I

BASICS

1

CLOUD COMPUTING

The U.S. National Institute of Standards and Technology (NIST) defines cloud computing as follows:

> Cloud computing is a model for enabling ubiquitous, 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 [NIST-800-145].

This definition frames cloud computing as a "utility" (or a "pay as you go") consumption model for computing services, similar to the utility model deployed for electricity, water, and telecommunication service. Once a user is connected to the computing (or telecommunications, electricity, or water utility) cloud, they can consume as much service as they would like whenever they would like (within reasonable limits), and are billed for the resources consumed. Because the resources delivering the service can be shared (and hence amortized) across a broad pool of users, resource utilization and operational efficiency can be higher than they would be for dedicated resources for each individual user, and thus the price of the service to the consumer may well be lower from a cloud/utility provider compared with the alternative of deploying and

Reliability and Availability of Cloud Computing, First Edition. Eric Bauer and Randee Adams.

operating private resources to provide the same service. Overall, these characteristics facilitate outsourcing production and delivery of these crucial "utility" services. For example, how many individuals or enterprises prefer to generate all of their own electricity rather than purchasing it from a commercial electric power supplier?

This chapter reviews the essential characteristics of cloud computing, as well as several common characteristics of cloud computing, considers how cloud data centers differ from traditional data centers, and discusses the cloud service and cloud deployment models. The terminologies for the various roles in cloud computing that will be used throughout the book are defined. The chapter concludes by reviewing the benefits of cloud computing.

1.1 ESSENTIAL CLOUD CHARACTERISTICS

Per [NIST-800-145], there are five essential functional characteristics of cloud computing:

1. on-demand self service;
2. broad network access;
3. resource pooling;
4. rapid elasticity; and
5. measured service.

Each of these is considered individually.

1.1.1 On-Demand Self-Service

Per [NIST-800-145], the essential cloud characteristic of "on-demand self-service" means "a consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider." Modern telecommunications networks offer on-demand self service: one has direct dialing access to any other telephone whenever one wants. This behavior of modern telecommunications networks contrasts to decades ago when callers had to call the human operator to request the operator to place a long distance or international call on the user's behalf. In a traditional data center, users might have to order server resources to host applications weeks or months in advance. In the cloud computing context, on-demand self service means that resources are "instantly" available to service user requests, such as via a service/resource provisioning website or via API calls.

1.1.2 Broad Network Access

Per [NIST-800-145] "broad network access" means "capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs)." Users expect

to access cloud-based services anywhere there is adequate IP networking, rather than requiring the user to be in a particular physical location. With modern wireless networks, users expect good quality wireless service anywhere they go. In the context of cloud computing, this means users want to access the cloud-based service via whatever wireline or wireless network device they wish to use over whatever IP access network is most convenient.

1.1.3 Resource Pooling

Per [NIST-800-145], the essential characteristic of "resource pooling" is defined as: "the provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand." Service providers deploy a pool of servers, storage devices, and other data center resources that are shared across many users to reduce costs to the service provider, as well as to the cloud consumers that pay for cloud services. Ideally, the cloud service provider will intelligently select which resources from the pool to assign to each cloud consumer's workload to optimize the quality of service experienced by each user. For example, resources located on servers physically close to the end user (and which thus introduce less transport latency) may be selected, and alternate resources can be automatically engaged to mitigate the impact of a resource failure event. This is essentially the utility model applied to computing. For example, electricity consumers don't expect that a specific electrical generator has been dedicated to them personally (or perhaps to their town); they just want to know that their electricity supplier has pooled the generator resources so that the utility will reliably deliver electricity despite inevitable failures, variations in load, and glitches.

Computing resources are generally used on a very bursty basis (e.g., when a key is pressed or a button is clicked). Timeshared operating systems were developed decades ago to enable a pool of users or applications with bursty demands to efficiently share a powerful computing resource. Today's personal computer operating systems routinely support many simultaneous applications on a PC or laptop, such as simultaneously viewing multiple browser windows, doing e-mail, and instant messaging, and having virus and malware scanners running in the background, as well as all the infrastructure software that controls the keyboard, mouse, display, networking, real-time clock, and so on. Just as intelligent resource sharing on your PC enables more useful work to be done cost effectively than would be possible if each application had a dedicated computing resource, intelligent resource sharing in a computing cloud environment enables more applications to be served on less total computing hardware than would be required with dedicated computing resources. This resource sharing lowers costs for the data center hosting the computing resources for each application, and this enables lower prices to be charged to cloud consumers than would be possible for dedicated computing resources.

