Realtime publishers

The Definitive Guide™ To

Cloud Computing

sponsored by

IBM®

Dan Sullivan

Copyright Statement

Chapter 2: Demystifying Cloud Computing

The term "cloud computing" has become a shorthand way of describing a wide range of different computing services. When describing their cloud offering, a vendor might focus on the ability to rapidly provision instances of virtual machines to run applications of your choice. Another vendor might use the term "cloud" when promoting a new way to license and run the vendor's applications on the vendor's servers. Of course, there are any number of definitions in between.

The goal of this chapter is to demystify cloud computing by defining a set of common characteristics that should be included in any cloud service that could be considered ready for enterprise use. The common characteristics, as we shall see, still leave plenty of room for different types of cloud computing. We will examine several types of cloud services and the advantages and disadvantages of each. The chapter concludes with a discussion of different cloud delivery models that range from public to private clouds.

A Note on Terminology

As noted in the first chapter, the types of computing services we are describing represent an evolution of information technology and service delivery. The elements of cloud computing are not radically new, but we are using and deploying them in new ways. This can sometimes lead to confusion in terminology.

Consider, for example, the term "provisioning." In the past, provisioning a server almost always meant that a physical server was acquired, configured, and deployed to an organization's network. The term still has that meaning, but it is not the only way the term is used when describing cloud computing. Provisioning can also mean creating an instance of a virtual machine, for example, to run a job in the cloud for some period of time after which the virtual machine is shut down.

The reason we use the same term for different processes is that both apply to making a computing resource available to a specific task. The key differences are tied to physical versus virtual servers, the duration for which the server is assigned to a specific task, and the time required to make the server available. (These difference underlie the efficiencies cloud computing introduces; however, before we can realize those efficiencies, we need to be clear about all the variables that are at work with services delivery. This chapter will make those variables clear.)

Throughout this chapter and the rest of this book, we will use explicit descriptions, distinguishing, for example, provisioning a physical server from provisioning an instance of a virtual machine. The text will also distinguish models of persistent storage when discussing databases. Relational databases are alive and well in clouds, but they are by no means the only database model available. "Systems management" is another term that is adapting to accommodate new tasks that application administrators are expected to handle when working with clouds.

Describing fundamental characteristics of cloud computing is a step to demystifying this new way of delivering services.

Searching for a Common Definition: 3 Fundamental Elements of Cloud Computing

Reasonable people can disagree about precise definitions of new technologies. We will forgo well-constrained definitions of cloud computing and instead consider three characteristics that are required to deliver the types of services most of us have come to expect from cloud computing:

- Massive scalability
- Ability to easily allocate cloud resources
- A service management platform

There are other characteristics, such as security, that are entailed within these three and will be discussed shortly. Massive scalability, the ability to easily allocate cloud resources, and a service management platform are essential constituents of a cloud computing service.

Massive Scalability

Massive scalability is the ability to rapidly allocate large amounts of computing resources on demand. This is not scalability in the sense of purchasing hundreds of servers, waiting for them to be delivered, configured, and deployed. Massive scalability in cloud computing is the ability to deliver significant resources in a matter of minutes, not days or weeks.

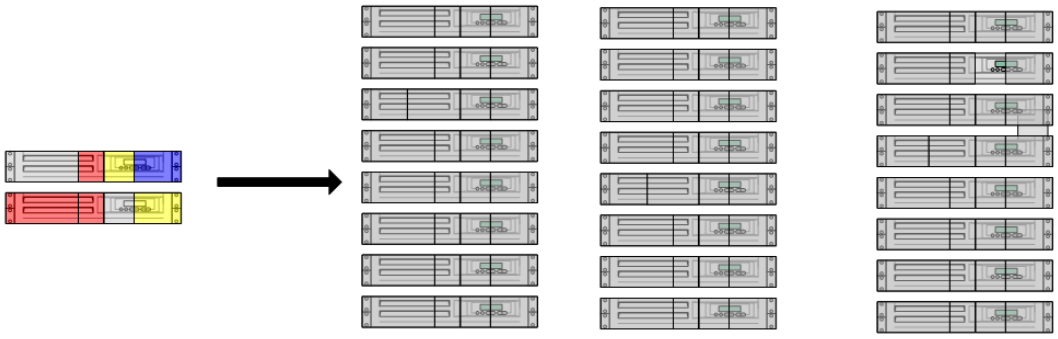

Figure 2.1: Massive scalability provides the ability to rapidly increase the amount of allocated cloud resources as needed for a job.

Three types of resources should be available:

- Computing resources
- Storage resources
- Network bandwidth

Computing Resources

Computing resources are the means to process information. If there were a single workhorse in cloud computing, this would be it. Computing resources are provisioned for a cloud computing task in different ways, depending on the cloud model. At minimum, there is a smallest unit of computing resource that is allocated. This could be, for example, a virtual machine equivalent to an x64 architecture, 2GHz CPU dual-core processor with 32GB of memory and 300GB of local storage. Specifications such as this should be considered a logical specification. The virtual machine running jobs could be hosted on any of a number of physical implementations. This is one of the advantages of cloud computing: The details of the physical implementation are abstracted so that the consumer of cloud services does not have to concern themselves with such details.

Abstracting computing services can also lead to more efficient delivery. For example, a cloud provider can:

- Vary the amount of hardware running at any time according to demand—During periods of peak demand, many large servers may be running while during low demand periods, only the most energy-efficient servers are kept powered on.
- Run jobs in different data centers to better allocate work load—This functionality is constrained to some degree by business requirements. For example, businesses subject to European Union (EU) privacy directives may require that all personal information on EU customers be kept in countries that meet a minimum level of privacy protections.
- Execute workloads on physical servers that minimize the distance between the compute resources and the storage resources

Cloud service providers all abstract some level of implementation details, but that level can vary significantly. Consider a few different scenarios.

