Realtime
publishers

The Definitive Guide to Cloud Acceleration

Dan Sullivan

sponsored by

CDNetworks

Copyright Statement

Chapter 1: Cloud Computing and Challenges to Delivering Services

Cloud computing is an increasingly popular way to use computing and storage technologies, and it is changing the way businesses deliver services. As with any innovation, you have to adapt your methods and procedures to take full advantage of the new technology. This guide examines how cloud computing and the architecture of the Internet shape service delivery, the challenges presented to reaching a global customer base, and techniques for accelerating content delivery. This chapter begins with an overview of cloud computing as well as key considerations for delivering services through the cloud.

Common Definition of the Cloud

Cloud computing is a model of delivering computing, storage, network and/or infrastructure in a shared manner that allows for on-demand scalability, self-service, and typically a pay-for-service pricing model.

Scalability

Scalability implies the ability to shift the amount of computing and storage as needed to meet current needs. For example, if a business experiences a spike in demand for one of its Web applications, the business might need to bring additional servers online to respond to all requests in an acceptable time.

In a cloud, these additional servers are already physically present in a data center. A cloud operating system (OS) is typically in place to deploy virtual images to additional servers and reconfigure load balancers, if required, to include the additional servers in an application cluster (see Figure 1.1).

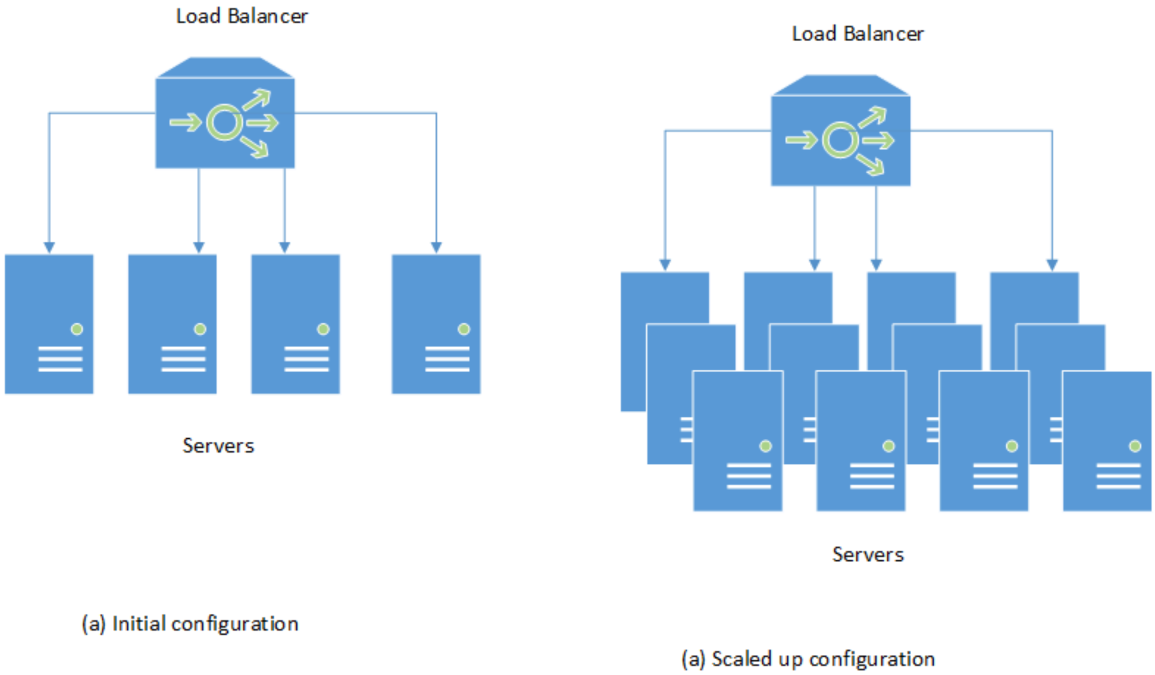

Figure 1.1: Clouds provide for rapid scalability.

Scalability implies the ability to rapidly downsize resources as well. In the given example, when the spike in traffic subsides, some of the servers would be released from the cluster and returned to the pool of cloud resources for other applications or customers to use as needed.

Storage services are treated in an analogous way in cloud computing. As more storage is required, it is allocated from a shared pool of storage resources. When it is no longer needed, storage is returned to the pool for others to use.

Self-Service

Prior to the advent of cloud computing, when an application administrator needed to scale computers to an application cluster or upgrade a server, it meant submitting requests to systems administrators and possibly provisioning additional hardware. Cloud computing platforms provide end users with the ability to provision servers and storage as needed through a cloud administration interface (see Figure 1.2).

Typically, these interfaces allow users to specify:

- The size of virtual machines to deploy
- The number of virtual machines
- The location of the data center to deploy the virtual machines
- The virtual image to deploy to each server

As clouds are virtualized computing resources, cloud providers can offer a wide range of machine configurations. For example, a small server might include 1 core, 2GB of memory, and 200GB of local storage, while a higher-end server might include 8 cores, 32GB of memory, and 1TB of local storage. Cloud users can choose the optimal configuration based on costs and requirements. CPU and memory-intensive applications might require a large and more costly server, while another application could be more cost effectively run on a number of low CPU/low memory virtual machines.

Cloud providers also maintain a catalog of virtual images. These can include a variety of OSs and preconfigured applications. If business analysts frequently work with a set of ad hoc reporting, statistical analysis, and visualization tools, the cloud provider can deploy a virtual image with these applications installed and configured so that they are readily available when needed.

