considered readiness of an organization to move to a cloud architecture in terms of technical issues, such as the use of Web application architectures and standardization on platforms and application stacks. We have also examined the alignment of business strategy with cloud services in terms of workload analysis and value metrics. We now turn our attention to a few issues related to longer-term management of cloud services. These are:

- The role of private, public, and hybrid cloud services
- Planning for growth
- Long-term management issues

These issues, as we shall see, are strongly influenced by demand for cloud services.

### Role of Private, Public, and Hybrid Cloud Services

There are three broad modes of delivery for cloud services: private, public, and hybrid. A private cloud is deployed and managed by an organization for its own internal use. The organization controls all aspects of cloud implementation, management, and governance. One of the most significant advantages of this approach is that data never leaves the control of its owner. This reduces the risk that an outside party will gain access to private or confidential data. Depending on the implementation and management details, private clouds may be more cost effective as well. For example, a business may have significant investment in servers that can be redeployed in the cloud, lowering the initial costs.

A public cloud is one that is managed by a third party that provides services to its customers. The primary advantage is low startup costs on the part of customers and minimal management overhead, at least with respect to basic cloud services. Businesses will still need to manage their workloads, allocate chargebacks, and so on.

Choosing between public and private cloud implementations is not an all-or-nothing proposition. Hybrid clouds, or the combination of private and public implementations to run business services, have emerged as a third alternative. Consider the economic benefits. There may be a point, however, at which the benefit of adding servers to a private cloud is not sufficient to offset the costs of adding them. For example, the distribution of workloads may entail a number of peak periods where demand exceeds the capacity of the private cloud. These peaks may be regular short periods (for example, at the end of the month when accounts are closed and data warehouses and data marts are updated and many reports are generated) or they may be more unpredictable periods of high demand.

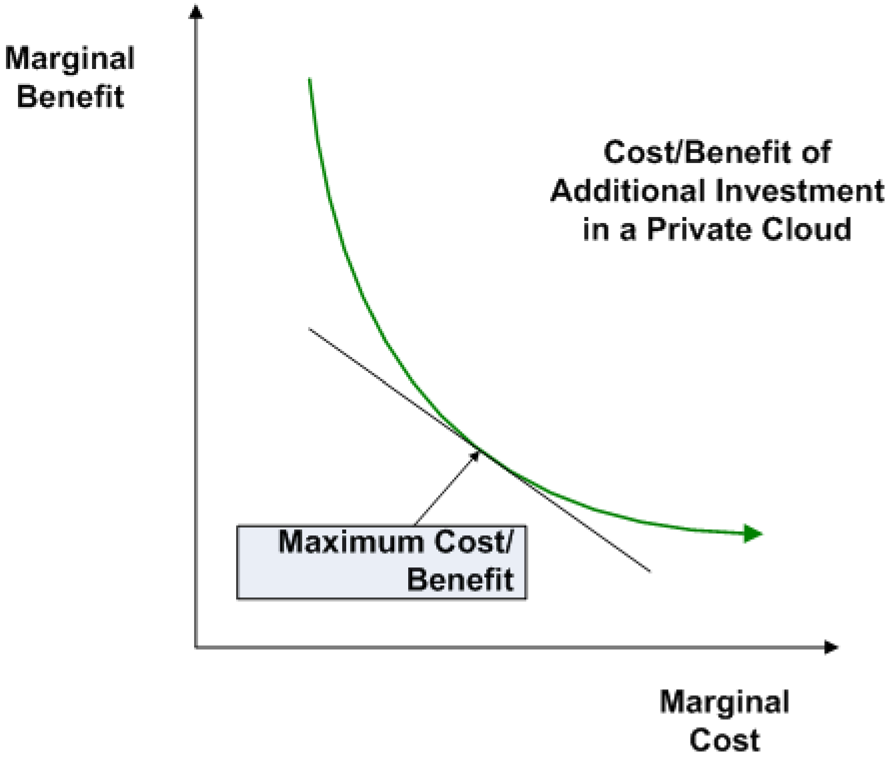


**Figure 7.5: The cost of adding and maintaining additional cloud resources eventually reaches a point where the costs outweigh the benefits. At this point, a hybrid cloud approach may be the most cost-effective option.**

### Planning for Growth

If successful, a cloud is likely to grow both in terms of underlying infrastructure and in terms of the number of services provided by the cloud. In the case of private clouds, growth in infrastructure can occur internally by adding servers, storage, and ancillary equipment as needs demand or by adopting a hybrid cloud approach.

Growth in services will put a different kind of management burden on cloud providers. In particular, cloud providers will need to plan for:

- Expansion in the number of OSs and application stacks that may be supported
- Growing demand for custom virtual machine images to accommodate specialized requirements
- A growing base of cloud consumers with widely different needs
- Emerging categories of users, such as long-term cloud consumers who need continuously running servers, users with intermittent but regularly scheduled needs for servers, users who will take advantage of the cloud for occasional needs, or spot users who will use the cloud only during off-peak hours if the cost is lower at those times.

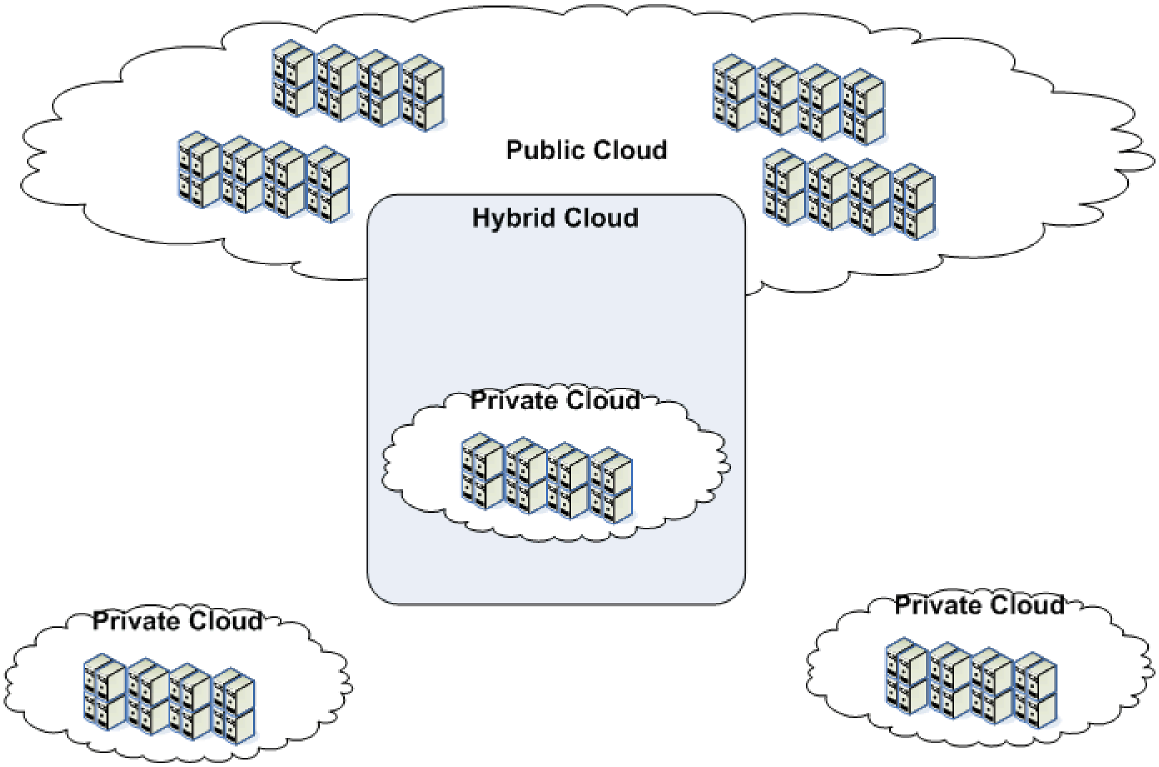


**Figure 7.6: Using public cloud services in a hybrid cloud configuration during peak demand periods may be the most cost-effective way of meeting the demand for peak capacity.**

These different factors will help shape management and pricing policies. A market pricing model, for instance, may be introduced to more evenly distribute the workload in cases where there are periods of high and low demand. Peak pricing could be instituted during high-demand periods and lower prices during low-demand periods. Another option is to use an auction model in which cloud consumers specify the price they are willing to pay for a resource; the cloud allocates resources to the highest bidder, then the next lower bidder, and so on until all resources are allocated.

There are many ways to manage and price services; an important point to remember is that the policies and methods used in the early days of cloud adoption may not be the best option in later stages. Following past practices because "that is the way we've always done it" is not always a recipe for success.

### Long-Term Management Issues

In the planning stages for adopting a cloud, it is important to consider some of the long-term management issues that cloud providers will face. These include both service and infrastructure issues:

- Maintaining the security and integrity of virtual machine images
- Monitoring, detecting, and blocking unauthorized uses of the cloud
- Planning for high availability and disaster recovery, possibly with multiple sites for a private cloud or with the use of a hybrid cloud approach
- Managing identity, authentication, and authorization mechanisms
- Handling physical configuration of the cloud and power consumption
- Acknowledging the potential for rapid, significant rise in demands, for example with the greater use of instrumentation and data collection

These are broad issues that will continue to evolve over time. In addition to these, there are several long-term issues and responsibilities that warrant more detailed consideration.

## Planning for Centralizing Resources

Cloud computing gains many of its advantages from centralizing resources, management, and governance. During the planning stage, it is important to begin formulating policies and practices that support centralization. This can come in several forms:

- Standardizing to reduce complexity
- Streamlining service management
- Virtualizing physical resources

These various forms of centralization are important individually, but they also reinforce and support the realization of each other.

### Standardizing to Reduce Complexity

Standardization reduces complexity, especially in the cloud. When we use the "one server for one application" approach to delivering services, there is less need for standardization than in cloud models. That is not to stay standardization is unimportant; it is important, but the degree of standardization required to realize benefits is not as great as it is with cloud computing.

Take for example a sales department that runs a small data mart. The department had hired an analyst who had worked with open source reporting tools in the past and persuaded the department manager to use those tools as well even though the business had standardized on a commercial tool suite. The department is responsible for building and maintaining its data mart, and the group functions well with it. Centralized IT is not responsible for maintaining sales' department's system and does not object to it. (We will ignore the security implications of this decision for the moment). Now picture this application moving to the cloud.

A virtual machine image would have to be created and maintained in the service catalog of the cloud. Centralized IT management would be responsible for deploying and maintaining the image. As it is in the catalog, other users might make use of it. The user base might grow to the point that IT must spend significant time to learn the tool in order to provide support. What started as an isolated instance of using non-standard software slowly shifts to becoming an institutionalized, supported application.

Standardization is a key method of reducing complexity. The goal of standardization is to meet all functional requirements with a minimal set of computing components. Once requirements are met, adding components adds to complexity—that is, the number of interacting components that need to be maintained and adapted to function with other components—without adding to the goal of meeting requirements. In the previous example about data mart reporting, a non-standard system was used when the enterprise standard solution would have worked. The result was additional complexity with no additional benefit. Such situations should be avoided when deploying a cloud.

### Streamline Service Management

One of the benefits of centralization is that by delivering services at large scales, it pays to invest in optimizing those services. A fast food chain that serves millions of sandwiches a year will optimize every aspect of the production, preparation, and delivery of those products. Similarly, the fact that hundreds or thousands of users will repeatedly invoke the same standardized set of services demands attention to streamlining and optimizing the delivery and management of those services.

In order to streamline service management, we need applications in place that reduce the manual labor and complexity of workflows required to implement management processes. In particular, service management should include:

- Support for discovering services provided in the cloud through detailed and up-to-date metadata about services
- Virtual machine images that are designed to support services, such as report generation, and not just OSs and application stacks, such as Linux with a statistical analysis package installed
- Management reporting that allows cloud consumers to track and optimize their own use of cloud resources

- Ability to provide timely support for cloud consumers in cases where there are problems executing jobs in the cloud
- Utilization analysis reports to give those responsible for managing cloud services the information they need to detect trends and analyze varying patterns of resource utilization

One of the factors that supports the ability to streamline service management is the ability to virtualize cloud infrastructure.

### Virtualizing Physical Resources

The final aspect of centralizing resources we will consider is the need to virtualize physical resources. As we have encountered repeatedly within our discussion of cloud computing, the ability to virtualize computing and storage services are at the foundation of the efficiencies provided by the cloud model. The key physical resources that should be virtualized are servers and storage.

Setting up a set of virtual machines on a single server is straightforward: install a hypervisor and create virtual machine instances based on OS(s) of choice. Scaling virtualization to a large number of servers requires management software that can manage multiple hypervisor clients from a single console.

Storage services also need to be virtualized so that they appear to cloud consumers to be a single storage device. Virtual machine instances in the cloud, for example, should be able to address storage space on the cloud SAN(s) without having to manage implementation details. Ideally, the same management console that is used to control servers in the cloud will support management and administration of storage resources.

Computing and storage clouds hide many of the implementation details that go into building and maintaining a large IT infrastructure. By standardizing services, streamlining service management, and virtualizing physical resources, cloud providers enable the technical resources needed by users to leverage cloud services. Those same users, however, also require attention to business considerations.

## Committing to SLAs

Business managers may look at cloud services and find the lower costs, greater control, and potential for scaling business processes compelling reasons to use cloud services. These reasons are often not enough, though. It is not sufficient for a cloud to work well today; it needs to work well for as long as users need it. This is why we have SLAs. SLAs are standard in IT, and it is no surprise that they are used with cloud services. Rather than focus just on the availability of a specific application, cloud SLAs may be more general and apply to capacity commitments, network infrastructure, storage infrastructure, and availability and recovery management. These SLAs are closely coupled to the infrastructure of the cloud, but the primary concern is on the business commitments cloud providers make to their customers.

### Capacity Commitments

A capacity commitment in an SLA outlines the number and types of server capacity that will be available for use when the cloud consumer attempts to use them. Several factors should be considered when making capacity commitments:

- The total infrastructure planned for a private cloud
- The ability to acquire additional resources (compute and storage) as needed through a hybrid cloud
- Changes in pricing models if hybrid resources are used
- A commitment to the percent of time the capacity will be available
- Length of time the capacity will be available without interruption once the capacity is provisioned

The workload analysis performed earlier in the planning process can help to understand the capacity commitments a cloud provider can make given a particular number of servers and storage capacity.

### Network Infrastructure

Network service commitments are especially important when there are high levels of data exchange in and out of the cloud. Service commitments will be limited by the network capacity of Internet service providers (ISPs) and the ability to distribute networking load across multiple ISPs. Cloud service providers are limited by the service level commitment they receive from their ISPs; however, by combining network services from multiple providers, a cloud provider can improve total throughput and availability.

### Storage Infrastructure

Storage SLAs take into account several factors:

- Amount of storage available for use
- Backup services, if any
- Availability commitments, including percent of time storage services will be available
- Throughput commitments

When considering the amount of storage available for use, take into account the need for redundant storage to improve performance and availability. These can significantly reduce the total amount of storage available for direct use by cloud consumers.

### Availability and Recovery Management

Another popular topic for SLAs is recovery management. The redundancy of servers in the cloud ensures that the failure of a single server in the cloud will not disrupt an operation. The service can be started again on another server. From a service level perspective, cloud providers may be able to commit to high levels of availability in terms of having servers available to run applications. One must account for the fact, though, that when a server fails and another is started in its place, there may be data loss depending on how the application is written. If the application writes state information to cloud storage, another instance of the application can recover from the last point at which state information was written to the disk. If the application depends on maintaining state information in memory, the recovery point would be earlier. A final set of issues that falls under the penumbra of business drivers is compliance requirements.

## Compliance Requirements and Cloud Services

Compliance requirements tend to focus on preserving the integrity of data, especially financial data, and protecting the privacy of confidential information. One of the greatest impediments to adopting public cloud computing is concern about protecting the integrity and confidentiality of data once it leaves the corporate-controlled network. Private clouds retain data within corporate firewalls where it will be subject to internal controls. The assumption behind this reasoning is that governance procedures that protect data in non-cloud infrastructure are sufficient to protect the same data in the cloud. This may be true for the most part, but the cloud introduces additional factors that should be considered:

- Applications running in a virtual machine might write data to local disks. When the virtual machine shuts down, all data written by it should be overwritten.
- Authorizations assigned to users for non-cloud resources should be respected in the cloud. For example, if data moves from a dedicated file server to cloud storage, the same restrictions on access should apply.
- Practices employed as part of compliance efforts, such as routine vulnerability scanning, will have to be adapted to scan machine images in the service catalog rather than just instances running at a particular point in time on a given set of servers

Reporting is another essential part of compliance. It is not sufficient to be in compliance; one must often be able to demonstrate one is in compliance. Again, existing procedures might need to be modified to accommodate reporting on cloud procedures that support compliance. For example, each time a virtual machine instance is shut down, a record may be logged indicating local data has been overwritten to prevent the next user from scanning local storage for residual data.