The (Near) Raw Iron Approach

One cloud provider allows consumers to select a type of virtual machine (types vary by number of cores, amount of memory, and so on) and the virtual image to run on that machine. There may be several operating systems (OSs) to choose from as well as a variety of application stacks. This model has the advantage of giving cloud consumers a wide range of options but at the cost of additional configuration responsibilities. For example, a cloud consumer may have the option to configure and run a particular statistical analysis package on a preferred version of Linux with this provider, but she is also responsible for tuning and patching this image.

The Server Role Approach

A second cloud provider may limit the range of options in return for a simplified deployment model. Rather than allow customers to build their own virtual machine images, the cloud provider may offer a small set of preconfigured images designed for specific roles, such as load balancing, running a Web server, or providing application services. Under this approach, cloud consumers could define the number Web servers they need and the number of application servers required without having to concern themselves with OS or application stack details.

The API Approach

Another approach a cloud vendor may provide is a general computing platform that abstracts even basic distinctions such as Web servers and application servers. Under this model, cloud consumers develop applications that use a cloud provider's application programming interface (API), which might include, for example, functions for:

- Defining data structures
- Creating associative arrays (key-value pairs)
- Specifying queries
- Implementing transactions
- Utilizing task queues

When the application is run, the cloud consumer need only specify the number of servers to dedicate to the task. By limiting the range of options for implementing an application, the cloud consumer has fewer systems management issues to address. As Figure 2.2 shows, there is a tradeoff between flexibility and systems management responsibility.

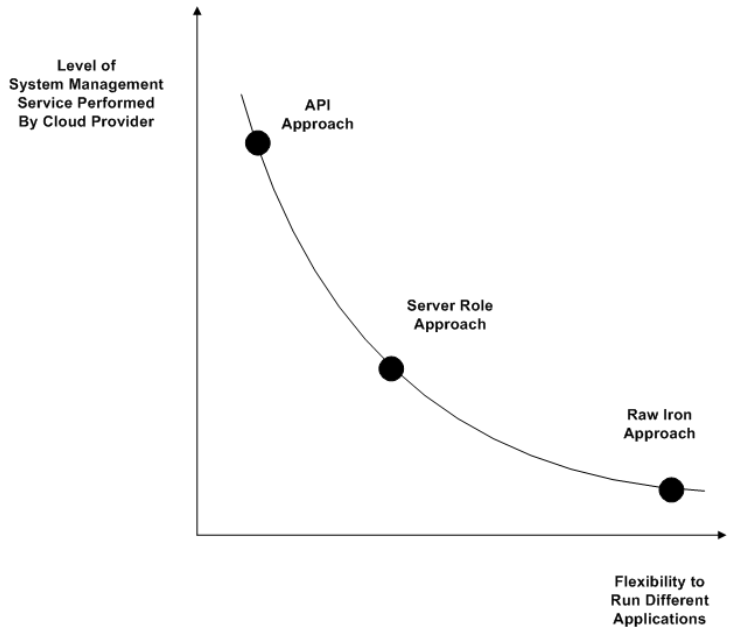

Figure 2.2: Cloud consumers have a range of options that balance different levels of flexibility with the need for systems management tasks.

Storage Resources

Massive scalability implies the ability to persistently store raw data and computed results. For the cloud consumer, the amount of storage needed at any point in time should rapidly scale to meet demand. As with computing resources, storage resources should be allocated as needed and there should always be storage available.

Cloud storage is available in a few forms:

- As file-based storage
- As block-based storage in which arbitrary large objects are stored
- As relational storage in which data is maintained in relational database structures

Data stored in the cloud is manipulated in much the same way as it is in non-cloud architectures with some minor differences. Block-based storage may be accessed via URL. Relational data is queried the same in or out of the cloud, but database administrators will have less to manage with regards to the physical allocation of space and replication of data for high availability.

Network Resources

The ability to move data from compute to storage resources must scale along with those resources. Within the cloud, the network capacity and infrastructure is defined and managed by the cloud provider. Providers can reasonably plan for moving data from servers to storage arrays or replicating data between storage devices. The situation changes when data has to move into or outside of the cloud.

Cloud service providers are more constrained in their ability to deliver network scalability because of dependence on the outside networks. Cloud consumers transfer data into and out of the cloud using whatever network services they have acquired. This may or may not be sufficient for the volumes of data that need to be transferred. In response, some cloud providers offer “sneakernet to the cloud” services: physical storage devices are shipped to the cloud provider where they are uploaded to the cloud.

Part of optimizing cloud-based services is determining the best way to move data into and out of the cloud and minimizing transfers outside the cloud. The network bottleneck is one reason to generate, process, and store data in the cloud as much as possible.

Massive scalability is a fundamental characteristic of cloud computing. Cloud providers offer different approaches to providing computing resources that tradeoff between flexibility in applications that can run in the cloud with demands on cloud consumers to manage system resources. Similarly, massive storage scalability is fundamental to cloud computing. At this point in time, networking resources outside the cloud are a potential bottleneck to moving data to and from the cloud.

Ability to Easily Allocate Cloud Resources

Cloud computing can significantly reduce the need for systems administration support by providing easy-to-use tools for allocating cloud resources. One of the advantages of abstracting many implementation details is that it allows for greater automation of the cloud resource provisioning. As noted earlier, cloud providers offer different levels of abstraction of services, but in all cases, the provider should offer tools that enable application administrators the ability to adjust the usage as demand dictates.