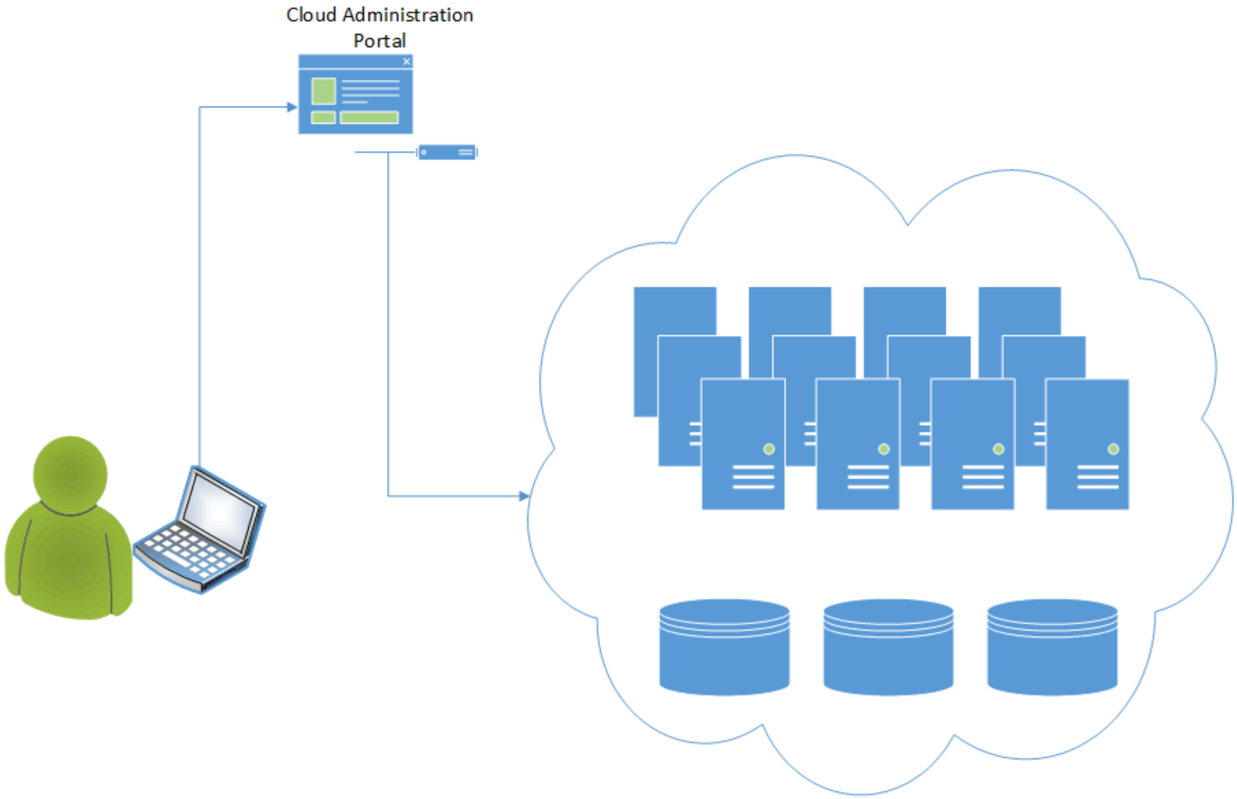

Figure 1.2: Self-service allows non-IT users to configure their own computing and storage resources.

Pay-for-Service Model

Another distinguishing feature of cloud computing is the pay-for-service model. Instead of buying dedicated hardware for an application, application managers now have the option of essentially renting resources when those resources are needed, and paying for only what is used.

Servers are typically billed in hour or minute time increments. The per-unit-of-time charge will vary with the virtual machine configuration and can range from pennies to dollars per hour per machine. Storage is usually charged based on the amount of storage used and the length of time data is stored.

Differences with Pre-Cloud Architectures

In many ways, cloud computing is not a new technology but rather a new way of using existing technologies. The building blocks of clouds—commodity hardware, virtualization platforms, widely used OSs and applications, and networking infrastructure—were all in use prior to the development of cloud computing. In spite of the similar components, there are significant differences between cloud computing architectures and pre-cloud architectures.

Pre-cloud architectures often suffered from under utilization. Systems designers would understandably configure servers for peak capacity so that applications would remain responsive under heavy but expected loads. In other cases, applications would be deployed to dedicated servers to keep them isolated from other applications and allow for OS configuration specifically tuned for that one application. A disadvantage of these approaches was that the business was paying for computing capacity it often did not use. Server virtualization helped to reduce underutilization while maintaining OS isolation, see Figure 1.3; however, virtualization was managed by systems administrators, unlike the self-service approach of cloud computing.

Prior to the cloud, there was less sharing of computing resources. Hardware is often purchased for a specific project or department, so it tends to be dedicated to that need. There are few incentives to share the resource or the cost of maintaining it. Cloud computing platforms track utilization and allow businesses to charge back to users for the resources they use. Having a charge-back system is less a technical advance than an organizational one. Now businesses can easily account for and bill for shared resources.