## Summary

Planning for cloud services is a multifaceted process that begins with assessing readiness for the cloud and aligning business strategy with cloud computing services. It also requires preparation for managing cloud services and planning for centralized resources. In addition, it entails a number of business-oriented concerns, such as SLAs and support for compliance efforts. To facilitate the planning process, a pre-implementation checklist is provided that summarizes the key points of this chapter.

| **Pre-Implementation Checklist** | |
|---|---|
| **Assessing Readiness for Cloud Computing** | Determine whether applications are designed to use a Web application architecture, service bus architecture, or n-tier architecture |
| | Assess ability to provide for self-service management of computing and storage services |
| | Standardize on platforms and application stacks |
| **Aligning Business Strategy with Cloud Computing Services** | Analyze workloads |
| | Determine value metrics with respect to labor, hardware, and software |
| **Preparing to Manage Cloud Service** | Understand the roles of private, public, and hybrid clouds and their utility for business requirements |
| | Plan for growth in demands for services |
| | Assess long-term management issues |
| **Committing to SLAs** | Perform capacity planning with respect to service level commitments |
| | Analyze capacity of network infrastructure |
| | Analyze capacity of storage infrastructure |
| | Formulate reasonable commitments with respect to availability and recovery management |
| **Meeting Compliance Requirements** | Determine security requirements for preserving the integrity and confidentiality of data |
| | Adapt reporting requirements to address compliance implementation issues introduced by the cloud |

# Chapter 8: Roadmap to Cloud Computing: The Implementation Phase

One of the most challenging IT tasks is to implement a new systems architecture. By definition, we are introducing a new way of delivering services; at the same time, we are often required to maintain existing services. It is analogous to repairing your car while driving it. The first step in the cloud adoption process is to develop a comprehensive plan that begins with assessing readiness for cloud computing, aligning business processes with cloud services, planning for centralized resources, and committing to service level agreements (SLAs). We described this first step in detail in the previous chapter; in this chapter, we shift focus from planning onto the actual implementation of the plan.

Many planning issues are common to both public and private clouds, but the implementation details are more complex in the case of private cloud computing. This chapter will address how to implement a private cloud and will include discussion of hybrid and public cloud issues as well. The structure of the discussion is divided into five core subtopics:

- Establishing a private cloud
- Transitioning compute and storage services to a cloud
- Completing a post-implementation checklist
- Managing cloud services
- Extending a private cloud with public services

By the end of the chapter, we will have outlined some of the fundamental issues that should be considered during the implementation phase in order to begin deploying cloud services within an organization.

## Establishing a Private Cloud

A private cloud begins with the deployment of hardware, networking, and software services. Throughout this book, we have often discussed the business services, software architecture issues, and other logical design considerations. All of those logical choices ultimately depend on lower-level services that in turn rely on an IT infrastructure that includes:

- Private cloud hardware
- Networking
- Application stacks

Deploying a cloud begins down in the infrastructure.

### Deploying Hardware for a Private Cloud

Many of the hardware issues we have to address in a private cloud are familiar to those with data center experience. They tend to cluster around

- Server-level issues, such as the number of servers and amount of network equipment and how they are deployed and configured
- Environmental concerns, such as space, power, and cooling
- Redundancy to prevent single points of failure

#### Servers and Network Equipment

Servers in a private cloud are housed in one or more data centers. There must be adequate space within the data centers for the server units. The number of servers in a cloud can grow incrementally quite easily but the physical space for housing them may not. Data centers should be sized according to initial space requirements as well as for foreseeable growth.

Servers are often rack-mounted in industry-standard 19-inch rack cabinets. These are typically configured to allow easy access to both the front and back of the cabinets. Cabling is run through racks to improve cable management; space required for an organized cable distribution system must also be taken into account when sizing the data center. Distances between components should be minimized in order to minimize cable lengths, but more importantly, the data center equipment should be organized in a logical fashion to support maintainability.

**Data Center Standards**

Standards for configuring data centers have been established by the Telecommunications Industry Association (TIA). For more guidance on configuring a data center, see the TIA-942 Data Center Standards Overview by ADC.

### Environmental Issues

Servers and networking equipment depend on environmental infrastructure to keep functioning, especially:

- Power
- Cooling
- Fire prevention
- Physical security

External power generators will typically supply electrical power to a data center. Key considerations are reliability and adequate supply of power. To prevent a single point of failure in the power supply system, a backup power system can be used. Uninterruptable power supplies can use batteries to supply power immediately in the case of a power failure while diesel generators are started. The generators are designed to supply power for longer periods of time.

Cooling is another factor that must be taken into account when designing a data center for a private cloud. Servers and other electrical equipment dissipate heat into the environment and the temperature in a data center will rise unless the center is cooled. Humidity control is also a concern because too much moisture in the air can result in condensation on electrical equipment. Air conditioning is the common method for cooling but alternatives, such as using outside air, are in use as well.

> **Tips on Energy Efficiency for Data Centers**
>
> See The Quick Start Guide to Increase Data Center Energy Efficiency by US General Services Administration and the US Department of Energy for tips on reducing the costs and environmental impact of operating a data center.

Fire prevention equipment includes active controls such as smoke detectors, sprinkler systems, and fire suppression gaseous systems. Passive controls, such as firewalls, can also be used to contain fires to one part of the data center.

The physical integrity of the data center must be protected with access controls to prevent unauthorized access. Guards, access control badges, and surveillance cameras are all used to protect data centers.

### Redundancy and Avoiding Single Points of Failure

Redundancy is found at multiple levels in a data center, from dual power supplies in air conditioning units all the way up to duplicate data centers. At the lowest level, redundancy is built-in to the components we deploy as single components, such as servers, air conditioners, and disk arrays. At mid-levels, we incorporate redundant components or backup systems in a data center. A second air conditioning unit is an example of the former; an uninterruptable power supply is an example of the latter.

At the top level, we duplicate entire data centers. This is obviously a costly option but has a number of advantages. Multiple data centers with similar infrastructures can act as backups for each other. If one data center is hit with a natural disaster, the other data centers can carry the workload of the downed data center. This kind of disaster recovery configuration requires a well-defined plan before the disaster. For example, data needs to be replicated between data centers in a timely manner.

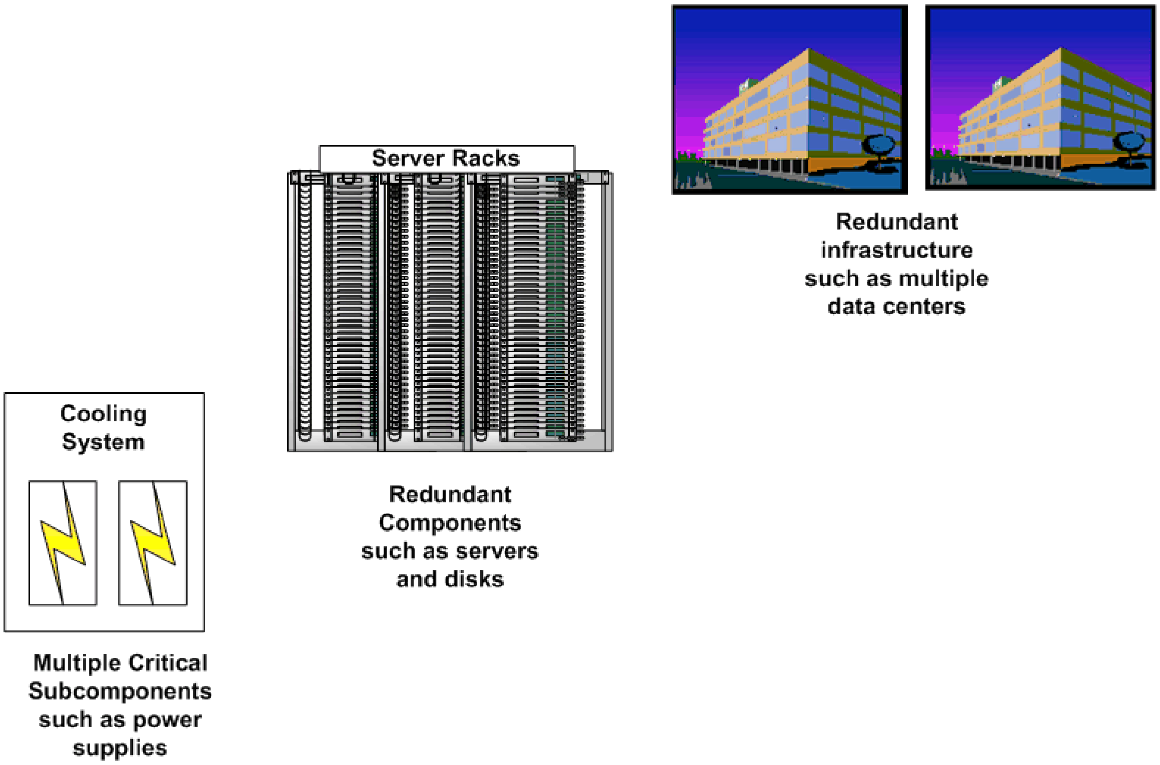

Server Racks
Redundant infrastructure such as multiple data centers
Cooling System
Redundant Components such as servers and disks
Multiple Critical Subcomponents such as power supplies

**Figure 8.1: Redundancy is used at multiple levels to avoid single points of failure that could shut down a single component or an entire business process.**

We may do this any way to ensure high availability even without regard for disaster recovery situations. For example, if a disk array fails in one data center or network traffic to that data center is unusually high, other data centers with the replicated data can respond to service requests for that data.

It should be noted that this process is not the same as backups. Backups are copies of data at a point in time and preserved from some period of time. Data replication copies data and overwrites existing data in some cases. If an application error corrupts a database in one data center, that database will eventually be replicated to other data centers unless the problem is discovered in time. A backup would allow the business to recover from the data corruption; replication may not.

In addition to compute and storage infrastructure, we need to deploy sufficient networking resources to meet the demand generated by cloud computing.

## Deploying Network Services for a Private Cloud

Business services delivered through the cloud will determine network bandwidth, latency, and reliability requirements. The network architecture selected for a private cloud will determine how those requirements are met. As with compute and storage hardware, redundant components such as routers and switches are important for avoiding a single point of failure. They also contribute to high availability by enabling load balancing across network devices.

Even with redundant devices on the corporate network, we still face a risk of losing network services on the internetwork between data centers and other corporate offices. Providing redundant links over the wide area network (WAN) is an obvious solution but there is a significant drawback: cost.

Consider a private cloud that uses two data centers and supports WAN connections between the data centers and for corporate offices. Figure 8.2 depicts a fully redundant WAN.

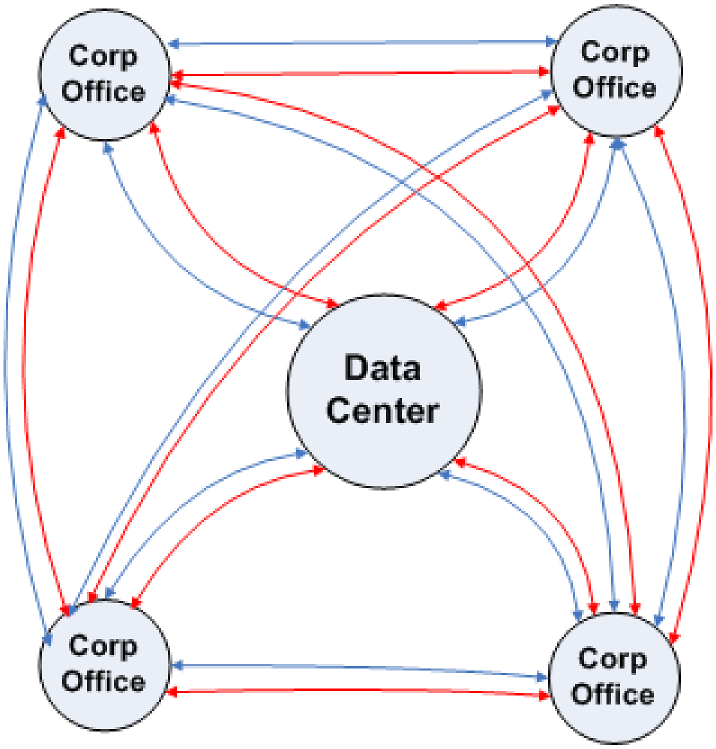


**Figure 8.2. A fully redundant network requires two or more links between each interlinked network.**

In this simple example of one data center and four corporate offices (five endpoints), there are a total of 20 WAN links. If we increase the number of data centers to two and add four more corporate offices (10 endpoints), we would need a total of 90 links. The number of links in a fully redundant network grows according to the formula: n(n -1) where n is the number of endpoints. This architecture can become cost prohibitive quite quickly.

An alternative approach is to use a mesh design in which each endpoint in the WAN has links to two or more other endpoints. If any single link fails, the endpoints can communicate using the other WAN link. Figure 8.3 shows an example of a mesh network that provides multiple routes between any two endpoints. Note, that Figure 8.3 depicts a network with 10 endpoints but uses only 18 WAN links.

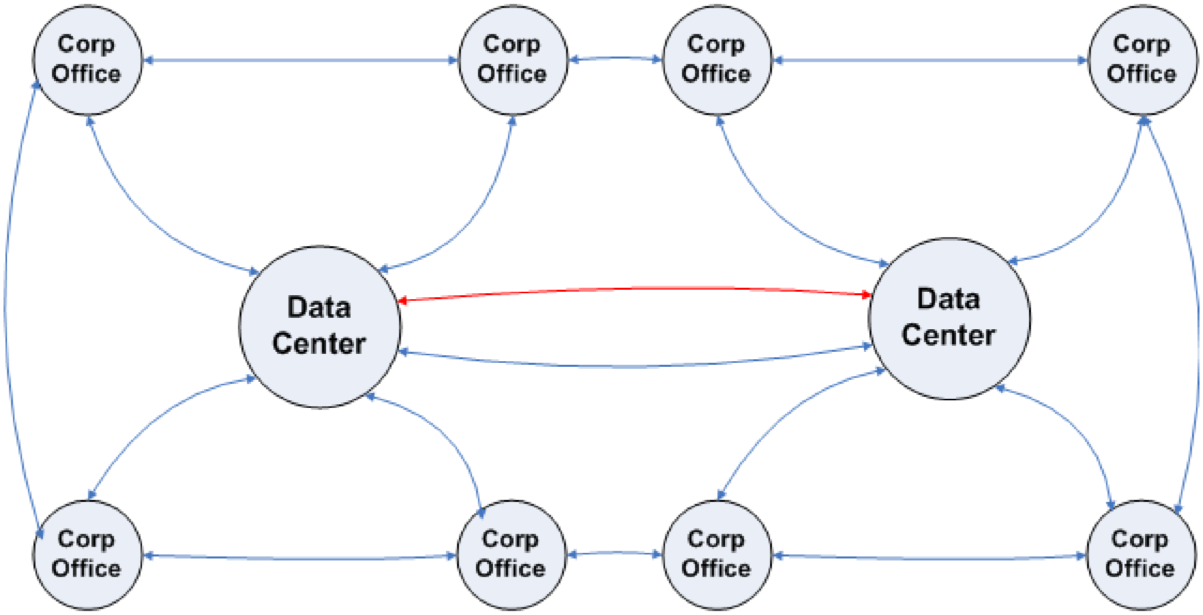


**Figure 8.3: A mesh network architecture provides redundancy with fewer links than a fully redundant design.**

## Providing Application Stacks

In addition to deploying hardware and networking services, we need to provide for and manage application stacks within a private cloud. This requires support for at least three elements: cloud management services, management policies, and management reporting.

