

Realtime
publishers

The Definitive Guide™ To

Cloud Computing

sponsored by



Dan Sullivan

Chapter 10: Key Steps in Establishing Enterprise Cloud Computing Services 185

- Aligning Business Drivers with Cloud Services 187
 - Understanding Business Objectives 187
 - Cloud Computing Enables Innovation 188
 - Accommodating Varying Demand for Services 190
 - Improving Existing Processes 190
 - Identifying Weaknesses in Existing IT Service Delivery..... 192
 - Prioritizing Initiatives..... 193
- Planning for Transition to Cloud Computing 193
 - Assessing the Current State of Readiness 194
 - Identifying the Differences Between Current Infrastructure and the Infrastructure to Deploy for the Cloud..... 195
 - Determining the Best Cloud Model for Your Requirements 196
 - Planning for Long-Term Management and Stability 197
- Implementing a Cloud Infrastructure..... 198
 - Implementing a Private Cloud 198
 - Adapting Public Cloud Services 200
 - Using a Hybrid Private-Public Cloud..... 201
- Managing and Maintaining a Cloud 202
 - Operational Issues..... 202
 - Business Management Issues..... 203
- Summary 204

Copyright Statement

© 2010 Realtime Publishers. All rights reserved. This site contains materials that have been created, developed, or commissioned by, and published with the permission of, Realtime Publishers (the “Materials”) and this site and any such Materials are protected by international copyright and trademark laws.

THE MATERIALS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. The Materials are subject to change without notice and do not represent a commitment on the part of Realtime Publishers or its web site sponsors. In no event shall Realtime Publishers or its web site sponsors be held liable for technical or editorial errors or omissions contained in the Materials, including without limitation, for any direct, indirect, incidental, special, exemplary or consequential damages whatsoever resulting from the use of any information contained in the Materials.

The Materials (including but not limited to the text, images, audio, and/or video) may not be copied, reproduced, republished, uploaded, posted, transmitted, or distributed in any way, in whole or in part, except that one copy may be downloaded for your personal, non-commercial use on a single computer. In connection with such use, you may not modify or obscure any copyright or other proprietary notice.

The Materials may contain trademarks, services marks and logos that are the property of third parties. You are not permitted to use these trademarks, services marks or logos without prior written consent of such third parties.

Realtime Publishers and the Realtime Publishers logo are registered in the US Patent & Trademark Office. All other product or service names are the property of their respective owners.

If you have any questions about these terms, or if you would like information about licensing materials from Realtime Publishers, please contact us via e-mail at info@realtimepublishers.com.

[**Editor's Note:** This eBook was downloaded from Realtime Nexus—The Digital Library for IT Professionals. All leading technology eBooks and guides from Realtime Publishers can be found at <http://nexus.realtimepublishers.com>.]

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 an 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.