Consider a simple example. A marketing analyst has just acquired several large data sets on product sales over the past several months. This is a onetime task and the analyst needs to aggregate the data for business reporting as well as run some statistical analysis programs over each data set. Outside the cloud, the analyst would need to perform several time-consuming steps:

- Find a department server with availability and convince the owner to allow the jobs to run on that server.
- Next, assuming a server is found, the analyst would then submit a ticket to systems administrators to install the necessary analysis software.
- When that is done, which could be a few hours to a few days depending on the IT support backlog, the analyst would need to upload the data. If the data is compressed, additional storage will be required to store both the compressed and decompressed files until the decompress operation is complete.
- Run the analysis jobs. This is a compute-intensive job, so the time to complete it will depend on the number of CPU resources available. If the analyst was provided with a virtual server running on a host with several other virtual machines, the workloads on the other virtual machines can adversely impact the data analysis job.

The same process in the cloud is significantly less arduous.

- Select a virtual machine image to run on cloud servers from a catalog of images. These can range from OS-only images to complete development or analysis environments.
- Specify the number of the virtual instances to run. In some cases, cloud vendors may offer options on the size of servers (for example, small, midsize, high-end), in which case, the size would need to be specified as well. As multiple servers are available, the analysis job can be subdivided into smaller jobs and run in parallel.
- Load the data into cloud storage and decompress if necessary.
- Run the analysis jobs.

These steps would be performed in a Web browser using a resource management interface by the analyst. There is no need for specialized IT support, no need to search for a server with available capacity, and no need to allocate disk space to a file system. The combination of massive scalability and easy-to-use interface to allocate resources provides two of the three core elements of cloud computing. The ability to manage services is the other.

Service Management Platform

Once we move beyond simple scenarios like the one previously described and start to consider enterprise-scale management issues, the need for a services management platform becomes clear. A cost-effective cloud service will offer a management platform that supports four aspects of service management:

- Support for automated provisioning and deprovisioning of resources
- Self-service interface
- A service catalog of standardized services
- Policy definition and enforcement

Support for automated provisioning and deprovisioning and the self-service interface were covered in the previous section, so we will focus our attention on the other topics here.

Service Catalog of Standardized Services

A service catalog introduces consistency and reusability to the cloud. A catalog includes virtual machine images that can run within the cloud with minimal setup on the part of the cloud consumer. These images capture design patterns that have worked well in other use cases.

For example, a basic Web server service might include the latest version of the Apache Web server, a fully patched and hardened Linux OS, and a properly configured firewall. Another image in the service catalog could provide an extraction, transformation, and load (ETL) application for use with data warehousing applications. With the ability to instantiate a fully functional ETL system in a matter of minutes using a self-service interface, the barriers to entry to business intelligence and data analytics is significantly reduced.

Policy Definition and Enforcement

A service management platform can ensure that operations in the cloud comply with organization policies. Technical policies can address issues such as:

- Authentication and authorization required to use resources
- Resource limits, such as the number of concurrent virtual servers a user can have instantiated during peak load periods
- Pre-instantiation checks, such as ensuring images are properly patched before executing or virtual machines use currently approved versions of supported OSs

Organizational policies can be enforced as well. These include:

- Adjusting the cost of using resources according to demand—This could be implemented with a policy of peak load pricing or bidding based spot pricing.
- Prioritizing workloads in the event sufficient resources are not available during peak demand periods
- Controlling the number of instances of a particular application that is running at any one time—This is would be used to ensure compliance with software licensing agreements

A service management platform is essential to reducing labor costs associated with delivering information services. It enables non-IT professionals to allocate resources they need when they need them while still ensuring organization policies are followed.

A Cloud by Any Other Name

Cloud computing has the potential to significantly reduce costs and improve the delivery of business services. It is no wonder vendors would want to offer something in this area. Simply calling a service offering a "cloud" is not enough, at least for *The Definitive Guide to Cloud Computing*. This guide has and will continue to argue that cloud computing entails massive scalability, easy to allocate resources, and a service management platform that includes a service catalog. These three elements are essential to offering a viable cloud computing service in an enterprise.

