Realtime
publishers

The Definitive Guide™ To

Cloud Computing

sponsored by

IBM®

Dan Sullivan

Chapter 8: Roadmap to Cloud Computing: The Implementation Phase 145
Establishing a Private Cloud 146
Deploying Hardware for a Private Cloud 146
Servers and Network Equipment 146
Environmental Issues 147
Redundancy and Avoiding Single Points of Failure 147
Deploying Network Services for a Private Cloud 149
Providing Application Stacks 150
Cloud Management Services 150
Cloud Management Policies 152
Cloud Management Reporting 153
Migrating Compute and Storage Services to a Private Cloud 154
Prioritizing Based on Business Drivers 154
Business Driver #1: Cost 154
Business Driver #2: Computing Resources 155
Reallocating Servers 156
Deploying Cloud-Enabling Applications 156
Testing and Quality Control 157
Deploying Management Applications 157
Migrating End User Applications 158
Building Virtual Machine Images 158
Migrating Data to Cloud Storage 158
Migrating Access Privileges to the Cloud 158
Post-Implementation Checklist 159
Managing Cloud Services 161
Integrating Service Management with the Cloud 161
Service Catalog Management 162

Service Level Management 162

Availability Management 163

Service Validation and Release Management 163

Usage Tracking and Accounting Services 163

Capacity Planning 164

Extending a Private Cloud with Public Services 164

Summary 165

Copyright Statement

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.

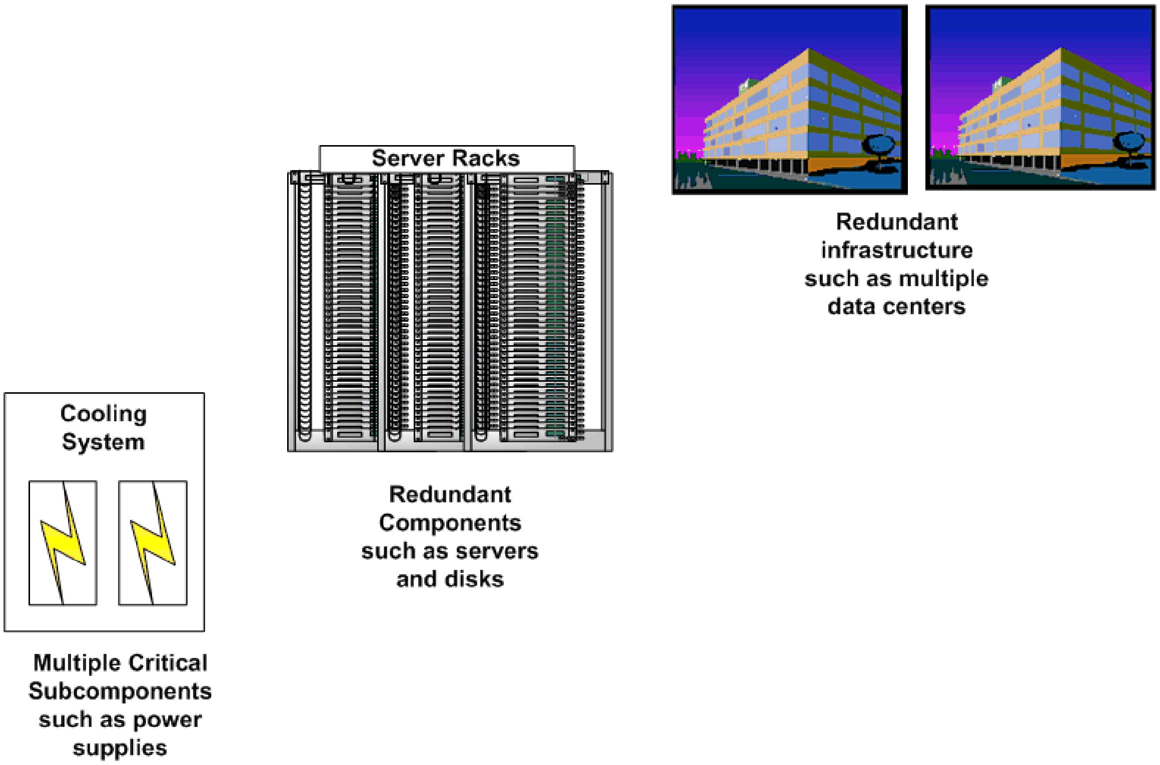

Server Racks
Redundant infrastructure such as multiple data centers
Cooling System
Redundant Components such as servers and disks
Multiple Critical Subcomponents such as power supplies

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.

