1

The Cloud at Your Service

Cloud computing is a way to use and share hardware, operating systems, storage, and network capacity over the Internet. Cloud service providers rent virtualized servers, storage, and networking components to consumers to use on demand, as needed. The advantages are many. Cloud service providers can constantly update services and hardware, upgrade features, and add new ones, so consumers can stay on top of the technology wave without investing directly in infrastructure and, in some cases, applications. That frees companies to focus on their specific businesses and avoid much of the cost of purchasing, maintaining, upgrading, and troubleshooting hardware and software. In the cloud provisioning model, consumers subscribe to services on a pay-as-you-go basis or over a fixed period of time, most commonly by the month or year, and they pay only for what they use. Consumers also have the option of hosting their own cloud environment for security and compliance reasons.

Although the underlying concepts of cloud computing are not new, with the advent of new tools and technologies, these concepts have matured, and the cloud now can provide services efficiently, securely, and at a massive scale. From a hardware perspective, there is Infrastructure as a Service (IaaS). From a software angle, we have applications or Software as a Service (SaaS). And for blended offerings, there is Platform as a Service (PaaS). The benefit of all these approaches is that they reduce the enterprise's total cost of owning and maintaining computing resources.

This chapter discusses cloud computing in general and presents a reference architecture for cloud computing. The concepts of virtualization and workload deployment are common to all platforms from any vendor.

Cloud Computing

This book follows the National Institute of Standards and Technology (NIST) definition of cloud computing: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." There are five widely recognized essential characteristics of cloud computing.

- Broad network access—First and foremost, all cloud-related services must be available and accessible ubiquitously over the network via standard mechanisms that promote use by heterogeneous thick or thin clients. Clients can range from workstations and laptops to tablets and smart devices.
- On-demand self-service—Consumers must be able to unilaterally and automatically provision services as needed, without requiring human interaction. The services can range from simple email applications to applications that require server time and network storage. On-demand services are exemplified by Google's Gmail and Amazon's Amazon Web Services (AWS).
- Resource pooling—The provider's computing resources, such as memory, processing, storage, network bandwidth, and even virtual machines (VMs), must be pooled to serve multiple consumers using a multi-tenant model. The multi-tenant model dynamically assigns and reassigns location-transparent physical and virtual resources based on consumer demand.
- Rapid elasticity—The basic difference between traditional computing and cloud computing is the provisioning capability. In cloud computing, resources and services must be able to rapidly and automatically scale up when needed and be released or scaled down when they are not. To the consumer, these services and resources usually appear to be unlimited and can be appropriated in any quantity at any time.
- Metering of services—Because cloud computing employs a pay-per-use model, resource usage must be able to be measured, controlled, and reported transparently to both the provider and consumer of the service. Thus, cloud systems must have a metering capability that can control and optimize resource use.

In addition to these five essentials, the Cloud Security Alliance (*www.cloudsecurityalliance.org*) advocates a sixth characteristic of cloud computing:

- Multi-tenancy—Implicit in the way cloud computing hosts tenants is a need for policy-driven enforcement, segmentation, isolation, governance, service levels, and chargeback/billing models for different consumer constituencies. Multi-tenancy comes into play when consumers utilize a provider's public cloud service offerings.

These six characteristics lend themselves to certain service delivery models, the main ones being IaaS, PaaS, and SaaS. Figure 1.1 shows the relationship between these and other cloud service layers. A provider can choose to offer one or more of these services.

IaaS

IaaS deals primarily with hosting the environment. Infrastructure as a Service is the new way to use and share hardware, operating systems, storage, and network capacity over the Internet. Instead of using slices and logical partitioning, consumers can

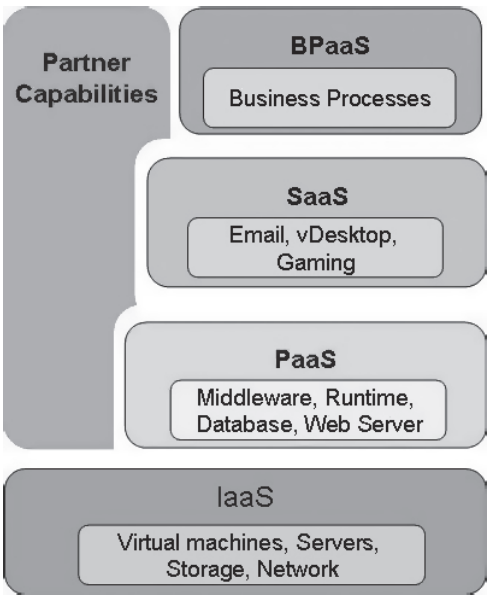

Figure 1.1: Cloud service layers

now rent virtualized servers, storage, and networking components. The consumer can provision processing, storage, networks, and other fundamental computing resources, then deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure, but does have control over the operating system, storage, and deployed applications, and possibly limited control of select networking components, such as switches in the datacenter and ports.