### Cloud Management Services

Cloud management services can be thought of as another layer in the software application stack. We have applications that run inside application servers that run inside an operating system (OS), and OSs that run as virtual machines within hypervisors. This layered approach continues in the cloud with cloud management software that carries out basic cloud operations:

- Starting and stopping virtual machine instances
- Providing access to network storage systems from virtual machines running in the cloud
- Managing cloud storage services
- Tracking usage information for accounting and billing

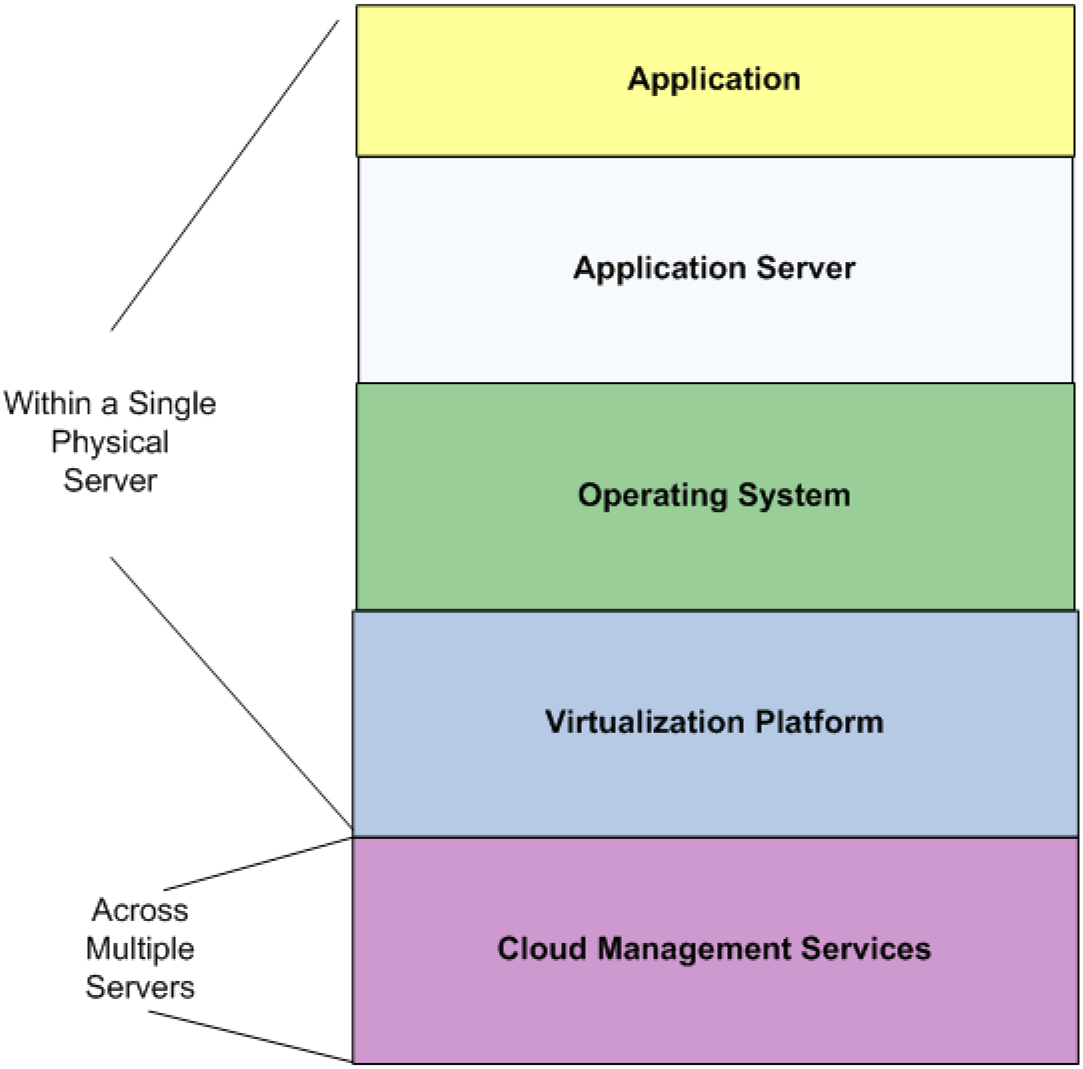
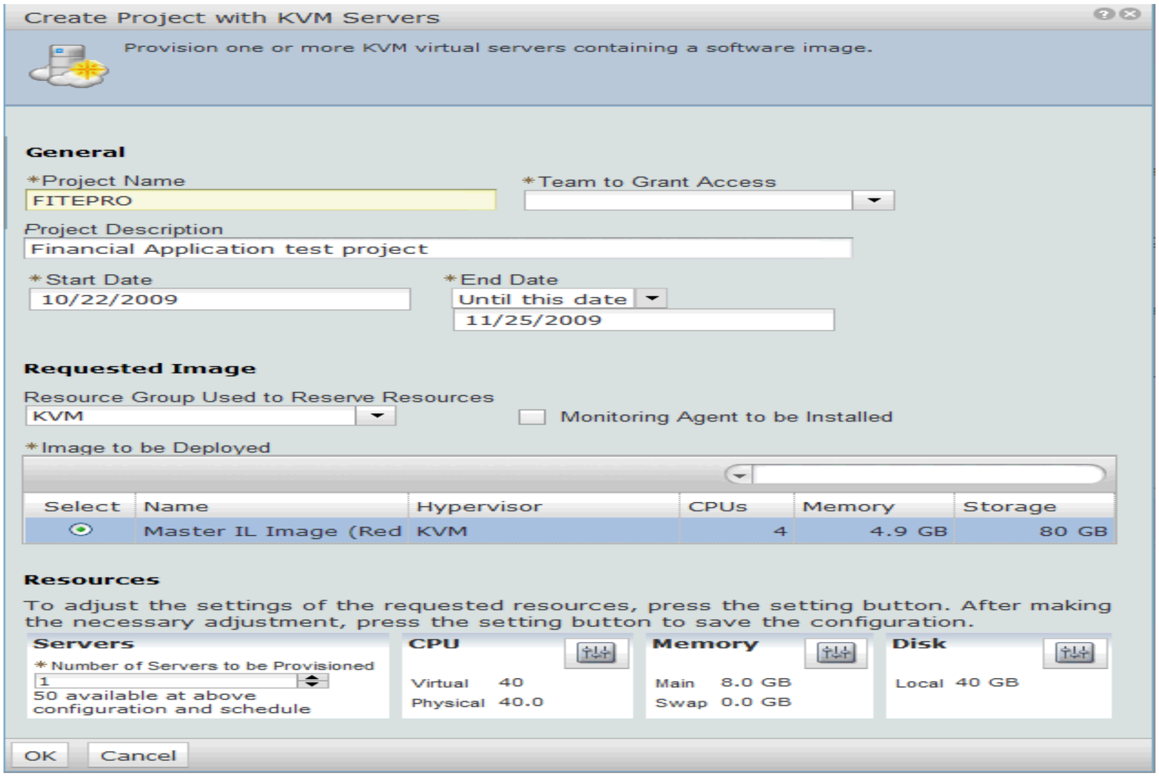


**Figure 8.4: The conventional application stack is extended in the cloud to include cloud management services below virtualization services.**

Cloud management services must accommodate several types of needs:

- Clustering groups of servers to support high-performance computing needs for tight coupling of applications running on different servers
- A service catalog, which is a repository of virtual machine images that may be run in the cloud
- Access controls on cloud services, such as the ability to start and shut down instances or add images to the service catalog
- Storage abstractions for persistent storage after virtual machine instances are shut down



Create Project with KVM Servers

Provision one or more KVM virtual servers containing a software image.

**General**

*Project Name: FITEPRO

*Team to Grant Access:

Project Description: Financial Application test project

*Start Date: 10/22/2009

*End Date: Until this date 11/25/2009

**Requested Image**

Resource Group Used to Reserve Resources: KVM

□ Monitoring Agent to be Installed

*Image to be Deployed

| Select | Name | Hypervisor | CPUs | Memory | Storage |
|---|---|---|---|---|---|
| ● | Master IL Image (Red | KVM | 4 | 4.9 GB | 80 GB |

**Resources**

To adjust the settings of the requested resources, press the setting button. After making the necessary adjustment, press the setting button to save the configuration.

**Servers**

*Number of Servers to be Provisioned: 1

50 available at above configuration and schedule

**CPU**

Virtual 40
Physical 40.0

**Memory**

Main 8.0 GB
Swap 0.0 GB

**Disk**

Local 40 GB

**Figure 8.5: Cloud management services include applications to allow users to provision their own virtual machines as needed without assistance from IT support personnel.**

## Cloud Management Policies

Cloud management policies specify how cloud resources are governed. Computing cloud architectures evolved from earlier IT architectures, so there are not necessarily new types of polices; instead, we have extensions to existing policies (for the most part). At minimum, a private cloud should assess current policies and make modifications as needed to accommodate:

- Privileges and limits on the number, types, and durations of use of virtual machines a single project can provision
- Access control policies with regard to provisioning virtual machines and storage allocations
- Backup services
- Limits on SLAs and the cost of different SLAs
- Data retention and data destruction policies

Policies are in place to ensure cloud service consumers can plan their use of the cloud according to enterprise-wide constraints. Policies also serve cloud providers who need to maintain compliance with internal requirements and SLAs as well as external regulations.

## Cloud Management Reporting

A system of reporting on cloud operations must also be in place early in the deployment phase. Cloud service providers will need management reports that describe key performance indicators of the cloud:

- Server utilization
- Storage utilization
- Network bandwidth and latency
- Security incident reports
- Service support tickets
- Service catalog inventory and summary descriptions

Ideally, these reports are available for aggregate measure across the enterprise as well as by important dimensions, such as time, department or line of business, data center, user location, and so on.

Cloud service consumers will also look for management reports but with an emphasis on managing their own use of the cloud. Typical reports in this category include:

- Number and type of servers used and the duration of each use by job or project
- Amount of storage allocated by job or project
- CPU utilization rates
- Images and software used, especially if charge backs are applied for software licenses
- Summary reports on jobs scheduled and time required to complete jobs and total cost by job

Cloud management reports should help cloud providers more efficiently deliver cloud services as well as help cloud consumers more efficiently support their business services and workflows.

Establishing a private cloud is a multistep process. Hardware must be deployed with consideration for physical infrastructure, such as power, cooling, and physical security, as well as architectural issues, such as redundancy and failover. Network services are essential to delivering cloud services. As the number of data centers and remote sites grows, the cost of point-to-point dedicated networks quickly becomes prohibitive. Networks will have to be designed with enough redundancy to provide robust networking but not so much that the costs outweigh the benefits. Application stacks must also be deployed with particular attention to cloud management services, management policies, and management reporting.

## Migrating Compute and Storage Services to a Private Cloud

So far in this chapter we have discussed aspects of deploying hardware, network services, and applications in a private cloud. We now turn our attention to a more detailed look at the sequence of events that are needed to establish such deployments. There are several steps in the transition to a cloud infrastructure:

- Prioritizing steps based on business drivers
- Reallocating servers
- Deploying cloud-enabling applications
- Testing and ensuring quality control
- Deploying management applications
- Migrating end user applications

This list is roughly the order in which the steps are executed during the migration.

### Prioritizing Based on Business Drivers

Before we start redeploying servers and moving applications off their current host servers, we need to formulate a plan. That plan should be shaped by the business drivers that motivated the move to a cloud architecture in the first place. There are several types of business drivers, and they should all be considered when formulating the plan.

#### Business Driver #1: Cost

Clouds can deliver services more efficiently than can dedicated servers in many cases. (We described the reasons for this in detail throughout this book and will not repeat them here.) A typical example of a lower-cost cloud-based delivery is when a single server is dedicated to an application that uses only a fraction of the computing resources of the server. Multi-core processors running on servers with significant amounts of memory can support compute-intensive operations, but many business operations never fully utilize the capabilities of servers.

Servers dedicated to file transfer, collaboration, and content management, for example, typically make little demand on server resources. Utilization can improve if the server uses virtualization to run multiple guest OSs with different services, but even this may not fully utilize the server's capabilities. Four lightweight services running on a high-end server are better than one service but can still leave CPU cycles wasted.

In a cloud, this problem is mitigated by adding virtual machines to servers as long as there are resources available to support another instance. In the case of a server running four OS instances but still has CPU cycles available, another instance can be added by the cloud management software. Of course, one could add another instance to a virtualized server without cloud management software but doing so would require an IT support person, which would drive up the cost.

### Business Driver #2: Computing Resources

Another major driver for utilizing a cloud is the ability to provision computing resources on demand. If a data warehouse must perform complex extraction, transformation, and load (ETL) operations every night, a cloud is an ideal way to do so. Source systems can send their input data streams to multiple servers, which perform record-level transformations and data quality control checks. These servers can then pipe their output to another set of servers that receive data based on some criteria, such as geographic location. The secondary set of servers aggregate data by region, and they, in turn, pipe their output to another server for the level of data aggregation.

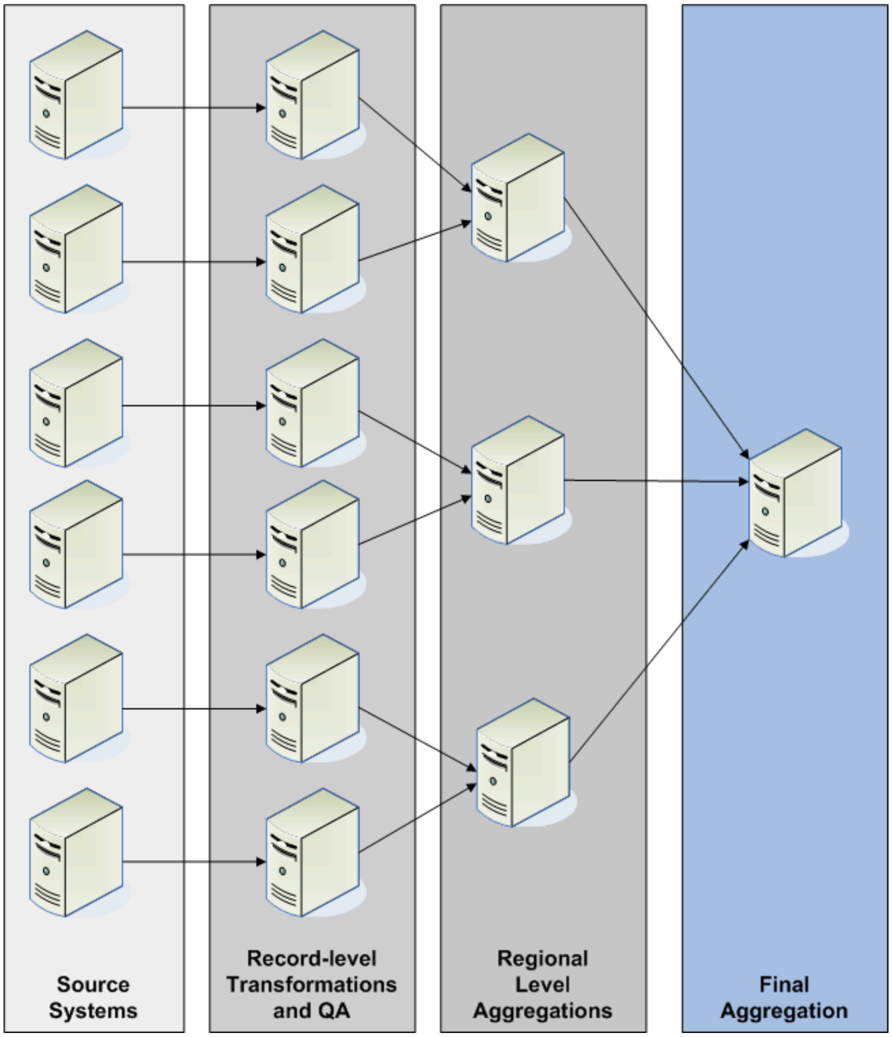

Source Systems
Record-level Transformations and QA
Regional Level Aggregations
Final Aggregation

**Figure 8.6: Many business processes, such as data warehouse ETL operations, can make use of multiple servers for relatively short periods of time.**

During the prioritization step, we should itemize both business processes and servers and determine (1) how the business process would benefit from flexible server allocation and (2) the relative utilization of the server. Business processes that use a server at fairly constant levels, such as those dedicated to transaction processing on a continuous stream of input, are less likely to benefit from flexible allocation. Business processes that experience high variability in resource demand are good candidates for early migration to the cloud. Servers that run at near capacity would be only marginally more productive in a cloud configuration, but those that are underutilized could be better utilized in the cloud.

### Reallocating Servers

Reallocating servers is not as simple as it may sound at first. Even once the order of redeployment is determined based on business drivers, we need to ensure that services that are currently provided by servers continue to be available as needed. For example, we might determine that several dozen servers hosting Web servers, small databases, collaboration servers, and several other department-level services will all be assigned to the cloud. To do so, we need to:

- Migrate applications to other servers, perhaps in the cloud if some are already available or to virtual hosts on servers dedicated to the migration process
- Back up data from the current production servers and restore to the transitory server hosting the application
- Delete data and applications from the server and install virtualization a platform and any cloud-specific applications
- Physically connect the server to the cloud network segments and attach the server to network storage

If the applications running on the servers prior to reallocation will be running in the cloud, virtual machine images must be added to the service catalog to support those applications.

### Deploying Cloud-Enabling Applications

After servers are physically allocated to the cloud and configured to use cloud networking services and cloud storage, the next step is to configure software for the servers. The servers will run virtual machine hypervisors and integrate with cloud-level management software for deploying virtual machine images. Depending on the type of cloud management software, servers might run different hypervisors, such as VMware products, Xen, or KVM.

