Realtime publishers

The Definitive Guide™ To

Cloud Computing

sponsored by

IBM®

Dan Sullivan

Introduction to Realtime Publishers

by Don Jones, Series Editor

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

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

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

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

Until then, enjoy.

Don Jones

Copyright Statement

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

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

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

Overview

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

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

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

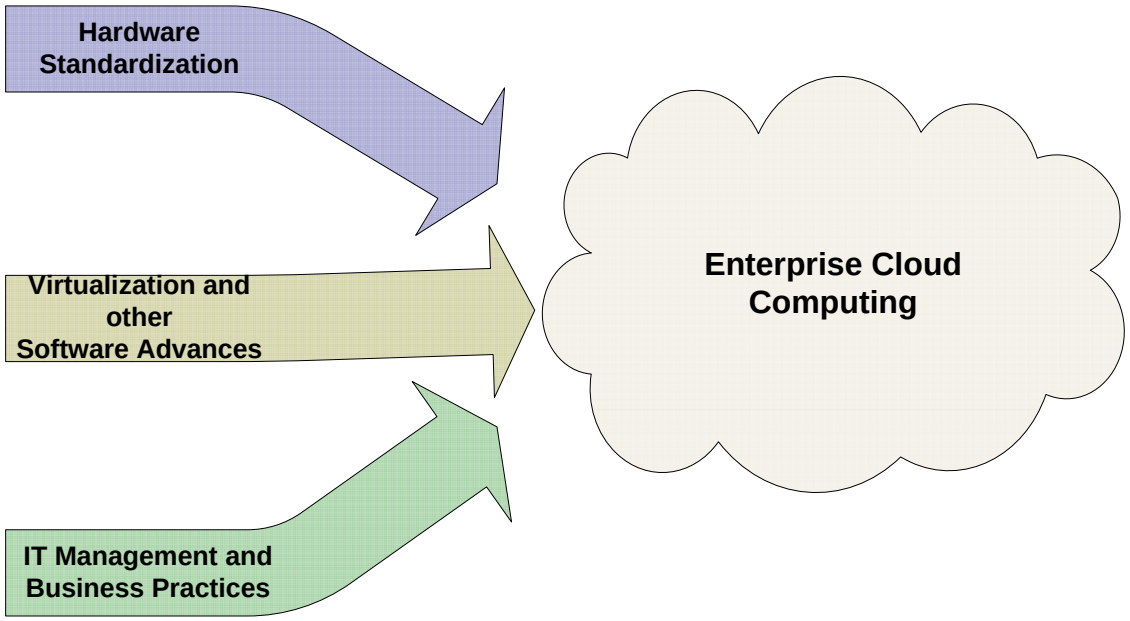

Figure 1.1: Enterprise cloud computing is the product of the confluence of advances in three distinct areas: server hardware standardization, virtualization and other software advances, and IT management and practices. Without all three, enterprise cloud computing would not be possible.

