What Exactly *Is* the Capacity of Ethernet?

The subject of Ethernet throughput is still a matter of great controversy, despite the fact that Ethernet is 25 years old. LAN traffic tends to come in bursts, and this tends to influence the perceived capacity substantially, so different people see different sides of the same technology. Let's look at the capacity of shared, Switched, and full-duplex Switched Ethernet. Note that the only difference between a shared and a switched (half-duplex) connection is the number of users on the network: A switched connection is limited to only two nodes!

Shared Ethernet Capacity

Ethernet's throughput capability depends on several different variables, so this question has no simple answer. The most important variable is the user, and the user's particular traffic pattern. Shared Ethernet is built on the premise that network traffic occurs in bursts. Different users will require network access at very different times and will transmit for different lengths. In this way, the capacity can be used to its fullest for the brief point in time that a particular user accesses the network. We'll find the random nature of bursts in typical network access very difficult to include, so we will assume a somewhat random, predictable traffic pattern from all users.

The maximum throughput then depends on essentially two variables:

- *Number of users*—The more users you have on a LAN, the more contention the channel will have and the more collisions occur. The maximum throughput is reached with just two nodes on a LAN, also known as a *switched connection*. This assumes that both nodes are capable of generating the maximum frame rate. The lowest throughput will be reached with the maximum number of users: 1024 for Ethernet.
- *Frame size*—As discussed previously, Ethernet throughput declines for smaller frames. This is intuitive because larger frames are more efficient. Larger frames provide the added benefit of providing less of an opportunity for channel contention and therefore collisions.

Figure 7.4 illustrates the relationship between the number of nodes, frame size, and throughput.

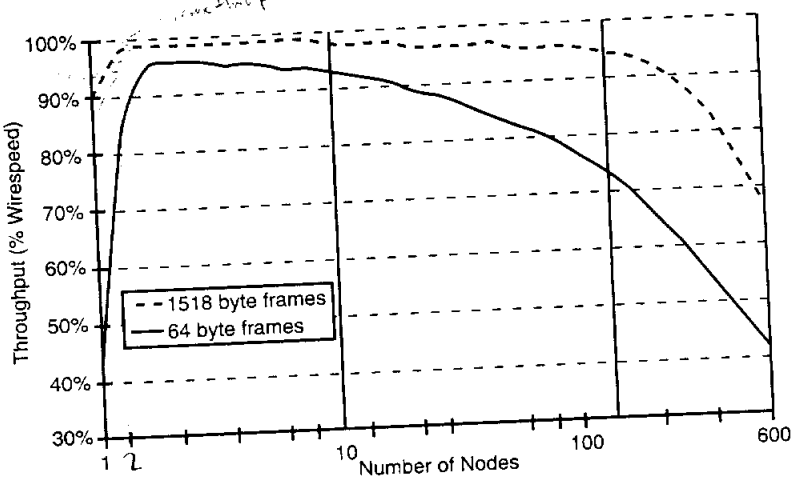

FIGURE 7.4 *Ethernet throughput as a function of node numbers plotted for minimum and maximum frame size.*

Bob Metcalfe, the original inventor of Ethernet, and David Boggs calculated the capacity of Ethernet as a function of frame length. They calculated the worst-case capacity of Ethernet to be 37% for minimum frames and maximum number of nodes: 64 bytes per 1024 nodes. With a maximum frame length of 1518 bytes, Metcalfe and Boggs calculated that this would increase to 93%. Practical tests have confirmed these calculations. Unfortunately, some people have taken the 64 bytes per 1024 nodes number, *the absolute worst-case capacity*, to be the *maximum* throughput for Ethernet in all situations. As discussed, most Ethernet traffic is

bimodal, so an average frame size of around 1000 bytes is probably more realistic than 64 bytes! Most networks also don't have 1024 nodes on them, so in reality Ethernet has a capacity that far exceeds the 37%. Some real-world networks have seen utilizations of over 90%, which may sound very surprising.

The fact that Ethernet utilization can reach over 90% may surprise you. Most people will not run an Ethernet network at this kind of utilization because users will encounter unacceptable delays. Some users may find fixed delays, also known as *latency* or *response time*, acceptable, but variable delays, also known as *jitter*, are more problematic. For example, if one file transfer takes fractions of a second whereas a later, similar transfer takes several seconds, users will complain ("Why is the LAN so slow this afternoon?"). Protocol stacks also expect acknowledgment frames within a certain time window. If the acknowledgment frame does not arrive within that time, the protocol will assume the frame was discarded and will attempt a retransmission. Certain newer applications, such as audio or video, are unsuitable for transmission over a network with variable delays.