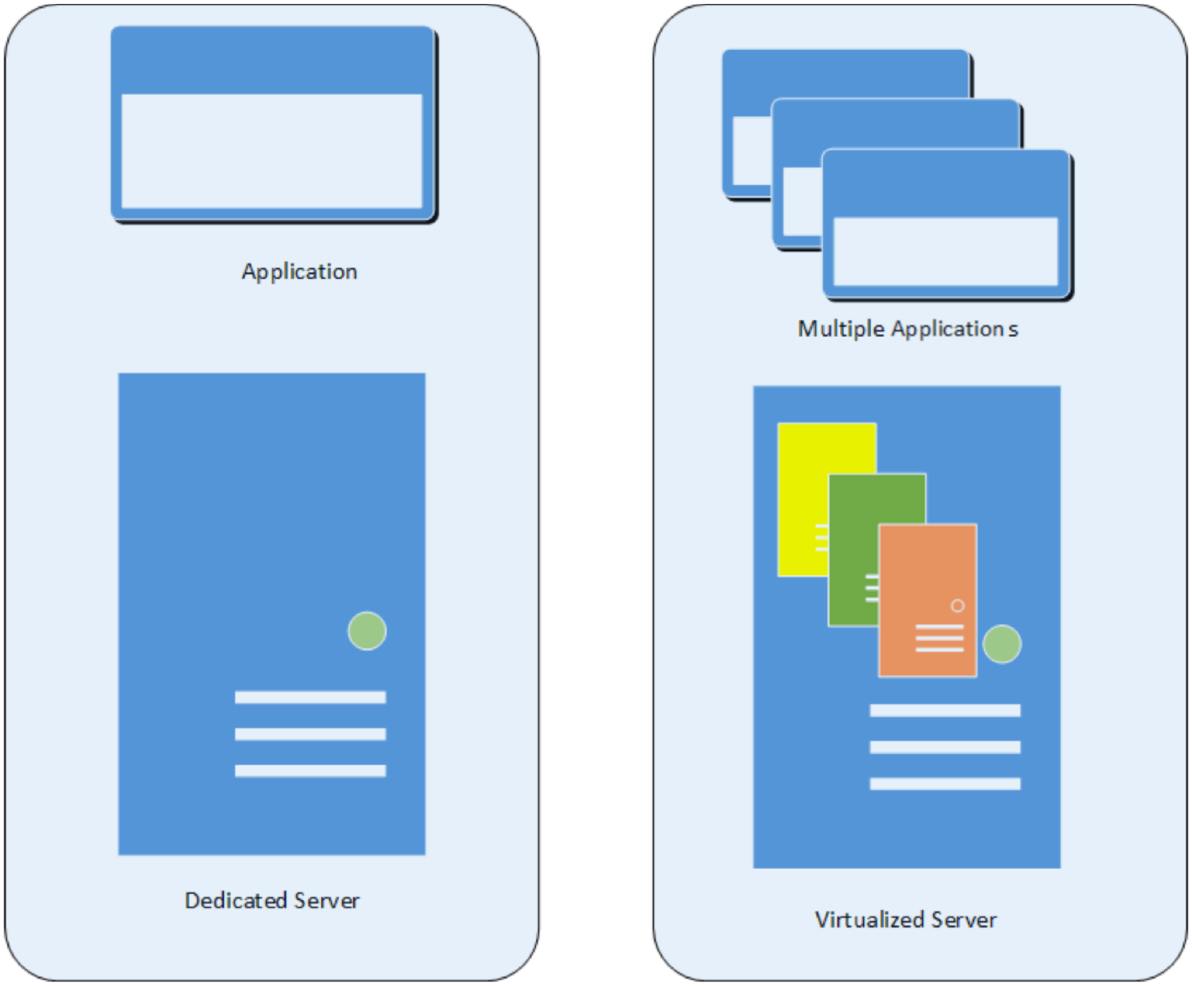

Figure 1.3: Prior to virtualization, it was common practice to dedicate a physical server to a single application or task. Virtualization allows for multiple applications to run on a single server while still maintaining OS isolation.

As previously mentioned, common characteristics of cloud computing include scalability, self-service administration, and pay-for-service charges. This combination of features has enabled more efficient use of computing and storage services and underlies more innovative use of computing resources. Starting with these three essential characteristics of cloud computing, three distinct deployment models have emerged.

Categorizing Clouds

Cloud computing services can be categorized according to who is granted access to the cloud and by the types of services offered by the cloud.

Cloud Access Models

Clouds can be categorized according to who is granted access. Three typical access models are:

- Public cloud
- Private cloud
- Hybrid cloud

Each of these deployment models has its benefits and drawbacks.

Public Clouds

Public clouds are essentially open to any user. Many cloud providers are well known in the IT industry and include Amazon, Microsoft, Google, IBM, HP, and Rackspace. One of the advantages of a public cloud is the low barrier to entry: virtually anyone with a credit card can set up an account and provision resources.

Also, public cloud providers have the advantage of specializing in cloud services offerings. They realize economies of scale, can invest in specialists to design and maintain their infrastructure, and can raise the capital required to deploy substantial cloud services. Common characteristics of public cloud providers include:

- Maintain multiple data centers
- Have redundant networks
- Have sufficient compute and storage resources to meet demand
- Provide standard service level agreements (SLAs)

Public cloud providers distinguish themselves more on specialized services than on price. For example, a cloud provider might offer a high-performance computing cluster designed with high-speed network interconnects for low latency and flash drives for improved I/O performance. In other cases, a provider might offer a low-cost storage service for archiving, private networks for added security, or accounting and billing services tailored to enterprise customers.

Although public clouds may offer a combination of commodity and specialized services, they do not always meet the needs of enterprise customers. For example, some public cloud offerings might not meet the requirements of industry regulations such as the Payment Card Industry Data Security Standard (PCI DSS). Retailer businesses and others using payment cards would not be able to run applications or store data subject to PCI DSS in those clouds and still remain in compliance.

Some businesses may not allow confidential or sensitive data to reside on servers or storage systems outside of corporate control due to concerns about data leaks and loss of confidentiality. However, data can be readily encrypted before it leaves corporate control. Depending on jurisdiction, businesses may be required to keep confidential and private information within the jurisdiction or within a partner jurisdiction with equivalent privacy protections.

Although the benefits of public cloud computing are well understood, for some business cases, a private cloud may be a more appealing option.

Private Clouds

Private clouds are controlled by organizations behind their firewalls and limit access to the cloud to organization members or partners. Large businesses and governments can have the need for and resources to build and maintain private clouds. Fortunately, businesses do not need to start from scratch to build a private cloud; IT vendors offer cloud computing packages that include the hardware and software required for a private cloud.

The single most significant benefit of a private cloud is that the organization deploying it maintains full control:

- Determining who has access to cloud resources
- Defining policies and procedures for allocating cloud resources
- Specifying charge-backs for services
- Implementing specialized software services, for example, a message queue, or hardware, such as flash storage devices
- Implementing monitoring and auditing procedures according to the organization's particular needs

The obvious drawbacks of private clouds are the capital expenditure to acquire the infrastructure and the ongoing costs of maintaining a private cloud. If resiliency is required for your business' cloud applications, you will probably need to maintain multiple data centers.

One option for private clouds is to locate your infrastructure in a third-party data center. This option affords some economies of scale and specialization of labor with regards to managing the physical infrastructure and redundant network services. The business still retains control over the computing and storage infrastructure, so many of the benefits of an on-premise private cloud remain in place.

