CHAPTER

# Data Center Evolution—Mainframes to the Cloud

# 2

The modern age of computing began in the 1950s when the first mainframe computers appeared from companies like IBM®, Univac, and Control Data. Communication with these computers was typically through a simple input/output (I/O) device. If you needed to compute something, you would walk to the computer room, submit your job as a stack of punch cards, and come back later to get a printout of the results. Mainframes later gave way to minicomputers like the PDP-11 from Digital Equipment Corporation (DEC), and new methods of computer networking started to evolve. Local area networks (LANs) became commonplace and allowed access to computing resources from other parts of the building or other parts of the campus. At the same time, small computers were transformed into servers, which "served up" certain types of information to client computers across corporate LANs. Eventually, servers moved into corporate data centers and they evolved from systems that looked like high-performance tower PCs into rack-mounted gear.

When the Advanced Research Projects Agency Network (ARPANET) gave birth to the internet, things started to get interesting. In order to provide web hosting services, dedicated data center facilities full of servers began to emerge. Initially, these data centers employed the same LAN networking gear used in the corporate data centers. By the end of the 1990s, Ethernet became the predominant networking technology in these large data centers, and the old LAN-based networking equipment was slowly replaced by purpose-built data center networking gear. Today, large cloud data center networks are common, and they require high-performance networks with special cloud networking features. This chapter will provide a brief history of the evolution of computer networking in order to give the reader a perspective that will be useful when reading the following chapters in this book.

## THE DATA CENTER EVOLUTION

Over the past 50 years or so, access to computer resources has come full circle from dumb client terminals connected to large central mainframes in the 1960s, to distributed desktop computing starting in the 1980s, to handhelds connected to large centralized cloud data centers today. You can think of the handheld device as a terminal receiving data computed on a server farm in a remote cloud data center, much like the terminal connected to the mainframe. In fact, for many applications, data processing is moving out of the client device and into the cloud. This section will provide an

overview of how computer networks have evolved from simple connections with large mainframe computers into today's hyper-scale cloud data center networks.

## Early mainframes

Mainframes were the first electronic computing systems used widely by businesses, but due to their high capital and operating costs, even large business or universities could afford only one computer at a given site. Because of the cost, time sharing became the mode of operation for these large computers. Client communication involved walking over to the computer center with a stack of punch cards or a paper tape, waiting a few hours, and then picking up a printout of the results. Later, teletype terminals were added, allowing users to type in commands and see results on printed paper. Originally, teletypes printed program commands on paper tape, which was manually fed into the computer. Later, teletypes were connected directly to the computer using proprietary communication protocols as shown in Figure 2.1. In the late 1960s, CRT terminals were becoming available to replace the teletype.

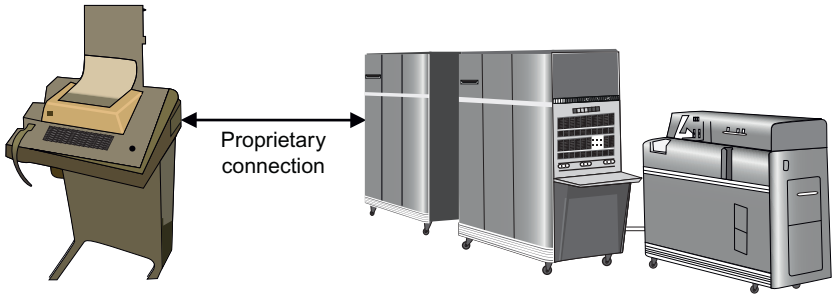


**FIGURE 2.1**

Mainframe client terminal connections.

## Minicomputers

In the late 1970s, integrated circuits from companies like Intel® were dramatically reducing the cost and size of the business computer. Companies such as DEC took advantage of these new chips to develop a new class of computing system called the minicomputer. Starting with the PDP-8 and then more famously the PDP-11, businesses could now afford multiple computing systems per location. I can remember walking through Bell Labs in the early 1980s where they were proudly showing a room full of PDP-11 minicomputers used in their research work. These computer rooms are now typically called enterprise data centers.

Around this same time, more sophisticated computer terminals were developed, allowing access to computing resources from different locations in the building or campus. By now, businesses had multiple minicomputers and multiple terminals accessing these computers as shown in Figure 2.2. The only way to efficiently connect

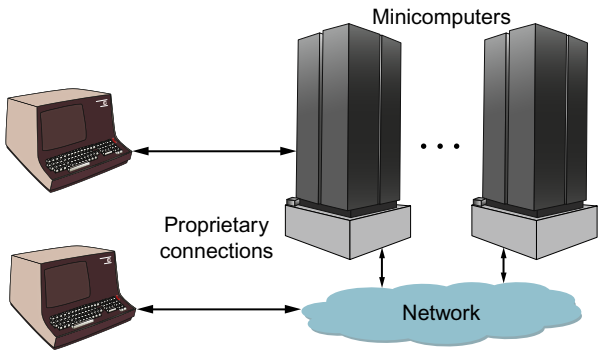


**FIGURE 2.2**

Minicomputer client terminal connections.

these was to build some sort of Local Area Network (LAN). This spawned a lot of innovation in computer network development which will be discussed in more detail in the next section.

## Servers

Around the late 1980s, IT administrators realized that there were certain types of information such as corporate documents and employee records that did not need the computing power of mainframes or minicomputers, but simply needed to be accessed and presented to the client through a terminal or desktop computer. At around the same time, single board computers were becoming more powerful and evolved into a new class of computers called workstations. Soon corporations were dedicating these single board computers to serve up information across their LANs. The age of the compute server had begun as shown in Figure 2.3.

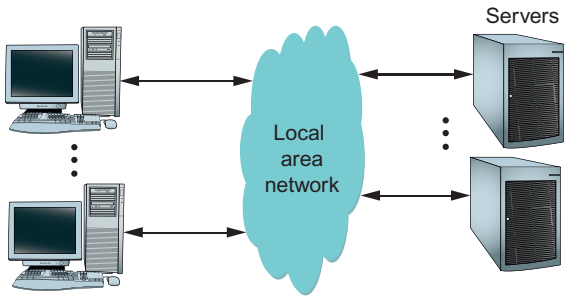


**FIGURE 2.3**

Early server network block diagram.

By the 1990s, almost all business employees had a PC or workstation at their desk connected to some type of LAN. Corporate data centers were becoming more complex with mixtures of minicomputers and servers which were also connected to the LAN. Because of this, LAN port count and bandwidth requirements were increasing rapidly, ushering in the need for more specialized data center networks. Several networking technologies emerged to address this need, including Ethernet and Token Ring which will be discussed in the next sections.

## Enterprise data centers

Through the 1990s, servers rapidly evolved from stand-alone, single board computers to rack-mounted computers and blade server systems. Ethernet emerged as the chosen networking standard within the data center with Fibre Channel used for storage traffic. Within the data center, the Ethernet networks used were not much different from the enterprise LAN networks that connected client computers to the corporate data center. Network administrators and network equipment manufacturers soon realized that the data center networks had different requirements compared with the enterprise LAN, and around 2006, the first networking gear specifically designed for the data center was introduced. Around that same time, industry initiatives, such as Fibre Channel over Ethernet (FCoE), were launched with the goal of converging storage and data traffic onto a single Ethernet network in the data center. Later in this chapter, we will compare traditional enterprise data center networks to networks specifically designed for the data center. Figure 2.4 shows a LAN connecting client computers to an enterprise data center that employs enterprise networking equipment.