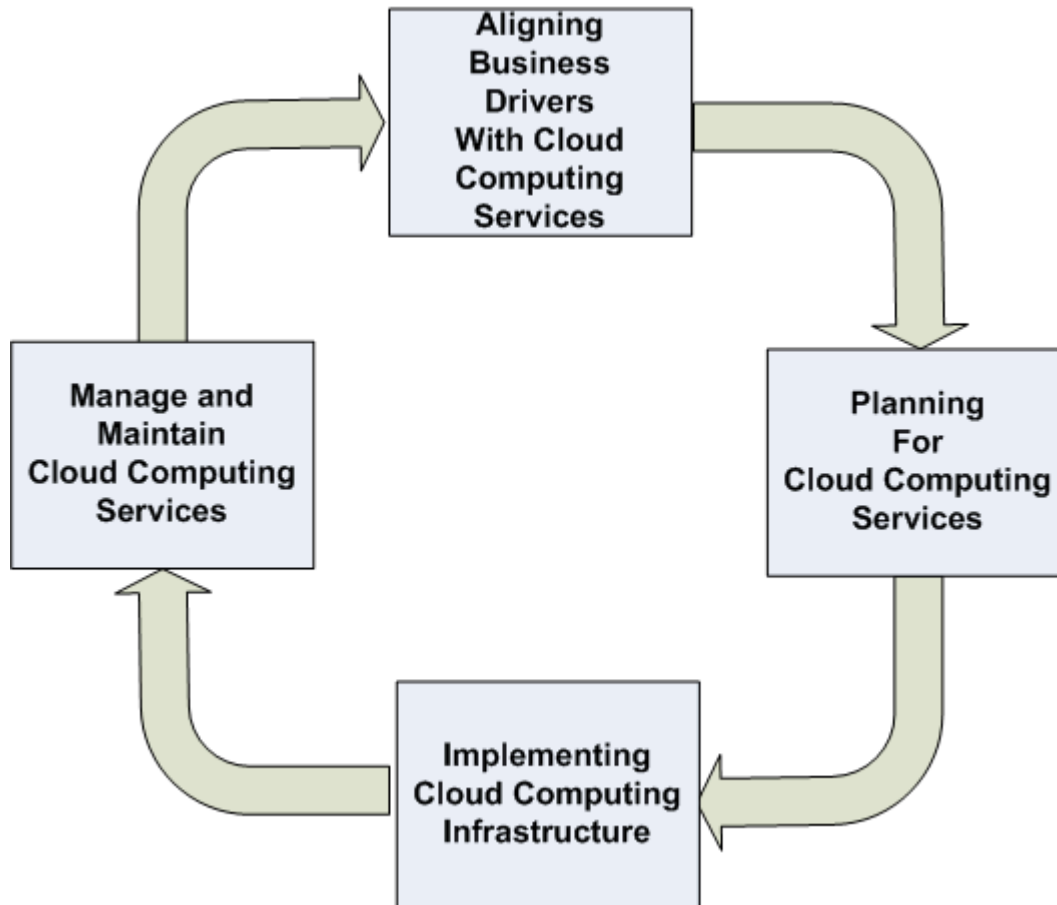


Figure 10.1: The key steps to transitioning to cloud services will persist in a similar form in an ongoing life cycle.

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.

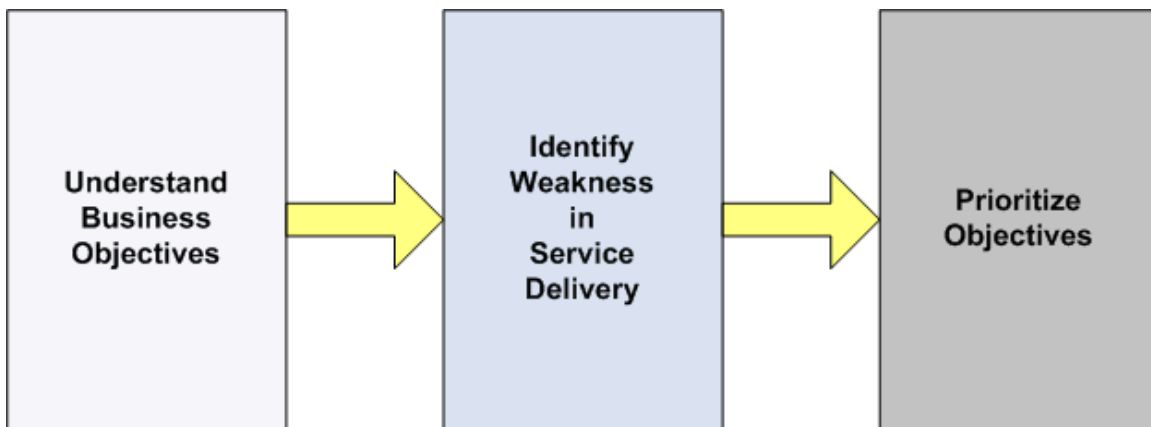


Figure 10.2: Aligning business objectives with cloud computing deployments is a three-step process.

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.

