1.1 Five main principles that define cloud computing

We can summarize the five main principles of cloud computing as follows:

- Pooled computing resources available to any subscribing users
- Virtualized computing resources to maximize hardware utilization
- Elastic scaling up or down according to need
- Automated creation of new virtual machines or deletion of existing ones
- Resource usage billed only as used

We assert, with very few notable exceptions called out later, that these five main principles are necessary components to call something *cloud computing*. They're summarized in table 1.1 with a brief explanation of each one for quick reference.

Table 1.1 The five main principles of cloud computing

Resource	Explanation
Pooled resources	Available to any subscribing users
Virtualization	High utilization of hardware assets
Elasticity	Dynamic scale without CAPEX
Automation	Build, deploy, configure, provision, and move, all without manual intervention
Metered billing	Per-usage business model; pay only for what you use

We'll now discuss these principles in concrete terms, making sure you understand what each one means and why it's a pillar of cloud computing.

1.1.1 Pooled computing resources

The first characteristic of cloud computing is that it utilizes pooled computing assets that may be externally purchased and controlled or may instead be internal resources that are pooled and not dedicated. We further qualify these pooled computing resources as contributing to a cloud if these resources are available to any subscribing users. This means that *anyone* with a credit card can subscribe.

If we consider a corporate website example, three basic operational deployment options are commonly employed today. The first option is the self-hosting option. Here,

companies choose not to run their own data center and instead have a third party lease them a server that the third party manages. Usually, managed hosting services lease corporate clients a dedicated server that isn't shared (but shared hosting is common as well). On this single principle, cloud computing acts like a *shared managed hosting service* because the cloud provider is a third party that owns and manages the physical computing resources which are shared with other users, but there the similarity ends.

Independent of cloud computing, a shift from self-hosted IT to outsourced IT resources has been underway for years. This has important economic implications. The two primary implications are a shift of capital expenses (CAPEX) to operational expenses (OPEX), and the potential reduction in OPEX associated with operating the infrastructure. The shift from CAPEX to OPEX means a lowering of the financial barrier for the initiation of a new project. (See the definition in section 3.1.)

In the self-hosted model, companies have to allocate a budget to be spent up front for the purchase of hardware and software licenses. This is a fixed cost regardless of whether the project is successful. In an outsourced model (managed hosting), the startup fees are typically equivalent to one month's operational cost, and you must commit to one year of costs up front. Typically, the one-year cost is roughly the same or slightly lower than the CAPEX cost for an equivalent project, but this is offset by the reduced OPEX required to operate the infrastructure. In sharp contrast, in a cloud model, there are typically no initial startup fees. In fact, you can sign up, authorize a credit card, and start using cloud services literally in less time than it would take to read this chapter. Figure 1.2 showcases side by side the various application deployment models with their respective CAPEX and OPEX sizes.

The drastic difference in economics that you see between the hosting models and the cloud is due to the fact that the cost structures for cloud infrastructures are vastly better than those found in other models. The reasons for the economies of scale are severalfold, but the primary drivers are related to the simple economics of volume. Walmart and Costco can buy consumer goods at a price point much lower than you or I could because of their bulk purchases. In the world of computing, the "goods" are computing, storage, power, and network capacity.

1.1.2 Virtualization of compute resources

The second of the five main principles of cloud computing has to do with virtualization of compute resources. Virtualization is nothing new. Most enterprises have been shifting much of their physical compute infrastructure to virtualized for the past 5 to 10 years. Virtualization is vital to the cloud because the

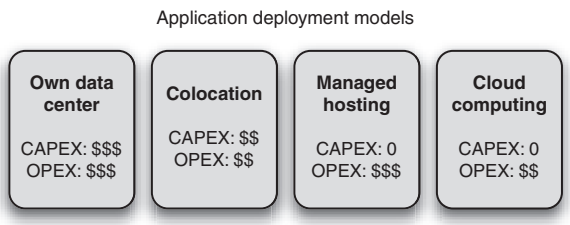

Figure 1.2 IT organizations have several alternatives for hosting applications. The choice of deployment model has different implications for the amount of CAPEX (up-front capital expenditure) and OPEX (ongoing operational costs). The number of $ signs represent the relative level of CAPEX and OPEX involved with the choice of deployment model.

scale of cloud infrastructures has to be enormous, based on thousands of servers. Each server takes up physical space and uses significant power and cooling. Getting high utilization out of each and every server is vital to be cost effective.

The recent technological breakthrough that enabled high utilization on commodity hardware—and which is the single biggest factor behind the cloud being a recent IT phenomenon—is virtualization where each physical server is partitioned into many virtual servers. Each one acts like a real server that can run an operating system and a full complement of applications.[1] Virtualized servers are the primary units that can be consumed as needed in the cloud. These virtualized servers constitute a large pool of resources available when required. But having such a large pool will work only if applications can use more or less of the pool as demands placed on the applications grow and shrink. As you'll see in chapter 4, the notion of a private cloud softens this first principal but keeps all the others.

1.1.3 Elasticity as resource demands grow and shrink

The fact that this large pool of resources exists enables a concept known as *elasticity*—the third of our five main principles. Elasticity is such a key concept in cloud computing that Amazon decided to name its cloud Amazon Elastic Compute Cloud.

Elasticity—a synonym for *dynamic scaling*—refers to the ability to dynamically change how much resource is consumed in response to how much is needed. Typical applications require a base level of resources under normal, steady-state conditions, but need more resource under peak load conditions.

In a non-cloud world, you would have to build sufficient capacity to not only perform adequately under baseline load conditions, but also handle peak load scenarios with sufficiently good performance. In the case of a self-hosted model, this means over-provisioning the amount of hardware for a given allocation. In the case of a managed hosting deployment, you can start with a small set of resources and grow as the requirements of the application grow. But provisioning for a new set of dedicated hardware resources takes weeks or, in many larger organizations, months. Having thousands of virtualized resources that can be harnessed and released in correlation to application demand would be useless if such allocation and freeing required manual intervention.

1.1.4 Automation of new resource deployment

The ability to automatically (via an API) provision and deploy a new virtual instance of a machine, and, equivalently, to be able to free or de-provision an instance, is our fourth principle of cloud computing. A cloud-deployed application can provision new instances on an as-needed basis, and these resources are brought online within minutes. After the peak demand ebbs, and you don't need the additional resources, these

[1] The rapid shift to multicore servers only strengthens the impact of virtualization. Each virtual machine with its operating system and full complement of applications can run on its own core simultaneously with all other virtual machines on the same physical server.

virtual instances can be taken offline and de-provisioned, and you will no longer be billed. Your incremental cost is only for the hours that those additional instances were in use and active.

1.1.5 Metered billing that charges only for what you use

The fifth distinguishing characteristic of cloud computing is a metered billing model. In the case of managed hosting, as we mentioned before, there typically is an initial startup fee and an annual contract fee. The cloud model breaks that economic barrier because it's a pay-as-you-go model. There is no annual contract and no commitment for a specific level of consumption.

Typically, you can allocate resources as needed and pay for them on an hourly basis. This economic advantage benefits not only projects being run by IT organizations, but also innumerable entrepreneurs starting new businesses. Instead of needing to raise capital as they might have in the past, they can utilize vast quantities of compute resources for pennies per hour. For them, the cloud has drastically changed the playing field and allowed the little guy to be on equal footing with the largest corporations.