The Moving Target that Is Cloud Computing

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

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

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

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

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

A Brief Introduction to Cloud Computing

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

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

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

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

A Massively Scalable Infrastructure

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

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

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

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

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

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

Rapid Allocation of Virtual Servers

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

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

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

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

Standard Hardware Platform

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

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

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

Persistent Storage in the Cloud

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

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

The Potential Network Bottleneck

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

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

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

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

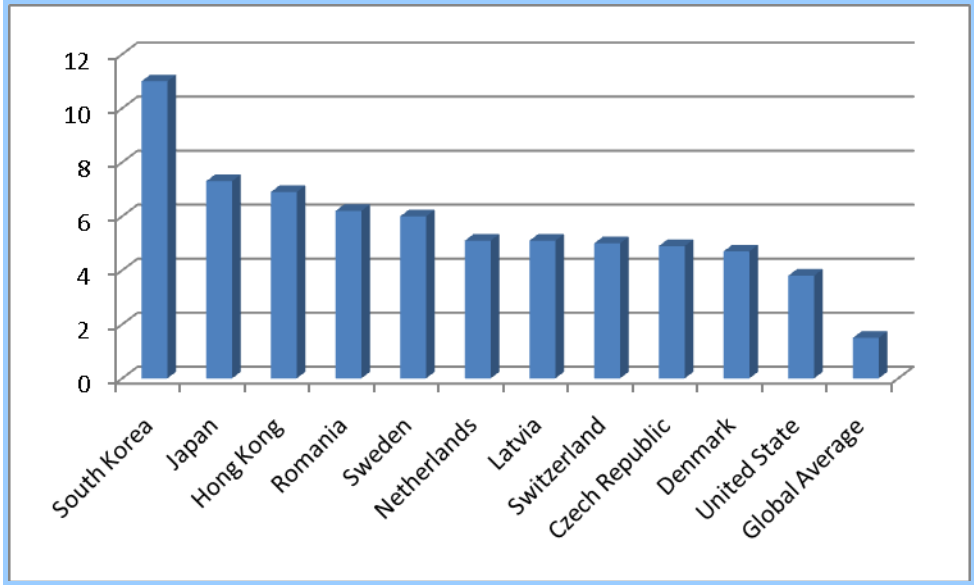

Figure 1.3: Average national broadband speeds (Mbps) vary widely by region (Source: The Akamai State of the Internet Report 2nd Quarter 2009. Volume 2 Number 2).

Universal Access

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

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

Fine-Grained Usage Controls and Pricing

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

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

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

Standardized Resources

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

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

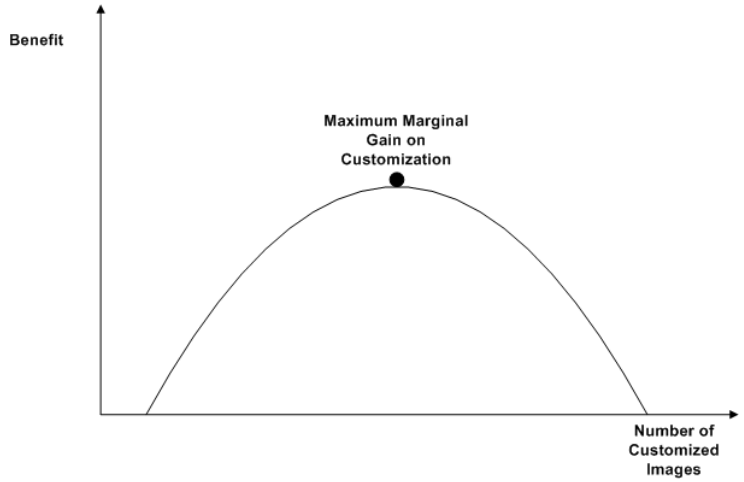

Figure 1.4: At some point, increasing customization of images incurs additional management costs and an associated decrease in marginal benefit.

Management Support Services

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

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

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

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

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

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

Drivers Behind Cloud Computing

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

A Better Way to Consume Services

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

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

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

Service-Oriented Architecture in the Cloud

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

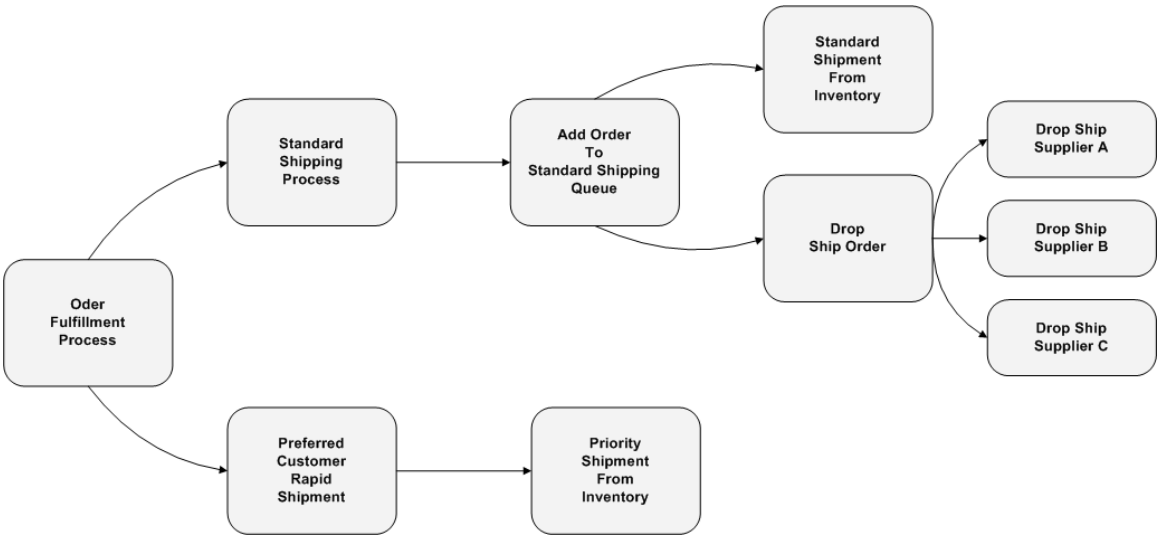

Figure 1.5: Services orchestration combines loosely coupled services in a flow of execution designed to complete a logical unit of work.