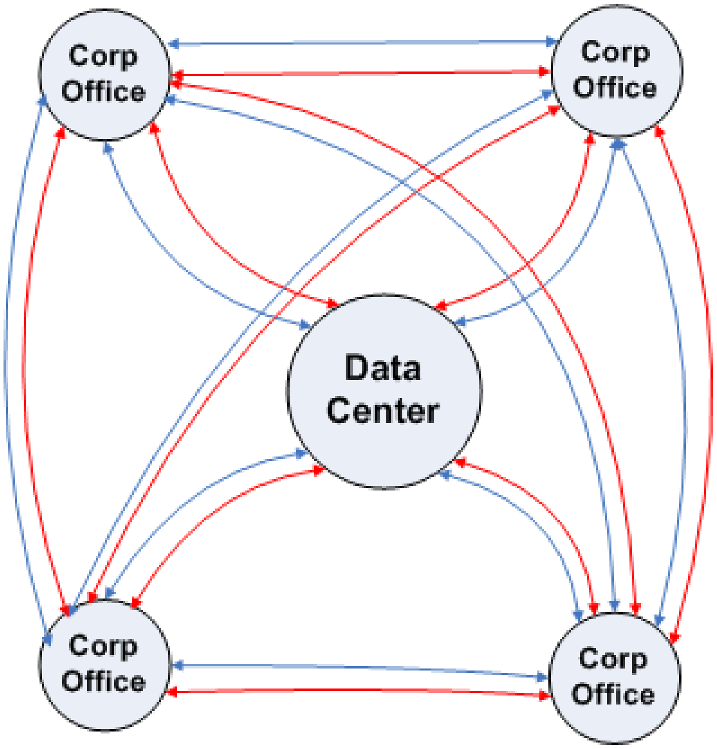


Figure 8.2. A fully redundant network requires two or more links between each interlinked network.

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.

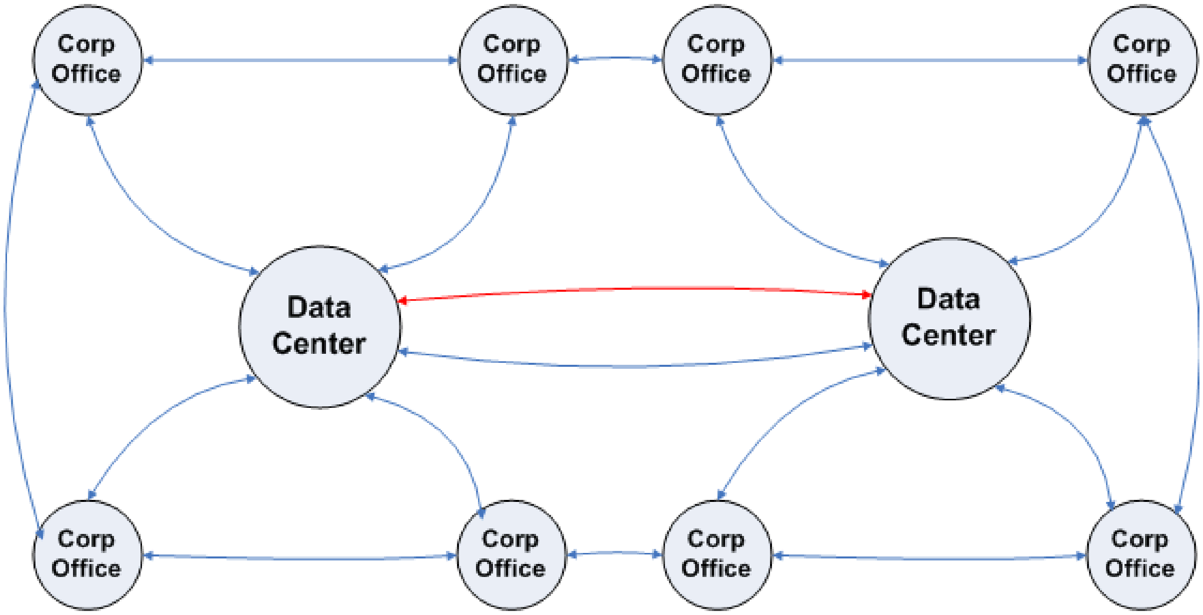


Figure 8.3: A mesh network architecture provides redundancy with fewer links than a fully redundant design.