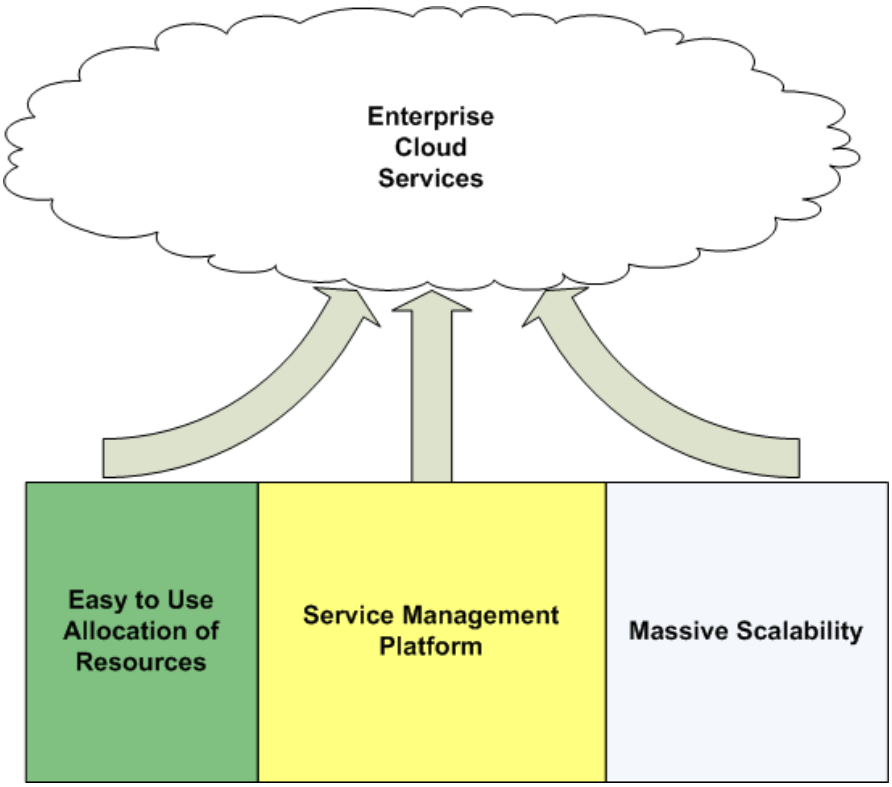

Figure 2.3: Cloud computing requires three fundamental elements to be effectively used in enterprise computing.

If any doubts remain, consider if any one of these three characteristics were missing. Without massive scalability, there would not be the resources required to meet fluctuating demand. Cloud consumers would have to have backup resources in place in case cloud resources were not available. Traditional service delivery models would continue to exist and undermine the cost benefits of cloud computing. Without easy provisioning, cloud consumers would still have to depend on IT support, creating the potential for backlogs and driving up labor costs. Without a service management platform, cloud consumers would not have a well-managed service catalog, the lack of which would drive up costs of creating and maintaining virtual machine images. IT support would not have a mechanism to enforce policies, leaving the potential to violate governance and compliance regulations. Billing and resource management would require more manual processes, driving up costs in turn.

Cloud computing lends itself to a wide array of services and service delivery models. As we will see in the next section, there are many ways to provide cloud services.

Different Types of Cloud Computing Services

Cloud computing can encompass a broad range of services, so it is not surprising to see a number of broad options emerging. These services range, in increasing order of specific type of service, to include:

- Infrastructure
- Platform services
- Application services

Each level of service meets a distinct set of needs.

Infrastructure Services

Infrastructure services deliver computing and storage services. This type of service is the one used as a model in the previous section describing the three defining characteristics of cloud computing. Here we will turn our attention to describing how this type of service can be used along with an example use case to show how cloud computing can significantly improve some types of service delivery.

Computing on Demand

The ability to provision computing resources for just about any computing requirement is valuable enough to drive the adoption of cloud computing even if none of the other types of cloud computing services were available. With computing on-demand services, organizations have the ability to allocate virtual machine resources for a variety of tasks:

- Executing proprietary workflows
- Meeting peak demand for computing
- Performing disaster recovery
- Running highly distributed applications

By allocating just basic computing services, cloud consumers can run proprietary workflows that do not depend on preconfigured services. A broad set of service images in a service catalog can provide a starting point for building proprietary workflows. For example, the service catalog would have virtual machine images with OSs and application servers, which users could instantiate and then add custom applications to complete the set of components needed for the workflow.

This type of cloud service also works well for accommodating peak demand periods for either standardized applications or proprietary workflows. Existing infrastructure may be sufficient for average loads, but during peak periods, such as the holiday shopping times in the retail industry, additional computing services may be needed for relatively short periods of time.

Maintaining a disaster recovery site can add significantly to the cost of providing a service. Even if a disaster recovery site is never used, businesses pay for the housing equipment, power to keep a minimal infrastructure running, and maintaining servers and other equipment. There may be marginal labor costs as well to maintain the site. An alternative, and one enabled by the computing on-demand model, is to use a cloud provider as a disaster recovery service. To do this, a business could:

- Maintain a set of virtual machine images that would run the business applications in the event of a disaster
- Maintain copies of data in cloud storage using an appropriate combination of backups and near-real-time replication
- Establish a plan for provisioning cloud services to meet disaster recovery requirements; for example, some services may be run on smaller, and therefore lower-cost, servers while in disaster recovery mode

Of course, as these requirements demonstrate, computing on demand can be closely coupled to storage on demand.

Storage on Demand

Storage on demand can provide file, block, or relational storage to meet a variety of requirements. In some cases, such as the need for offsite backup, the need for storage is fairly consistent. Cloud storage offers the ability to protect backups from site-specific damage but without the need to maintain another physical site. When dealing with multiple remote sites, copying backups to the cloud can be an appealing option rather than physically transporting tapes from those sites or maintaining additional disk storage at a data center to accommodate those backups.

Demand for storage can vary widely. For example, an accounting firm may have peak demand for 2 to 3 months prior to tax-filing deadlines when large amounts of data are coming into the firm. After the deadline, data can be archived and moved off disk, but without an option such as cloud-based storage on demand, the firm would have to maintain peak storage capacity all year. The wide potential for on-demand computing and storage can be demonstrated with a more generally applicable example as well.

Business Intelligence Use Case

Business intelligence reporting is driven by large volumes of up-to-date information. Collecting and processing this data can impose significant demands on computing and storage resources, especially when the ETL phase has to occur in a limited window of time. With on-demand computing and storage, data can be uploaded from multiple local sources simultaneously. That data is then aggregated at low and mid levels in parallel before being aggregated at a global level and finally stored in a cloud database for later report generation.