As in the old mainframe days, the service provider owns the equipment and is responsible for housing it, running it, and maintaining it. The characteristics of today's IaaS are

- self service via the Internet
- dynamic scaling, sometimes referred to as elasticity
- policy-based services
- utility computing billing

In the IaaS provisioning model, the consumer pays for the infrastructure on a per-use basis. A major benefit of cloud computing in general and IaaS in particular is that it allows companies to avoid up-front infrastructure costs and focus on business-related projects.

PaaS

As we move up the stack shown in Figure 1.1, we come to the computing platform, which includes the operating system, an execution engine, database, and Web server. Cloud vendors package these services and offer them as Platform as a Service.

PaaS concentrates on the building and deployment aspect of the cloud. Consumers can deploy applications to the cloud infrastructure using programming languages, libraries, services, and tools that the provider supports. Applications are typically middleware and can be consumer-created or acquired. The consumer does not manage or control the underlying cloud infrastructure such as the network, servers, or storage, but does have control over the deployed applications and possibly over configuration settings for the hosting environment.

PaaS resources can be used to run existing applications or develop and test new ones. Application developers have to worry only about developing their apps and running them on the platform as and when needed. Some vendors offer automatic scaling of computing and storage resources to match application demand.

Note: New applications created on cloud platforms are termed "born on the cloud."

Vendors of PaaS cloud services not only provide the infrastructure but also manage it and support client applications. Vendors can continually update services and add new features. PaaS providers can assist developers from the conception of original ideas to creating, testing, and deploying applications.

Vendors typically include the following resources and features in a PaaS offering:

- operating system
- database
- middleware
- server-side scripting
- compute nodes
- storage
- network access
- development and test tools
- maintenance and support

PaaS has several advantages for software developers because operating system complexities are hidden, while the resources and features are easily accessible and can be increased or decreased according to demand. Development teams may be geographically distributed, yet are still able to work together on software projects. Services can be obtained from diverse sources, some of which may be on-premises while others are hosted on remote systems across the globe. Enterprises can reduce setup and ongoing costs and eliminate unnecessary duplication of functions by using

infrastructure services from a single vendor rather than maintaining multiple hardware facilities on their own. And IT expenses can be minimized by sharing services and repositories and centralizing software development and test environments.

It's important to mention the subtle differences between PaaS and traditional distributed systems. Tasks such as continuous integration, build, deploy, and test are automated and repeatable. PaaS resources can be scaled up or conserved according to policies and need, and platform-neutral scripting makes it possible for applications to interact with the system firmware. In short, PaaS lets systems administrators focus on systems rather than servers, helps architects evaluate new technology quickly and directly, and enables IT developers to quickly develop and test projects.

On the downside, there are some pitfalls to be aware of. PaaS involves some risk of "lock-in" if offerings require proprietary service interfaces or development languages. And the flexibility of offerings might not meet the needs of users whose requirements evolve rapidly.

SaaS

Software as a Service is all about consumption. SaaS hosts software and applications on the cloud and delivers them to consumers as a service, sometimes referred to as "on-demand software." The consumer uses the provider's applications, which run on a cloud infrastructure. Although the applications are ubiquitously accessible, the consumer does not manage or control the underlying cloud infrastructure—not even the individual applications. The only aspects that the consumer has some control over are who can access the software and user-specific application configuration settings.

In most cases, users access SaaS via a thin client, such as a Web browser. SaaS is a common delivery model for everything from simple word processing software to enterprise resource planning, business process management, and customer relationship management software. Customer relationship management is the largest market for SaaS, but this delivery model is also used for business applications such as messaging, database management, and business management applications.

One of the main selling points of SaaS is the potential to reduce IT costs by outsourcing hardware and software maintenance and support to the SaaS provider. The pricing model for SaaS applications is typically a monthly or yearly flat fee per user, so the cost of the service is predictable and adjusted whenever users are added or removed.