1.1.4 Rapid Elasticity

[NIST-800-145] describes "rapid elasticity" as "capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out, and rapidly released

to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time."

Forecasting future demand is always hard, and there is always the risk that unforeseen events will change plans and thereby increase or decrease the demand for service. For example, electricity demand spikes on hot summer afternoons when customers crank up their air conditioners, and business applications have peak usage during business hours, while entertainment applications peak in evenings and on weekends. In addition, most application services have time of day, day of week, and seasonal variations in traffic volumes. Elastically increasing service capacity during busy periods and releasing capacity during off-peak periods enables cloud consumers to minimize costs while meeting service quality expectations. For example, retailers might experience heavy workloads during the holiday shopping season and light workloads the rest of the year; elasticity enables them to pay only for the computing resources they need in each season, thereby enabling computing expenses to track more closely with revenue. Likewise, an unexpectedly popular service or particularly effective marketing campaign can cause demand for a service to spike beyond planned service capacity. End users expect available resources to "magically" expand to accommodate the offered service load with acceptable service quality. For cloud computing, this means all users are served with acceptable service quality rather than receiving "busy" or "try again later" messages, or experiencing unacceptable service latency or quality.

Just as electricity utilities can usually source additional electric power from neighboring electricity suppliers when their users' demand outstrips the utility's generating capacity, arrangements can be made to overflow applications from one cloud that is operating at capacity to other clouds that have available capacity. This notion of gracefully overflowing application load from one cloud to other clouds is called "cloud bursting."

1.1.5 Measured Service

[NIST-800-145] describes the essential cloud computing characteristic of "measured service" as "cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and the consumer of the utilized service." Cloud consumers want the option of usage-based (or pay-as-you-go) pricing in which their price is based on the resources actually consumed, rather than being locked into a fixed pricing arrangement. Measuring resource consumption and appropriately charging cloud consumers for their actual resource consumption encourages them not to squander resources and release unneeded resources so they can be used by other cloud consumers.

1.2 COMMON CLOUD CHARACTERISTICS

NIST originally included eight common characteristics of cloud computing in their definition [NIST-B], but as these characteristics were not essential, they were omitted

from the formal definition of cloud computing. Nevertheless, six of these eight common characteristics do impact service reliability and service availability, and thus will be considered later in this book.

- *Virtualization.* By untethering application software from specific dedicated hardware, virtualization technology (discussed in Chapter 2, "Virtualization") gives cloud service providers control to manage workloads across massive pools of compute servers.
- *Geographic Distribution.* Having multiple geographically distributed data center sites enables cloud providers flexibility to assign a workload to resources close to the end user. For example, for real-time gaming, users are more likely to have an excellent quality of experience via low service latency if they are served by resources geographically close to them than if they are served by resources on another continent. In addition, geographic distribution in the form of georedundancy is essential for disaster recovery and business continuity planning. Operationally, this means engineering for sufficient capacity and network access across several geographically distributed sites so that a single disaster will not adversely impact more than that single site, and the impacted workload can be promptly redeployed to nonaffected sites.
- *Resilient Computing.* Hardware devices, like hard disk drives, wear out and fail for well-understood physical reasons. As the pool of hardware resources increases, the probability that some hardware device will fail in any week, day, or hour increases as well. Likewise, as the number of online servers increases, so does the risk that software running on one of those online server instances will fail. Thus, cloud computing applications and infrastructure must be designed to routinely detect, diagnose, and recover service following inevitable failures without causing unacceptable impairments to user service.
- *Advanced Security.* Computing clouds are big targets for cybercriminals and others intent on disrupting service, and the homogeneity and massive scale of clouds make them particularly appealing. Advanced security techniques, tools, and policies are essential to assure that malevolent individuals or organizations don't penetrate the cloud and compromise application service or data.
- *Massive Scale.* To maximize operational efficiencies that drive down costs, successful cloud deployments will be of massive scale.
- *Homogeneity.* To maximize operational efficiencies, successful cloud deployments will limit the range of different hardware, infrastructure, software platforms, policies and procedures they support.