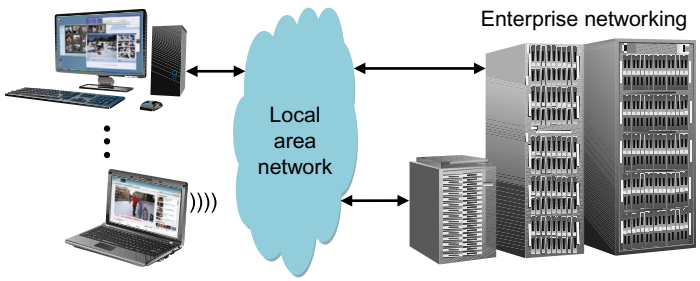


**FIGURE 2.4**

Enterprise data center networks.

## Cloud data centers

When I was in high school, I remember listening to the *Utopia* album from Todd Rundgren. This album had one side dedicated to a song called "The Ikon," which impressed upon me the idea of a central "mind" from which anyone could access any information they needed anytime they needed it. Well, we are definitely headed

in that direction with massive cloud data centers that can provide a wide variety of data and services to your handheld devices wherever and whenever you need it. Today, whether you are searching on Google, shopping on Amazon, or checking your status on Facebook, you are connecting to one of these large cloud data centers.

Cloud data centers can contain tens of thousands of servers that must be connected to each other, to storage, and to the outside world. This puts a tremendous strain on the data center network, which must be low cost, low power, and high bandwidth. To minimize the cost of these data centers, cloud service providers are acquiring specialized server boards and networking equipment which are built by Original Design Manufacturers (ODMs) and are tailored to their specific workloads. Facebook has even gone as far as spearheading a new server rack standard called the Open Compute Project that better optimizes server density by expanding to a 21-inch wide rack versus the old 19-inch standard. Also, some cloud data center service providers, such as Microsoft, are using modular Performance Optimized Data center modules (PODs) as basic building blocks. These are units about the size of a shipping container and include servers, storage, networking, power, and cooling. Simply stack the containers, connect external networking, power, and cooling, and you're ready to run. If a POD fails, they bring in a container truck to move it out and move a new one in. Later in this chapter, we will provide more information on the types of features and benefits enabled by these large cloud data centers. Figure 2.5 is a pictorial representation showing client devices connected through the Internet to a large cloud data center that utilizes specialized cloud networking features.

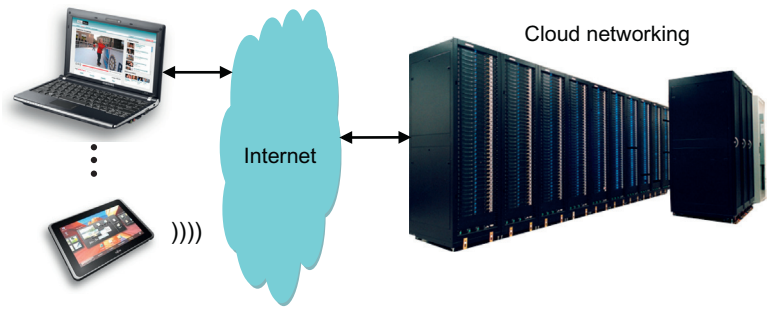


**FIGURE 2.5**

Cloud data center networks.

## Virtualized data centers

Many corporations are seeing the advantage of moving their data center assets into the cloud in order to save both capital and operating expense. To support this, cloud data centers are developing ways to host multiple virtual data centers within their physical data centers. But the corporate users want these virtual data centers to appear to them as private data centers. This requires the cloud service provider to offer isolated, multitenant environments that include a large number of virtual

machines and virtualized networks as shown in Figure 2.6. In this simplified view, we show three tenants that are hosted within a large cloud data center.

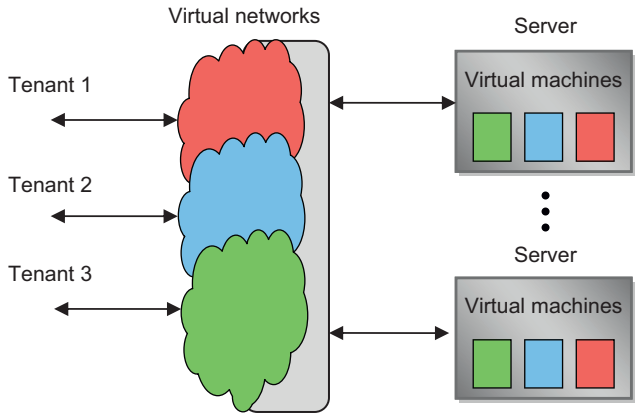


**FIGURE 2.6**

The virtualized data center.

Within the physical servers, multiple virtual machines (virtual servers) can be maintained which help maximize data center efficiency by optimizing processing resource utilization while also providing server resiliency. Within the network, tunneling protocols can be used to provide multiple virtual networks within one large physical network. Storage virtualization can also be used to optimize storage performance and utilization. In this book, we will not go very deep into storage virtualization and only describe virtual machines in the context of data center networking. But we will dive deeper into some of the network tunneling standards that are employed for these multitenant environments.

## COMPUTER NETWORKS

In the last section, we went through a brief history of enterprise computing and the evolution toward cloud data centers. We also mentioned local area networking as a key technology development that eventually evolved into purpose-built data center networks. A variety of different network protocols were developed over the last 50 years for both LANs and wide area networks (WANs), with Ethernet emerging as the predominant protocol used in local area, data center, and carrier networks today. In this section, we will provide a brief history of these network protocols along with some information on how they work and how they are used. For completeness, we are including some protocols that are used outside the data center because they provide the reader with a broader view of networking technology. Ethernet will be covered separately in the following section.

## Dedicated lines

As you may have guessed, the initial methods used to communicate with mainframe computers were through dedicated lines using proprietary protocols. Each manufacturer was free to develop its own communication protocols between computers and devices such as terminals and printers because the end customer purchased everything from the same manufacturer. These were not really networks per se, but are included in this chapter in order to understand the evolution to true networking technology. Soon, corporations had data centers with multiple computer systems from different manufacturers along with remote user terminals, so a means of networking these machines together using industry standard protocols became important. The rest of this section will outline some of these key protocols.

## ARPANET

The ARPANET was one of the first computer networks and is considered to be the father of today's internet. It was initially developed to connect mainframe computers from different universities and national labs through leased telephone lines at the astounding rate of 50Kbit per second. To put it into today's terms, that's 0.00005Gbps. Data was passed between Interface Message Processors (IMPs), which today we would call a router. Keep in mind that there were only a handful of places you could route a message to back then, including universities and research labs.

ARPANET also pioneered the concept of packet routing. Before this time, both voice and data information was forwarded using circuit-switched lines. Figure 2.7 shows an example of the difference between the two. In a circuit-switched network, a connection path is first established, and data sent between point A and B will always take the same path through the network. An example is a phone call where a number is dialed, a path is set up, voice data is exchanged, the call ends, and then the path is taken down. A new call to the same number may take a different path through the network, but once established, data always takes the same path. In addition, data is broken up into fixed sized cells such as voice data chucks before it is sent through the network (see the section "SONET/SDH").

