Realtime
publishers

The Definitive Guide™ To

Cloud Computing

sponsored by

IBM®

Dan Sullivan

Copyright Statement

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

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

Chapter 3: Enabling Business Innovation by Using Cloud Computing

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

The chapter is divided into three main sections:

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

Launching a New Business Service

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

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

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

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

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

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

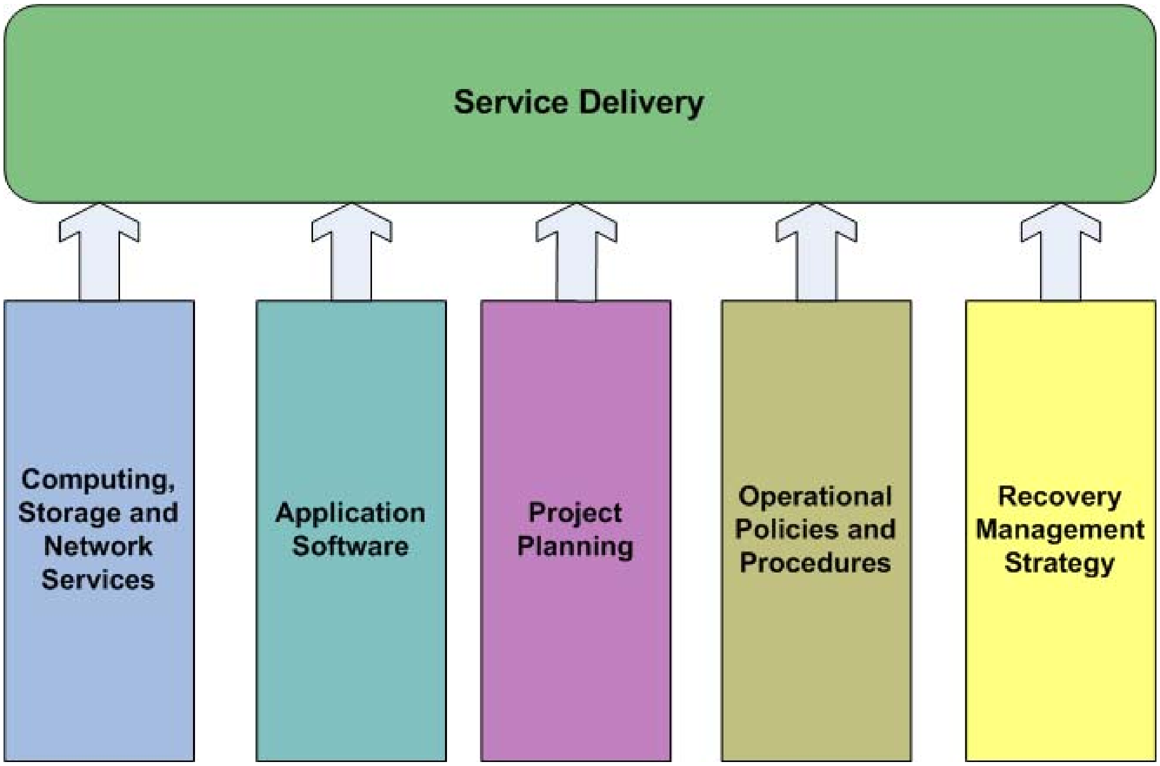

Service Delivery
Computing, Storage and Network Services
Application Software
Project Planning
Operational Policies and Procedures
Recovery Management Strategy