1.3 BUT WHAT, EXACTLY, IS CLOUD COMPUTING?

Fundamentally, cloud computing is a new business model for operating data centers. Thus, one can consider cloud computing in two steps:

1. What is a data center?
2. How is a cloud data center different from a traditional data center?

1.3.1 What Is a Data Center?

A data center is a physical space that is environmentally controlled with clean electrical power and network connectivity that is optimized for hosting servers. The temperature and humidity of the data center environment are controlled to enable proper operation of the equipment, and the facility is physically secured to prevent deliberate or accidental damage to the physical equipment. This facility will have one or more connections to the public Internet, often via redundant and physically separated cables into redundant routers. Behind the routers will be security appliances, like firewalls or deep packet inspection elements, to enforce a security perimeter protecting servers in the data center. Behind the security appliances are often load balancers which distribute traffic across front end servers like web servers. Often there are one or two tiers of servers behind the application front end like second tier servers implementing application or business logic and a third tier of database servers. Establishing and operating a traditional data center facility—including IP routers and infrastructure, security appliances, load balancers, servers' storage and supporting systems—requires a large capital outlay and substantial operating expenses, all to support application software that often has widely varying load so that much of the resource capacity is often underutilized.

The Uptime Institute [Uptime and TIA942] defines four tiers of data centers that characterize the risk of service impact (i.e., downtime) due to both service management activities and unplanned failures:

- *Tier I.* Basic
- *Tier II.* Redundant components
- *Tier III.* Concurrently maintainable
- *Tier IV.* Fault tolerant

Tier I "basic" data centers must be completely shut down to execute planned and preventive maintenance, and are fully exposed to unplanned failures. [UptimeTiers] offers "Tier 1 sites typically experience 2 separate 12-hour, site-wide shutdowns per year for maintenance or repair work. In addition, across multiple sites and over a number of years, Tier I sites experience 1.2 equipment or distribution failures on an average year." This translates to a data center availability rating of 99.67% with nominally 28.8 hours of downtime per year.

Tier II "redundant component" data centers include some redundancy and so are less exposed to service downtime. [UptimeTiers] offers "the redundant components of Tier II topology provide some maintenance opportunity leading to just 1 site-wide shutdown each year and reduce the number of equipment failures that affect the IT operations environment." This translates to a data center availability rating of 99.75% with nominally 22 hours of downtime per year.

Tier III "concurrently maintainable" data centers are designed with sufficient redundancy that all service transition activities can be completed without disrupting

service. [UptimeTiers] offers "experience in actual data centers shows that operating better maintained systems reduces unplanned failures to a 4-hour event every 2.5 years. . . ." This translates to a data center availability rating of 99.98%, with nominally 1.6 hours of downtime per year.

Tier IV "fault tolerant" data centers are designed to withstand any single failure and permit service transition type activities, such as software upgrade to complete with no service impact. [UptimeTiers] offers "Tier IV provides robust, Fault Tolerant site infrastructure, so that facility events affecting the computer room are empirically reduced to (1) 4-hour event in a 5 year operating period. . . ." This translates to a data center availability rating of 99.99% with nominally 0.8 hours of downtime per year.

1.3.2 How Does Cloud Computing Differ from Traditional Data Centers?

Not only are data centers expensive to build and maintain, but deploying an application into a data center may mean purchasing and installing the computing resources to host that application. Purchasing computing resources implies a need to do careful capacity planning to decide exactly how much computing resource to invest in; purchase too little, and users will experience poor service; purchase too much and excess resources will be unused and stranded. Just as electrical power utilities pool electric power-generating capacity to offer electric power as a service, cloud computing pools computing resources, offers those resources to cloud consumers on-demand, and bills cloud consumers for resources actually used. Virtualization technology makes operation and management of pooled computing resources much easier. Just as electric power utilities gracefully increase and decrease the flow of electrical power to customers to meet their individual demand, clouds elastically grow and shrink the computing resources available for individual cloud consumer's workloads to match changes in demand. Geographic distribution of cloud data centers can enable computing services to be offered physically closer to each user, thereby assuring low transmission latency, as well as supporting disaster recovery to other data centers. Because multiple applications and data sets share the same physical resources, advanced security is essential to protect each cloud consumer. Massive scale and homogeneity enable cloud service providers to maximize efficiency and thus offer lower costs to cloud consumers than traditional or hosted data center options. Resilient computing architectures become important because hardware failures are inevitable, and massive data centers with lots of hardware means lots of failures; resilient computing architectures assure that those hardware failures cause minimal service disruption. Thus, the difference between a traditional data center and a cloud computing data center is primarily the business model along with the policies and software that support that business model.