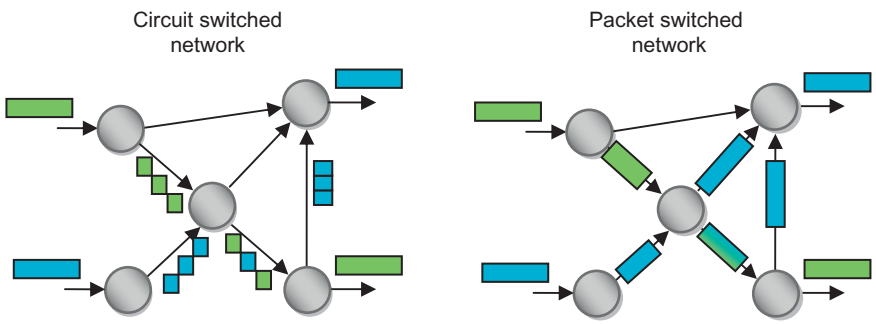


**FIGURE 2.7**

Circuit switched verse packet switched network.

ARPANET established the new concept of packet switching, in which variable sized packets are used that can take various paths through the network depending on factors such as congestion and available bandwidth, because no predefined paths are used. In fact, a given exchange of information may take multiple different paths. To do this, each packet was appended with a network control protocol (NCP) header containing information such as the destination address and the message type. Once a node received a packet, it examined the header to determine how to forward it. In the case of ARPANET, the IMP examined the header and decided if the packet was for the locally attached computer or whether it should have been passed through the network to another IMP. The NCP header was eventually replaced by the Transmission Control Protocol/Internet Protocol (TCP/IP), which will be described in more detail below.

## TCP/IP

With the ARPANET in place, engineers and scientists started to investigate new protocols for transmitting data across packet based networks. Several types of Transmission Control Protocol (TCP) and Internet Protocol (IP) standards were studied by universities and corporate research labs. By the early 1980s, a new standard called TCP/IP was firmly established as the protocol of choice, and is what the internet is based on today. Of course, what started out as a simple standard has evolved into a set of more complex standards over the years; these standards are now administered by the Internet Engineering Task Force (IETF). Additional standards have also emerged for sending special types of data over IP networks; for example, iSCSI for storage and iWARP for remote direct memory access, both of which are useful in data center networks. Figure 2.8 shows a simplified view some of the high-level functions provided by the TCP/IP protocol.

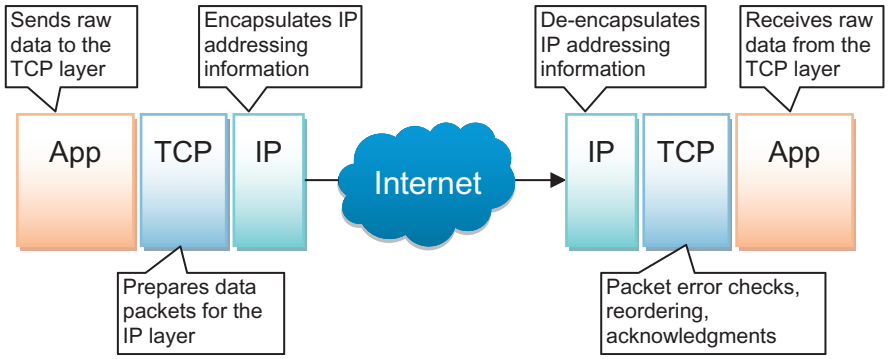


**FIGURE 2.8**

High-level TCP/IP functions.

The application hands over the data to be transmitted to the TCP layer. This is generally a pointer to a linked list memory location within the CPU subsystem.

The TCP layer then segments the data into packets (if the data is larger than the maximum packet size supported), and adds a TCP header to each packet. This header includes information such as the source and destination port that the application uses, a sequence number, an acknowledgment number, a checksum, and congestion management information. The IP layer deals with all of the addressing details and adds a source and destination IP address to the TCP packet. The Internet shown in the figure contains multiple routers that forward data based on this TCP/IP header information. These routers are interconnected using layer 2 protocols such as Ethernet that apply their own L2 headers.

On the receive side, the IP layer checks for some types of receive errors and then removes the IP address information. The TCP layer performs several transport functions including acknowledging received packets, looking for checksum errors, reordering received packets, and throttling data based on congestion management information. Finally, the raw data is presented to the specified application port number. For high-bandwidth data pipes, this TCP workload can bog down the CPU receiving the data, preventing it from providing satisfactory performance to other applications that are running. Because of this, several companies have developed TCP offload engines in order to remove the burden from the host CPU. But with today's high-performance multicore processors, special offload processors are losing favor.

As you may have gathered, TCP/IP is a deep subject and we have only provided the reader with a high-level overview in this section. Much more detail can be found online or in various books on networking technology.

## Multi-Protocol Label Switching

When a router receives a TCP/IP packet, it must look at information in the header and compare this to data stored in local routing tables in order to determine a proper forwarding port. The classic case is the 5-tuple lookup that examines the source IP address, destination IP address, source port number, destination port number, and the protocol in use. When packets move into the core of the network and link speeds increase, it becomes more difficult to do this lookup across a large number of ports while maintaining full bandwidth, adding expense to the core routers.

In the mid-1990s, a group of engineers at Ipsilon Networks had the idea to add special labels to these packets (label switching), which the core routers can use to forward packets without the need to look into the header details. This is something like the postal zip code. When a letter is traveling through large postal centers, only the zip code is used to forward the letter. Not until the letter reaches the destination post office (identified by zip code) is the address information examined. This idea was the seed for Multi-Protocol Label Switching (MPLS) which is extensively used in TCP/IP networks today. This idea is also the basis for other tunneling protocols such as Q-in-Q, IP-over-IP, FCoE, VXLAN, and NVGRE. Several of these tunneling protocols will be discussed further in later chapters in this book.

Packets enter an MPLS network through a Label Edge Router (LER) as shown in Figure 2.9. LERs are usually at the edge of the network, where lower bandwidth

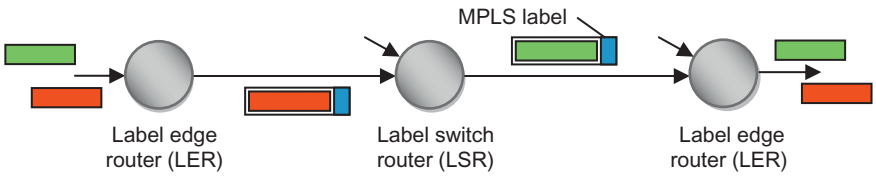


**FIGURE 2.9**

MPLS packet forwarding.

requirements make it easier to do full header lookups and then append an MPLS label in the packet header. Labels may be assigned using a 5-tuple TCP/IP header lookup, where a unique label is assigned per flow. In the core of the network, label switch routers use the MPLS label to forward packets through the network. This is a much easier lookup to perform in the high-bandwidth network core. In the egress LER, the labels are removed and the TCP/IP header information is used to forward the packet to its final destination. Packets may also work their way through a hierarchy of MPLS networks where a packet encapsulated with an MPLS header from one network may be encapsulated with another MPLS header in order to tunnel the packet through a second network.