## Testing and Quality Control

Testing is an essential part of cloud deployment. At this point, servers are allocated, cloud storage is in place, and necessary controllers are deployed. The goal of this step is to test and exercise the cloud configuration before opening it for production work. The test plan should include several steps that ensure:

- Virtual machine hypervisors are installed and running correctly on all servers
- Virtual machine instances can be started and stopped as expected
- Cloud management software correctly starts specified machine images on the correct number of servers
- All servers can read and write from cloud storage
- LDAP or other directory services are in place and function correctly on all servers
- Security policies are implemented correctly; for example, all data on local storage is deleted when a virtual machine instance is shut down

After testing these individual elements of cloud functionality, we can move on to performance testing. This type of testing should be driven by the SLAs we expect to support. When it comes to performance, more is always better, at least in theory; however, there are costs associated with marginal improvements in performance. During performance testing, we want to verify that:

- Virtual machine instances start and are available for use in an acceptable amount of time
- Read and write operations to cloud storage are performing as expected
- Large numbers of parallel operations, such as starting instances or writing to storage, are performed in an acceptable amounts of time
- Network latency and bandwidth are sufficient to meet SLAs

During testing, we also want to ensure that usage and accounting information is tracked correctly.

## Deploying Management Applications

As noted earlier, management applications are needed for both cloud providers and cloud consumers. These may both be hosted on cloud controller infrastructure, such as servers dedicated to collecting usage data and generating reports and data services. At this point, we also need to implement policies and procedures for basic operations, such as startup and shutdown of virtual machine instances, recording usage information for accounting purposes, monitoring server and network utilization, and ensuring supporting operations, such as replicating data between data centers, is functioning as expected. When the cloud infrastructure is in place and functioning properly, the next step is to migrate end user applications to the cloud.

## Migrating End User Applications

Migrating end user applications is a three-step process:

- Building virtual machine images with necessary application stacks
- Migrating data to cloud storage
- Migrating access control privileges and directory information to the cloud.

### Building Virtual Machine Images

Building virtual machine images is a straightforward task, but we must be careful to analyze application dependencies to ensure all necessary supporting software is in place. Also, different configurations of an application may require different versions of supporting libraries, so we may need to support several versions of similar images. Applications may have different configurations depending on how the application is used, and this could also warrant having multiple versions. For example, a Java application server may be configured differently if we expect heavy, moderate, or light use. Rather than expect the user to adjust configurations each time a virtual machine instance is created, we could store different versions so that the user can choose the appropriate one as needed.

### Migrating Data to Cloud Storage

Migrating data to the cloud is another process that sounds simple but has some potential challenges. There are different ways of storing data in the cloud. One option is to use block storage in which data is written to logical blocks on cloud storage; another option is to use a relational database management system (RDBMS) to manage data in the cloud. The second option has similar functionality to RDBMSs that run on dedicated servers but without having to manage some of the lower-level storage issues, such as tablespace file placement. Some changes may be required in applications to make use of cloud block storage, so we should review an application storage scheme before migrating it to the cloud.

### Migrating Access Privileges to the Cloud

Applications that run on dedicated servers often make use of LDAP directories or Active Directory (AD) to store and serve information about users, resources, and privileges. This information has to be migrated to the cloud infrastructure and adjusted as needed in the cloud.

Adjustments range from mapping access controls to specific servers and directories (for example, user AJones has read and write privilege to \\server1\directoryA) to the comparable location in the cloud storage. Additional data may also be required, such as limits on the number of virtual instances a user may start at any one time, the maximum time those servers can run, accounting information for charge backs, and so on.

Transitioning compute and storage services to the cloud is a multistep process that begins with prioritizing services to migrate to the cloud based on business drivers and moves through reallocating servers, deploying cloud-enabling applications, testing and quality control, deploying management applications, and finally migrating end users applications. There are many steps to the process; the following post-implementation checklist summarizes the key steps.

## Post-Implementation Checklist

| | **Topic Area** | **Notes** |
|---|---|---|
| **Deploying Hardware for Private Cloud** | | |
| | Servers and network equipment | Establish data center infrastructure |
| | Environmental issues | Power, cooling, physical security, fire suppression |
| | Avoiding single points of failure | Is redundancy used for critical components, systems, and data centers? |
| **Deploying Network Services for Private Cloud** | | |
| | Network capacity | Is network bandwidth and latency sufficient for SLA? |
| | Redundancy | Are redundant routes implemented in a cost-effective manner? |
| **Deploying Application Stacks for Private Cloud** | | |
| | Cloud management services | Provisioning virtual machines and storage |
| | Policies | Privileges, access controls, backups, data retention policies |
| | Management reporting | Cloud provider and cloud consumer reporting |
| **Prioritizing Based on Business Drivers** | | |
| | Cost drivers | Which servers can be most efficiently redeployed in the cloud? |
| | Compute drivers | Which business services require significant computing resources? Which processes need regular peak capacity significantly in excess of more common workloads? |

| | | |
|---|---|---|
| **Reallocating Servers** | | |
| | Migrating applications | Plan migration and switch over |
| | Backup data | Consider how to synchronize data if existing application continues to run during migration |
| | Initialize servers for cloud | Wipe existing data on server, physically move servers to data center |
| | Physically connect servers to cloud infrastructure | Establish connections to other servers, storage, and network |
| **Deploying Cloud-Enabling Applications** | | |
| | Deploying hypervisors | Install low-level software for OS and virtual machine functions |
| | Server-specific monitoring applications | Enable server monitoring services |
| **Testing and Quality Control** | | |
| | Server-based functional testing | Does the server function as expected with regard to starting and stopping virtual machine instances? Writing to and reading from cloud storage? Use network services? |
| | Performance testing | Do servers function as expected under significant loads? Test for both compute and I/O loads |
| **Migrating End User Applications** | | |
| | Building virtual machine images | Build service catalog with images as needed to meet the full range of application requirements |
| | Migrating data to cloud storage | Copy application data to cloud and verify applications function properly with regard to cloud storage |
| | Migrating access control information | Update LDAP or other services in the cloud that store authentication and authorization data |

## Managing Cloud Services

After the transition period when infrastructure is migrated to a cloud configuration, our attention shifts to more operational and maintenance-oriented considerations:

- Service management integration with the cloud
- Usage tracking and accounting services
- Capacity planning

These are business operations that likely existed well before cloud computing was introduced, so it is usually a matter of extending these business processes to function with the cloud.

### Integrating Service Management with the Cloud

Service management is a set of practices that orient IT operations around customers' needs and business processes rather than around technology. Throughout this book, we have had a decidedly technology-centric focus, but that should not be construed as meaning cloud computing cannot be customer focused. Actually, by streamlining the delivery of computing and storage services, cloud computing actually improves customer service and supports the objectives of service management.

There are different ways of implementing service management. One of the most formal and well-known approaches is the IT Infrastructure Library (ITIL), which advocates a broad and fairly structured approach to service management. There are many elements in the ITIL framework and service management in general, but we will only consider:

- Service catalog management
- Service level management
- Availability management
- Service validation and release management

There are other aspects of service management that are relevant to cloud computing but are outside the scope of this chapter; these include risk management, financial management, and supplier management.

> **ITIL v3**
>
> For more information about the ITIL framework and other service management issues, see http://www.itil-officialsite.com/home/home.asp.

### Service Catalog Management

Service catalogs are sets of business and support services available from IT departments. Before we go any further, it should be noted that the term "service catalog" has two similar meanings, and it is important to distinguish them here. A service catalog in the service management sense is an abstract description of the set of services available from information technology providers. We also use the term "service catalog" to describe a repository of virtual machine images that are available for use in the cloud. In this section, we will always refer to the latter as the "service catalog repository" to avoid confusion.

Business services are made available through the cloud when they are added to the cloud's service catalog repository. We have discussed the service catalog repository from a technology perspective with topics such as ensuring software dependencies are accommodated in images, images are maintained as part of patch and vulnerability management, and so on. In terms of service management, we should think of virtual machine images as vehicles for delivering service. This perspective requires us to think more in terms of the following:

- Are the services that cloud consumers expect available in the catalog?
- Is meta data associated with virtual machine images sufficient for users to find the services they need and to distinguish among similar images?
- Are software license restrictions properly accounted for in the way virtual machine images are made available?

Other business services are not necessarily tied to virtual machine images run in the cloud. Support services, such as ticketing systems for incident and problem management, are part of the service catalog in the management sense of the term.

### Service Level Management

Service level management is the practice of managing commitments to cloud users. These commitments are usually documented in SLAs. Requirements are defined in SLAs, and Quality of Service (QoS) metrics are usually associated with these requirements. In the cloud, SLAs may include requirements around:

- Number and type of virtual machine instances that will be available at regular times and for some length of time
- The duration from requesting a set of virtual servers to the time they are available
- Percentage of time other requirements, such as guaranteed number of servers, will be met
- Availability of software packages in the service catalog repository

The details of SLA metrics will be slightly different with a cloud, but the framework is essentially the same to that which we use in non-cloud environments.

### Availability Management

Availability management is the process of ensuring compute and storage resources are available as needed to meet SLAs. One of the advantages of cloud computing is that it eases availability management.

In an environment with servers dedicated to particular tasks, we often use replication to keep standby servers ready to take over in case of a failure. In a cloud, servers do not have identities and the software they run is a function of the virtual machine image loaded on to them by an end user. Failure of a single server or even 10 servers in a cloud can be managed by instantiating the images that were running on the failed servers on other cloud servers. Assuming data on the failed servers is persisted in cloud storage, the new instances of the applications will have access to data.

### Service Validation and Release Management

Service validation and release management are procedures for testing and deploying new services to the cloud. As with availability management, this task is easier in the cloud than in a dedicated service environment. Designing, testing, and validating applications in the cloud is similar to designing, testing, and validating in a dedicated server environment. The advantages stem from the fact that a new release can be deployed as another virtual machine image in the service catalog repository. If there is a problem with the new release, the old version is easily run without the challenges of reinstalling software on a dedicated server.

Service management is a business practice used to control the delivery of IT services. Cloud computing does not eliminate the need for this kind of management but does require adaptations and, in some cases, makes it easier to execute these management operations.

### Usage Tracking and Accounting Services

There is an old saying that if you cannot measure it, you cannot manage it. This is especially true in the cloud. With large numbers of users running a wide array of applications across a large number of servers, one will need an efficient method for tracking use. The ideal tracking system will:

- Function seamlessly as part of the instantiation process when virtual machines are started or when storage is allocated
- Collect and maintain fine-grained detail about use; for example, at the user and image level
- Allow project or department-level charging
- Feed data directly into financial reporting systems

Adapting current charge back systems may require some work to allow for automated transactions indicating when instances are started or storage is allocated. These operations are largely self-service steps in the cloud (whereas they are not in dedicated server environments).

### Capacity Planning

Capacity planning is yet another service management process that is familiar to many IT professionals. The principles are the same with cloud architectures, but once again, this process is just a bit less challenging in a cloud environment. Forecasting growth with dedicated servers often requires planning for peak capacity in multiple applications, departments, and business units. In the cloud, we can manage to aggregate trends. We can ask questions—such as how many physical servers will be needed to support all SLAs—rather than asking how many servers will be needed to support Department A, Service B, and so on.

We manage cloud services much as we manage any service provided by IT. Service management practices, usage tracking and accounting, and capacity planning are all well-established practices. They will continue to be needed when managing a cloud but, fortunately, with little bit less difficulty.

## Extending a Private Cloud with Public Services

As flexible as a private cloud is, there are limits. At some point, the costs of adding more servers or storage to a private cloud will outweigh the benefits. Public cloud providers can realize economies of scale that are not available to most private cloud providers. Of course, private clouds continue to have their benefits, such as the ability to control the infrastructure on which private and confidential data resides. Businesses may find that the optimal solution is to combine private and public clouds to realize the benefits of both.

In cases where additional compute and storage resource are provided by public cloud providers, it is imperative that security controls are in place to protect information that leaves the organization. For example, you might need to encrypt data as it is transmitted to public cloud servers, and store it in an encrypted form on cloud storage. Also, you might need to set a policy that no data is written to local storage of a virtual machine running in the private cloud to prevent any possibility of a later user of that device having the ability to restore data that previously resided on the disk.

Policies should be in place that define the acceptable use cases of public cloud services, including the types of data that can be sent to private cloud servers and the types of applications that can be run in the private cloud. A proprietary process or analysis procedure that instantiates significant intellectual property, for example, is a good candidate for keeping out of public cloud services. Hybrid clouds that combine the benefits of private and public clouds can improve the efficiency, cost effectiveness, and capabilities of a private cloud, but hybrid clouds must be used in a way that does not violate policies or the interests of the business.

## Summary

Establishing a private cloud is a multistep process. Hardware must be procured or re-assigned, network services provisioned, and software configured for use in the cloud. Transitioning services to the cloud requires that we carefully plan other steps, including prioritizing based on business drivers, deploying applications, implementing quality controls, and deploying management applications. Many existing IT processes, such as service management and capacity planning, can be readily adapted to the cloud. Finally, it may be beneficial to consider the use of a hybrid cloud to take advantage of the economies of scale of public clouds while maintaining the control advantages of a private cloud.

# 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.

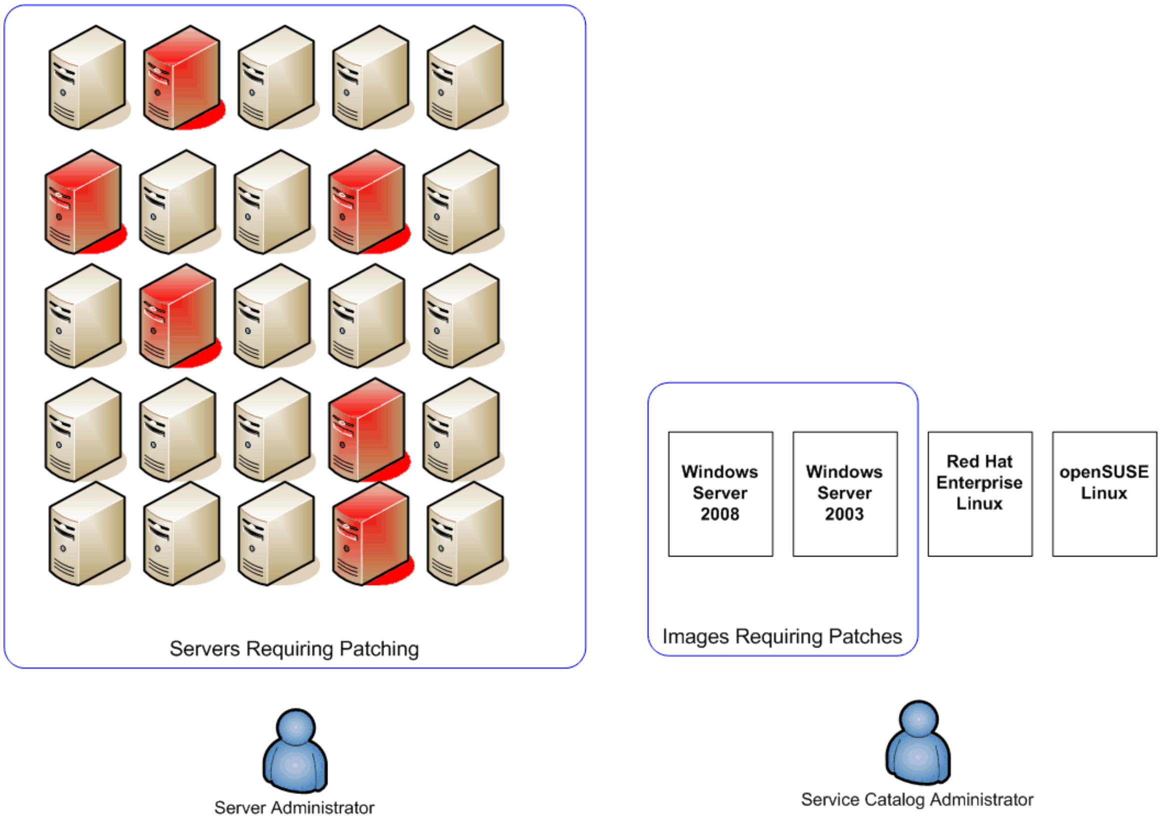


**Figure 9.1: The need to apply security patches is the same with or without a cloud; however, the execution can be less problematic when working with a service catalog rather than individual servers where the patch may fail for different reasons.**

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.