1.4 SERVICE MODELS

NIST defines three service models for cloud computing: infrastructure as a service, platform as a service, and software as a service. These cloud computing service models logically sit above the IP networking infrastructure, which connects end users to the

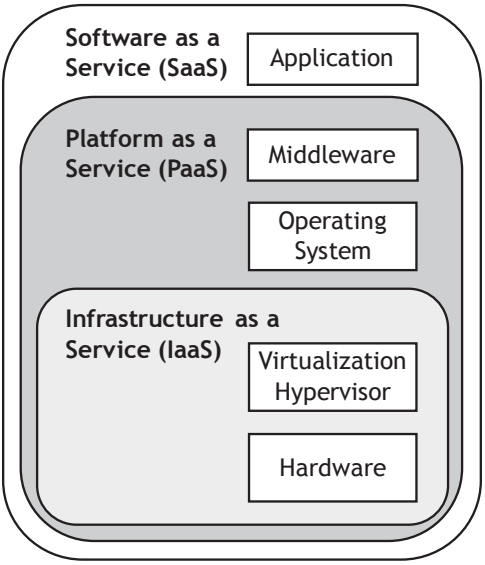

Figure 1.1. Service Models.

applications hosted on cloud services. Figure 1.1 visualizes the relationship between these service models.

The cloud computing service models are formally defined as follows.

- *Infrastructure as a Service (*IaaS*).* "[T]he capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls)" [NIST-800-145]. IaaS services include: compute, storage, content delivery networks to improve performance and/or cost of serving web clients, and backup and recovery service.
- *Platform as a Service (*PaaS*).* "[T]he capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations" [NIST-800-145]. PaaS services include: operating system, virtual desktop, web services delivery and development platforms, and database services.
- *Software as a Service (*SaaS*).* "[T]he capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The consumer does not manage or control the underlying cloud infrastructure including

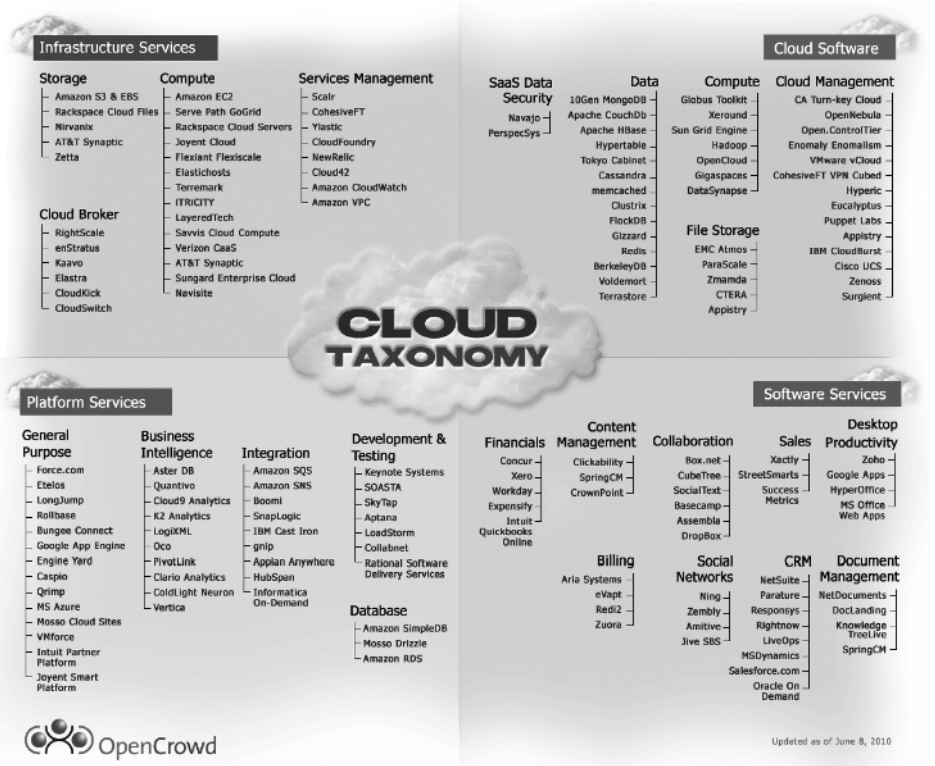

Figure 1.2. OpenCrowd's Cloud Taxonomy.