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

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

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

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

New Services Under a Traditional IT Service Model

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

Scenario 1: Tutorial Videos for the DIY Customer

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

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

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

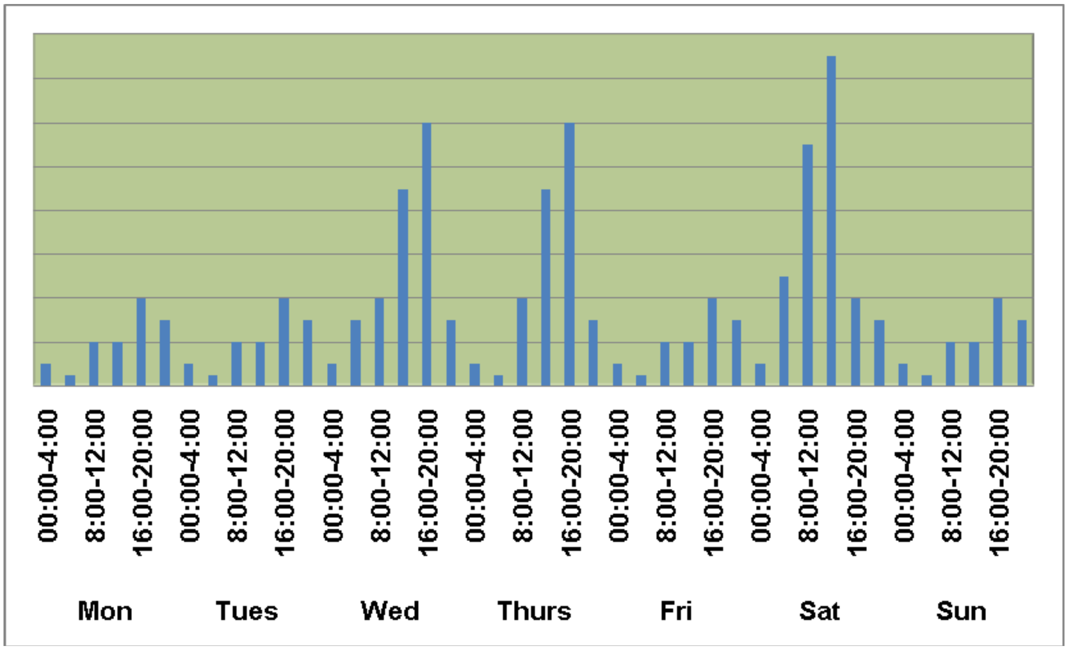


Figure 3.2: Service demand will vary widely by day of week and time of day. (Times are relative to the time zone of the data center hosting the service).

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

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

The major components they decide on include:

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

Figure 3.3 shows the configuration.

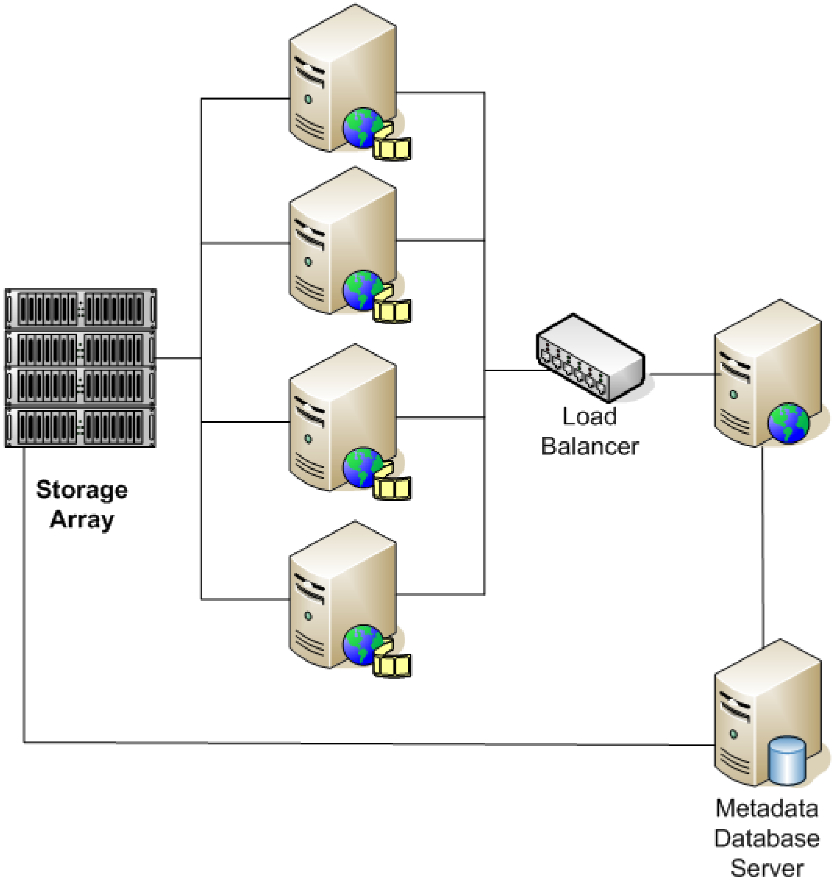


Figure 3.3: The video tutorial service requires hardware to meet peak demand even though the average demand is significantly less.