## SONET/SDH

Early telephone systems used manually connected patch panels to route phone calls. Soon, this evolved into mechanical relays and then into electronic switching systems. Eventually, voice calls became digitized, and, with increased bandwidth within the network, it made sense to look at ways to combine multiple calls over a single line. And why not also transmit other types of data right along with the digitized voice data? To meet these needs, Synchronous Optical Network (SONET) was created as a circuit-switched network originally designed to transport both digitized DS1 and DS3 voice and data traffic over optical networks. But to make sure all data falls within its dedicated time slot, all endpoints and transmitting stations are time synchronized to a master clock, thus the name Synchronous Optical Network. Although the differences in the standards are very small, SONET, developed by Telcordia and American National Standards Institute (ANSI), is used in North America, while Synchronous Digital Hierarchy (SDH), developed by the European Telecommunications Standards Institute, is used in the rest of the world.

At the conceptual level, SONET/SDH can be depicted as shown in Figure 2.10. SONET/SDH uses the concept of transport containers to move data throughout the network. On the left of the figure, we have lower speed access layers where packets are segmented into fixed length frames. As these frames move into the higher bandwidth aggregation networks, they are grouped together into containers and these containers are grouped further into larger containers as they enter the core network. An analogy would be transporting automobiles across the country. Multiple automobiles from different locations may be loaded on a car carrier truck. Then multiple car carrier trucks may be loaded onto a railroad flatcar. The SONET/SDH frame transport

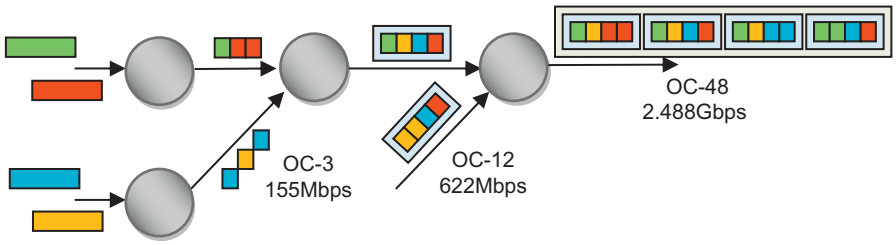


**FIGURE 2.10**

SONET/SDH transport.

time period is constant so the data rates are increased by a factor of four at each stage (OC-3, OC-12, OC-48. . .). Therefore, four times the data can be placed within each frame while maintaining the same frame clock period. Time slot interchange chips are used to shuffle frames between containers at various points in the network and are also used extensively in SONET/SDH add-drop multiplexers at the network edge.

SONET/SDH has been used extensively in telecommunication networks, whereas TCP/IP has been the choice for internet traffic. This led to the development of IP over SONET/SDH systems that allowed the transport of packet based IP traffic over SONET/SDH networks. Various SONET/SDH framer chips were developed to support this including Asynchronous Transfer Mode (ATM) over SONET, IP over SONET, and Ethernet over SONET devices. But several factors are reducing the deployment of SONET/SDH in transport networks. One factor, is that most of all traffic today is packet based (think Ethernet and IP phones). Another factor is that Carrier Ethernet is being deployed around the world to support packet based traffic. Because of these and other factors, SONET/SDH networks are being slowly replaced by carrier Ethernet networks.

## Asynchronous Transfer Mode

In the late 1980s, ATM emerged as a promising new communication protocol. In the mid-1990s, I was working with a group that was developing ATM over SONET framer chips. At the time, proponents were claiming that ATM could be used to transfer voice, video, and data throughout the LAN and WAN, and soon every PC would have an ATM network interface card. Although ATM did gain some traction in the WAN with notable equipment from companies like Stratacom (acquired by Cisco) and FORE Systems (acquired by Marconi), it never replaced Ethernet in the LAN.

The ATM frame format is shown in Figure 2.11. This frame format shows some of the strong synergy that ATM has with SONET/SDH. Both use fixed size frames along with the concept of virtual paths and virtual channels. ATM is a circuit-switched technology in which virtual end-to-end paths are established before transmission begins. Data can be transferred using multiple virtual channels within a virtual path, and multiple ATM frames will fit within a SONET/SDH frame.

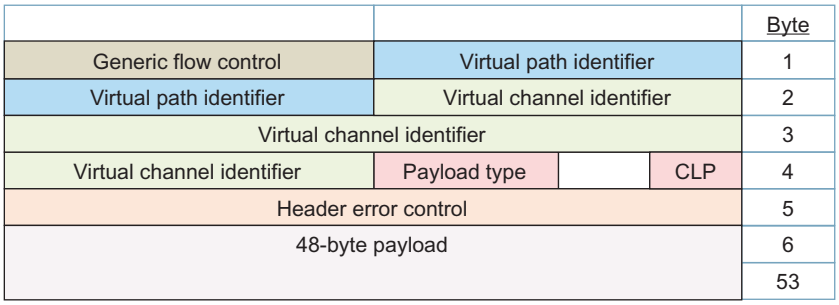


**FIGURE 2.11**

Asynchronous Transfer Mode frame format.

The advantage of using a fixed frame size is that independent streams of data can easily be intermixed providing low jitter, and fixed frames also work well within SONET/SDH frames. In packet based networks, a packet may need to wait to use a channel if a large packet is currently being transmitted, causing higher jitter. Because most IT networks use variable sized packets, as link bandwidths increase it becomes more difficult to segment and reassemble data into 53-byte frames, adding complexity and cost to the system. In addition, the ATM header overhead percentage can be larger than packet based protocols, requiring more link bandwidth for the same effective data rate. These are some of the reasons that ATM never found success in the enterprise or data center networks.

## Token Ring/Token Bus

So far in this section, we have been describing several network protocols mainly used in telecommunication and wide area networks. We will now start to dig into some network protocols used within the enterprise to interconnect terminals, PCs, mainframes, servers, and storage equipment.

One of the earliest local area networking protocols was Token Ring, originally developed by IBM in the early 1980s. Token Bus is a variant of Token Ring where a virtual ring is emulated on a shared bus. In the mid-1980s, Token Ring ran at 4Mbps, which was increased to 16Mbps in 1989. Both speeds were eventually standardized by the IEEE 802.5 working group. Other companies developing Token Ring networks included Apollo Computer and Proteon. Unfortunately, IBM network equipment was not compatible with either of these companies' products, segmenting the market.

In a Token Ring network, empty information frames are continuously circulated around the ring as shown in Figure 2.12. In this figure, when one device wants to send data to another device, it grabs an empty frame and inserts both the packet data and destination address. The frame is then examined by each successive device, and if the frame address matches a given device, it takes a copy of the data and sets the token to 0. The frame is then sent back around the ring to the sending device as an

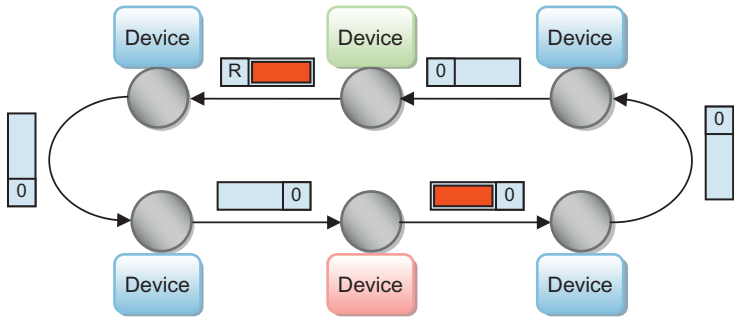


**FIGURE 2.12**

Token Ring network.

acknowledgment, which then clears the frame. Although this topology is fine for lightly loaded networks, if each node wants to continuously transmit data, it will get only 1/*N* of the link bandwidth, where *N* is the number of nodes in the ring. In addition, it can have higher latency than directly connected networks. Because of this and other factors, Token Ring was eventually replaced by Ethernet in most LAN applications.