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

Differentiated Levels of Service

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

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

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

More Efficient Delivery of Services

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

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

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

Management Infrastructure

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

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

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

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

Optimization of Workloads Across Shared Infrastructure

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

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

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

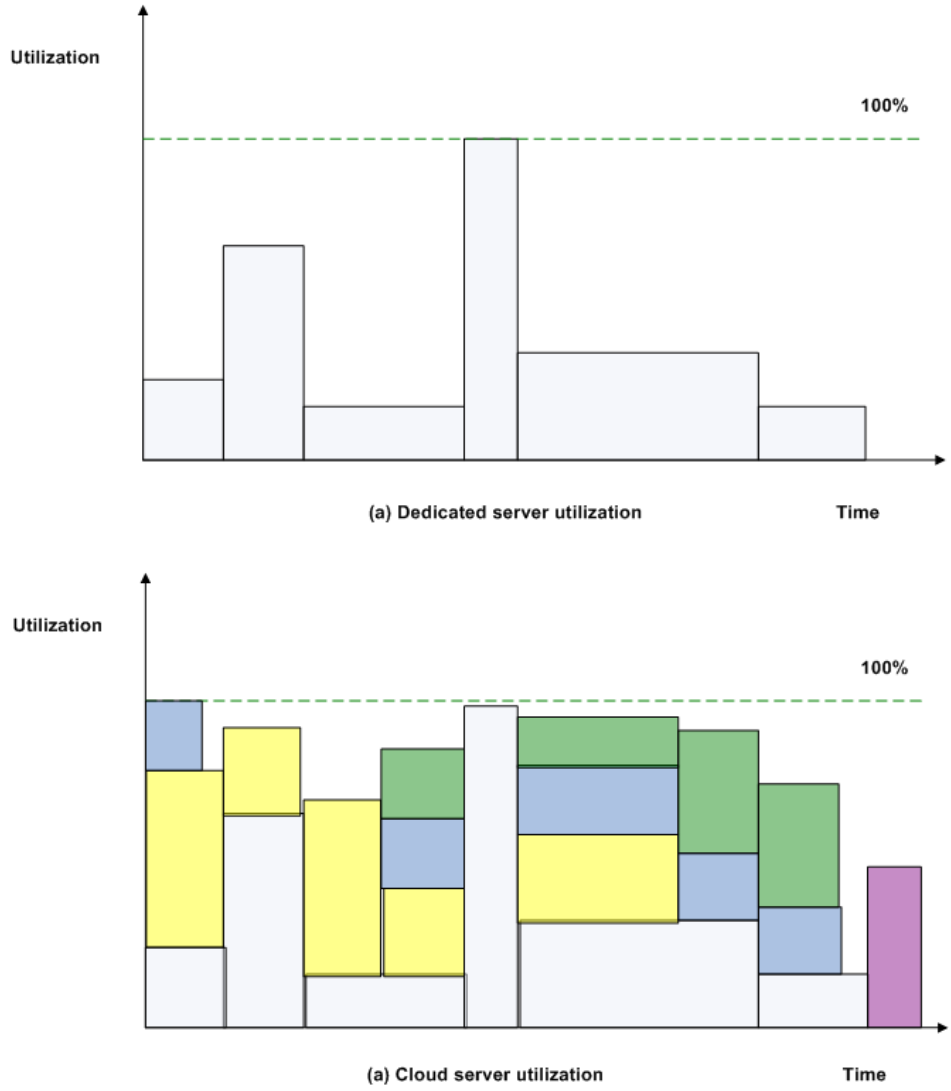

Figure 1.6: In the cloud, server utilization can be significantly higher when workloads are distributed and optimized over available servers.

Self-Service Management

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

Monitoring

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

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

Improving the User Experience through Cloud Computing

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

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

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

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

Changing Economics of IT

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

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

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

Reducing Capital Expenditures

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

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

Efficiently Allocating Resources

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

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

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

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

Rapidly Delivering IT Services

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

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

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

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

Aligning Business Strategy and IT

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

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

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

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

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

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

Summary

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

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

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

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

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

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

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

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.