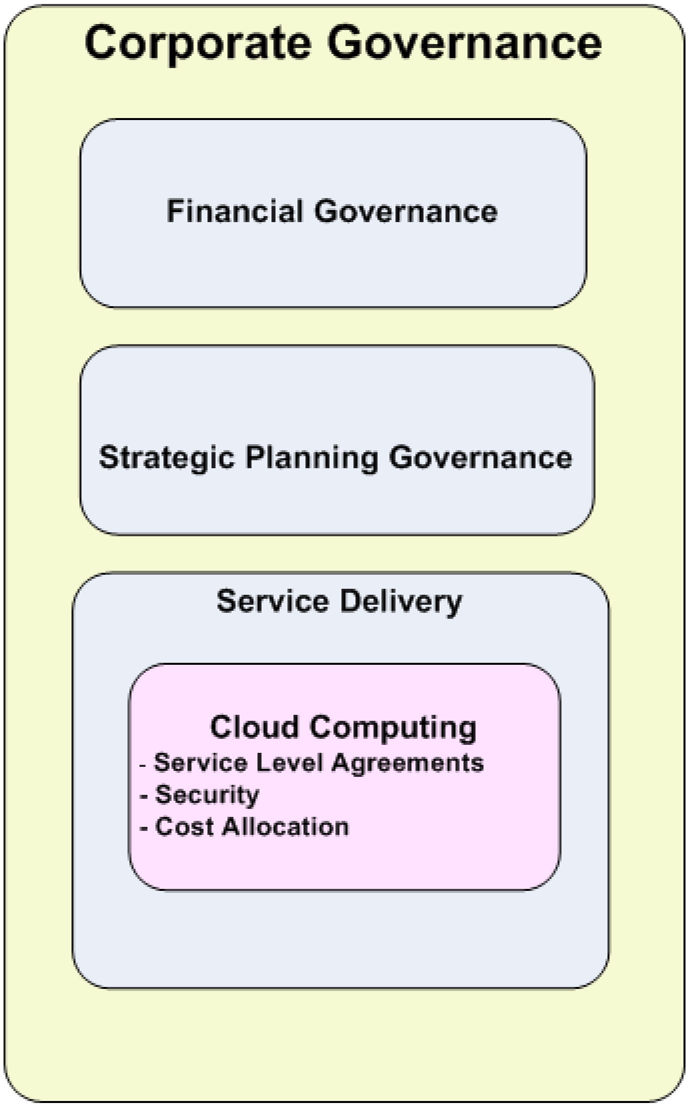


**Figure 9.2: The hierarchy of corporate governance subsumes cloud computing governance.**

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.

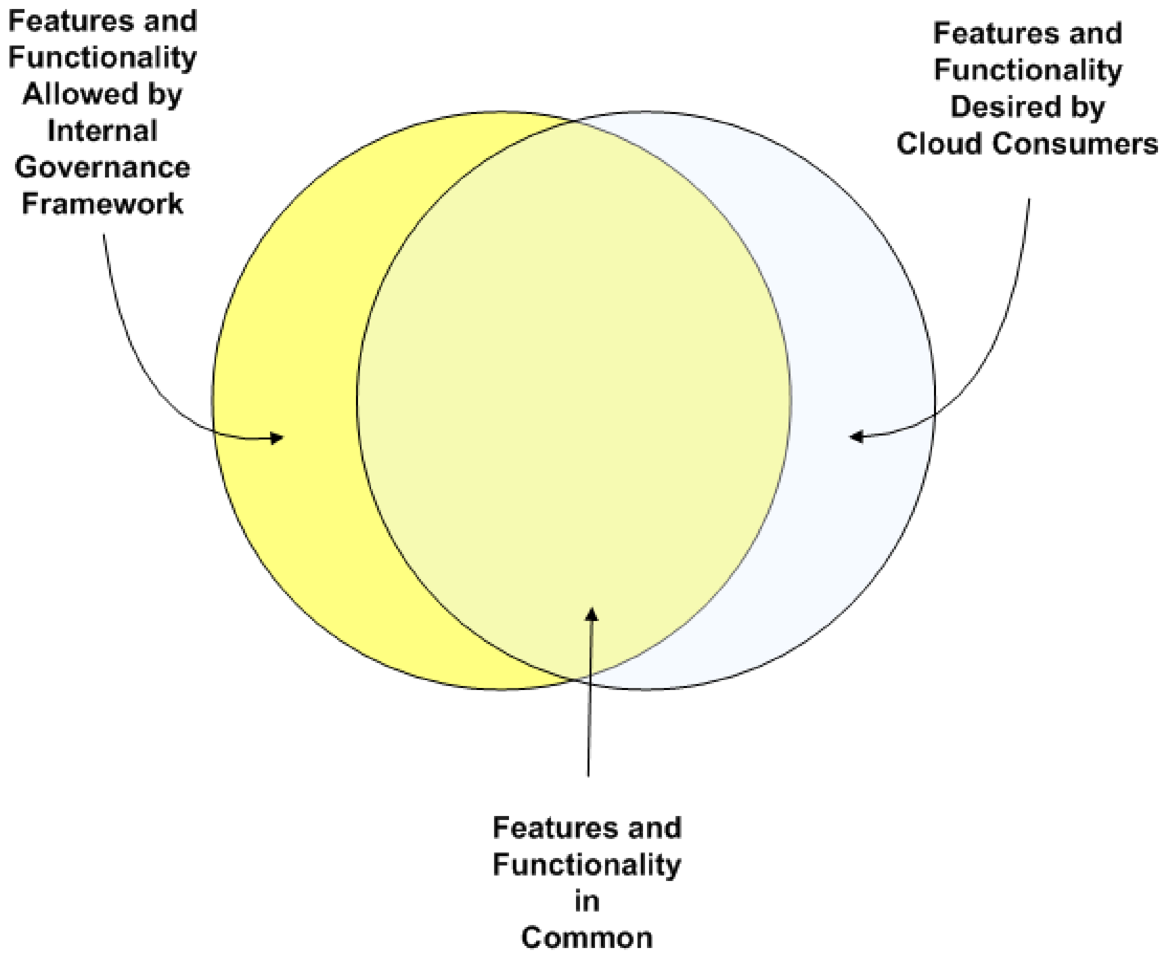

Features and Functionality Allowed by Internal Governance Framework
Features and Functionality Desired by Cloud Consumers
Features and Functionality in Common

**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.

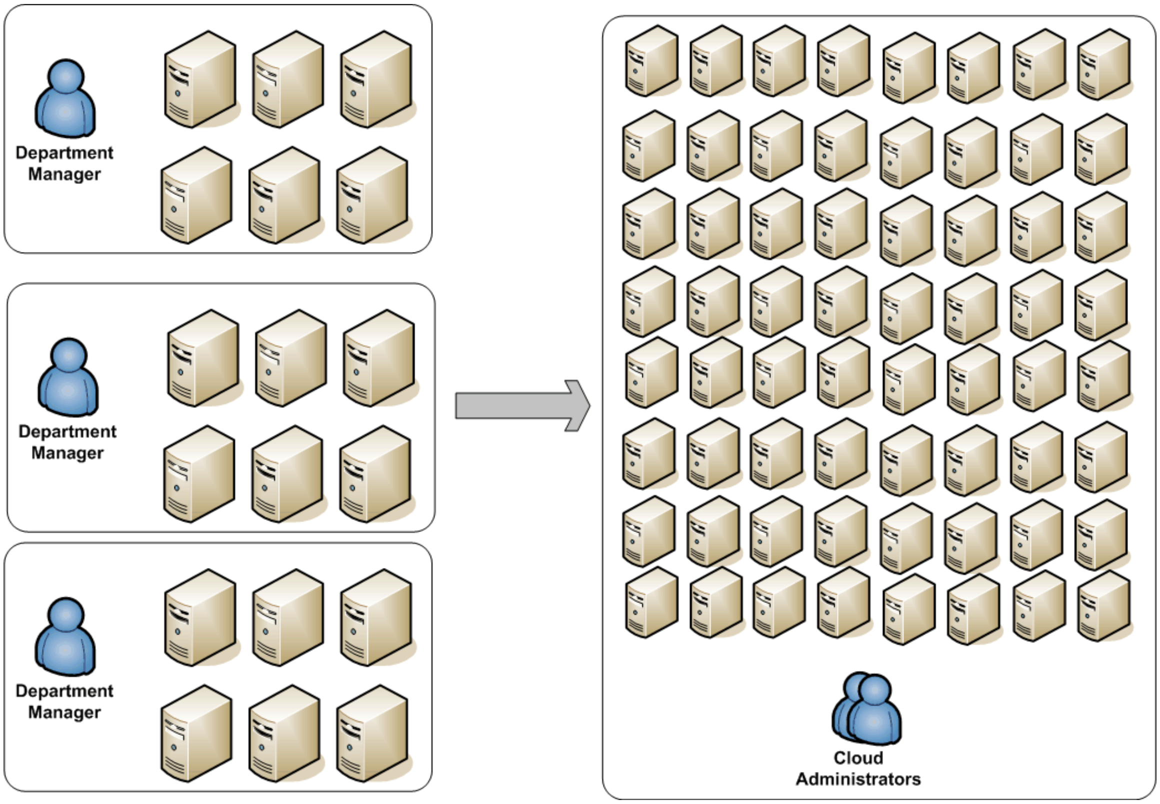


**Figure 9.4: With the adoption of cloud computing, the scope of capacity planning shifts from individual applications and departments to a centralized cloud service provider.**

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.

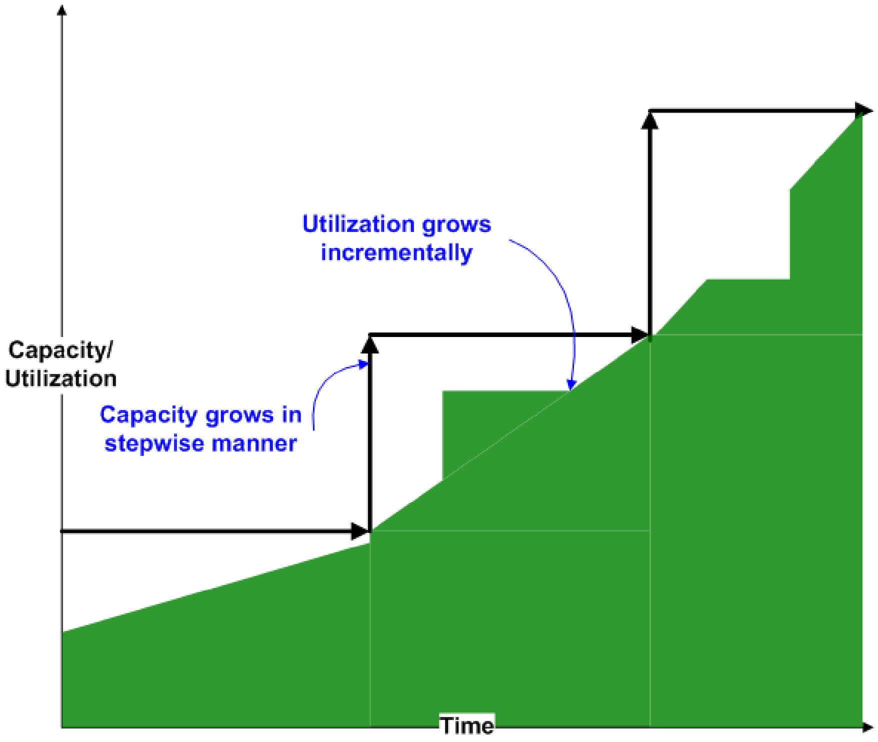


**Figure 9.5: Capacity is often acquired in bulk, giving a stepwise growth in capacity. Utilization tends to grow incrementally, although there may be spikes or temporary drops in utilization.**

## 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.

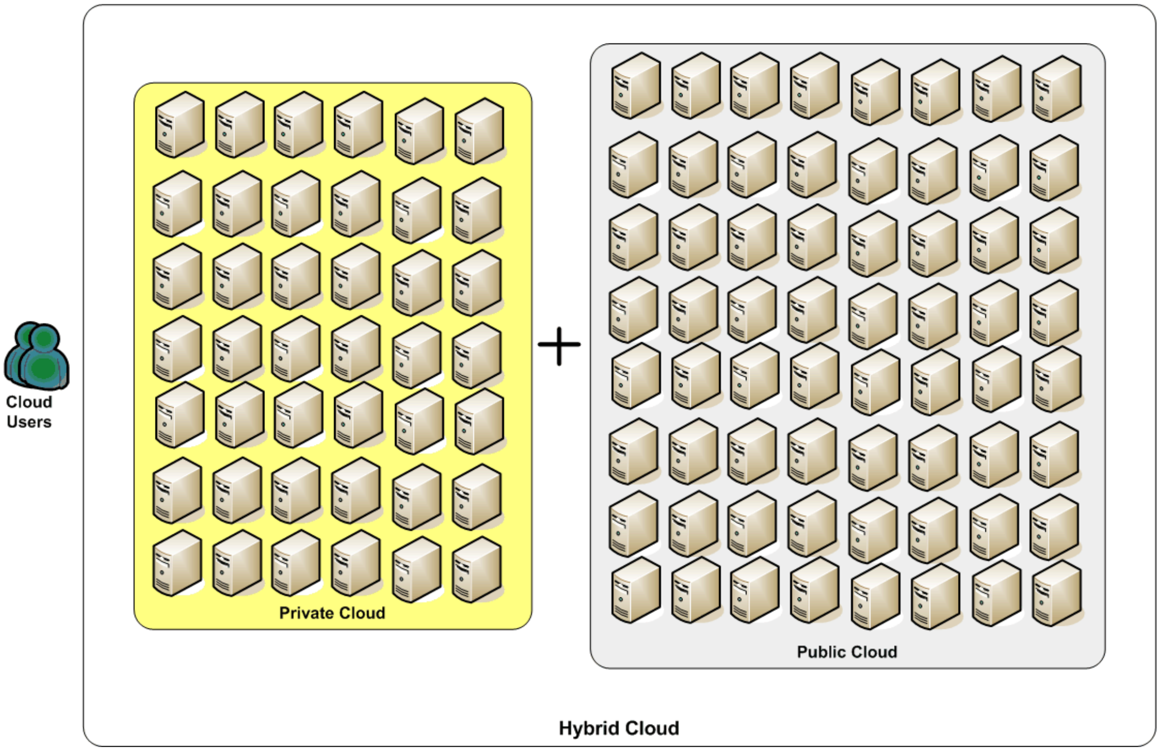


**Figure 9.6: Hybrid clouds appear to users to be functionally equivalent to private clouds. Private cloud administrators hide the implementation details from end users.**

## 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.

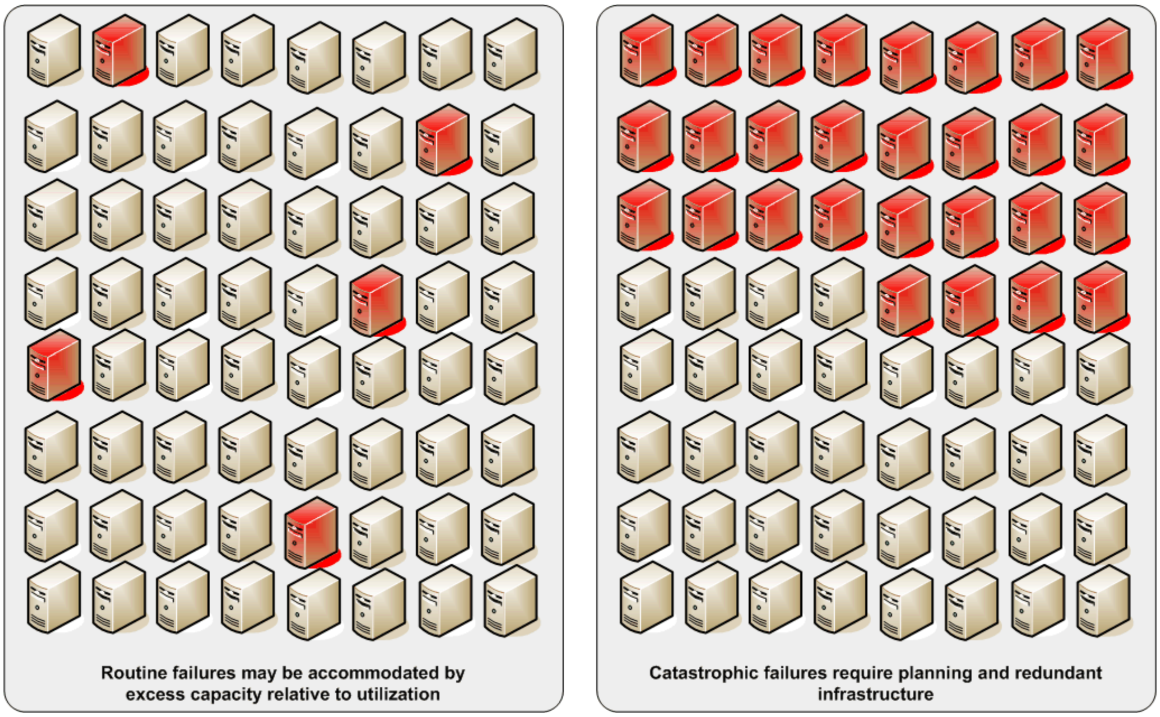

Routine failures may be accommodated by excess capacity relative to utilization
Catastrophic failures require planning and redundant infrastructure

**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.

# Chapter 10: Key Steps in Establishing Enterprise Cloud Computing Services

Adopting cloud computing technology in an enterprise can produce substantial improvements in service delivery and cost control. That is, if it is done right. The driving force behind the use of any technology should be a business imperative. For that reason, the first key step in establishing enterprise cloud services is to understand the business objectives that can be served by the technology.