Community Clouds: Part Public, Part Private

The commonly used public-private dichotomy does not cover all options with regards to cloud access models. The community cloud, sometimes referred to as a gated community model, has characteristics of both private and public clouds. Community cloud providers screen potential customers before granting them access to clouds.

This setup is designed to ensure that only legitimate organizations that meet the vendor's criteria can make use of the community cloud. For example, a community cloud provider specializing in healthcare might accept only healthcare provider and insurers as customers. This model allows the vendor to specialize services to their target market, such as providing more in-depth auditing information to meet Health Insurance Portability and Accountability Act (HIPAA) compliance regulations.

Hybrid Clouds

A hybrid cloud, as the name implies, is a combination of private and public clouds. The model was developed by the desire for the benefits of both private and public clouds. In a hybrid cloud, jobs and data that need to stay within the corporate network can run on the private cloud while other jobs and data can be shifted to a public cloud provider, as Figure 1.4 shows. This approach can reduce the demand for private cloud resources and therefore reduce the capital expenditure needed to establish a private cloud.

Maintaining a hybrid cloud introduces challenges not encountered with the other models. If the cloud OSs running in the private and public clouds are not compatible, you might find yourself maintaining two catalogs of virtual images as well as two access control systems. Accounting and billing might also require different systems and create additional work to integrate. Using the same cloud OS—for example, OpenStack—in both the public and private clouds can reduce integration challenges. Compatible cloud OSs, such as the Amazon AWS platform and Eucalyptus, are not the same but use common APIs that can reduce the challenges to implementing a hybrid cloud.

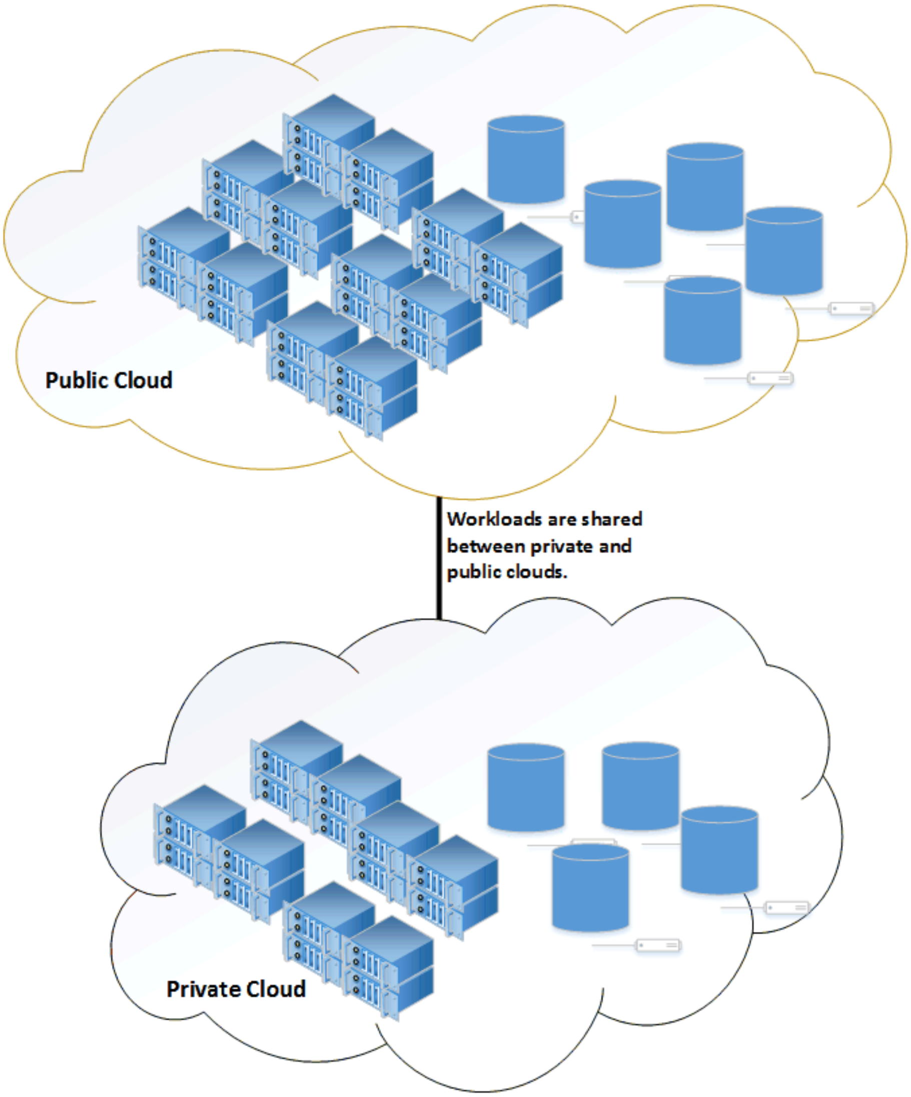

Figure 1.4: Hybrid clouds combine private and public clouds and allow for workloads to move between the two.

Public, private, and hybrid clouds can all be used to deploy services for the benefit of customers, partners, and employees. The choice of the most appropriate access model will vary according to security, compliance, performance, and cost constraints.

In addition to categorizing clouds by access model, it is common to distinguish public clouds by the types of services offered.

Cloud Service Models

Clouds are often grouped into one of three service categories:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)

These categories offer increasing levels of specialization and reduced levels of management overhead.

Infrastructure as a Service

IaaS clouds offer access to virtual servers, storage, and related services. Cloud users provision virtual servers and storage as needed, and manage all aspects of the infrastructure at the OS level and above (see Figure 1.5). This option gives users substantial control over the size of virtual servers used, the software installed, and the way storage systems are utilized.

This model also imposes the most responsibility on the cloud users. For example, software engineers using a public cloud for development would need to select an appropriate-size machine, load a virtual image with an appropriate OS, install additional tools if needed, and configure persistent storage.

IaaS solutions are good choices when you need to maximize control over the OS, applications, and storage options. Alternatively, if you need less control over the infrastructure, a PaaS cloud may be a suitable option.