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

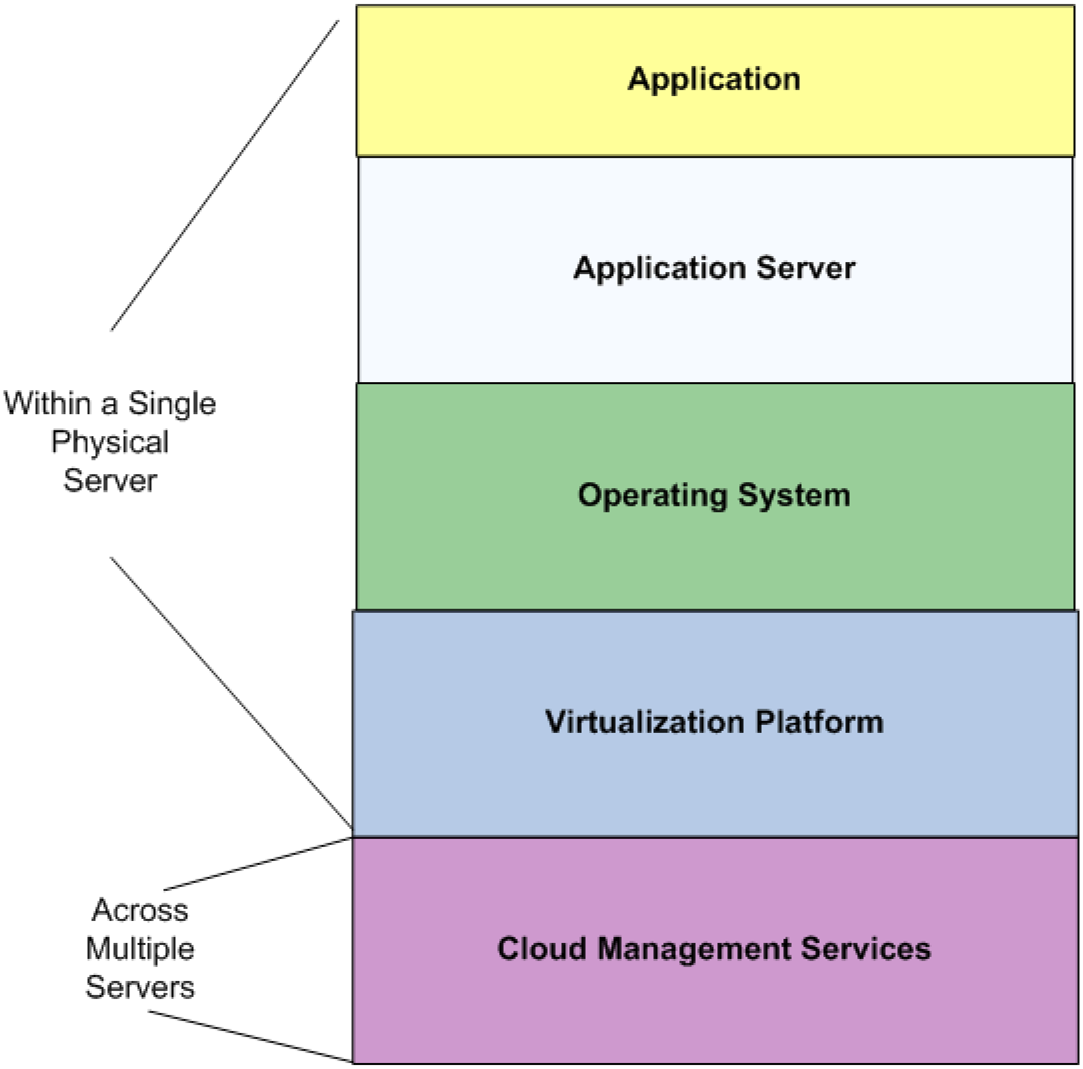
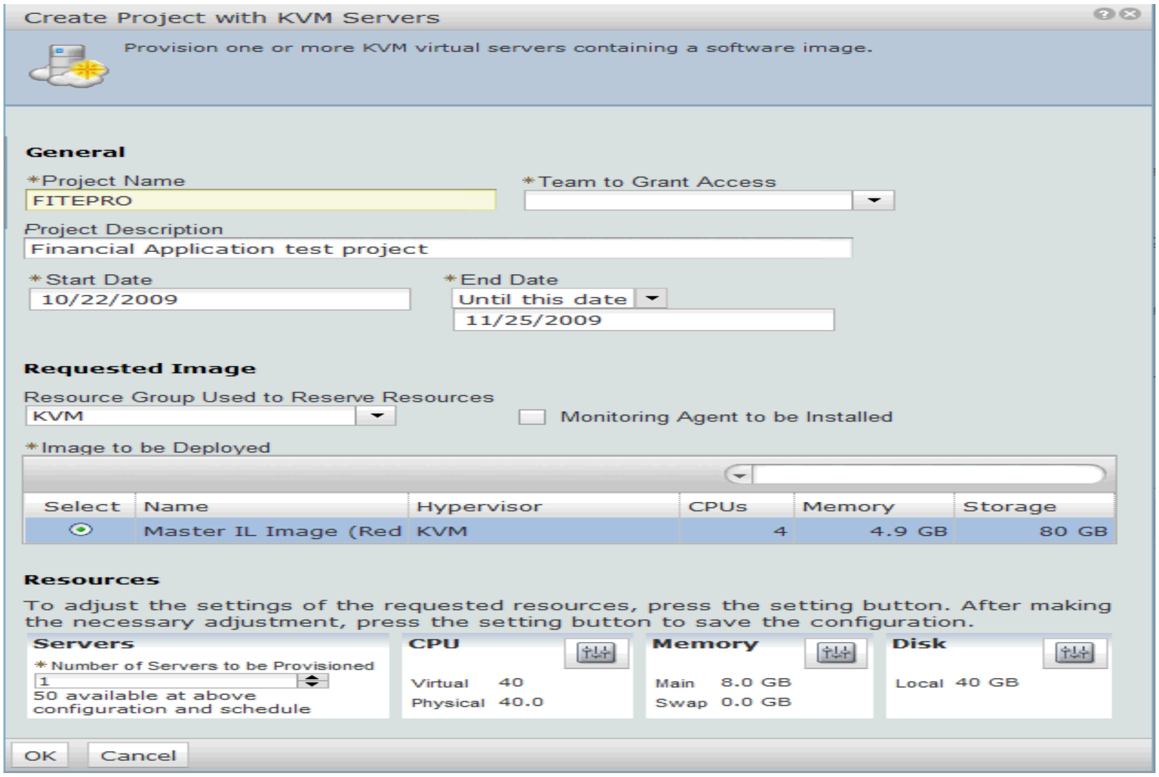