In this, the final chapter of *The Definitive Guide to Cloud Computing,* we begin by examining how to align business drivers with cloud services. This process includes understanding business objectives, identifying weaknesses in existing IT service delivery, and prioritizing the multiple objectives that can be served by cloud computing.

Once we have established what we want to achieve with cloud computing in the enterprise, we move on to the second key step in the process: planning. The planning phase requires a combination of business and technical knowledge that typically requires a team of professionals from across the organization. Some of the issues we must address at this stage are assessing the current state of readiness, determining the best cloud model for a given set of requirements, and planning for long-term management and sustainability.

The implementation phase follows the planning stage. The details of this phase will vary depending on whether a business decides to adopt a private cloud model, a public cloud service model, or a hybrid setup. Later in the chapter, we will examine issues that should be considered in each case, such as reallocating server hardware when implementing a private cloud or establishing service level agreements (SLAs) with a cloud provider when a public cloud service is used.

The fourth key step in establishing cloud computing services is to develop a maintenance model. Maintenance has both technical and business dimensions. Technical issues include establishing procedures to monitor services, identifying and correcting failed services, and maintaining proper patch levels of software underlying cloud services. The business side of maintenance focuses on tasks such as establishing value metrics and planning for adequate capacity.

As Figure 10.1 depicts, implementing cloud computing in the enterprise will introduce can ongoing life cycle that mirrors many of the steps we follow to establish cloud computing services. For example, business drivers will change over time. New services will be rolled out. Strategic initiatives will be launched. Service offerings will be curtailed as the business shifts its focus to new opportunities. The process of aligning cloud computing services with business drivers is not a one-time operation. Cloud computing services may be regularly adjusted to meet incrementally changing requirements, but the services should also undergo periodic comprehensive review to ensure that the provided policies and procedures in place continue to meet the long-term objectives of the organization.

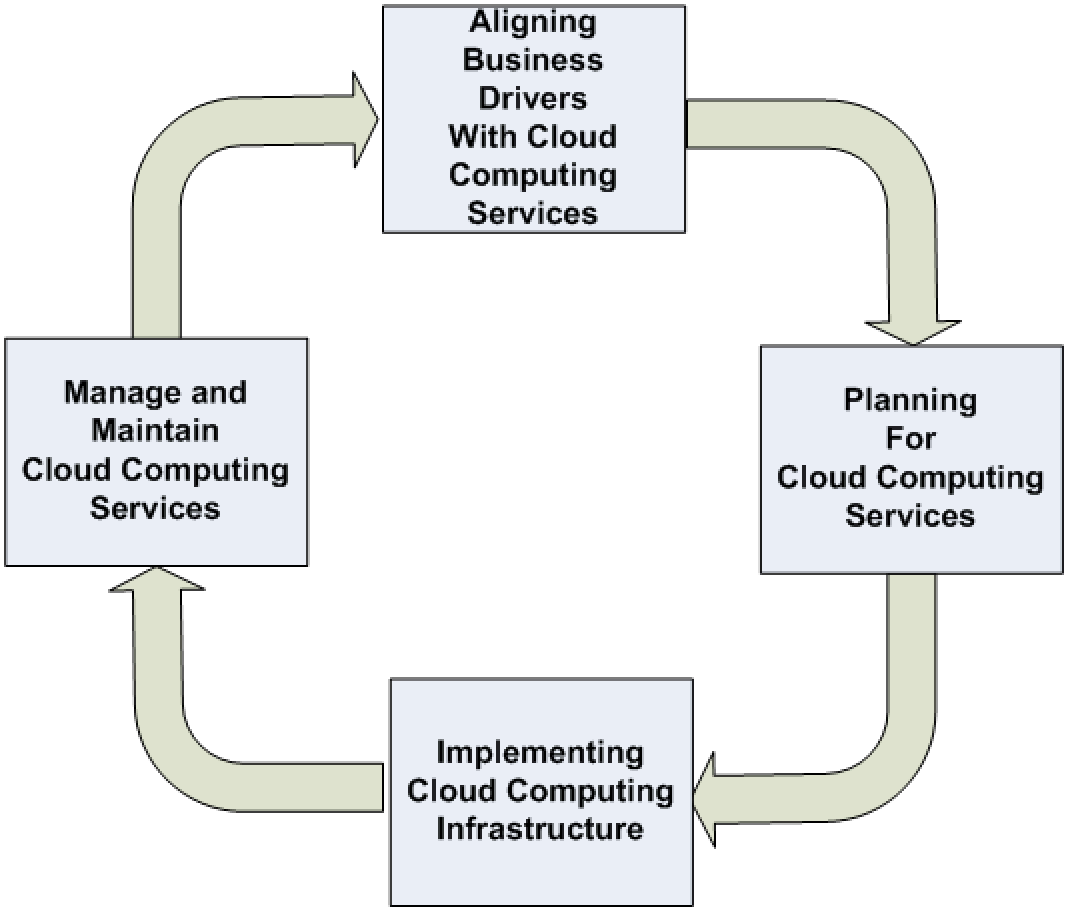

Aligning Business Drivers With Cloud Computing Services
Planning For Cloud Computing Services
Implementing Cloud Computing Infrastructure
Manage and Maintain Cloud Computing Services

**Figure 10.1: The key steps to transitioning to cloud services will persist in a similar form in an ongoing life cycle.**

Just as aligning business objectives and technology is part of the cloud computing life cycle, so are the other stages outlined here. Planning logically follows from strategic assessments, implementations follow planning, and maintenance follows implementation. Getting the cloud computing life cycle started in the right way will help establish the framework for the ongoing job of adjusting and adapting cloud services to the dynamic needs of the enterprise.

## Aligning Business Drivers with Cloud Services

Throughout this book, we have discussed the characteristics of cloud computing, delved into some of the technical details, and discussed the advantages and disadvantages of various models of cloud computing. These are obviously important considerations, but they are not the only ones. In fact, the most fundamental question we can ask with regards to cloud computing is "Why?"

Cloud computing, or any technology, is not an end in itself. Technology is deployed to serve a business purpose. To reduce the risk of misusing or misapply cloud computing in an organization, we are well served by undertaking three tasks early in the cloud computing adoption process:

- Understanding business objectives
- Identifying weaknesses in existing IT service delivery
- Prioritizing initiatives

Together these three tasks help to keep the focus, and therefore the benefits, of cloud computing on business needs in a way that maximizes the return on investment.

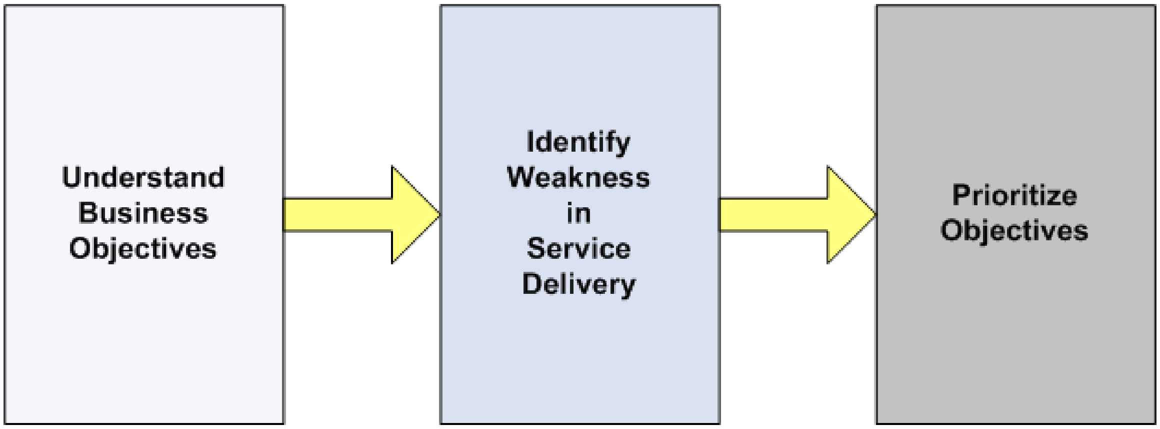

Understand Business Objectives
Identify Weakness in Service Delivery
Prioritize Objectives

**Figure 10.2: Aligning business objectives with cloud computing deployments is a three-step process.**

### Understanding Business Objectives

At the most coarse level, business objectives can be categorized into two types: developing new products, services, and capabilities and improving existing processes. New services that are especially well positioned to take advantage of cloud computing services are those that are compute- or storage-intensive. Cloud computing can enable innovation not practical under other IT models. For example, consider a manufacturing firm that produces customized machine parts.

### Cloud Computing Enables Innovation

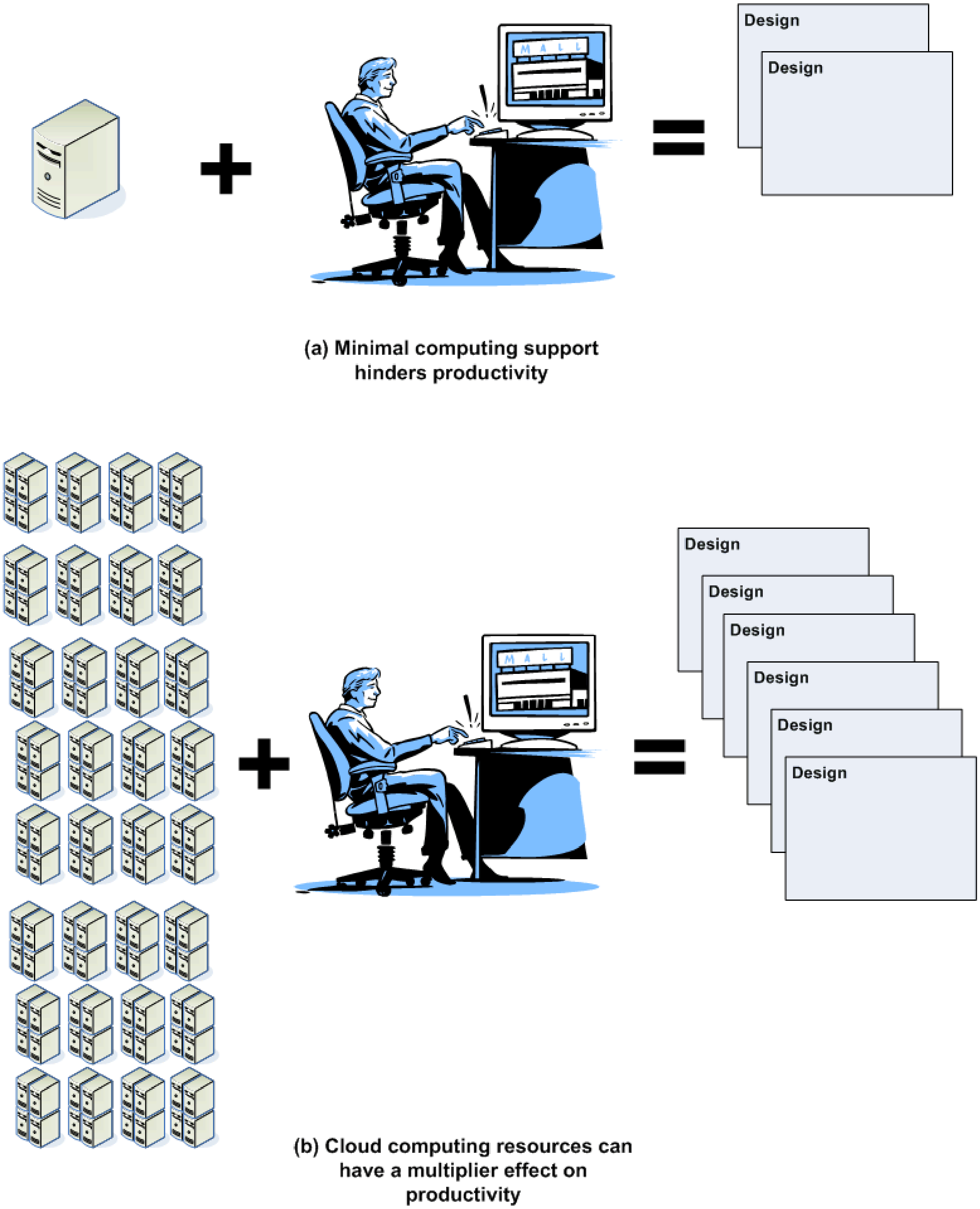
Customers continue to use the manufacturer because of the company's high-quality parts even though the time required to define the requirements for new parts is longer than most customers want. The manufacturer is well aware of its customers' time constraints but it has decided not to sacrifice quality for speed. The fundamental problem is that highly-skilled engineers are required to do the design work and the manufacturer cannot carry too many of these professionals.

Engineers could be more productive if they could better leverage the capabilities of computer aided design (CAD) software, but the kinds of analysis they need are compute-intensive. The manufacturer also does not have the IT expertise to implement and maintain a high-performance computing environment with clusters of high-end servers. Using public cloud services, the manufacturer could run the compute-intensive CAD software in the cloud as needed, freeing engineers to work on additional design problems. The combination of innovative software and cloud computing resources allows the engineers to off-load automatable design tasks.

When you are examining business objectives and assessing the opportunities for offering new services, consider several factors about workflows that make them candidates for cloud computing services.

- Is your ability to deliver the service limited by available computing or storage resources?
- Can some parts of labor-intensive processes be automated?
- Can a workflow be changed to automate 80% of the workload while leaving the other 20% for employees?

Existing workflows may not obviously lend themselves to cloud computing but re-engineered forms of the same workflow may be more amenable to automation.



(a) Minimal computing support hinders productivity

(b) Cloud computing resources can have a multiplier effect on productivity

**Figure 10.3: Cloud computing can enable increase productivity through the innovative treatment of existing applications.**

### Accommodating Varying Demand for Services

Another factor that may hold back a business initiative is uncertainty about demand. Demand may be low at first but expected to grow. There may be uncertainty about the rate of growth, especially during downturns in the business cycle. This type of uncertainty may be enough to derail an otherwise promising plan. On-demand computing and storage can help in just this type of situation.

Pilot projects can be readily started using only cloud resources. Not a single server needs to be purchased. Eliminating the procurement process saves not only money but also time. If a pilot project is successful, the service can be rolled out to larger groups of customers and cloud resources can be scaled accordingly. Spikes in demand or temporary (or even prolonged) downturns in demand are readily accommodated by adjusting the level of cloud resources allocated to the service. With no significant capital investment required to start such a project, there is greater freedom to experiment with new business services. The potential to apply innovative application of existing services and to experiment and quickly implement new services are two of the key types of business opportunities that should be considered when trying to understand how to leverage cloud computing and align it with business objectives.

### Improving Existing Processes

Another key type of business objective is cost control. This can take on several forms:

- Inefficient business processes and workflows
- Below-expected productivity from professional staff
- Poor utilization of IT resources
- Prolonged time to complete IT processes, such as deploying hardware or patching software

Inefficient business processes and insufficient productivity of professional staff can be addressed using the methods described earlier in the discussion on innovation. The other cost control areas require further elaboration.

IT resources, such as servers and storage arrays, are costly investments. Well-run businesses will work to get a reasonable return on that investment. Technical issues, however, can get in the way. One of the most significant problems is low utilization of servers, especially when they are dedicated to a single business process. Chapter 1 analyzed this problem and showed how cloud computing more efficiently allocates computing resources, shown in Figure 10.4 (which first appeared in Chapter 1 as Figure 1.6).