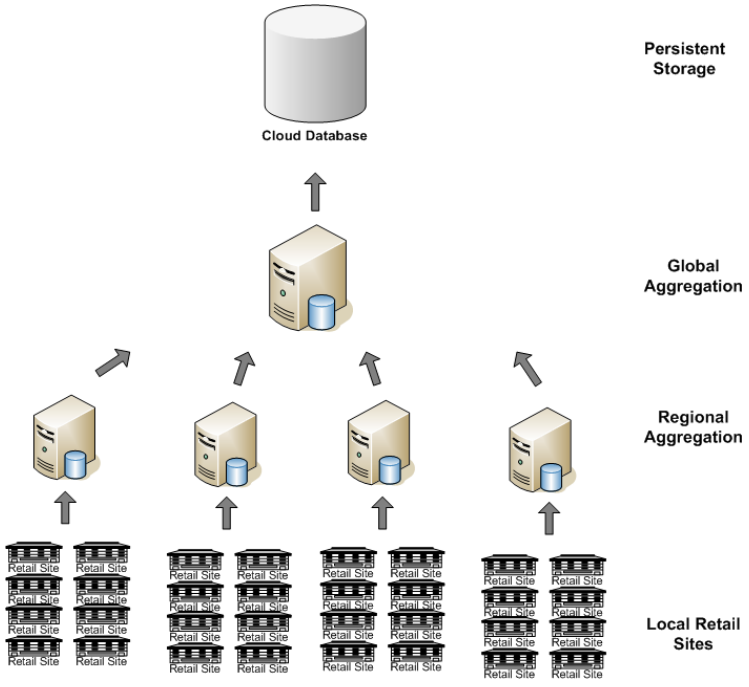

Figure 2.4: With on-demand computing and storage, time-critical operations like aggregating data for business intelligence reporting, can be done in parallel. This ensures the job completes within the allotted time.

Running the aggregation operations in parallel allows the process to complete faster than if done sequentially. Running the operations with cloud resources eliminates the need for maintaining dedicated servers that would otherwise be underutilized.

Infrastructure services are an appropriate delivery model when organizations require basic computing and storage resources. When those needs include components commonly found in application stacks, the platform services delivery model may be a better fit.

Platform Services

Platform-based cloud services deliver higher-level services than the infrastructure-based model offers. Platform-based services include tools for designing, developing, and deploying applications using a set of supported application components, such as relational databases and application security services that span multiple layers of the application stack.

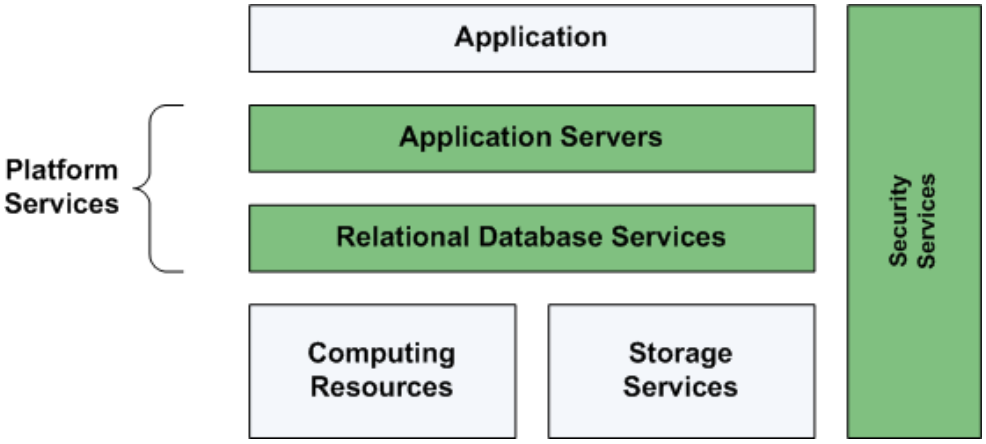

Figure 2.5: Platform services (in green) provide application development components built on lower-level cloud services.

Relational Database Services

Relational databases are the data backbone of most enterprise applications. Since the later 1970s, relational data models have offered significant advantages over other database frameworks. Continuous improvement in relational database management systems have allowed relational databases to keep up with growing and changing demands for managing persistent data. One of the latest advances is the ability to host relational databases in a cloud.

To avoid any confusion, it is worth noting that there are two ways one could host a database in the cloud. One method is suitable for small projects with short life spans, the other takes advantage of cloud infrastructure for a more scalable solution.

A Simple Relational Database System in the Cloud

The first method basically transfers the same approach to database management we typically use outside the cloud and applies it in the cloud. Under this method, a database administrator provisions a virtual server and installs the database management system on that server using local disk storage for database files. This approach may be suitable for limited needs but is not a general solution for persistent relational storage in the cloud.

One drawback is that local storage is allocated to a user's virtual machine instance only as long as the instance is running. One of the advantages of the cloud is that virtual machine instances are started and stopped as needed. Unless the instance hosting the database is kept running, the database will be lost. Another drawback is that the versions of relational database management systems running on typical enterprise servers are not designed to take advantage of cloud storage services based on allocating blocks or buckets of storage for arbitrary data. Although this is one way to use relational databases in the cloud, it is not what is generally considered a relational database service.

Relational Database Services Optimized for the Cloud

Relational database services for the cloud take advantage of the scalability of compute and storage resources of the cloud. As one might expect, relational database services attend to a number of low-level implementation details that are typically the responsibility of a database administrator. For example, within the cloud, database administrators do not have to concern themselves with:

- Managing disk space
- Specifying how to distribute low-level data structures, such as tablespaces, across multiple disks to optimize performance
- Monitoring I/O patterns to detect bottlenecks in disk operations
- Replicating data to ensure high availability since persistent data is typically written to multiple locations within cloud storage

Of course, this does not mean the end to database administrators as we know them any more than cloud computing is putting an end to systems administration. Database administrators working with relational database services can focus more attention on the logical aspects of database design:

- Defining schemas
- Optimizing indexes
- Tuning stored procedures and triggers
- Creating views and other abstractions to better support application development

Also, expect cloud providers to support the three fundamental characteristics of cloud computing with respect to relational databases: massive scalability, easy to allocate resources, and a service management platform.