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

Scenario 2: Advanced Analytics for Auto Insurance Premium Calculations

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

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

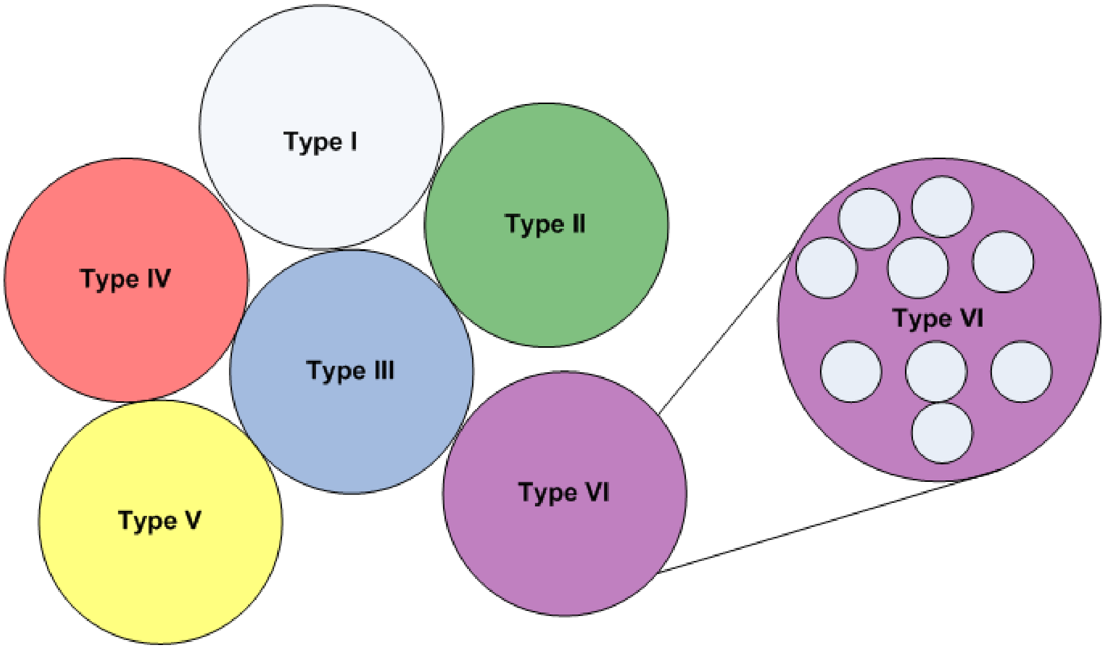


Figure 3.4: Finer-grained clustering of customers can create a competitive advantage by allowing more precise and accurate premium pricing.