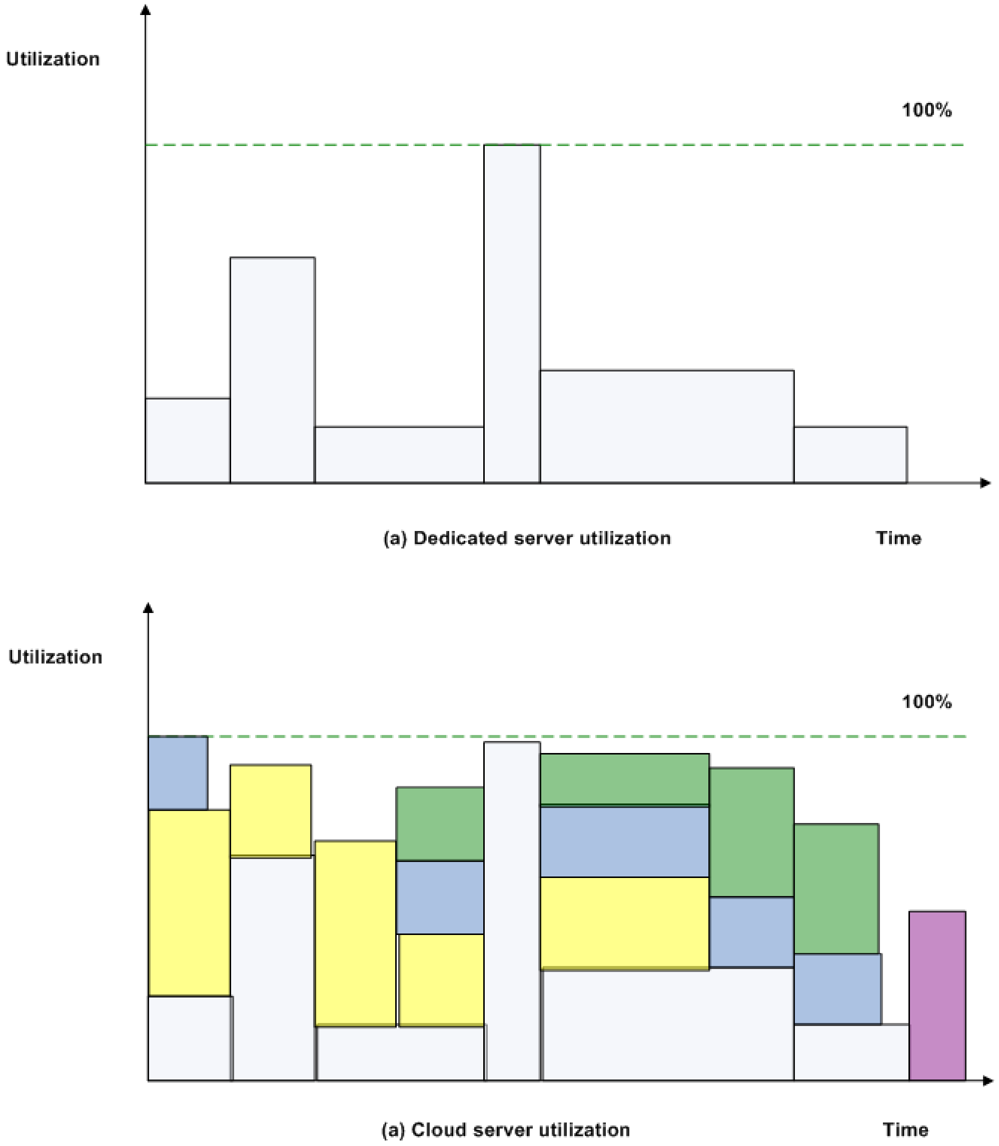


**Figure 10.4: Cloud computing more efficiently utilizes computing resources than dedicating servers to single operations that have widely varying levels of demand.**

Also consider the cost of IT support staff when evaluating business drivers behind a move to cloud computing. The combination of a number of features of the cloud delivery model often makes it a cost-effective approach. The most important of these features are:

- Virtualized servers
- Standardized catalog of software and services
- Self-service allocation and management
- Cloud management applications for IT administrators

This combination of features allows fewer IT professionals to support a larger number of users and more hardware resources than would be possible under dedicated server/dedicated systems administrator approaches.

### Identifying Weaknesses in Existing IT Service Delivery

IT departments have policies and procedures for delivery services. When new hardware is procured, there is a procedure to follow. When new applications are brought online, there are procedures to follow. The list could go on to include policies and procedures that describe how to implement security controls, software maintenance, network management, and systems monitoring and auditing. Any one of these areas can represent a weakness in the ability to deliver IT services.

Consider an example: A line of business wants to deploy a new service that will require several servers and a commonly used application stack. Everything the department wants is well within the ability of the IT department to support but still there are problems:

- The time required to review the server orders and verify the configurations are correct
- Determination of whether additional licenses are required to run the application stack
- Identification of IT staff to perform the installation and systems administration tasks
- Determination of where the hardware will be located and assurance that there is sufficient power, network connections, and other infrastructure to support the new servers

If this same new application were deployed in the cloud, we would still have to address these same issues, but we could do it more efficiently. Servers would not have to be ordered just for this application. A license management scheme (for example, site licenses) would presumably already be in place for cloud-based applications. The installation process would be reduced to ensuring the correct images are available in the service catalog. Application administrators would start virtual servers running the necessary applications on an as-needed basis. Hardware would be in place, so questions about infrastructure would not arise. Implementation issues such as these put a drag on innovation or improvement to existing processes. By identifying steps in IT processes that hinder other business operations, we can better understand where we can apply cloud computing to avoid those issues.

### Prioritizing Initiatives

Chapter 4 outlined common high-priority objectives that are worth repeating:

- Controlling costs
- Expanding market share in mature industries
- Expanding into new markets in growth industries
- Improving customer service
- Improving customer retention
- Increasing cross selling

The last step in understanding business drivers for adopting cloud computing is prioritizing all the ways we improve business operations. We can prioritized based on the value of supporting innovation, reducing the barriers to introducing new services, improving IT service delivery, and reducing the staff required to maintain a particular level of service delivery. Each of these implies either a direct cost, such as labor costs, or opportunity costs, such as those associated with delays in releasing new products and services.

Aligning business initiatives with cloud computing services is the essential first step in adopting cloud computing. By understanding business drivers, identifying weaknesses in existing processes, and prioritizing among all the potential ways to leverage cloud computing, a business will be in a firm position to take on the challenging task of planning for a transition to cloud computing.

## Planning for Transition to Cloud Computing

The planning phase of the cloud transition is primarily focused on technical issues:

- Assessing the current state of readiness
- Indentifying the differences between current infrastructure and the infrastructure to deploy for the cloud
- Determining the best cloud model for your requirements
- Planning for long-term management and stability

Not surprisingly, the first step in getting to where we want to go is to understand where we are.

## Assessing the Current State of Readiness

Cloud computing takes advantage of a particular style of application architecture. The closer we are to that style when we begin, the better off we are. Three elements of this style are Web application architecture, self-management of compute and storage services, and standard platforms and application stacks. These elements were described in detail in Chapter 7, so they will be only briefly described here.

Web application architecture is decentralized and depends on multiple processes running on multiple servers. A simple three-tiered model includes a server for persistent storage, which is usually a relational database; a middle tier of an application server, such as a Java J2EE application server or a .NET application; and a client tier providing a user interface (UI). A common variation on this model is to have several application servers providing services to a Web server that coordinates those services for a client interface (see Figure 10.5).

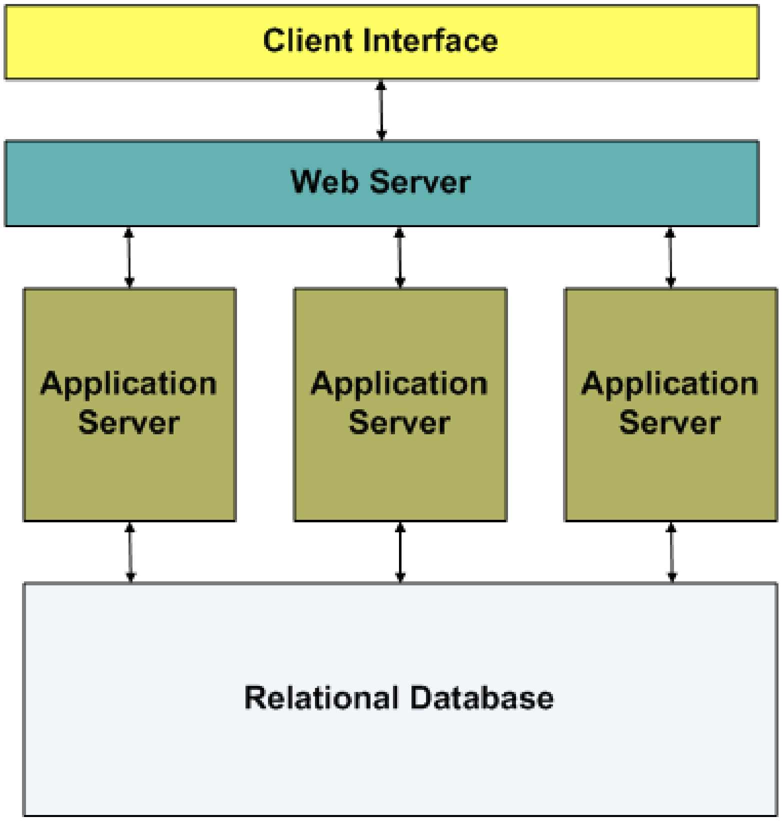


**Figure 10.5: A common decentralized amenable to cloud computing model uses a multi-tier stack to implement applications.**

The more centralized an application, the more difficult it is to take advantage of the cloud. For example, in the application architecture shown in Figure 10.5, if any of the application servers reaches capacity, another instance of that application server could be instantiated to assist with the load. Centralized applications do not offer the opportunity to scale parts of the application like that.

One of the cost control benefits of cloud computing is the ability to offer self-service management to cloud consumers. This setup removes high-cost IT professionals from common tasks such as starting instances of virtual machines or allocating storage for an application. The software required to implement self-service can be deployed in the next phase of the transition process, but cloud consumers should be in a position to take advantage of self-service features when they arrive.

Another factor to consider is how standardized your application stacks are. Are departments running a wide range of applications and different platforms? Do you support three or four major relational databases? Are departments running different versions of Windows and Linux operating systems (OSs)? The answers to these questions will give you some indication of how standardized your organization is with respect to application stacks. The transition to cloud computing can be an opportunity to prune the set of supported applications. This will further improve the cost benefits of cloud computing by reducing the demand for patching, licensing management, and support services related to different applications.

### Indentifying the Differences Between Current Infrastructure and the Infrastructure to Deploy for the Cloud

Cloud services can run on commodity hardware. They can also run on specialized hardware assuming virtualization services are available. What set of hardware servers, storage, and network equipment is available in your organization? The optimal set of infrastructure components is a function of several factors. On the one hand, if hardware is in place, it seems logical to use it; on the other hand, the greater the diversity in equipment, the greater the administration and overhead costs. Some things to consider with regard to assessing what you have and what you would like for hardware infrastructure include:

- The capacity of servers to support multiple virtual instances, including processor speed and memory capacity
- The ability of servers to run software in the services catalog
- The range of support skills required to maintain the infrastructure
- Power consumption and cooling requirements

The goal is to provide needed cloud services at the lowest cost. This requires us to consider the full range of expenses, from the cost of new hardware to the cost of maintaining power and cooling for older hardware that may require more support than newer hardware. The best combination of new and existing infrastructure is a function of your resources, environment, and requirements. There is no single answer or simple formula for determining the optimal solution.

## Determining the Best Cloud Model for Your Requirements

As we have described throughout this guide, there are three models for delivering cloud services: private, public, and hybrid. Which is the best option for you?

A private cloud is suitable for enterprises that have the infrastructure, support skills, and management framework to maintain such an architecture. We use the term infrastructure broadly, to include not only IT hardware but physical infrastructure such as data centers, redundant power supplies, and multiple high-speed Internet connections. IT professionals running a private cloud will be required to manage large numbers of similarly configured servers, multiple disk arrays, a complex array of network management systems, and robust security controls. A management system must be in place as well to implement cost recovery, capacity planning, service delivery, licensing negotiations, and other administrative capabilities.

These are significant barriers to adopting a private cloud model, but there are advantages as well. Your organization has complete control over the service catalog, who is allowed to use cloud resources, and the ability to monitor all cloud services. The fact that data and applications would not have to reside outside the corporate firewalls can be a substantial advantage from a compliance perspective.

A public cloud has several advantages:

- Minimal capital expenditures
- Ability to maintain existing infrastructure in its current configuration, allowing for a period of time in which both existing and new cloud-based instances are used
- Possibly lower costs per unit of computing service or storage because of the economies of scale
- Less management overhead for day-to-day operations but potentially more overhead for negotiating, monitoring, and enforcing SLAs

The potential drawbacks of a private cloud include the need to move sensitive data outside the corporate infrastructure, the potential costs of transmitting large volumes of data over the network, and the delays in moving data into the cloud by shipping storage devices (done in some cases to reduce upload costs).

A hybrid cloud can offer the advantages of both the private and public cloud. Sensitive information can be maintained in a private cloud while other data is moved to the public cloud. Existing infrastructure can be readily redeployed to a cloud while older or less amenable hardware is not. Initial capital expenditures may be reduced because peak loads in the private cloud can be accommodated by allocating resources in a public cloud.

Once again, there is no solution that is optimal for all cases. The advantages and disadvantages of each model must be weighed against the business requirements and constraints.

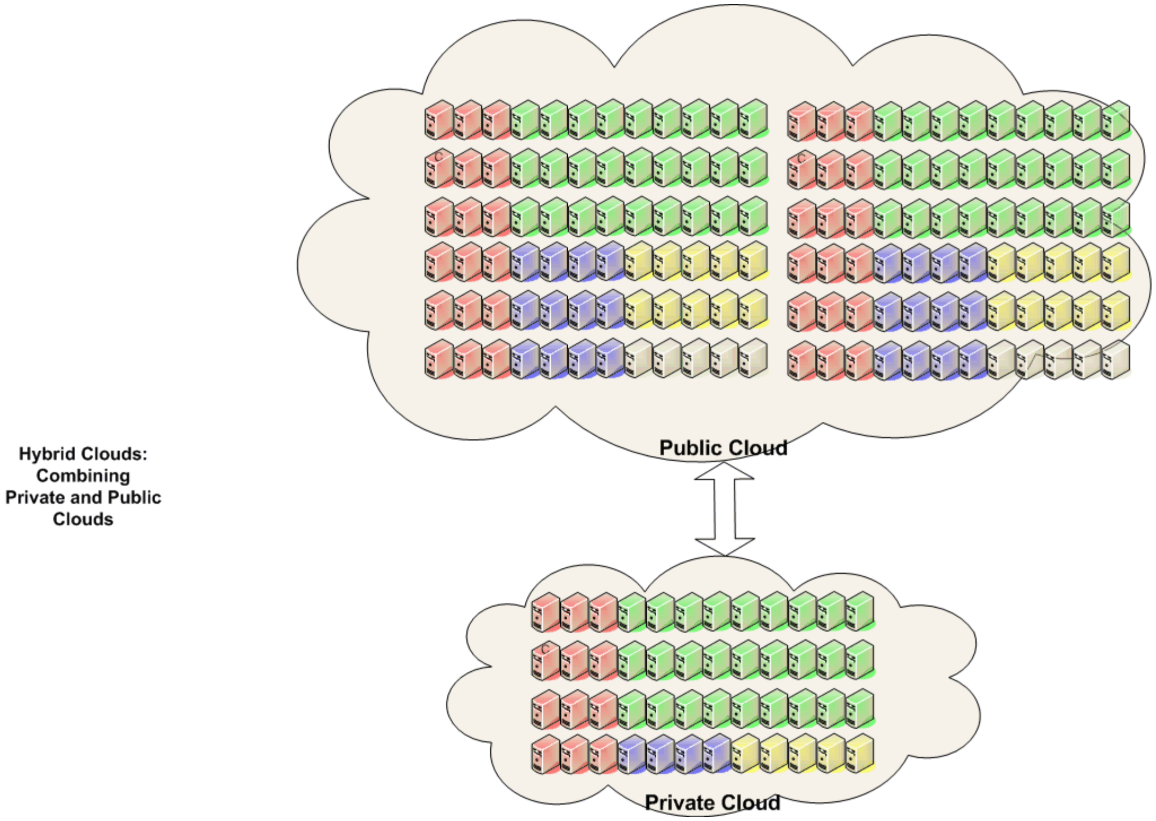


**Figure 10.6: A combination of private and public clouds can enable an organization to realize the benefits of both.**

## Planning for Long-Term Management and Stability

Implementing a computing and storage cloud is a long-term proposition that requires attention to a number of areas in addition to those already mentioned. In particular, we need to plan for security, disaster recovery, and maintenance of physical infrastructure.

Security considerations include protecting physical infrastructure as well as logical access to services and resources. Cloud data centers will require the same types of physical protections as one would find in any large data center. Access to infrastructure should be limited to those with legitimate needs. The site should be monitored and security procedures audited. Fire suppression equipment should be in place. Logical access controls begin with identity management. Policies should be in place defining who has access to various cloud resources, such as servers and applications. Licensing restrictions should be taken into consideration as well. Policies and procedures should define how authentication and authorizations are granted, monitored, and revoked.

Long-term management includes planning for disaster. Maintaining multiple data centers may be a reasonable strategy for some private cloud users but not others. The costs can be prohibitive. One alternative is to use a public cloud for disaster recovery purposes, although there are still issues regarding confidentiality and compliance.