## Ethernet

Ethernet was introduced in 1980 and standardized in 1985. Since then, it has evolved to be the most widely used transport protocol for LANs, data center networks, and carrier networks. In the following section, we will provide an overview of Ethernet technology and how it is used in these markets.

## Fibre Channel

Many data centers have separate networks for their data storage systems. Because this data can be critical to business operations, these networks have to be very resilient and secure. Network protocols such as Ethernet allow packets to be dropped under certain conditions, with the expectation that data will be retransmitted at a higher network layer such as TCP. Storage traffic cannot tolerate these retransmission delays and for security reasons, many IT managers want to keep storage on an isolated network. Because of this, special storage networking standards were developed. We will describe Fibre Channel networks in more detail in Chapter 8 which covers storage networking.

## InfiniBand

In the early 1990s, several leading network equipment suppliers thought they could come up with a better networking standard that could replace Ethernet and Fibre Channel in the data center. Originally called Next Generation I/O and Future I/O, it soon became known as InfiniBand. But like many purported world beating

technologies, it never lived up to its promise of replacing Ethernet and Fibre Channel in the data center and is now mainly used in high-performance computing (HPC) systems and some storage applications. What once was a broad ecosystem of suppliers has been reduced to Mellanox® and Intel (through an acquisition of the InfiniBand assets of QLogic®).

InfiniBand host channel adapters (HCAs) and switches are the fundamental components used in most HPC systems today. The HCAs sit on the compute blades which are interconnected through high-bandwidth, low-latency InfiniBand switches. The HCAs operate at the transport layer and use verbs as an interface between the client software and the transport functions of the HCA. The transport functions are responsible for in-order packet delivery, partitioning, channel multiplexing, transport services, and data segmentation and reassembly. The switch operates at the link layer providing forwarding, QoS, credit-based flow control and data integrity services. Due to the relative simplicity of the switch design, InfiniBand provides very high-bandwidth links and forward packets with very low latency, making it an ideal solution for HPC applications. We will provide more information on high performance computing in Chapter 10.

## ETHERNET

In the last section, we described several popular communication protocols that have been used in both enterprise and carrier networks. Because Ethernet is such an important protocol, we will dedicate a complete section in this chapter to it. In this section, we will provide a history and background of Ethernet along with a high-level overview of Ethernet technology including example use cases in carrier and data center networks.

### Ethernet history

You can make an argument that the Xerox® Palo Alto Research Center (PARC) spawned many of the ideas that are used in personal computing today. This is where Steve Jobs first saw the mouse, windows, desktop icons, and laser printers in action. Xerox PARC also developed what they called Ethernet in the early to mid-1970s.

The development of Ethernet was inspired by a wireless packet data network called ALOHAnet developed at the University of Hawaii, which used a random delay time interval to retransmit packets if an acknowledgment was not received within a given wait time. Instead of sharing the airwaves like ALOHAnet, Ethernet shared a common wire (channel). By the end of the 1970s, DEC, Intel, and Xerox started working together on the first Ethernet standard which was published in 1980. Initially, Ethernet competed with Token Ring and Token Bus to connect clients with mainframe and minicomputers. But once the IBM PC was released, hundreds of thousands of Ethernet adapter cards began flooding the market from companies such as 3Com and others. The Institute of Electrical and Electronic Engineers (IEEE) decided to standardize Ethernet into the IEEE 802.3 standard which was completed in 1985.

Initially Ethernet became the *de facto* standard for LANs within the enterprise. Over the next two decades, Ethernet port bandwidth increased by several orders of magnitude making it suitable for many other applications including carrier networks, data center networks, wireless networks, industrial automation, and automotive applications. To meet the requirements of these new markets, a wide variety of features were added to the IEEE standard, making Ethernet a deep and complex subject that can fill several books on its own. In this book, we will focus on how Ethernet is used in cloud data center networks.

## Ethernet overview

Ethernet started as a shared media protocol where all hosts communicated over a single 10Mbps wire or channel. If a host wanted to communicate on the channel, it would first listen to make sure no other communications were taking place. It would then start transmitting and also listen for any collisions with other hosts that may have started transmitting at the same time. If a collision was detected, each host would back off for a random time period before attempting another transmission. This protocol became known as Carrier Sense Multiple Access with Collision Detection (CSMA/CD). As Ethernet speeds evolved from 10Mbps to 100Mbps to 1000Mbps (GbE), a shared channel was no longer practical. Today, Ethernet does not share a channel, but instead, each endpoint has a dedicated full duplex connection to a switch that forwards the data to the correct destination endpoint.

Ethernet is a layer 2 protocol compared to TCP/IP which is a layer 3 protocol. Let's use a railroad analogy to explain this. A shipping company has a container with a bar code identifier that it needs to move from the west coast to the east coast using two separate railway companies (call them Western Rail and Eastern Rail). Western Rail picks up the container, reads the bar code, loads it on a flatcar and sends it halfway across the country through several switching yards. The flat car has its own bar code, which is used at the switching yard to reroute the flat car to the destination. Half way across the country, Eastern Rail now reads the bar code on the container, loads it onto another flatcar, and sends it the rest of the way across the country through several more switching yards.

In this analogy, the bar code on the container is like the TCP/IP header. As the frame (container) enters the first Ethernet network (Western Rail), the TCP/IP header is read and an Ethernet header (flatcar bar code) is attached which is used to forward the packet through several Ethernet switches (railroad switching yards). The packet may then be stripped of the Ethernet header within a layer 3 TCP/IP router and forwarded to a final Ethernet network (Eastern Rail), where another Ethernet header is appended based on the TCP/IP header information and the packet is sent to its final destination. The railroad is like a layer 2 network and is only responsible for moving the container across its domain. The shipping company is like the layer 3 network and is responsible for the destination address (container bar code) and for making sure the container arrives at the destination. Let's look at the Ethernet frame format in Figure 2.13.

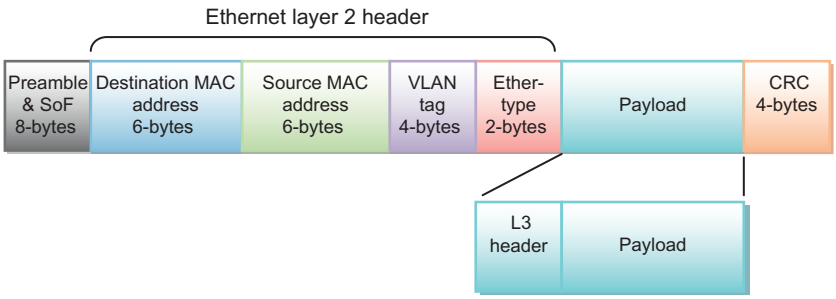


**FIGURE 2.13**

Ethernet frame format.