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

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

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

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

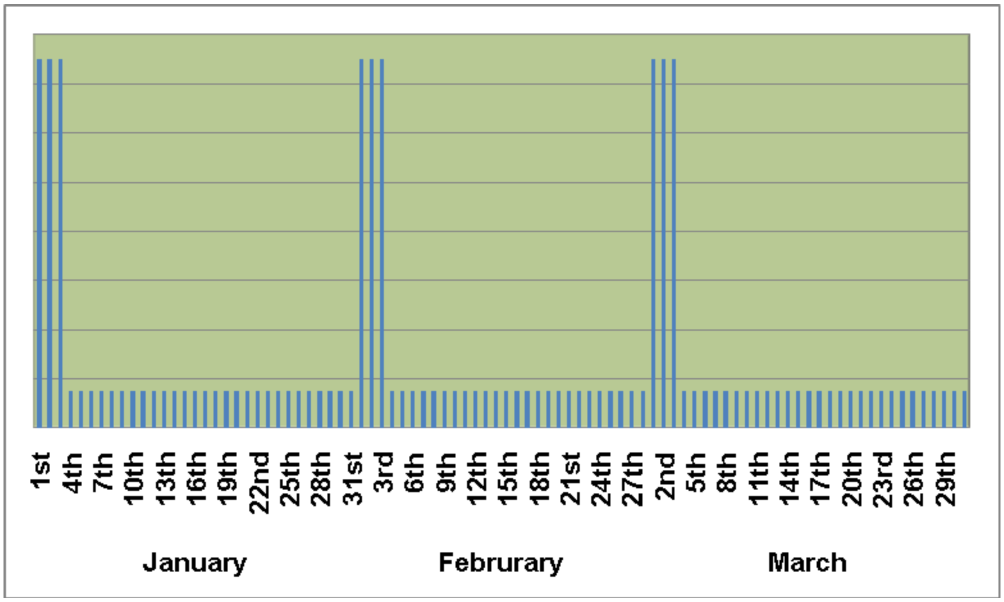


Figure 3.5: Analytic operations have fairly predictable demand patterns that include significant periods of peak demand followed by analysis operations.

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

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

New Services Under the Cloud Computing Model

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

Scenario 1: Tutorial Videos in the Cloud

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

Scenario 2: Advanced Analytics in the Cloud

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

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

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

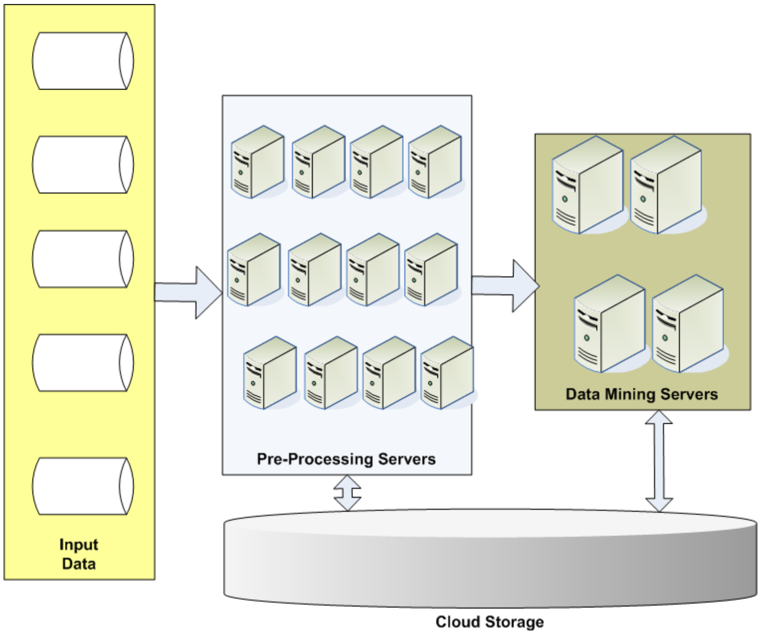


Figure 3.6: In the cloud, servers can be allocated to do task as long as needed and released at which point other servers are instantiated for the next step in the workflows. Service providers only pay for when they are using compute and storage resources.