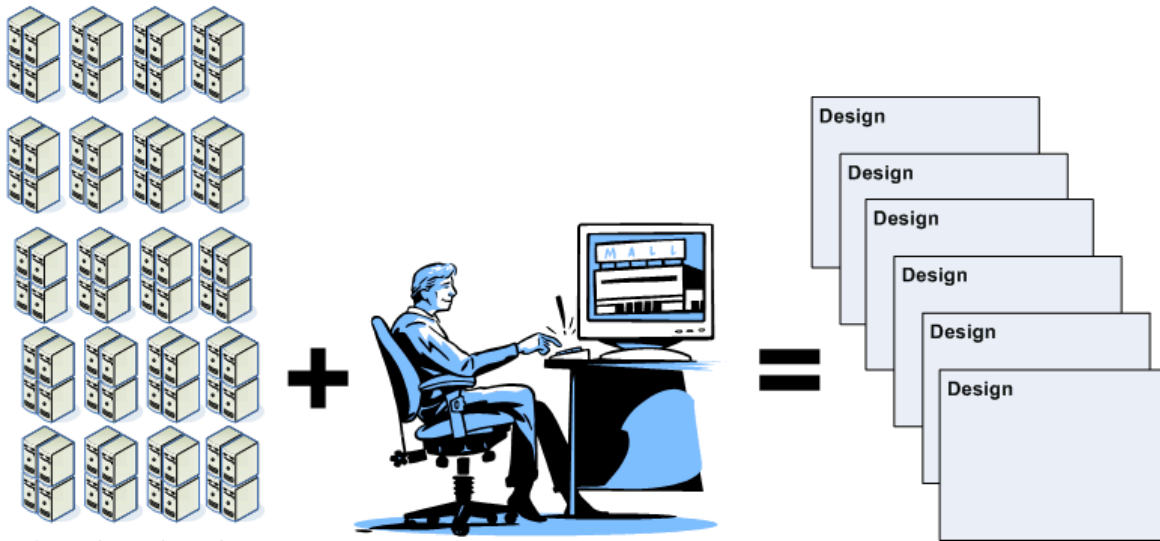
When you are examining business objectives and assessing the opportunities for offering new services, consider several factors about workflows that make them candidates for cloud computing services.

- 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).