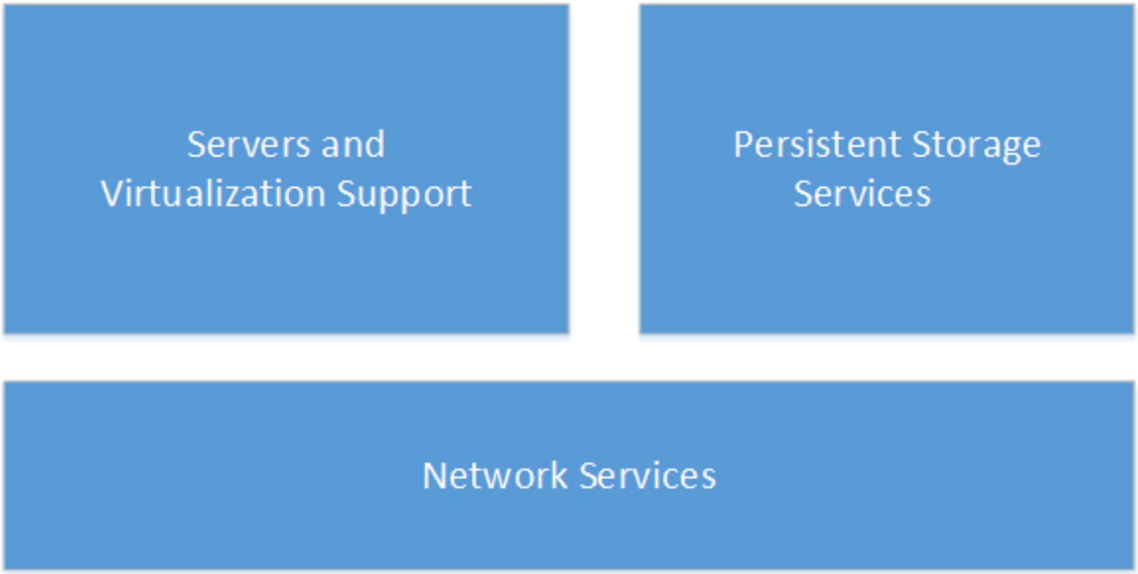

Figure 1.5: Infrastructure as a Service provides primarily computing, storage, and networking services.

Platform as a Service

PaaS clouds provide access to application services while alleviating the need for device management (see Figure 1.6). For example, a developer might use a PaaS cloud to run a large number of tests on a new software. The developer can choose the appropriate number of preconfigured servers and submit the job without needing to set up the servers themselves.

PaaS can also reduce the time required to set up and manage application stacks. Instead of setting up application and database servers, PaaS users can use the application and data management platforms provided by the PaaS cloud. Google App Engine, for example, allows software developers to run their Java or Python applications on Google infrastructure without the need to manage virtual machines. Microsoft Windows Azure cloud includes a relational database service, Azure SQL, which a business can use instead of managing its own Microsoft SQL Server instance. The lines between IaaS and PaaS are sometimes blurred, as IaaS providers offer services, such as databases and messaging services, as part of their IaaS services.

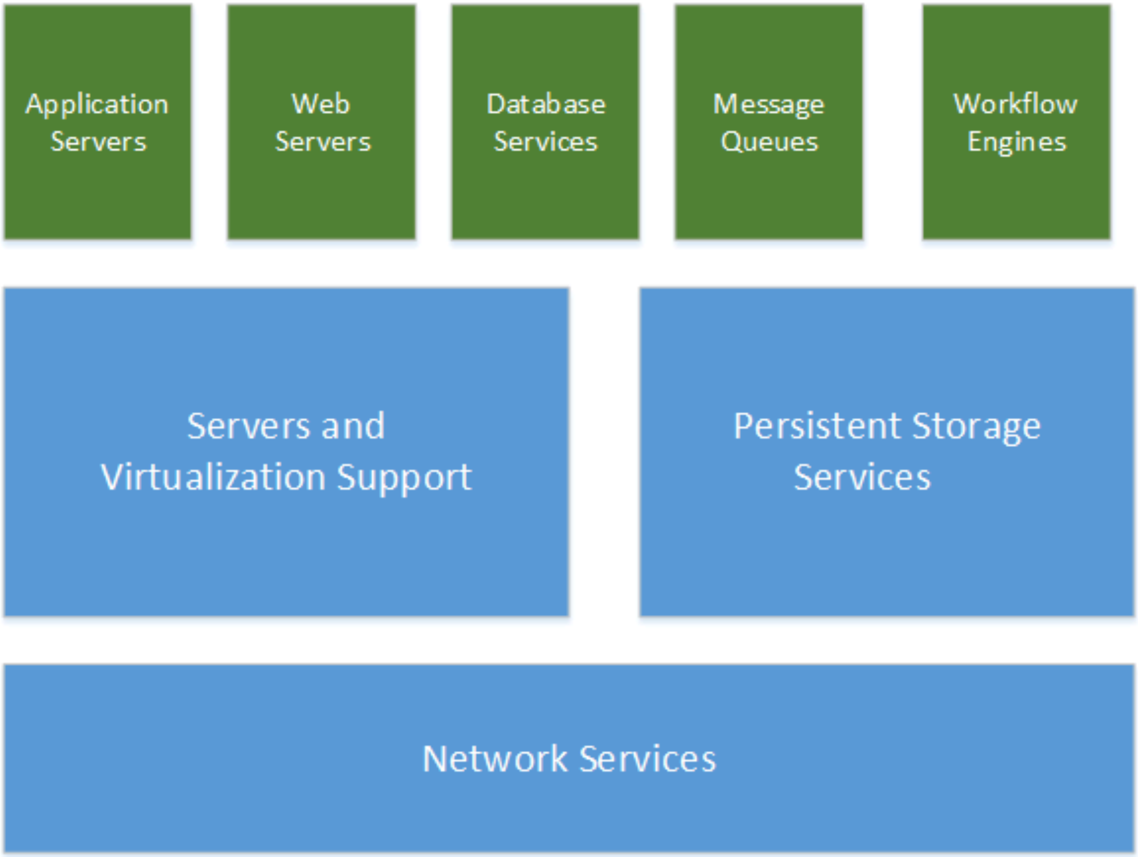

Figure 1.6: Platform as a Service extends the IaaS level of services to include application stack services.