The following is a description of the header fields shown in the figure. An interframe gap of at least 12 bytes is used between frames. The minimum frame size including the header and cyclic redundancy check (CRC) is 64 bytes. Jumbo frames can take the maximum frame size up to around 16K bytes.

- *Preamble and start-of-frame (SoF)*: The preamble is used to get the receiving serializer/deserializer up to speed and locked onto the bit timing of the received frame. In most cases today, this can be done with just one byte leaving another six bytes available to transfer user proprietary information between switches. A SoF byte is used to signal the start of the frame.
- *Destination Media Access Control (MAC) address*: Each endpoint in the Ethernet network has an address called a MAC address. The destination MAC address is used by the Ethernet switches to determine how to forward packets through the network.
- *Source MAC address*: The source MAC address is also sent in each frame header which is used to support address learning in the switch. For example, when a new endpoint joins the network, it can inject a frame with an unknown designation MAC. Each switch will then broadcast this frame out all ports. By looking at the MAC source address, and the port number that the frame came in on, the switch can learn where to send future frames destined to this new MAC address.
- *Virtual local area network tag (optional)*: VLANs were initially developed to allow companies to create multiple virtual networks within one physical network in order to address issues such as security, network scalability, and network management. For example, the accounting department may want to have a different VLAN than the engineering department so packets will stay in their own VLAN domain within the larger physical network. The VLAN tag is 12-bits, providing up to 4096 different virtual LANs. It also contains frame priority information. We will provide more information on the VLAN tag in Chapter 5.
- *Ethertype*: This field can be used to either provide the size of the payload or the type of the payload.
- *Payload*: The payload is the data being transported from source to destination. In many cases, the payload is a layer 3 frame such as a TCP/IP frame.
- *CRC (frame check sequence)*: Each frame can be checked for corrupted data using a CRC.

## Carrier Ethernet

With Ethernet emerging as the dominant networking technology within the enterprise, and telecom service providers being driven to provide more features and bandwidth without increasing costs to the end users, Ethernet has made significant inroads into carrier networks. This started with the metro networks that connect enterprise networks within a metropolitan area.

The Metro Ethernet Forum (MEF) was founded in 2001 to clarify and standardize several Carrier Ethernet services with the idea of extending enterprise LANs across the wide area network (WAN). These services include:

- *E-line*: This is a direct connection between two enterprise locations across the WAN.
- *E-LAN*: This can be used to extend a customer's enterprise LAN to multiple physical locations across the WAN.
- *E-tree*: This can connect multiple leaf locations to a single root location while preventing interleaf communication.

This movement of Ethernet out of the LAN has progressed further into the carrier space using several connection oriented transport technologies including Ethernet over SONET/SDH and Ethernet over MPLS. This allows a transition of Ethernet communication, first over legacy transport technologies, and, ultimately, to Ethernet over Carrier Ethernet Transport, which includes some of the following technologies.

Carrier Ethernet networks consist of Provider Bridge (PB) networks and a Provider Backbone Bridge (PBB) network as shown in Figure 2.14. Provider bridging utilizes an additional VLAN tag (Q-in-Q) to tunnel packets between customers using several types of interfaces. Customer Edge Ports (CEP) connect to customer equipment while Customer Network Ports (CNP) connect to customer networks. Provider

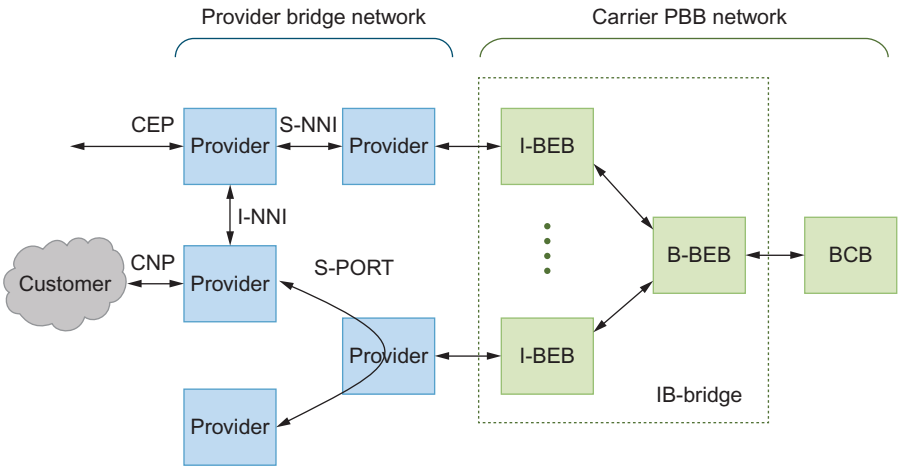


**FIGURE 2.14**

Carrier Ethernet block diagram.

equipment can be interconnected directly using an I-NNI interface, or tunneled through another provider network using an S-PORT CNP interface. Two service providers can be interconnected through an S-NNI interface. A fundamental limitation of Provider Bridging is that only 4096 special VLAN tags are available, limiting the scalability of the solution.

In the carrier PBB network, an additional 48-bit MAC address header is used (MAC-in-MAC) to tunnel packets between service providers, supporting a much larger address space. The I-component Backbone Edge Bridge (I-BEB) adds a service identifier tag and new MAC addresses based on information in the PB header. The B-component Backbone Edge Bridge (B-BEB) verifies the service ID and forwards the packet into the network core using a backbone VLAN tag. The Backbone Core Bridge (BCB) forwards packets through the network core.

As carrier networks migrate from circuit switching to packet switching technologies, they must provide Operation Administration and Maintenance (OAM) features that are required for robust operation and high availability. In addition, timing synchronization must be maintained across these networks. As Carrier Ethernet technology replaces legacy SONET/SDH networks, several new standards have been developed such as Ethernet OAM (EOAM) and Precision Time Protocol (PTP) for network time synchronization.

While Carrier Ethernet standards such as PB, PBB, and EOAM have been in development by the IEEE for some time, other groups have been developing a carrier class version of MPLS called MPLS-TE for Traffic Engineering or T-MPLS for Transport MPLS. The idea is that MPLS has many of the features needed for carrier class service already in place, so why develop a new Carrier Ethernet technology from scratch? The tradeoff is that Carrier Ethernet should use lower cost switches versus MPLS routers, but MPLS has been around much longer and should provide an easier adoption within carrier networks. In the end, it looks like Carrier networks will take a hybrid approach, using the best features of each depending on the application.

Data centers are connected to the outside world and to other data centers through technology such as Carrier Ethernet or MPLS-TE. But within the data center specialized data center networks are used. The rest of this book will focus on Ethernet technology used within the cloud data center networks.

## ENTERPRISE VERSUS CLOUD DATA CENTERS

Originally, servers were connected to clients and to each other using the enterprise LAN. As businesses started to deploy larger data centers, they used similar enterprise LAN technology to create a data center network. Eventually, the changing needs of the data center required network system OEMs to start developing purpose-built data center networking equipment. This section will describe the major differences between enterprise networks and cloud data center networks.

## Enterprise data center networks

If you examine the typical enterprise LAN, you will find wired Ethernet connections to workgroup switches using fast Ethernet (or 1Gb Ethernet) and wireless access points connected to the same workgroup switches. These switches are typically in a 1U pizza-box form factor and are connected to other workgroup switches either through 10Gb Ethernet stacking ports or through separate 10GbE aggregation switches. The various workgroup switches and aggregation switches typically sit in a local wiring closet. To connect multiple wiring closets together, network administrators may use high-bandwidth routers, which also have external connections to the WAN.