Source: Copyright 2010, Image courtesy of OpenCrowd, opencrowd.com.

network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings" [NIST-800-145]. SaaS applications include: e-mail and office productivity; customer relationship management (CRM), enterprise resource planning (ERP); social networking; collaboration; and document and content management.

Figure 1.2 gives concrete examples of IaaS, PaaS, and SaaS offerings.

1.5 CLOUD DEPLOYMENT MODELS

NIST recognizes four cloud deployment models:

- *Private Cloud.* "the cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise." [NIST-800-145]

- *Community Cloud.* "the cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise" [NIST-800-145].
- *Public Cloud.* "the cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services" [NIST-800-145].
- *Hybrid Cloud.* "the cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds)" [NIST-800-145].

Cloud service providers typically offer either private, community or public clouds, and cloud consumers select which of those three to use, or adopt a hybrid deployment strategy blending private, community and/or public clouds.

1.6 ROLES IN CLOUD COMPUTING

Cloud computing opens up interfaces between applications, platform, infrastructure, and network layers, thereby enabling different layers to be offered by different service providers. While NIST [NIST-C] and some other organizations propose new roles of *cloud service consumers*, *cloud service distributors*, *cloud service developers and vendors*, and *cloud service providers*, the authors will use the more traditional roles of suppliers, service providers, cloud consumers, and end users, as illustrated in Figure 1.3.

Specific roles in Figure 1.3 are defined below.

- *Suppliers* develop the equipment, software, and integration services that implement the cloud-based and client application software, the platform software, and the hardware-based systems that support the networking, compute, and storage that underpin cloud computing.
- *Service providers* own, operate, and maintain the solutions, systems, equipment, and networking needed to deliver service to end users. The specific service provider roles are defined as:
  - *IP network service providers* carry IP communications between end user's equipment and IaaS provider's equipment, as well as between IaaS data centers. Network service providers operate network equipment and facilities to provide Internet access and/or wide area networking service. Note that while there will often be only a single infrastructure, platform, and software service provider for a particular cloud-based application, there may be several different network service providers involved in IP networking between the IaaS

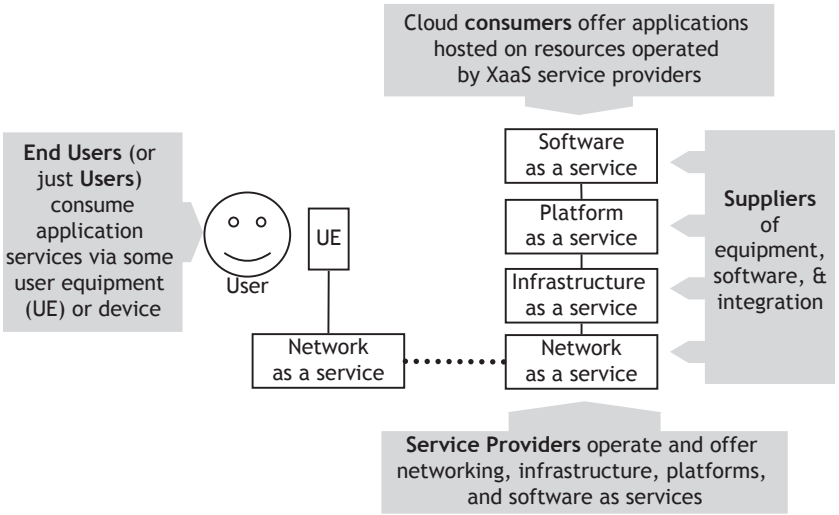

Figure 1.3. Roles in Cloud Computing.

service provider's equipment and end users' equipment. Internet service providers and Internet access providers are examples of network service providers. While IP networking service is not explicitly recognized in NIST's service model, these service providers have a crucial role in delivering end-to-end services to cloud users and can thus impact the quality of experience for end users.