Application Servers

Application component services provide middleware services in the cloud. Like relational databases, middleware applications, such as application servers and portal servers, can be optimized for the cloud. This ensures the components can take advantage of scalability, high-availability, and service management platforms provided in the cloud.

Security Services

Security is not a component one can isolate like a database or a messaging queue. Security is a product of specialized components, such as authentication and authorization services, as well as systems design. The fundamental principles of security are no different in the cloud than outside the cloud. We cannot, however, simply use the same security procedures in the cloud that we use outside the cloud anymore than we can simply run a database management system built for a single server in the cloud and expect cloud-like benefits.

Security services need to be embedded into cloud platform services and, at a minimum, include support for:

- Authentication
- Authorization
- Auditing and reporting
- Key management
- Security token management

Authentication and authorization are necessary to determine who is using a system and limiting what they are allowed to do. Auditing and reporting are required to ensure policies and procedures are enforced and to detect unauthorized activity as soon as possible. Key management and security token management are especially important in distributed systems where multiple systems depend on trusted identity management systems to perform authentication, authorization, and other security services on their behalf.

Above infrastructure services and platform services in the hierarchy of cloud services, we find applications.

Application Services

Today's complex enterprise applications are often built on application frameworks and design patterns, so it is not surprising to see support for these in the cloud. The frameworks vary but include components such as runtime libraries, development frameworks, and higher-level application components. The level of support for different frameworks will vary by cloud provider, especially if providers specialize in supporting one type of framework. In some cases, a cloud provider may offer a framework specifically designed for the cloud and not available in other architectures.

Even with variation in frameworks and programming languages, a number of application services may be available that allow programmers to take further advantage of what a cloud infrastructure has to offer. Two such services are messaging queues and support for highly distributed, parallel processing.

Messaging Queues

Messaging queues provide for asynchronous communication between processes running in the cloud. Messaging is useful for constructing workflows, implementing distributed transactions, and accommodating the failure of a component within a distributed system. Consider as an example a Web interface running on one server accepts requests from users. In a tightly coupled application, the interface may pass the request to one instance of a backend service and wait for a response. If the backend service is down, the application fails. In a loosely coupled design, the interface would submit the request to a queue. Any one of a number of instances of the backend service could read the request from the queue, respond to it, then delete the request. If a single instance of the backend server is down, the request can still be serviced. If one of the backend instances crashes while processing a request, another instance can still read the request because it is not deleted from the queue until the response is generated.

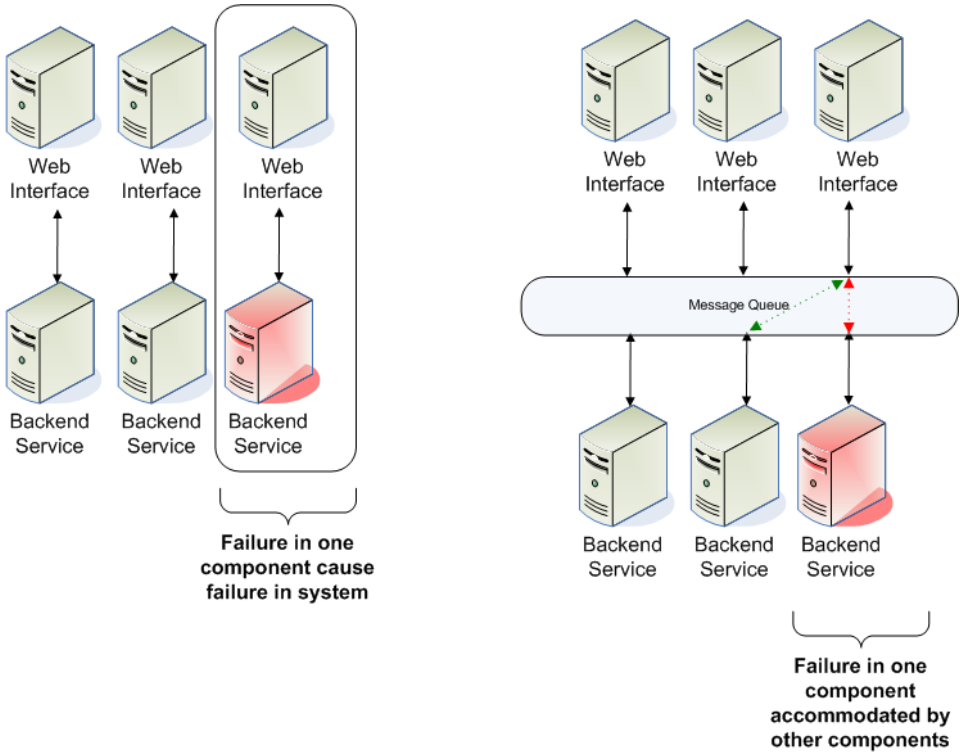

Figure 2.6: Tightly coupled systems are more likely to have single points of failure; messaging queues enable more robust application design.

Application services within the cloud also include higher-level components that enable enterprise application functionality.