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

Advantages of Doing Business with Cloud Computing

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

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

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

Time to Deploy Services

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

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

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

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

Cost Control and Ability to Scale to Demand

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

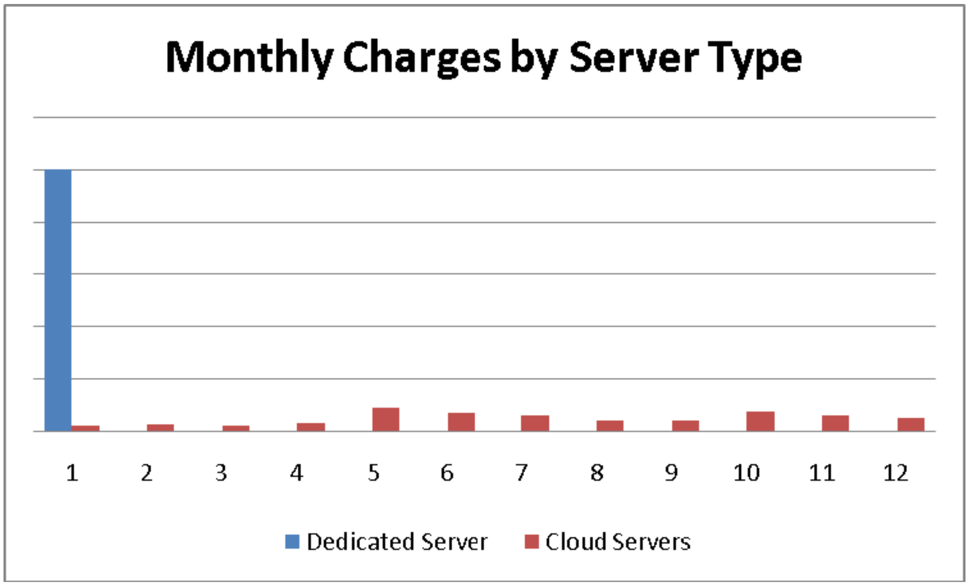


Figure 3.7: Dedicated servers incur high initial cost inline with anticipated peak demand. Cloud servers incur costs for actual use over time.

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

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

Adaptability of Resources

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

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

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

Source of ROI in the Cloud

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

Lowering Capital Costs with Cloud Computing

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

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

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

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

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

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

Lowering Operational Costs with Cloud Computing

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

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

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

On-Demand Provisioning

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

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

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

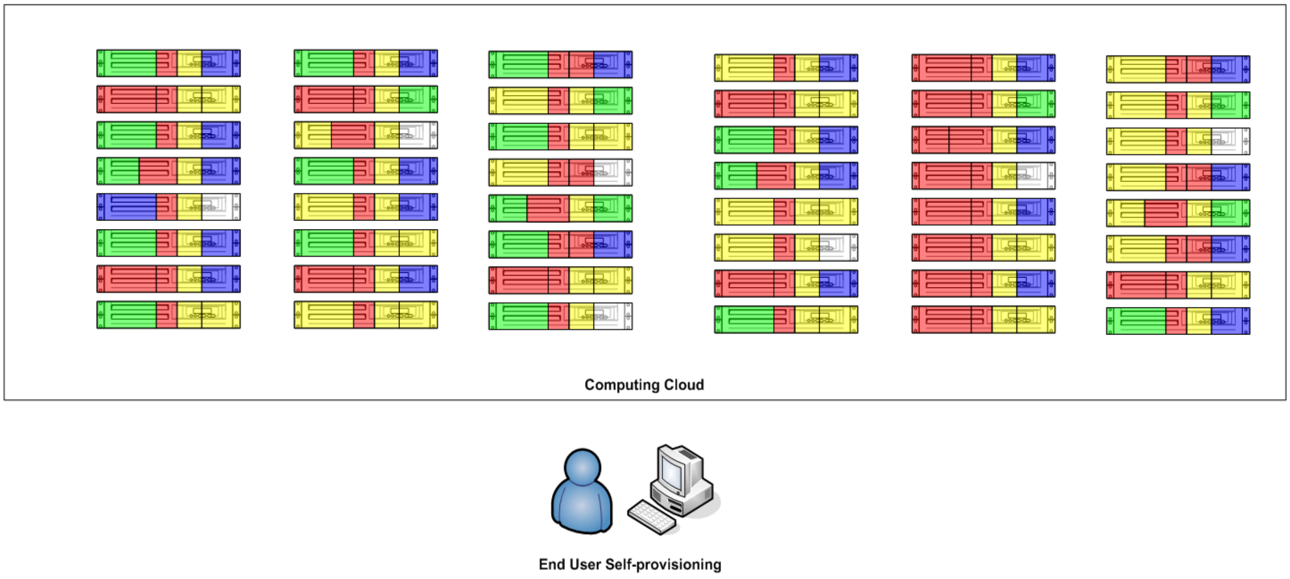


Figure 3.8: Self-provisioning allows cloud consumers to allocate and manage their own resources.