- *IaaS providers* "have control of hardware, hypervisor and operating system, to provide services to consumers. For IaaS, the provider maintains the storage, database, message queue or other middleware, or the hosting environment for virtual machines. The [PaaS/SaaS/cloud] consumer uses that service as if it was a disk drive, database, message queue, or machine, but they cannot access the infrastructure that hosts it" [NIST-C]. Most IaaS providers focus on providing complete computing platforms for consumers' VMs, including operating system, memory, storage, and processing power. Cloud consumers often pay for only what they use, which fits nicely into most companys' computing budget.
- *PaaS providers* "take control of hardware, hypervisor, OS and middleware, to provide services. For PaaS, the provider manages the cloud infrastructure for the platform, typically a framework for a particular type of application. The consumer's application cannot access the infrastructure underneath the platform" [NIST-C]. PaaS providers give developers complete development environments in which to code, host, and deliver applications. The development environment typically includes the underlying infrastructure, development tools, APIs, and other related services.

  - *SaaS providers* "rely on hardware, hypervisor, OS, middleware, and application layers to provide services. For SaaS, the provider installs, manages and maintains the software. The provider does not necessarily own the physical infrastructure in which the software is running. Regardless, the consumer does not have access to the infrastructure; they can access only the application"[NIST-C]. Common SaaS offerings include desktop productivity, collaboration, sales and customer relationship management, and documentation management.
- *Cloud consumers*, (or simply "consumers") are generally enterprises offering specific application services to end users by arranging to have appropriately configured software execute on XaaS resources hosted by one or more service providers. Cloud consumers pay service providers for cloud XaaS resources consumed. End users are typically aware only of the enterprise's application; the services offered by the various XaaS service providers are completely invisible to end users.
- *End users* (or simply *users*) use the software applications hosted on the cloud. Users access cloud-based applications via IP networking from some user equipment, such as a smartphone, laptop, tablet, or PC.

There are likely to be several different suppliers and service providers supporting a single cloud consumer's application to a community of end users. The cloud consumer may have some supplier role in developing and integrating the software and solution. It is possible that the end users are in the same organization as the one that offers the cloud-based service to end users.

1.7 BENEFITS OF CLOUD COMPUTING

The key benefit of cloud computing for many enterprises is that it turns IT from a capital intensive concern to a pay-as-you-go activity where operating expenses track usage—and ideally computing expenses track revenue. Beyond this strategic capital expense to operating expense shift, there are other benefits of cloud computing from [Kundra] and others:

- *Increased Flexibility.* Rapid elasticity of cloud computing enables resources engaged for an application to promptly grow and later shrink to track the actual workload so cloud consumers are better able to satisfy customer demand without taking financial risks associated with accurately predicting future demand.
- *Rapid Implementation.* Cloud consumers no longer need to procure, install, and bring into service new compute capacity before offering new applications or serving increased workloads. Instead, they can easily buy the necessary computing capacity "off the shelf" from cloud service providers, thereby simplifying and shortening the service deployment cycle.
- *Increased Effectiveness.* Cloud computing enables cloud consumers to focus their scarce resources on building services to solve enterprise problems rather

than investing in deploying and maintaining computing infrastructure, thereby increasing their organizational effectiveness.

- *Energy Efficiency.* Cloud service providers have the scale and infrastructure necessary to enable effective sharing of compute, storage, networking, and data center resources across a community of cloud consumers. This not only reduces the total number of servers required compared with dedicated IT resources, but also reduces the associated power, cooling, and floor space consumed. In essence, intelligent sharing of cloud computing infrastructure enables higher resource utilization of a smaller overall pool of resources compared with dedicated IT resources for each individual cloud consumer.

1.8 RISKS OF CLOUD COMPUTING

As cloud computing essentially outsources responsibility for critical IS/IT infrastructure to a service provider, the cloud consumer gives up some control and is confronted with a variety of new risks. These risks range from reduced operational control and visibility (e.g., timing and control of some software upgrades) to changes in accountability (e.g., provider service level agreements) and myriad other concerns. This book considers only the risks that service reliability and service availability of virtualized and cloud-based solutions will fail to achieve performance levels the same as or better than those that traditional deployment scenarios have achieved.