Figure 8.4: The conventional application stack is extended in the cloud to include cloud management services below virtualization services.

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.

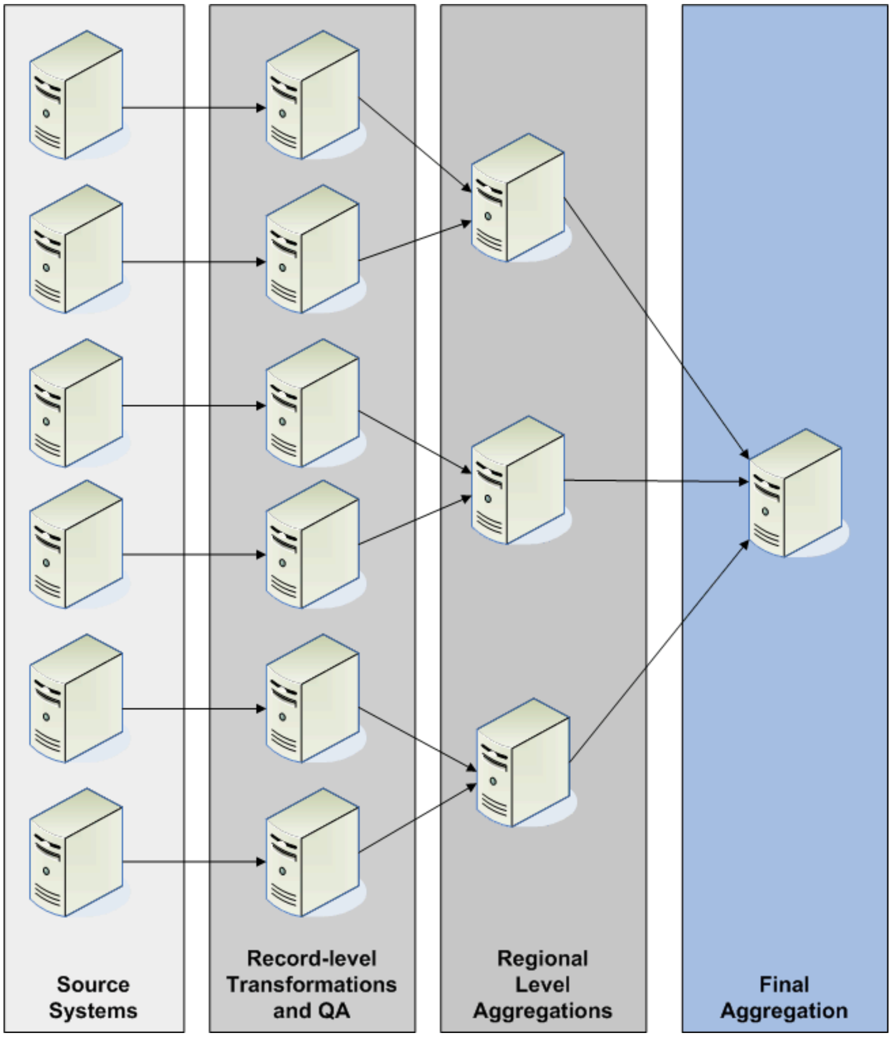

Source Systems
Record-level Transformations and QA
Regional Level Aggregations
Final Aggregation

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.

Download Additional eBooks from Realtime Nexus!