Software as a Service

The third category of cloud service type, SaaS, provides fully functional applications to end users. Applications as different as word processing and customer relationship management (CRM) are available from SaaS providers. A key advantage of the SaaS model is that users do not have to manage any part of the infrastructure. Some applications will require end users to configure access controls and program options and other application settings, but the SaaS provider manages all aspects of the computing, storage, and network infrastructure, as Figure 1.7 illustrates.

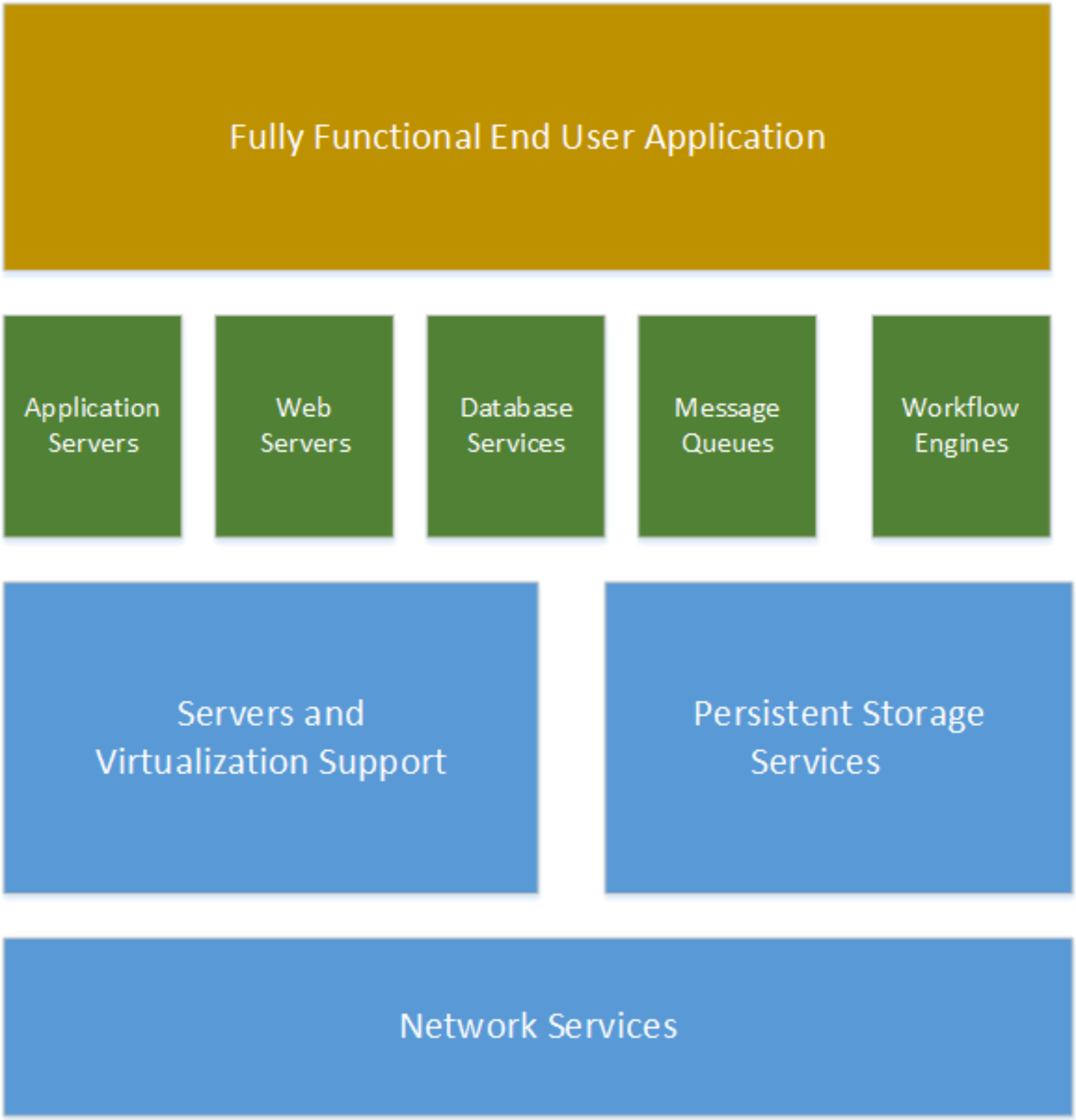
Fully Functional End User Application
Application Servers
Web Servers
Database Services
Message Queues
Workflow Engines
Servers and Virtualization Support
Persistent Storage Services
Network Services

Figure 1.7: Software as a Service provides turnkey applications that minimize the demands on end users to set up and configure the application.

SaaS has created opportunities for both SaaS consumers and SaaS providers. Users of SaaS services can reduce or eliminate the need to maintain specialized applications in-house or in a cloud. For example, an architecture firm using a SaaS for managing its financials can avoid having to run a financials package in-house and may be able to reduce the number of staff dedicated to supporting the financial package. SaaS providers have opportunities to create services that might not be efficiently implemented within a single organization. For example, a SaaS that provides HIPAA-compliant records management services could find a large market of small and midsize healthcare providers interested in their services. SaaS providers may implement their applications in public, private, or hybrid clouds.

Application Response Time and Benefits of Cloud Acceleration

Cloud computing and the global reach of the Internet has created opportunities for businesses to expand their markets and customer base. The scalability and elasticity of cloud computing allows businesses to grow their computing systems according to their business demand. This flexibility lessens the need to make capital expenditures for hardware that might be needed in the future. It also allows operators to make decisions about provisioning compute and storage services at a much more fine-grained level. If there is a peak demand for a day or two, then additional servers can be provisioned in the cloud. When demand then subsides, those servers can be released. Compute and storage elasticity are essential parts of maintaining quality of service. They are not, however, the only factors.