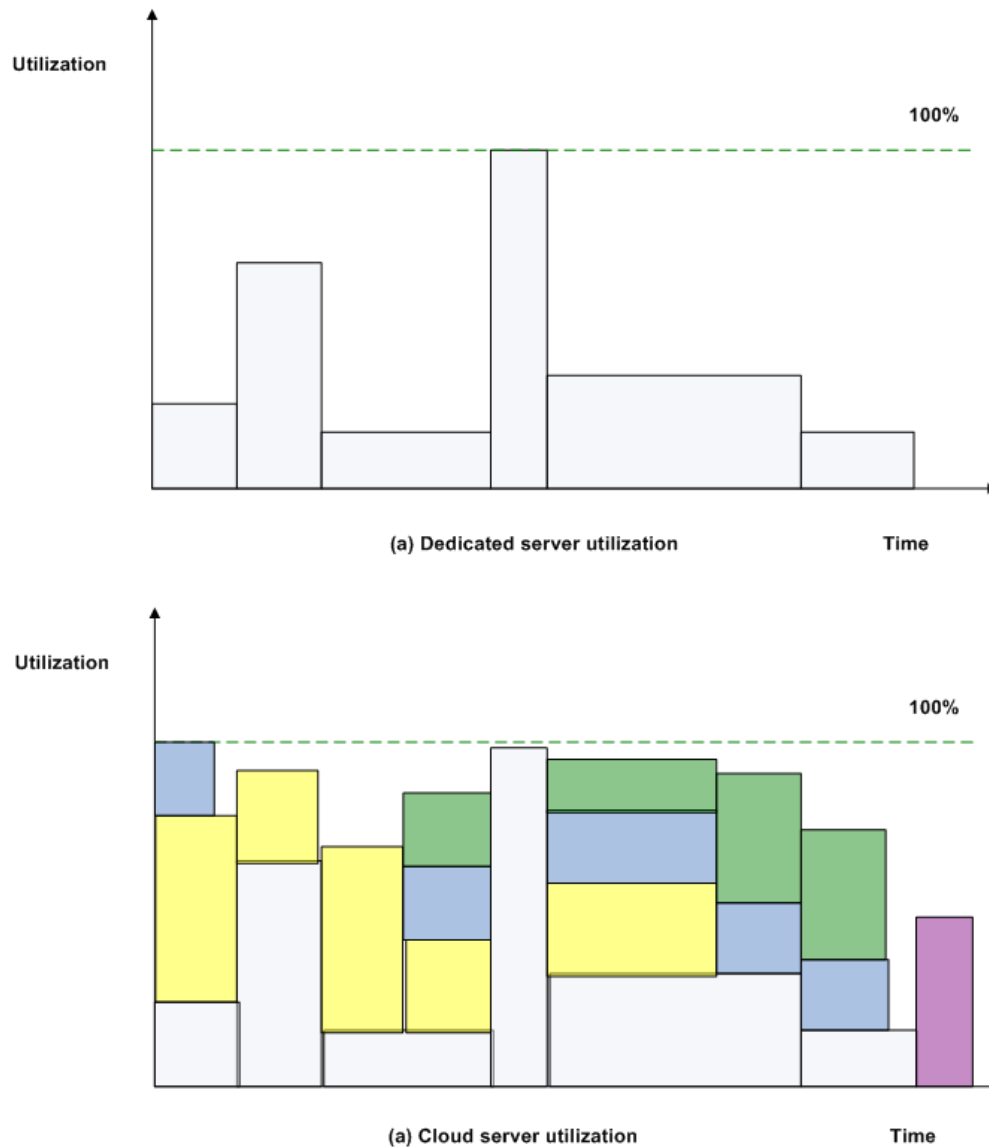


Figure 10.4: Cloud computing more efficiently utilizes computing resources than dedicating servers to single operations that have widely varying levels of demand.

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 prioritize 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
- Identifying 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).

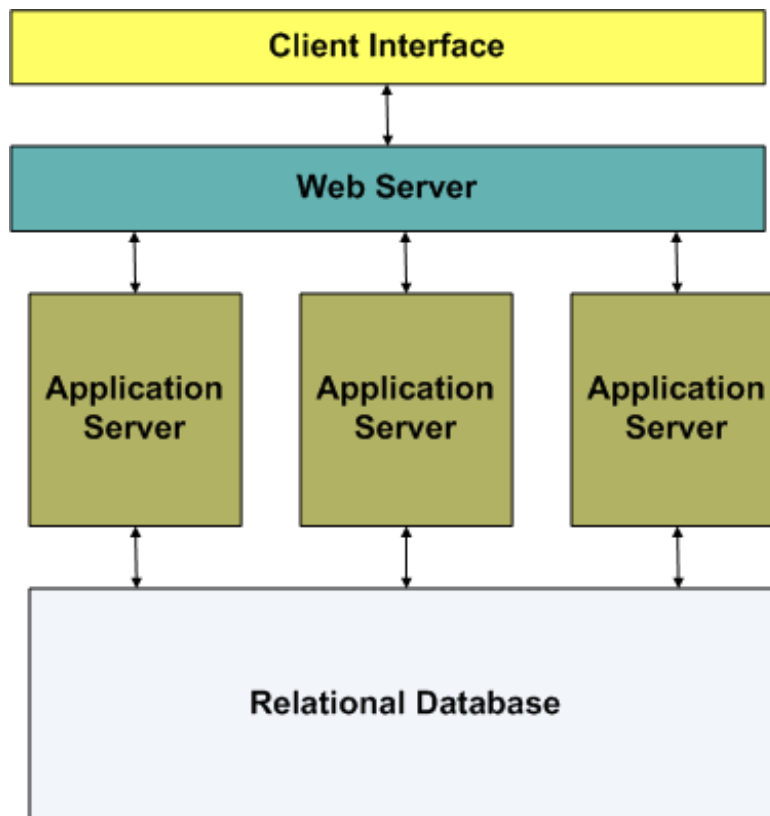


Figure 10.5: A common decentralized amenable to cloud computing model uses a multi-tier stack to implement applications.

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.

Identifying 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.