When enterprise system administrators started to develop their own high-density data centers, they had no choice but to use the same networking gear as used in the LAN. Figure 2.15 shows an example of how such an enterprise data center may be configured. In this figure, workgroup switches are repurposed as top of rack (ToR) switches with 1GbE links connecting to the rack servers and multiple 1GbE or 10GbE links connecting to the aggregation switches. The aggregation switches then feed a core router similar to the one used in the enterprise LAN through 10Gb Ethernet links.

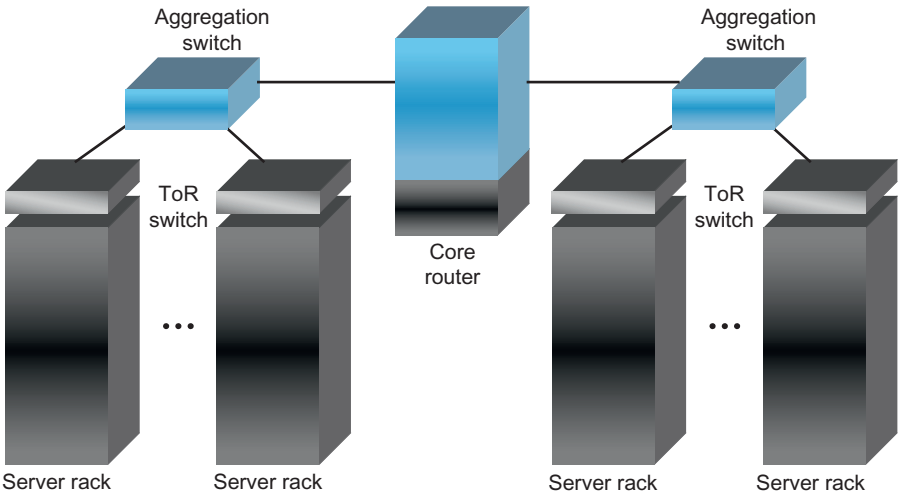


**FIGURE 2.15**

Enterprise data center network.

There are several issues with this configuration. First, packets need to take multiple hops when traveling between servers. This increases latency and latency variation between servers, especially when using enterprise networking gear that has relatively high latency, as latency is not a concern in the LAN. Second, enterprise networks will drop packets during periods of high congestion. Data center

storage traffic needs lossless operation, so, in this case, a separate network such as Fibre Channel will be needed. Finally, core routers are very complex and expensive given that they need to process layer 3 frames at high-bandwidth levels. In addition, enterprise equipment typically comes with proprietary and complex software that is not compatible with other software used in the data center.

## Cloud data center networks

Because of the issues listed above, and the cost of using more expensive enterprise hardware and software in large cloud data centers, network equipment suppliers have developed special networking gear targeted specifically for these data center applications. In some cases, the service providers operating these large cloud data centers have specified custom built networking gear from major ODMs and have written their own networking software to reduce cost even further.

Most data center networks have been designed for north-south traffic. This is mainly due to that fact that most data center traffic up until recently has been from clients on the web directly communicating with servers in the data center. In addition, enterprise switches that have been repurposed for the data center typically consist of north-south silos built around departmental boundaries. Now we are seeing much more data center traffic flowing in the east-west direction due to server virtualization and changing server workloads. Besides complexity, the problem with enterprise style networks is latency and latency variation. Not only is the latency very high for east-west traffic, it can change dramatically, depending on the path through the network. Because of this, data center network designers are moving toward a flat network topology as shown in Figure 2.16.

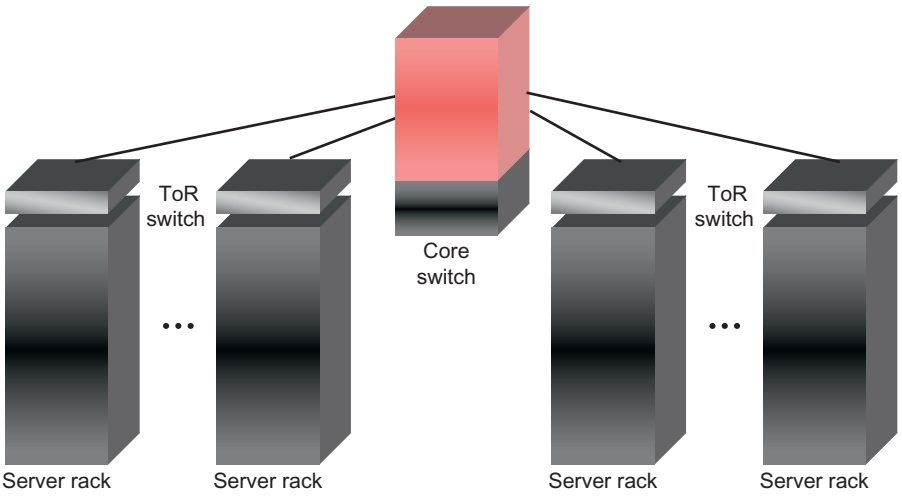


**FIGURE 2.16**

Cloud data center network.

By providing 10GbE links to the rack servers, the network can support the convergence of storage and data traffic into one network, reducing costs. As shown in the figure, ToR switches are used with high-bandwidth links to the core and the core routers have been replaced with simpler core switches with a larger number of ports allowing them to absorb the aggregation function, making this a "flatter" network. This type of network can better support all of the east-west traffic that is seen in large data centers today with lower latency and lower latency variation. In addition, by moving the tunneling and forwarding intelligence into the ToR switch, a simpler core switch can be developed using high-bandwidth tag forwarding much like an MPLS label switch router. More information on cloud data center network topologies will be presented in Chapter 4.

## MOVEMENT TO THE CLOUD

Enterprise data centers have continued to add more equipment and services in order to keep pace with their growing needs. Offices once dominated by paperwork are now doing almost everything using web-based tools. Design and manufacturing companies rely heavily on arrays of computing resources in order to speed their time to market. But now, many corporations are seeing the value of outsourcing their computing needs to cloud service providers. This section will describe some of the driving forces behind this transition, along with security concerns. We will also describe several types of cloud data centers and the cloud services they provide.

### Driving forces

Designing, building, and maintaining a large corporate data center is a costly affair. Expensive floor space, special cooling equipment, and high power demands are some of the challenges that data center administrators must face. Even with the advent of virtualized servers, low server utilization is a common problem as the system administrator must design for periods of peak demand. As an illustrative example, consider a small company doing large chip designs. Early in the design process, computing demands can be low. But as the chip design is being finalized, chip layout, simulation, and design verification tools create peak workloads that the data center must be designed to accommodate. Because of this, if a company is only developing one chip per year, the data center becomes underutilized most of the time.

Over the last 10 years or so, large data centers have become very common across the world. Some of this has been driven by the need to support consumers such as in the case of companies like Amazon, Google, and Facebook. And some of this has been driven by the need to support services such as web hosting. Building a large data center is not an easy task due to power cooling and networking requirements. Several internet service providers have become experts in this area and now deploy very efficient hyper-scale data centers across the world.