Distributed, Parallel Processing

One of the advantages of cloud architectures is access to a large number of servers. This introduces opportunities for performing operations in parallel that would normally have to be done sequentially when only a small, fixed number of servers are available. A programming paradigm known as map-reduce is one suitable-for-clouds method to implement parallel applications.

The basic idea behind map-reduce is that some problems are inherently parallel: Some steps in the computation can be done independently of other steps and the results of individual computations can be combined to produce the final result. The ETL example cited earlier highlights a problem with course-grained parallelism. That problem can be broken down into a small number (for example, on the order of 10) steps followed by an aggregation process to combine results. Other problems, especially those with large amounts of data, can be divided into even larger numbers of sub-problems.

Take for example, analyzing click-stream data. A business is analyzing patterns of activity on their e-commerce site to determine whether there are common characteristics shared across customer interactions in which the customer abandons his or her cart. The click stream data from the Web site contains information about what products the customer viewed, reviews that were read, and navigation paths taken to the point where a product was added to the cart. As one customer's activity is independent of others, this is a good candidate for highly parallel analysis.

A map-reduce approach to this problem could be defined as follows:

- Split the set of all click stream data by customer session
- Partition the customer sessions across 100 instances of the analysis program
- For each customer session, scan the click stream for the number of times each possible 3-page sequence pattern occurs; to simplify the pattern, look for types of pages, such as product details, reviews, search results—this is the map phase
- Combine the results of each map phase to produce the aggregate number of times each pattern occurred—this is the reduce phase

A key advantage of this approach is that large volumes of click stream data can be analyzed much faster in parallel than sequentially, thereby creating the possibility for greater amounts and more in-depth analysis of customer interaction behavior.

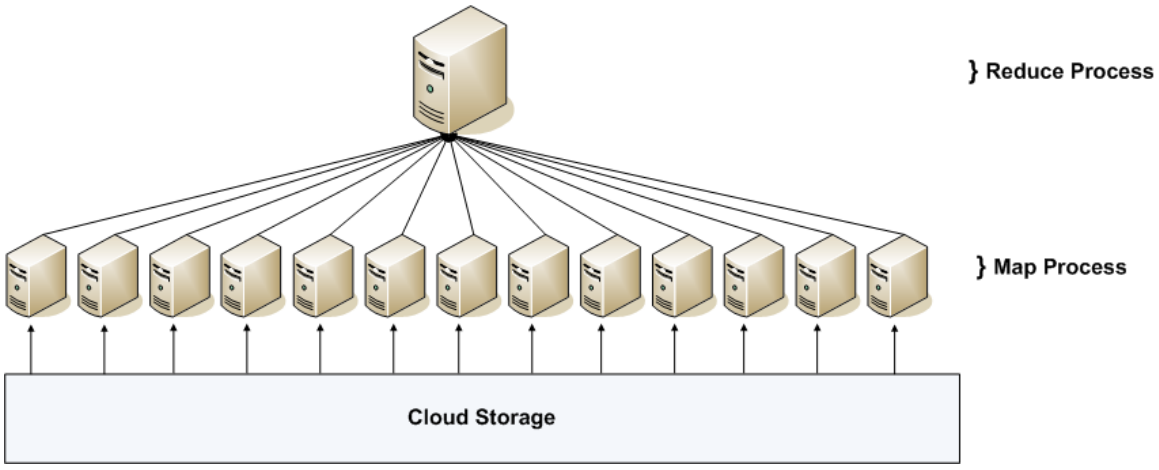

Figure 2.7: Map-reduce is a parallel programming framework that works well with cloud computing and storage.

Application middleware, such as application servers; design patterns, such as the use of messaging queues for asynchronous communication across multiple processes within a cloud; and programming frameworks, like map-reduce that exploit the parallel capabilities of a cloud, are all enabling components for delivering enterprise applications in the cloud. As cloud technology adoption grows, we can expect to see more enterprise applications being offered directly to cloud service consumers.

Applications and Business Services

Providing application and business services from the cloud presents an opportunity to consolidate those services. The beneficial features of cloud computing, such as flexible scalability and a service management framework, can enable organizations to reduce the number of separate instances of applications running throughout the enterprise.

Consolidating Enterprise Applications

Consider a few common types of enterprise applications:

- Customer relationship management (CRM)
- Enterprise resource planning (ERP)
- Business intelligence

Each of these types of applications can have broad reach throughout a business. With the commonly used "one server/one application" approach that has been used for years, businesses may find themselves limited to how many users they can support with these applications.

For example, consider a company that runs a CRM application on a server sufficient for current needs as well as some moderate growth. The company then merges with another business that also needs CRM support. The IT staff of the new company will have to determine whether a single server can support the newly merged enterprise or multiple instances of the system will have to be run. The latter option can lead to fragmentation and arbitrary divisions that in turn can lead to organizational problems down the road.

Let's assume the business decides that running two instances of the CRM application is the more cost-effective alternative. The customers are divided geographically with North America, South America, and Southeast Asia customers in one instance, and Europe, Middle East, Africa, and other Asia customers in the second instance. A host of questions arise:

- How should customers in global, transnational companies be divided?
- Will regional subdivisions of customers be separated?
- How costly and time consuming will it be if the allocation of customers has to be re-arranged to align with new business strategy?
- What is required to support a federated identity management system so that users in one system can access the other system as needed?

Similar questions can be asked about ERP systems; instead of customers though, the questions would focus on budgets, inventories, financial projections, and accounting issues.

In the case of business intelligence, fragmentation can occur around tools and procedures. Enterprise-scale data warehouses may have dedicated database administrators who are able to tune and manage complex database management systems. Departments with more limited requirements may build locally managed data marts employing easier-to-use databases and reporting tools. This may be the most expeditious approach in the short run but over time it can lead to duplicated data, increased software licensing costs, and redundant administration costs.