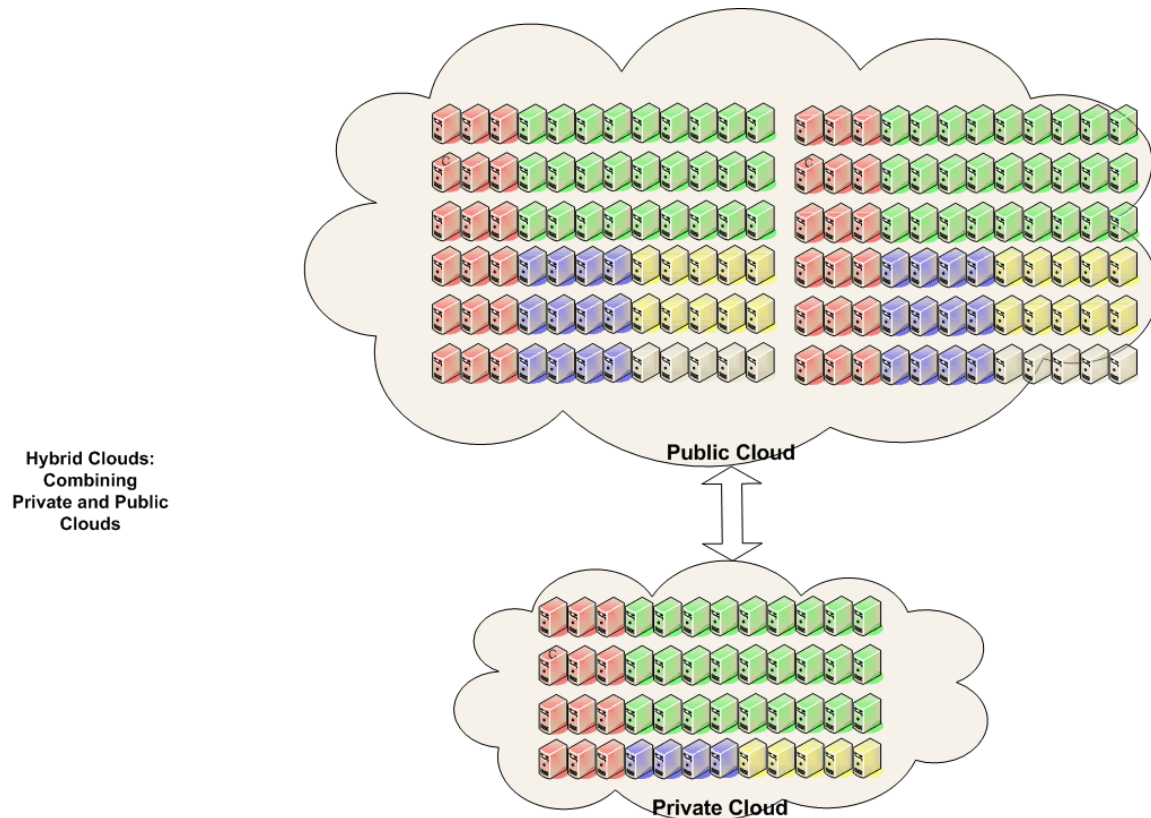


Figure 10.6: A combination of private and public clouds can enable an organization to realize the benefits of both.

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.

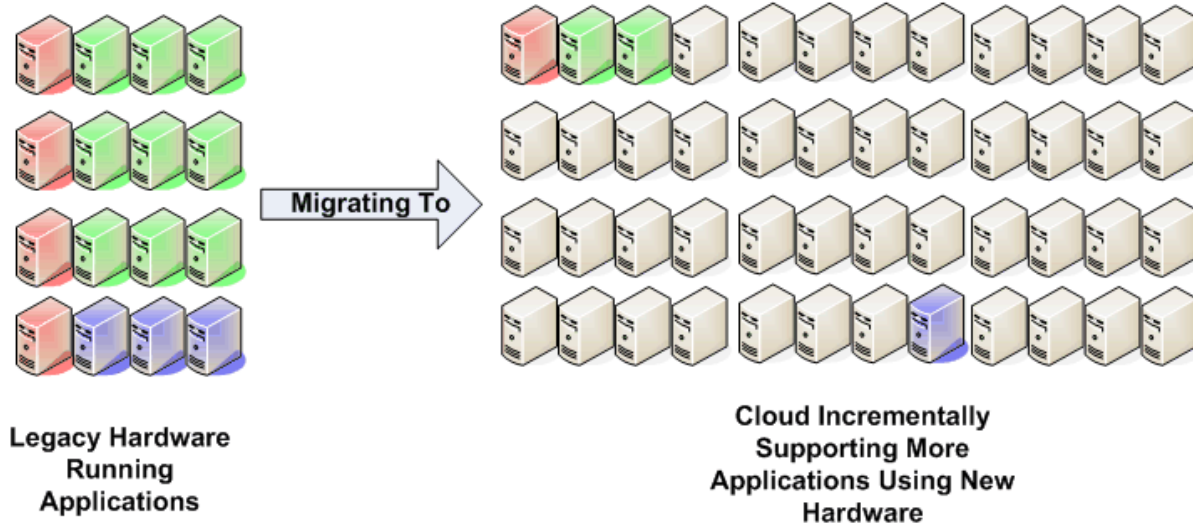


Figure 10.7: When new hardware is deployed in the cloud, applications can migrate directly to the cloud.

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.