Maintaining the physical infrastructure of a cloud is an ongoing operation. With large numbers of servers and disks, it is reasonable to expect regular equipment failures. Even with long mean times between failures, when we are dealing with thousands of pieces of equipment, parts will fail. Services, such as power and Internet access, will fail as well. Backup power supplies and redundant Internet providers should be used.

A useful rule of thumb for managing cloud computing and the services it can provide is to assume that change and innovation are inherent. New equipment and applications will be added while others are retired. Equipment will fail. Power will go down. New business requirements will emerge. Cloud computing, like the business environment it serves, is dynamic.

## Implementing a Cloud Infrastructure

Analyzing business drivers can be challenging because of complex, interdependent goals and objectives. Planning can be difficult because one has to merge both business requirements and technical constraints in a way that serves business objectives. The next stage of the process, implementation, is difficult primarily for technical reasons. The specific challenges will vary depending on the type of cloud model that is being used: private, public, or hybrid.

### Implementing a Private Cloud

The key tasks to implementing a private cloud center on deploying hardware and establishing operations. Three such tasks are:

- Reallocating and deploying servers
- Establishing software and application management procedures
- Implementing a management framework

Reallocating servers must be done carefully to avoid disrupting existing business services. When new hardware is used for cloud deployments, the transition is relatively straightforward, as depicted in Figure 10.7. Applications can continue to run on legacy hardware as long as needed as those same applications are moved to the cloud.

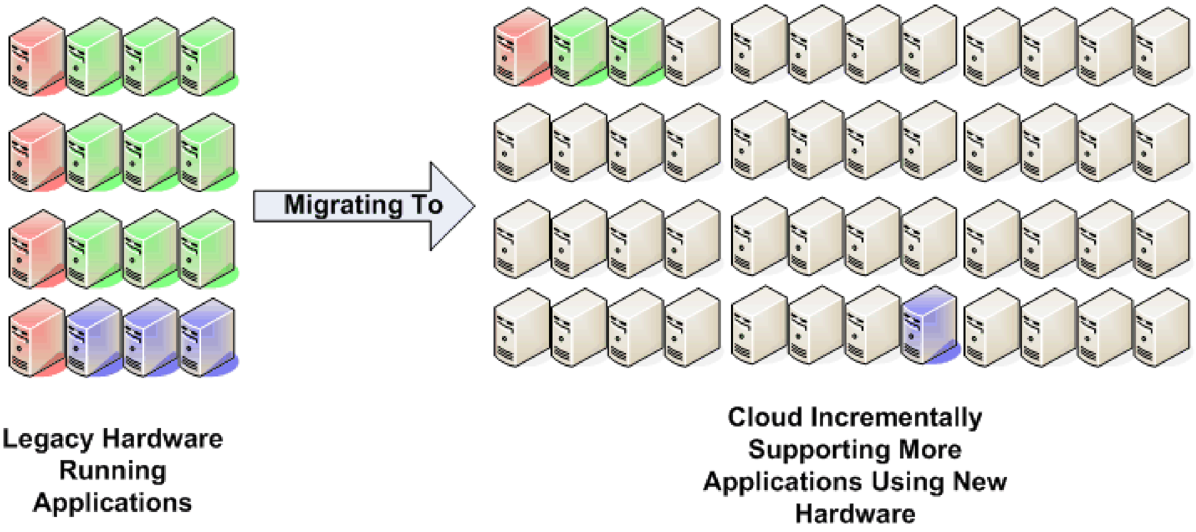

Migrating To
Legacy Hardware Running Applications
Cloud Incrementally Supporting More Applications Using New Hardware

**Figure 10.7: When new hardware is deployed in the cloud, applications can migrate directly to the cloud.**

When existing hardware is redeployed to the cloud, the migration is less direct. A basic challenge is to keep services available while migrating hardware from an application-centric use of servers to a cloud computing model. One way to handle this challenge is to migrate applications from their dedicated servers to a set of virtual machines running on servers temporarily allocated to support the migration. This approach works when servers dedicated to applications are not using the full capacity of servers. Applications are temporarily hosted on transition servers while hardware is migrated to the cloud. Once the hardware, software, and supporting cloud services are in place, applications can begin running in the cloud.

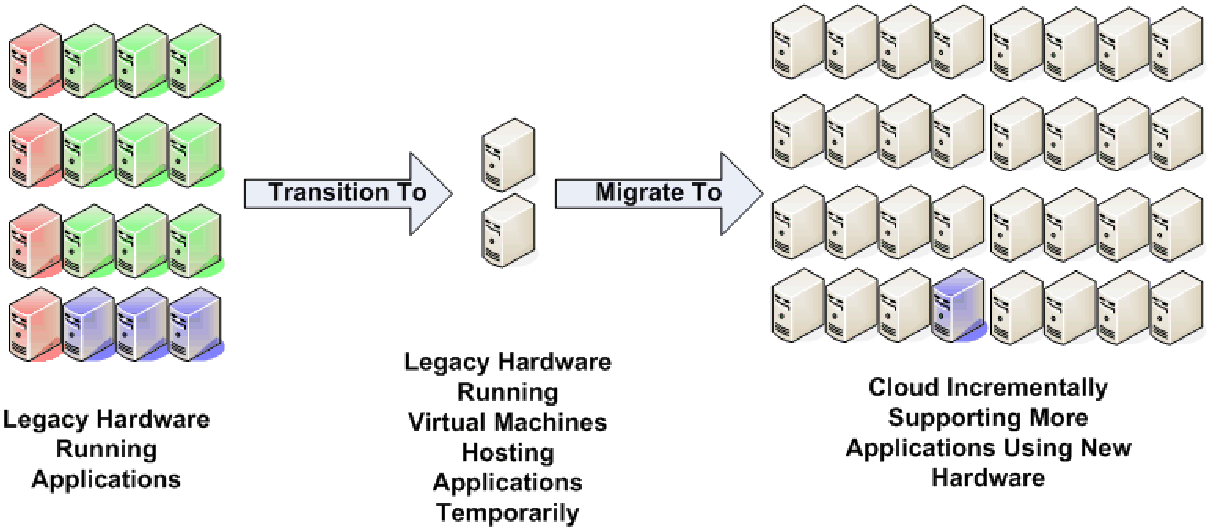

Transition To
Migrate To
Legacy Hardware Running Applications
Legacy Hardware Running Virtual Machines Hosting Applications Temporarily
Cloud Incrementally Supporting More Applications Using New Hardware

**Figure 10.8: Applications may be hosted on transition virtual servers in cases where existing hardware is to be redeployed to the cloud.**

Management procedures must be established for maintaining the diverse array of software that will be used in the cloud. These include establishing policies and procedures for:

- Adding and removing applications from the service catalog
- Patching images in the service catalog
- Controlling the use of licensed applications to ensure their use is in compliance with licenses
- Performing security reviews, such as vulnerability and malware scans on images in the service catalog

A private cloud also requires a management framework for non-software management issues. A number of essential management tasks should be in place before the cloud is widely used in the enterprise:

- Tracking compute and storage usage for billing and cost recovery purposes
- Monitoring performance and load for capacity planning
- Auditing patterns of use and access as part of security review procedures

Introducing public cloud services brings with it a different set of implementation tasks.

### Adapting Public Cloud Services

Using a public cloud relieves a business of many of the implementation tasks associated with private clouds. There is no need to transition hardware or redeploy servers. No service catalogs to establish and manage. No low-level billing infrastructure to put in place. Instead the focus tends to be more on defining SLAs and reviewing compliance and security issues.

SLAs are essentially contracts between a business and a cloud provider. SLAs are important for clarifying what services are expected, the cost of such services, the quality of these services, and compensation for failure to meet agreements. SLAs with public cloud providers can include agreements about many factors:

- The number and types of servers that will be available for use at any time
- Restrictions on the number of virtual or dedicated servers that may be allocated in a single request
- Minimum and maximum storage usage
- Guaranteed bandwidth into and out of data centers used by the public cloud
- Security controls and procedures
- Audit and monitoring responsibilities of the provider and the business customer
- Compute and storage rates, billing periods, and so on
- Individual and aggregate demand reports

Several of the topics addressed in SLAs are security oriented. Clearly, a top priority for most businesses using public cloud services is ensuring that private, sensitivity, and confidential data is protected. This will require a combination of secure communications between the cloud data center and user sites; secure, probably encrypted persistent data storage in the cloud; access controls on private images or applications stored and run in the cloud; and verification that cloud software is routinely patched and scanned for vulnerabilities and malware.

## Using a Hybrid Private-Public Cloud

A hybrid private-public cloud delivers the benefits of both models of cloud computing. It also brings with it the responsibilities of both that we just described—and a bit more. The combined resources of a private and public cloud may appear to be seamlessly integrated from the users' perspective but there are operational differences. Only data and applications that are deemed safe to store or run in a public cloud should be made available outside the private cloud.

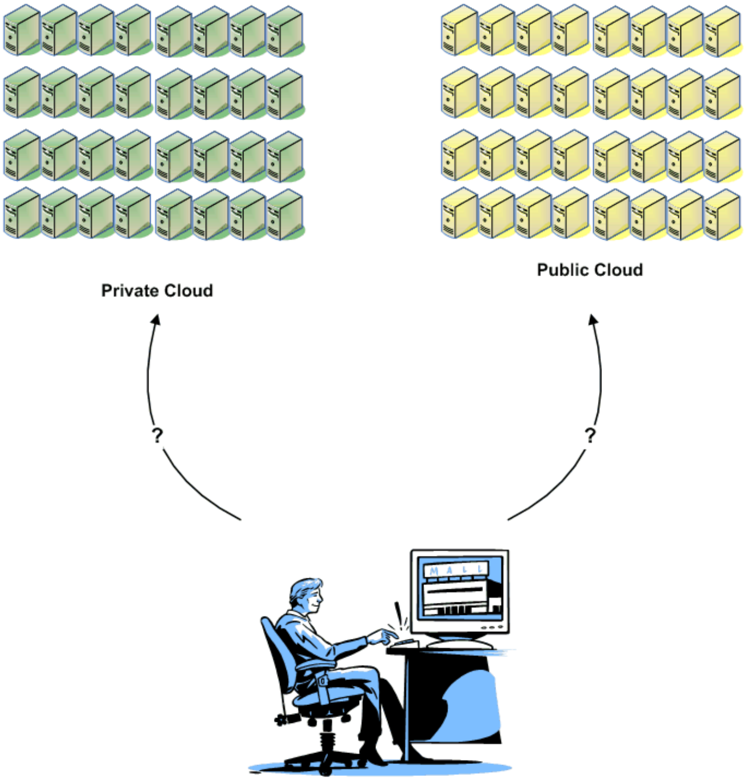


**Figure 10.9: If users are given a choice of where to run applications in a hybrid cloud, policies and incentives should be in place to promote the optimal balance from an enterprise perspective.**

Users of cloud services should also be made aware of any cost differences between the private and public clouds. For example, will the IT department charge an additional fee on top of the public cloud provider's charges to cover the overhead of managing the hybrid cloud? Also consider whether rules or cost structures should be in place to incentivize users to use private cloud services before turning to the public cloud. This is especially important if cost recovery pricing is used and assumptions are made about the level of utilization in the private cloud. The last key area to address for the long-term maintenance of an enterprise cloud is, in fact, maintenance.

## Managing and Maintaining a Cloud

The tasks of managing and maintaining a cloud computing environment can be broken down into operational issues and business management issues.

### Operational Issues

Once hardware is deployed, management infrastructure is deployed, applications are installed, and security controls have been put in place, a cloud is ready to use. After that, we are in maintenance mode. At this point, new business requirements will arise and will be accommodated in an incremental manner. There will still be business analysis, planning, and implementation tasks as described earlier in the discussion about the cloud computing life cycle (see Figure 10.1). On a day-to-day basis, some of the most important operational tasks will be:

- Monitoring
- Fault detection and correction
- Systems maintenance

Cloud administrators will have to routinely monitor several attributes of a cloud. Utilization of servers and storage capacity should be regularly monitored. This data is useful for short-term management, for example, when additional servers have to be brought online during periods of peak demand, as well as for long-term capacity planning. The images run from the service catalog also need to be monitored. Systems administrators should know which applications are used most frequently, especially when licensing costs are an issue. This information is also useful for prioritizing patching, security scans, and upgrade planning. Monitoring should also include security monitoring, such as user activity, suspicious events—such as authentication failures or repeated unauthorized access attempts, and scanning of inbound and outbound network traffic.

Hardware follows the rule of large numbers: with a sufficiently large number of devices, some of those devices will fail and instances of failure will be more frequent for a cloud than for a single server. The logic is simple: the probability of a server failing is the probability of Server A failing plus the probability of Server B failing plus the probability of Server C failing, and so on. In a private cloud, systems administrators will need to detect faults in servers and storage devices and be able to take those devices offline. In the case of a failed server, applications running on the failed server will need to be moved to another server. When a storage device fails, read and write operations should be able to continue using redundant copies of the data that was lost. Aggregate data about failure rates of devices can be collected over time and provide a baseline for predicting rates of failures.

System maintenance is a rather generic term for a broad set of tasks that one needs to perform to keep applications running as expected. The set includes managing user identities, establishing access controls, patching software, scanning images for malware and vulnerabilities, and other tasks we had prior to moving to a cloud model. Changing architectures does not change the need for basic system management tasks.

## Business Management Issues

Long-term business management issues of supporting a cloud infrastructure can be as varied as the technical issues, ranging from establishing value metrics to ensuring continuity of services in the event of a disaster. At the most basic level, organizations adopt cloud computing because it will improve the ability of the business to meet its objectives. That is the idea when the process gets started, but how do you know whether the implementation is succeeding or if you are anywhere near realizing the benefits expected? A set of value metrics need to be in place to measure the value of the cloud. These value metrics can include generic measures such as return on investment (ROI) or more specific ones such as

- Reduction in time to release a new product or service
- Number of CPU hours utilized for delivery business services
- Utilization rate of storage in the enterprise
- Transaction processed per unit of computing and storage resource
- Reduction in IT support costs per server

Some value metrics should measure technical aspects, such as server utilization rates, but others should clearly measure the business value of the cloud, such as ROI. The former helps systems administrators and IT managers drive efficiencies in the cloud; the latter ensures that these are worth the effort from a business perspective.

Capacity planning requires a close coupling of business planning and technology management. Operational data about server, network, and storage utilization, numbers of business operations supported by the cloud, and number of users and their distribution in the company are vital for capacity planning. For example, if a product design group is a major user of cloud services and the company is about to acquire another firm that will significantly increase the size of the product design group, the cloud management team needs to know. If a Web application development team at a national home improvement retailer plans to provide a large number of "do it yourself" videos on the Web site and significantly increase network utilization, the cloud team should be prepared. These realities provide examples where creating and maintaining lines of communications between different parts of a business are important to the long-term effectiveness of an IT service.

Long-term planning also requires attention to disaster recovery. If we assume a disaster could strike and disable a data center, we need to be able to answer the question, what happens then? If we have geographically distributed data centers with redundant storage and additional computing resources, we can move operations to one or more alternative data centers. Although servers may be able to fail over fairly seamlessly and redundant copies of data can be made available, the programs running in the failed data center may not be as robust. For example, an application that runs for extended periods of time without writing state information to persistent storage may have to restart its processing from the beginning of a job rather than recover mid-stream. When planning for disaster recovery, we must consider details from the lowest implementation level, such as the availability of power and cooling systems, to high-level design issues, such as how applications manage state information.

Maintenance and long-term management issues in cloud environments are similar to those found in other IT environments. Fortunately, many of the best practices and management techniques that have evolved over the years are relevant and applicable today, albeit with some slight tuning for the unique characteristics of the cloud.

## Summary

Cloud computing is changing the way we deliver business services. The cloud architecture allows for more efficient utilization of infrastructure, a more efficient delivery mechanism for services, and an improved user experience. By aligning business objectives with the capabilities of cloud computing, businesses can realize faster time to market, reduced IT support costs, and more effective use of capital for investments.

Cloud computing is characterized by its massive scalability, easy-to-use provisioning services, and a service management platform. These may be delivered privately within the corporate boundaries, publicly through a third-party provider, or as a combination of the two. There are different levels of cloud services, such as infrastructure providers, platform services, and application services. These services can be deployed according to business needs, and lead to improved ability to deliver current services and introduce new services without undoing encumbrance from having to deploy complex IT infrastructure.

*The Definitive Guide to Cloud Computing* has presented a comprehensive overview of cloud computing with a focus on identifying steps needed to successfully deploy cloud computing in your business. Technical details of cloud computing will change, but the analysis and management principles are based on the IT industry's prior experience with other architectures and service delivery models. The valuable lessons learned deploying and managing mainframes, client-server applications, and first-generation are applicable to the cloud, with of course, some adaptation.