Another cloud service layer that is garnering attention is Business Process as a Service, or BPaaS. A type of business process outsourcing, BPaaS is any horizontal or vertical business process—transaction processing, for example—that is delivered via the cloud services model. BPaaS minimizes up-front capital expenditures and reduces operational expenses. Some argue that BPaaS is simply a specialized SaaS offering similar to desktop as a service, communication as a service, or database as a service (DBaaS). A good debate is a learning experience.

Responsibilities

Figure 1.2 is a different way of looking at what constitutes cloud services and specifically at who is responsible for the various components. In the traditional IT department, the organization buys all hardware and software, houses it in a data center, and is in charge of running the operations. In cloud jargon, this approach is known as on-premises systems, and the client either manages the resources in-house or outsources management to an IT operations company.

Notice that as you move from left (IaaS) to right (SaaS), responsibility shifts from the client to the cloud provider. Divesting operations that are not directly related to the business is a benefit to the client. A retail company can concentrate on making

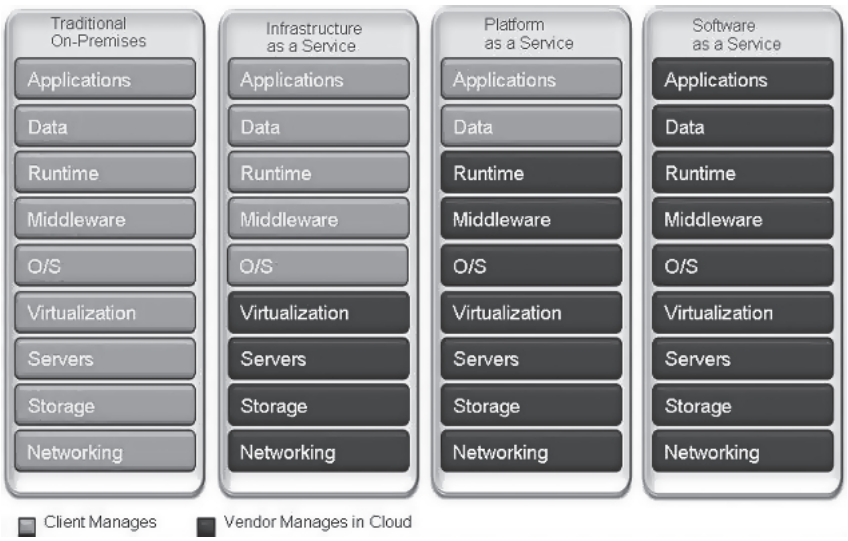

Figure 1.2: Separation of responsibilities

and marketing its products and not have to worry about managing its computing hardware or software, for example. In addition, by letting a cloud vendor manage things, the client is assured of getting and working with the latest in technology.

Although cloud computing proponents point to the cost savings and agility of such an environment, some IT organizations view it as a threat to their roles and a challenge to their expertise. Chapter 8 discusses how the cloud is changing IT roles and responsibilities.

Cloud Computing Reference Architecture

Reference architectures serve as a blueprint for architects and practitioners in the design of a solution or a "to-be" model. A Cloud Computing Reference Architecture (CCRA) assists in the design of public and private clouds. Specifically, it helps developers define scope and make architectural decisions, thus assisting teams in delivering consistent design and project results. Most reference architectures are created over time, based on experiences and real-world implementations.