Adverse Effects of Slow Application Response Time

From a customer's perspective, the quality of an application is determined in part by its responsiveness. Applications that appear to run slowly are problematic from a user's perspective and can lead to user dissatisfaction and lost revenue. A number of studies have demonstrated a correlation between application response time and discontinued use of a Web-based application. According to a study by the Aberdeen Group, a 1-second delay in page load times can result in:

- 11% fewer page views
- 16% decrease in customer satisfaction
- 7% loss in conversions

Another set of findings published by KissMetrics reveals that:

- 73% of mobile device users report encountering Web sites that were slow to load
- 47% of consumers expect Web pages to load in 2 seconds or less
- 40% abandon sites that take more than 3 seconds to load
- 79% of shoppers who are dissatisfied with the site's performance are less likely to buy from that site again

Clearly, the responsiveness of an application can have a direct impact on customer satisfaction, loyalty, and ultimately revenue.

Improving Application Response Time

Many factors contribute to application responsiveness, such as the way the application code is written, the way the database has been designed, and network throughput and latency.

Software-based Options

One way to improve performance is to tune application code. This task can include:

- Selecting more efficient algorithms
- Analyzing code to identify time-consuming functions
- Re-writing database queries to reduce the amount of data returned
- Tuning database design by implementing additional indexes and other measures to reduce I/O operations performed by the database

Improving software can yield significant improvements in some cases, but these improvements can be costly and may require more time than other options to implement.

Hardware Options

The cloud also allows businesses to implement a well-known but sometimes questionable practice of "throwing more hardware at the problem." Rather than review and revise code, it might be faster to simply scale up the servers that are running the code. One could scale vertically by deploying the application to a server with more cores and memory and faster storage devices. Alternatively, applications that lend themselves to distributed workloads can scale horizontally. This action entails adding additional servers to a load-balanced cluster and allowing the load balancer to distribute the work among more servers.

Both of these scenarios can help improve performance, assuming there are no bottlenecks outside the servers (for example, the time required to perform I/O operations on a storage array). If I/O performance is a problem, you might be able to improve performance by switching to faster storage technology.

Network Issues and Cloud Acceleration

Although tuning application code and database design can often improve the throughput of servers, they do not always improve application response time. Network latency, or the time delay in sending data between two networked devices, cannot be improved by tweaking algorithms on the server or optimizing database queries. Within a data center, cloud providers may offer higher performance networking infrastructure for specialized tasks, such as high-performance computing. These specialized jobs may run on clusters with 10Gb Ethernet while most common jobs run on servers interconnected with slower, interfaces. For data that is sent outside the data center and over the Internet, additional measures are required to reduce latency.

Cloud Acceleration

In this guide, the term *cloud acceleration* refers to cloud techniques for improving the overall responsiveness of an application by reducing the time it takes to deliver content to an end user. Without going too deeply into technical details in this chapter, it is worth noting that cloud acceleration can be implemented with a combination of content delivery networks for distributing content around the globe and reduced network traffic using specialized optimization.

Challenges to Cloud Acceleration

The remainder of this guide will delve into the technical details of cloud acceleration techniques; for now, this chapter will briefly examine four challenges to implementing cloud acceleration:

- Scalability and geographic reach
- Redundancy
- Consolidation of services
- Cost

Each of these challenges must be addressed to successfully implement a cloud acceleration solution.

Scalability and Geographic Reach

Networking is constrained by physics as well as engineering. We will never tweak the laws of physics to improve the speed with which we can transmit signals. Although an organization can improve the engineering of its networking hardware, the business is still dependent on the infrastructure used by Internet service providers (ISPs) around the globe.

Content delivery networks (CDNs) compensate for network limitations by maintaining copies of data around the globe and responding to user requests for content by using the closest facility to the end user and providing the best path between endpoints. A customer in Amsterdam, for example, might be served from content stored in a data center in Paris, while a customer in Shanghai receives the same content from a data center in Singapore (see Figure 1.8).

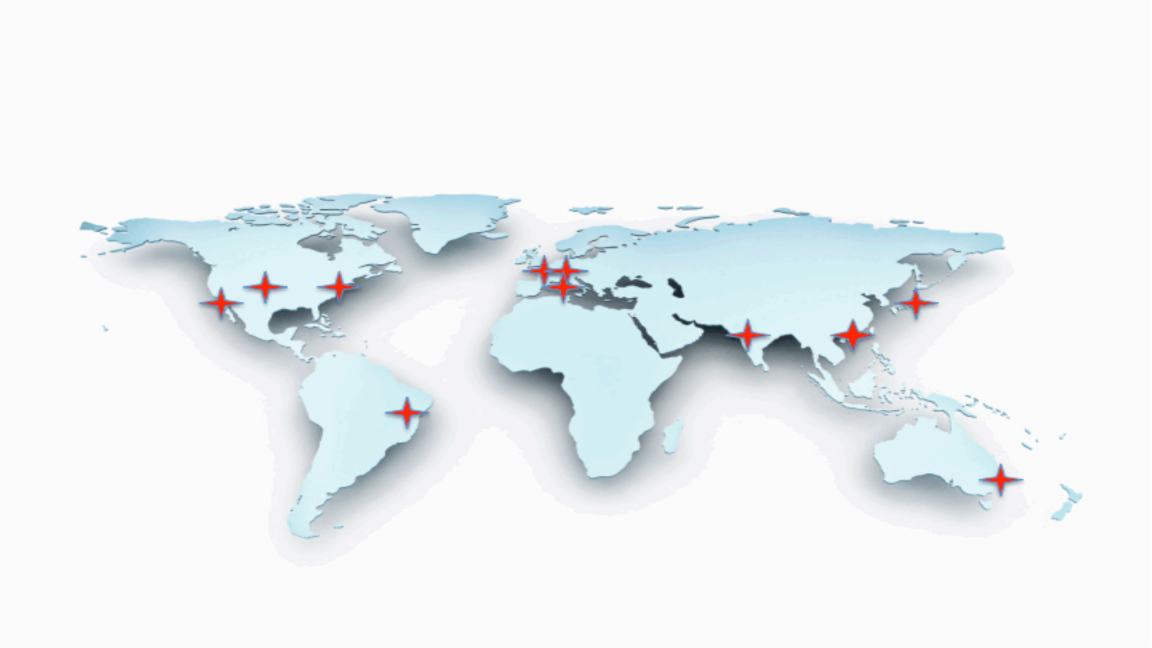

Figure 1.8: Global data centers are essential for geographically distributing replicated content.

