2.1.2 Ensuring high server utilization in the cloud with virtualization

Virtualization, following the car analogy, is the suspension. It provides the high server utilization you need. It smoothes out the variations between applications that need barely any CPU time (they can share a CPU with other applications) and those that are compute intensive and need every CPU cycle they can get. Virtualization is the single-most revolutionary cloud technology whose broad acceptance and deployment truly enabled the cloud computing trend to begin. Without virtualization, and the 60-plus percent server utilization it allows, the economics of the cloud would not work.

> **VIRTUALIZATION** For this book, we're interested primarily in *platform* virtualization. Platform virtualization is a technique to abstract computer resources such that it separates the operating system from the underlying physical server resources. Instead of the OS running on (that is, directly using) hardware resources. The OS interacts instead with a new software layer called a *virtual machine monitor* that accesses the hardware and presents the OS with a virtual set of hardware resources. This means multiple virtual machine images or instances can run on a single physical server, and new instances can be generated and run on demand, creating the basis for elastic computing resources.

As we discussed earlier, virtualization isn't new at all. IBM mainframes used time-sharing virtualization in the '60s to enable many people to share a large computer without interacting or interfering with each other. Previously, constraints of scheduling dedicated time on these machines required you to get all your work for the day done in that scheduled time slot. The concept of virtual memory, introduced around 1962, although considered pretty radical, ultimately freed programmers from having to constantly worry about how close they were to the limits of physical memory. Today, server virtualization is proving equally dramatic for application deployment and scaling. And it's the key enabler for the cloud. How did this happen?

The average server in a corporate data center has typical utilization of only 6 percent.[5] Even at peak load, utilization is no better than 20 percent. In the best-run data centers, servers only run on average at 15 percent or less of their maximum capacity. But when these same data centers fully adopt server virtualization, their CPU utilization increases to 65 percent or higher. For this reason, in a few short years, most corporate data centers have deployed hundreds or thousands of virtual servers in place of their previous model of one server on one hardware computer box. Let's see how server virtualization works to make utilization jump this dramatically.

HOW IT WORKS

Server virtualization transforms or *virtualizes* the hardware resources of a computer—including the CPU, RAM, hard disk, and network controller—to create a fully functional virtual machine that can run its own operating system and applications like a physical computer. This is accomplished by inserting a thin layer of software directly on the computer hardware that contains a virtual machine monitor (VMM)—also called a *hypervisor*—that allocates hardware resources dynamically and transparently. Multiple guest operating systems run concurrently on a single physical computer and share hardware resources with each other. By encapsulating an entire machine, including CPU, memory, operating system, and network devices, a virtual machine becomes completely compatible with all standard operating systems, applications, and device drivers. You can see the virtual machine architecture for VMware on the x86 in figure 2.3.

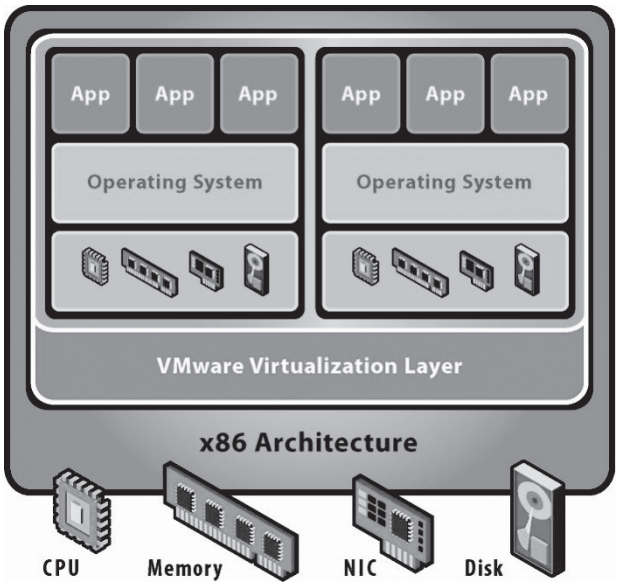

Figure 2.3 Virtual machine architecture using VMware as an example. The virtualization layer is what interfaces directly with all hardware components, including the CPU. That layer then presents each guest operating system with its own array of *virtual* hardware resources. The guest OS doesn't operate differently than it would if installed on the bare hardware, but now several instances of guest OSs with all their applications can share a single physical device and have higher effective utilization. Source: VMWare.

[5] McKinsey & Company, 2008 Data Center Efficiency report.

VIRTUALIZATION AS APPLIED TO THE CLOUD

When virtualization passed muster with enterprise architects and CIOs, it had arrived. It was all about saving money. Enterprises began seeing utilization of their hardware assets increase dramatically. It was easy to go from the typical 5 or 6 percent to 20 percent. They could get 65 percent utilization or better with good planning.

In addition to increased utilization and the associated cost savings, virtualization in corporate data centers set the stage for cloud computing in several interesting ways. It decoupled users from implementation; it brought speed, flexibility, and agility never before seen in corporate data centers; and it broke the old model of software pricing and licensing. Let's look at table 2.1 for more clarity.

Table 2.1 Impact of virtualization on corporate data centers

Benefit	Explanation
Decouples users from implementation	The concept of a virtual server forces users to not worry about the physical servers or their location. Instead, they focus on service-level agreements and their applications.
Decreases server provisioning from months to minutes	Getting a (physical) server requisitioned, installed, configured, and deployed takes larger organizations 60–90 days and some 120 days. In the virtual server model, it's literally minutes or hours from request to fully ready for application deployment, depending on how much automation has been put in place.
Breaks software pricing and licensing	No longer can the data center charge for an entire server or every server the software runs on. Instead, they have to charge for actual usage—a whole new model for IT.

Table 2.1 illustrates the services the cloud providers offer. We also see a growing recognition of and readiness for the cloud within the enterprise. This is because the model change that virtualization has already brought to enterprise IT has prepared companies to adapt more easily to the cloud computing model.

Let's look at a scenario that uses thousands of physical servers. Each one is virtualized and can run any number of guest OSs, can be configured and deployed in minutes, and is set up to bill by the CPU hour. The combination of cheap, abundant hardware and virtualization capability, coupled with automated provisioning and billing allows the huge economies of scale now achievable in the mega data centers to be harnessed through cloud computing. This is possible because of virtualization, much as car suspension systems enable vehicles to speed up without killing the occupants at every bump in the road.

But a powerful engine (data center) and a smooth suspension (virtualization) aren't enough. Following the vehicle analogy, you need a set of controls to start, stop, and steer the car; you need an API to control your cloud.