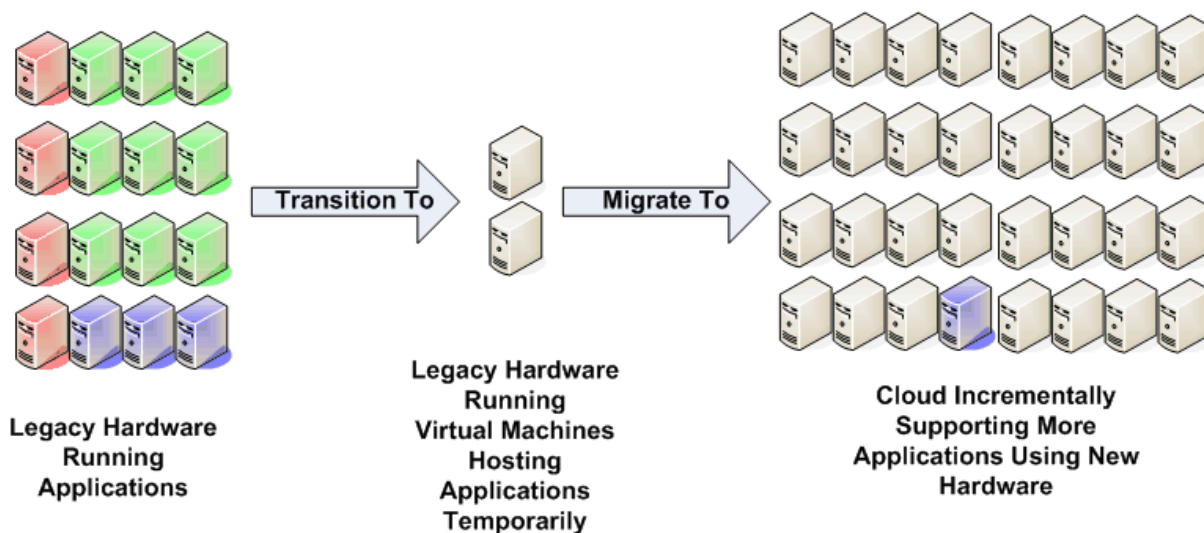


Figure 10.8: Applications may be hosted on transition virtual servers in cases where existing hardware is to be redeployed to the cloud.

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.

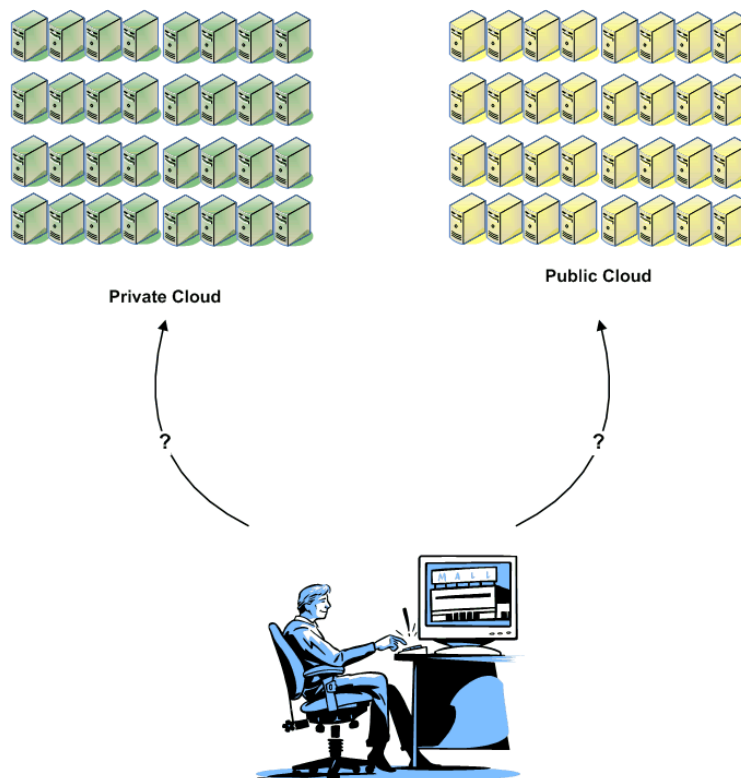


Figure 10.9: If users are given a choice of where to run applications in a hybrid cloud, policies and incentives should be in place to promote the optimal balance from an enterprise perspective.

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.

Download Additional eBooks from Realtime Nexus!

Realtime Nexus—The Digital Library provides world-class expert resources that IT professionals depend on to learn about the newest technologies. If you found this eBook to be informative, we encourage you to download more of our industry-leading technology eBooks and video guides at Realtime Nexus. Please visit <http://nexus.realtimepublishers.com>.