Businesses can deploy and maintain their own data centers or infrastructure within co-location facilities around the globe. Such a deployment would have to have sufficient global reach to respond to customers, employees, and business partners wherever they may be. These deployments would also have to include sufficient hardware to scale to meet the peak demands each data center would encounter.

Redundancy

Redundancy is another consideration. Hardware fails. Software crashes. Networks lose connectivity. If a data center were to fail, other data centers around the globe should be configured to respond to traffic normally handled by the failed site.

Redundancy also entails maintaining up-to-date copies of content. Replication procedures should be in place to ensure that content is distributed to all data sites in a timely manner.

Consolidation of Services and Costs

If a business is going to all the effort and cost to deploy cloud acceleration systems, it is best to capitalize on that investment by consolidating services and applications that can benefit. As with private clouds, there is the potential for significant capital investment to establish and maintain cloud acceleration infrastructure. Ongoing maintenance costs will add to the overall operational expenses of the organization as well.

> **Reference**
>
> Later chapters will examine options for addressing these challenges.

Key Considerations for Deploying Cloud Applications

Along with the technical challenges to implementing cloud acceleration technologies, it is important to consider other characteristics that influence how a business can improve application responsiveness. One factor that determines the optimal cloud acceleration technique is the use of generated versus reusable content. Reusable content, sometimes referred to as static content, can be replicated and sent from Web servers without additional processing by an application. Reusable content includes material such as information from product catalogs, documents, and general information Web site pages. Generated content is the result of some application process, such as querying a database to retrieve a customer's order history. Reusable content can be replicated to data centers around the globe; dynamically generated content cannot. Instead, dynamically generated content can benefit from optimization techniques that improve throughput and latency between data centers.

Other factors one must contend with when providing services on a large geographic scale are a function of the design of the Internet. For example, the Internet is comprised of multiple ISPs working together to route data as needed across different ISPs' networks. Congestion at the physical interconnection of networks can adversely impact application performance (see Figure 1.9). This and other issues that derive from the large-scale architecture of the Internet drives the need for multiple data centers in geographically dispersed arrangements.

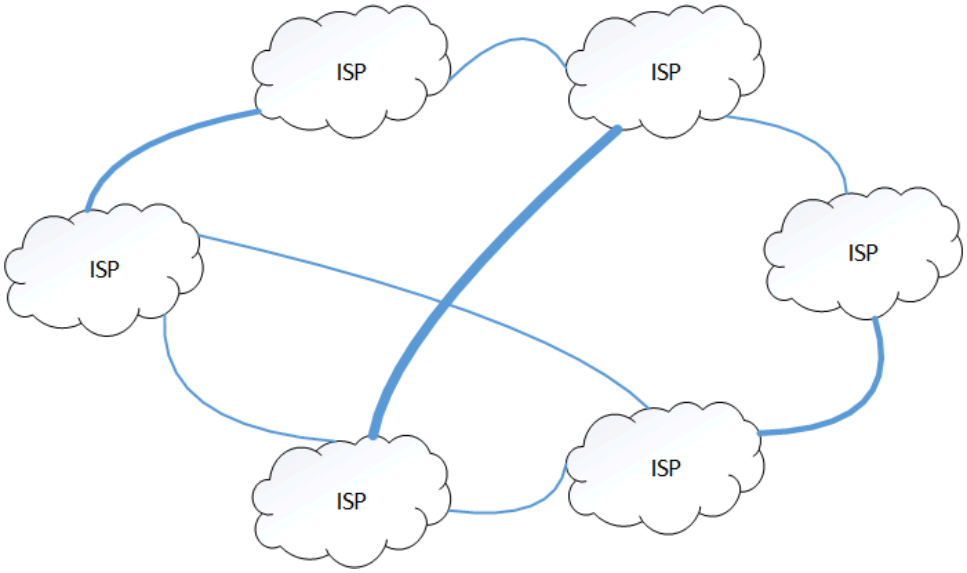

Figure 1.9: The rate of data exchange between ISPs will depend on multiple factors, including the topology of the network. Congestion at the links between ISPs can contribute to high latency in global Web applications.

In addition to differences in infrastructure, ISPs may have different business perspectives on linking with other ISPs. In the most basic scenario, ISPs view their relationships as reciprocal and pass traffic between ISPs without compensation. In other cases, one ISP may believe another ISP gains more from a peering relationship and therefore requires payment to accept traffic from and send traffic to the other ISP. Competition between ISPs can limit data exchange as well. Both technical and business considerations can affect the flow of your application traffic around the globe. Although most businesses cannot directly influence their ISP's business model and relationships with other ISPs, businesses can work around the limitations imposed by peering arrangements by using cloud acceleration techniques.

Cloud providers can also be a potential network bottlenecks. If their networking services are insufficient for an organization's needs and the provider's distribution of data centers is not enough to compensate for network congestion and latency issues, alternative cloud acceleration options may be required.

Summary

Cloud computing is creating opportunities for businesses to expand their reach to a global scale. The cost and complexity of deploying computing and storage services is lowered with cloud computing. There is also greater flexibility to adapt to new business opportunities by leveraging IaaS and PaaS platforms to create new applications and services. The increasing adoption of SaaS platforms also presents an opportunity for businesses to offer their services in a SaaS model. Businesses must pay particular attention to Web application performance for all customers regardless of those customers' locations. Adding servers and storage will improve some but not all aspects of application responsiveness. Cloud acceleration techniques may be required to ensure consistent and acceptable levels of performance for all application users.