Moving enterprise applications such as CRM, ERP, and business intelligence systems to the cloud can help reduce costs and improve the delivery of business services. With standardized virtual machine images and centralized cloud storage, additional compute resources can be brought online as demand for services grows. As data is consolidated in the cloud, we can avoid data fragmentation problems. Standardized virtual machine images deployed through a services management platform reduce the demand for specialized database and systems administration expertise in departments running local applications, such as data marts.

Managing Business Services and Workloads

As applications move to the cloud, there will be a need to manage according to service level agreements (SLAs) and other expectations for performance and availability. This will require both technical and management approaches to the problem.

On the technical side, application administrators will need to utilize performance reporting provided by the service management platform to ensure SLAs are met in cost-effective ways. Running multiple instances of an application and load balancing across those instances can help maintain performance and provide a level of reliability to the system.

On the management side, we need to be cognizant of utilization. There is no point running six instances of an application with an average server utilization of 25% when running three instances still leaves plenty of margin for spikes in demand without the need to instantiate another virtual machine image.

It is clear as we consider the different types of services, from infrastructure to platform to application services, there are many ways to leverage cloud services and the benefits generally arise from a set of common attributes.

Common Attributes of Cloud Service Models

The three defining characteristics of clouds—massive scalability, easy to allocate resources, and a service management platform—describe key architectural elements of computing and storage clouds. A consumer of cloud services may see a different set of attributes from their perspective:

- On demand self service—The ability to allocate, use, and manage computing, storage, application, and other business services at will without depending on IT support staff
- Ubiquitous network access—The ability to work with cloud resources from any point with Internet access; cloud service consumers are not dependent on being in corporate headquarters or in a data center to have access to an enterprise cloud
- Location independent resource pools—Compute and storage resources may be located anywhere that is network accessible; resource pools enable redundancy and reduce the risks of single points of failure
- Elastic scalability—Cloud consumers decide how much of any resource they utilize at any time; allocation is driven by immediate demand not the need to maintain capacity for peak demand
- Flexible pricing—Cloud providers typically charge with a "pay as you go" model; as cloud computing matures, we will likely see a variety of pricing models, including prices that vary by level of demand

We have described cloud services from an architectural view, in terms of services delivered, and from the perspective of a cloud consumer. One remaining dimension we should consider is the public/private cloud distinction.

Cloud Delivery Models

When cloud computing first emerged as a viable platform, the term generally applied to what we would now call a public cloud. As cloud computing expanded, so did the delivery models to the point where we have at least three distinct delivery models:

- Public clouds
- Private clouds
- Hybrid clouds

Public and private clouds have advantages and disadvantages; hybrid clouds attempt to capture the best of both worlds.

Public Clouds

Public clouds are computing and storage services that are open to any consumer. An immediate advantage of using a public cloud is that there is no upfront capital expenditure required of business users. Cloud consumers purchase computing and storage services as needed and pay as they go. There are likely costs associated with transferring data to and from the cloud, and these costs can easily grow beyond the cost of computing and storage for high-transfer rates. Another disadvantage is that businesses are dependent on the viability and reliability of the cloud provider. If there is a significant service outage, data and services will be inaccessible. Risk assessments and mitigation strategies are called for when working with any cloud, but they are especially necessary when critical business services are dependent on third parties.

Private Clouds

Private clouds are owned and operated by businesses for their internal use. This delivery model can be especially appealing when compliance, security, and other risks factor significantly when developing a cloud strategy. A key advantage of a private cloud is that the business is in control of the service: it can set pricing and policies, control access, and define its own service catalog of virtual machine images for use in the cloud. A private cloud does require capital expenditure to procure hardware and software for the cloud. A staff of IT professionals must also be available to administer and manage services. To realize the greatest benefit of the cloud architecture, multiple data centers will implement distributed storage and compute infrastructure. Capacity planning is also an issue. A business could find a successful private cloud creates demands that exceed current capacity. Expanding a private cloud can require substantial capital expenditure; a hybrid model could be a better alternative.

Hybrid Clouds

A hybrid cloud combines public and private clouds. A business that has implemented a private cloud can use public cloud resources as an extension of their own cloud. There are a few different ways to do so.

The two clouds could be separately managed service platforms. Policies are established to govern what kinds of jobs can run in the public cloud, and cloud consumers have the option to run and manage their jobs in the public cloud. This approach gives cloud consumers freedom to choose between two services. There may be cases where the public cloud is less expensive or can provide capacity unavailable on the private cloud.

Another way to manage the hybrid private-public cloud is to enable access to the public cloud from within the service management platform. The two services are still independent, but cloud consumers would have a single point of management.

Finally, the public cloud could be treated as an extension of the private cloud by implementing a virtual private network (VPN) in the public cloud. Under this model, a portion of the public cloud is treated as an extension of the private cloud.

As is so often the case in information technology, there is more than one way to deliver a service, and the best option in any situation is highly dependent on specific requirements.

Summary

Cloud computing is relatively young, but in the short time since its inception, it has managed to create a host of competing definitions, architectures, service models, and delivery methods. Across all of these varying ways of looking at cloud computing, we find common characteristics, including massive scalability, ease of allocating resources, and a service management platform. Building on this foundation, cloud providers can deliver a range of services, from infrastructure to platforms to applications and business services. No single delivery model meets all needs, but the combination of public, private, and hybrid clouds offer a range of options suitable for many business requirements.

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.