Starting in 2006, Amazon had the idea to offer web services to outside developers, who could take advantage of their large efficient data centers. This idea has taken root and several cloud service providers now offer corporations the ability to outsource some of their data center needs. By providing agile software and services, the cloud service provider can deliver on-demand virtual data centers to their customers. Using the example above, as the chip design is being finalized, external data center services could be leased during peak demand, reducing the company's internal data center equipment costs. But the largest obstacle keeping companies from moving more of their data center needs over to cloud service providers are concerns about security.

### *Security concerns*

In most surveys of IT professionals, security is listed as the main reason as to why they are not moving all of their data center assets into the cloud. There are a variety of security concerns listed below.

- Data access, modification, or destruction by unauthorized personnel.
- Accidental transfer of data between customers.
- Improper security methods limiting access to authorized personnel.
- Accidental loss of data.
- Physical security of the data center facility.

Data access can be controlled through secure gateways such as firewalls and security appliances, but data center tenants also want to make sure that other companies cannot gain accidental access to their data. Customers can be isolated logically using network virtualization or physically with dedicated servers, storage, and networking gear. Today, configuring security appliances and setting up virtual networks are labor intensive tasks that take time. Software defined networking promises to automate many of these tasks at a higher orchestration level, eliminating any errors that could cause improper access or data loss. We will provide more information on software defined networking in Chapter 9. Physical security means protecting the data center facility from disruption in power, network connections, or equipment operation by fire, natural disaster, or acts of terrorism. Most data centers today are built with this type of physical security in mind.

## Cloud types

Large cloud data centers can be dedicated to a given corporation or institution (private cloud) or can be shared among many different corporations or institutions (public cloud). In some cases, a hybrid cloud approach is used. This section will describe these cloud data center types in more detail and also list some of the reasons that a corporation may choose one over the other.

### *Private cloud*

Large corporations may choose to build a private cloud, which can be administered either internally or through an outside service, and may be hosted internally or at an external location. What sets a private cloud apart from a corporate data center is the efficiency of operation. Unlike data centers that may be dedicated to certain groups within a corporation, a private cloud can be shared among all the groups within the corporation. Servers that may have stayed idle overnight in the United States can now be utilized at other corporate locations around the world. By having all the corporate IT needs sharing a physical infrastructure, economies of scale can provide lower capital expense and operating expense. With the use of virtualized services and software defined networking, agile service redeployments are possible, greatly improving resource utilization and efficiencies.

### *Public cloud*

Smaller corporations that don't have the critical mass to justify a private cloud can choose to move to a public cloud. The public cloud has the same economies of scale and agility as the private cloud, but is hosted by an external company and data center resources are shared among multiple corporations. In addition, corporations can pay as they go, adding or removing compute resources on demand as their needs change.

The public cloud service providers need to develop data centers that meet the requirements of these corporate tenants. In some cases, they can provide physically isolated resources, effectively hosting a private cloud within the public cloud. In the public cloud domain, virtualization of compute and networking resources allows customers to lease only the services they need and expand or reduce services on the fly. In order to provide this type of agility while at the same time reducing operating expense, cloud service providers are turning to software defined networking as a means to orchestrate data center networking resources and quickly adjust to changing customer requirements. We will provide more details on software defined networking in Chapter 9 of this book.

### *Hybrid cloud*

In some cases, corporations are unwilling to move their entire data center into the public cloud due to the potential security concerns described above. But in many cases, corporations can keep sensitive data in their local data center and exploit the public cloud without the need to invest in a large data center infrastructure and have the ability to quickly add or reduce resources as the business needs dictate. This approach is sometimes called a hybrid cloud.

## Public cloud services

The public cloud service providers can host a wide variety of services from leasing hardware to providing complete software applications, and there are now several providers who specialize in these different types of services shown in Figure 2.17.

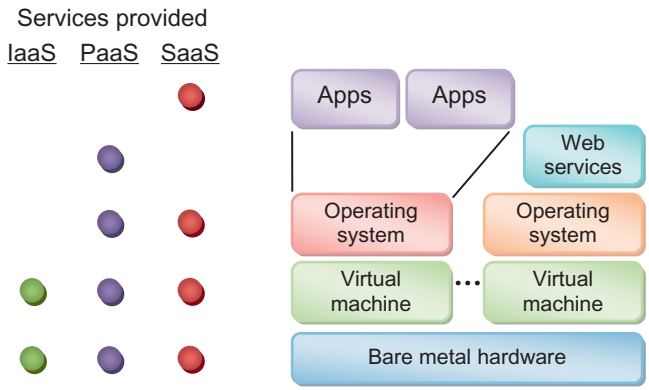


**FIGURE 2.17**

Services available from cloud service providers.

Infrastructure as a Service (IaaS) includes hardware resources such as servers, storage, and networking along with low-level software features such as hypervisors for virtual machines and load balancers. Platform as a Service (PaaS) includes higher layer functions such as operating systems and/or web server applications including databases and development tools. Software as a Service (SaaS) provides web-based software tools to both individuals and corporations. Figure 2.17 shows typical data center functional components along with the types of services provided by IaaS, PaaS, and SaaS. Some applications offered by large cloud service providers that we use every day are very similar to SaaS, but are not classified that way. For example Google Search, Facebook, and the App Store are applications that are run in large data centers, but are not necessarily considered SaaS.

### *Infrastructure as a Service*

With IaaS, the service provider typically leases out the raw data center building blocks including servers, storage, and networking. This allows the client to build their own virtual data center within the service provider's facility. An example of this is hosting a public cloud. The service provider may provide low-level software functions such as virtual machine hypervisors, network virtualization services, and load balancing, but the client will install their own operating systems and applications. The service provider will maintain the hardware and virtual machines, while the client will maintain and update all the software layers above the virtual machines. Some example IaaS providers include Google Compute Engine, Rackspace®, and Amazon Elastic Compute Cloud.

### *Platform as a Service*

This model provides the client with a computing platform including operating system and access to some software tools. An example of this is web hosting services, in which the service provider not only provides the operating system on which a

web site will be hosted, but also access to database applications, web development tools, and tools to gather web statistics. The service provider will maintain the operating system and tools, while the client will maintain their own database and web pages. The service provider can provide a range of services and a variety of hosting options with added hardware performance depending on expected web traffic volume. Some example PaaS providers include Windows Azure Cloud Services, Google App Engine, and a variety of web hosting companies.

### *Software as a Service*

Cloud service providers can also deliver software applications and databases to the end user through the SaaS business model. With this model, the end user pays a subscription fee or on a per-use basis for on-demand software services. The service provider maintains the data infrastructure, operating systems, and software, while the end user simply runs the applications remotely. This has the potential to reduce corporate IT operating costs by outsourcing the maintenance of hardware and software to the SaaS provider. Some example SaaS providers are Microsoft Office 360 and Google Apps. It's interesting how we have evolved from running time-sharing software on large mainframes to running applications on large cloud data centers. In both cases, clients use remote "terminals" to access centralized computer resources.

## REVIEW

In this chapter, we provided a brief history of how corporate data center resources have evolved from mainframes to the cloud. We also provided some background on key networking technologies that have been used over the years with a focus on Ethernet. We described the differences between enterprise data centers and cloud data centers along with the reasons that many corporations are moving their data centers into the cloud. Finally, we described several cloud services that are some of the driving forces behind the movement to the cloud. In the next chapter, we will introduce switch fabric technologies which are the fundamental building blocks of cloud data center networking equipment.