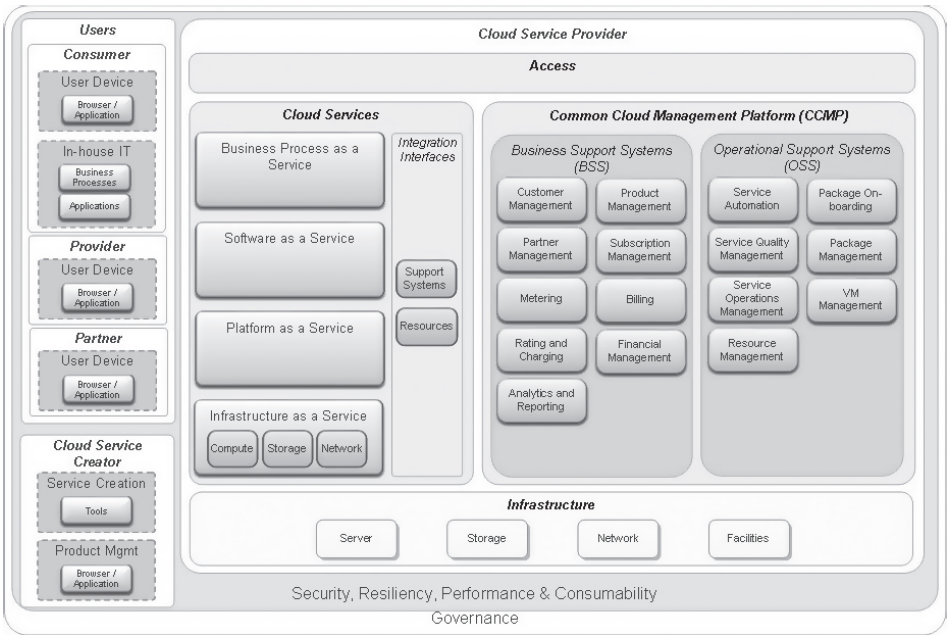

Figure 1.3: IBM's CCRA

At a high level, the major actors of a CCRA are a cloud service creator, a cloud service provider, and a cloud service consumer. The reference architecture may also include a cloud broker and cloud auditor. Figure 1.3 shows a pictorial representation of a CCRA that's based on the IBM® CCRA.

Although the cloud service provider seems to occupy the largest chunk of territory in Figure 1.3, service creators play a big role, and service consumers are also key actors. The ease with which consumers can access and use services is the hallmark of a successful implementation of the CCRA.

Cloud-Ready vs. Cloud-Centric Applications

The terms *cloud-ready* and *cloud-centric* describe the constitution of applications meant for the cloud. The distinction between these two terms is straightforward. When an existing application is deployed into either a private or public cloud, it is known as a cloud-ready or migrated-to-the-cloud application. New applications that are created specifically to run on the cloud are known as cloud-centric or born-on-the-cloud.

This distinction brings up a couple of nuances. Traditional applications need to be designed and developed in such a way that they can take advantage of capabilities provided by the cloud platform, be it at the PaaS layer or IaaS layer. Even if a traditional application runs on the cloud, it is still based on the old tenets of compute-limited software design. Applications designed specifically for the cloud will perform better in the cloud than those that are adapted to a cloud model.

Cloud-centric or born-on-the-cloud applications should be developed

- to take advantage of cloud computing features such as scalability, which is the ability to handle additional workloads in the future, and elasticity, which is the ability to dynamically add resources or give back resources according to demand;
- using the new breed of tools and runtimes that are more nimble and dynamic;
- with constant change and upgrade in mind, using a DevOps approach; and
- to operate in a multi-tenant model

Adapting and developing services for the cloud does not mean all traditional apps and tools have to be abandoned. However, new rules of application design must

be followed. Application isolation, security, and scalability have to be kept in mind when adapting traditional apps to take advantage of the cloud delivery model. These apps need to run without colliding with other apps in the shared infrastructure and should not break. Additionally, cloud application designers and developers should keep in mind that there is a strong requirement these days for mobile support for most customer-facing applications.

Cloudlets

In addition to the cloud terms discussed in this chapter, you may hear references to *community cloud*, which is a collaborative effort between several organizations to share infrastructure and resources. And you probably have heard that there are three basic types of cloud solutions: private cloud, public cloud, and hybrid cloud. The public cloud uses primarily IaaS. PaaS lends itself to private clouds, but it is moving to public clouds as well.

To quote an IBM executive, there are three business objectives that the cloud enables:

- Speed: enterprises can quickly obtain IT resources.
- Economics: capital expenditures are minimized.
- Empowerment: developers and users can access computing resources when needed.

One customer who has adopted cloud computing says, "We no longer plan for capacity because we have capacity on demand." Another observes, "Cloud is not just about infrastructure; it allows us to use ready-made applications in a cost-effective manner."

Of course, not everything about the cloud is rosy. IT architects face challenges regarding the lack of standardization, and CIOs need to consider whether a particular cloud service provider will still be in business five years from now. Nevertheless, in 2012 Gartner predicted that by 2016, 80 percent of Fortune 1000 enterprises will use some cloud computing services, and 20 percent of business will own no IT assets at all. Cloud computing and IT security remain at the top of the list of key priorities for 2015-2016 identified by IT decision-makers in Peak 10's annual U.S. market survey. And Forrester Research forecasts that the global market for cloud computing will grow to more than $241 billion by 2020.