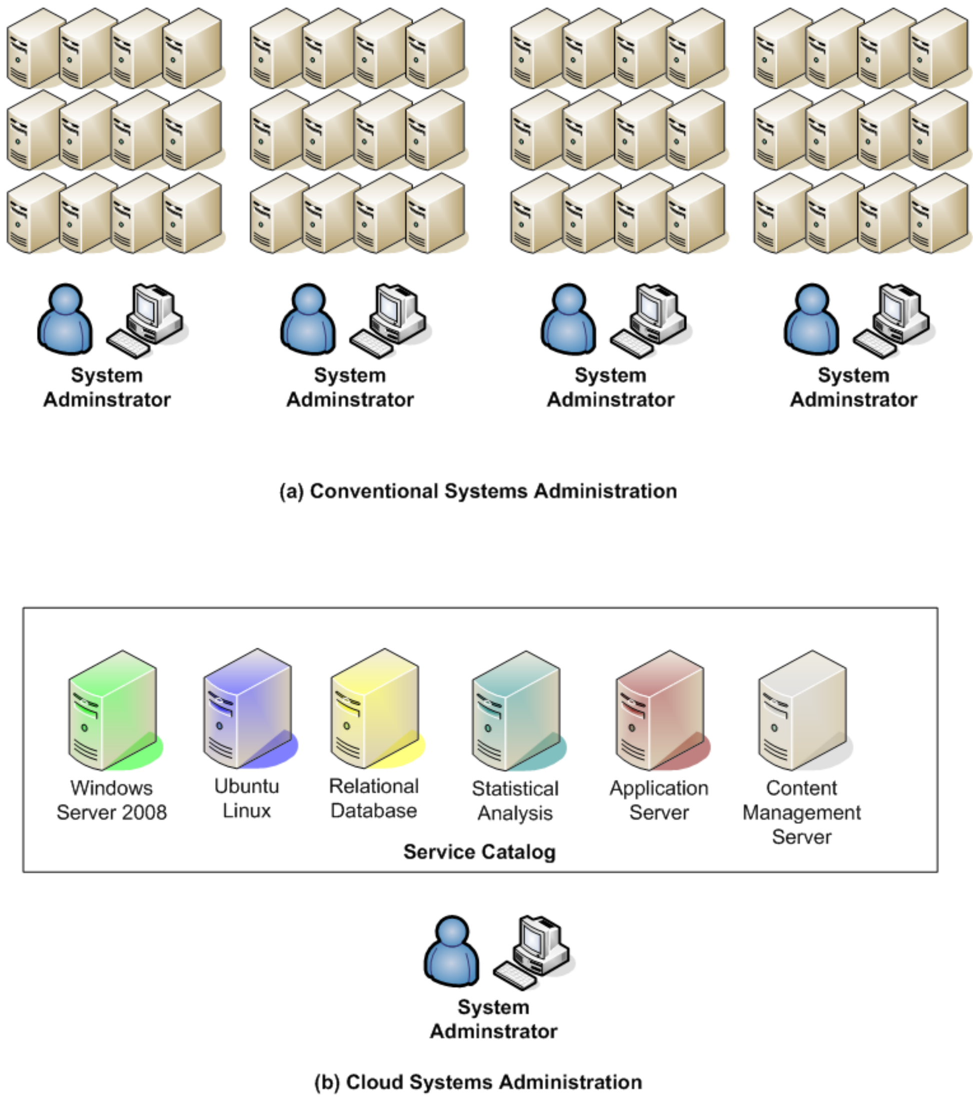
With a self-provisioning system, cloud consumers have access to management systems that allow them to specify the type and number of virtual instances to create. All the hardware in the cloud is managed centrally and virtual machine images are maintained in a service catalog, so cloud consumers do not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

Reducing Marginal Costs of Systems Administration

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

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

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



(a) Conventional Systems Administration

(b) Cloud Systems Administration

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

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

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

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

Standardization and Automation

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

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

Service Management Reporting

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

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

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

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

Assessing the Business Value of Cloud Services

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

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

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

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

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

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

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

Summary

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

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

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

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