The key to determining the maximum utilization for an Ethernet network is then to establish the level of delay and jitter you are willing to tolerate. Figure 7.5 illustrates this point.

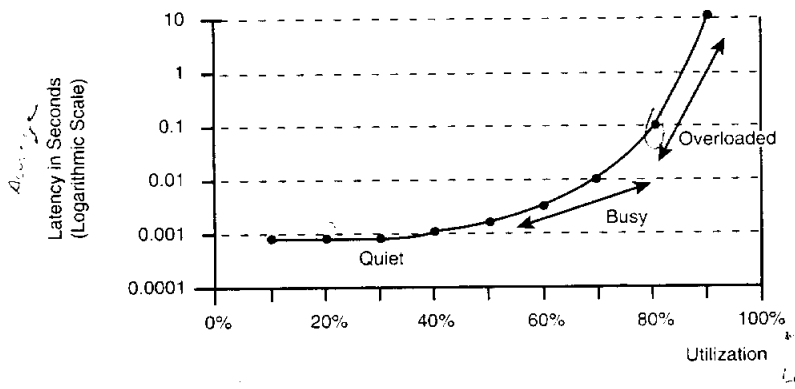

FIGURE 7.5 *A 10Mbps Ethernet response time measured as a function of utilization. Notice the three distinct operating regions. We recommend that you stay below 30% average utilization.*

Ethernet has essentially three distinct operating regions:

- *Light load (0%–50% utilization)*—When the network is running at less than 50%, the network is relatively quiet, with few collisions (less than 10%). The network is very responsive, and latencies are negligibly small, of the

order of 0.001 to 0.01 second. In this mode, Ethernet provides the capability to carry audio and video traffic because jitter is not noticeable. We recommend that your average utilization stay well below this 50% line. If you find that your average utilization is approaching 30%, you should think about upgrading because an average of 30% probably corresponds to peaks of 50% or higher. An average of 30% also leaves you some headroom and therefore time to plan your upgrade properly.

- *Moderate to heavy load (50%–80% utilization)*—In this level of utilization, the network is starting to show measurable delays, ranging from 0.01 to 0.1 second. This is still acceptable for regular file transfers or accessing an application database, but jitter is becoming an issue. It is acceptable for short bursts of traffic to push the network into this level of utilization, but you shouldn't operate a network in this mode permanently.
- *Saturation (80% and higher)*—Here, the network is showing large delays, sometimes exceeding a second. The network is very busy, and the situation worsens by what is known as the *capture effect*: Some nodes are transmitting long streams of frames, whereas others are waiting seconds. This is the no-go zone for Ethernet: You'd better do something fast. When Ethernet is saturated, it is also known as *congested* or *overloaded*.

Switched Ethernet Capacity

A half-duplex Switched Ethernet connection is still a shared-media network. This means that all of the previous discussion still applies. The benefit of a switched connection is that only two nodes compete for the channel, and therefore collisions can still occur, although they are much less frequent. Accordingly, a switched connection can reliably and permanently operate at utilization levels of over 80%. Of course, the middle-of-the-road situation occurs in which a LAN consists of just a handful of users. In this case, the average utilization should be less than 90% but can still be higher than the 50% limit recommended for large LANs. From Figure 7.4, we can determine the maximum throughput by reading off the value for two nodes. For 1000 byte frames, the throughput can reach 90% of wire speed!

Full-Duplex Switched Ethernet Capacity

Full-duplex (FDX) Ethernet provides performance capabilities that exceed those of Switched Ethernet. FDX Ethernet requires a switched two-node connection. FDX Ethernet turns off the CSMA/CD MAC instead of continuously sending Ethernet frames down a transmit and receive channel. The channel has no contention, and consequently collisions cannot take place. Under ideal circumstances, full-duplex can sustain utilization rates of close to 100% on each

channel. We recommend a maximum utilization rate of 95%. Adding up transmit and receive utilization then gives FDX Ethernet a limit of 190%.

In practice, the bimodal traffic flow discussed previously means that clients and servers are unlikely to exceed utilization rates of more than 130%. Full-duplex switch-switch connections, however, can reach the theoretical maximum utilization.

Improving the Ethernet MAC

Over the years, several attempts have been made to improve the Ethernet CSMA/CD MAC or increase frame size to improve the utilization level. Some of the proposed MAC changes were minor, and some proposals went as far as changing the access method altogether and doing away with collisions completely. The Gigabit Ethernet MAC also includes some very subtle changes that make the MAC just slightly different from the 10 and 100Mbps Ethernet variants. With the approval of 802.3z standard the IEEE for the first time officially modified the 25-year-old CSMA/CD Ethernet MAC algorithm. Let's look at the Gigabit Ethernet MAC in a bit more detail, as well as three other Ethernet improvements that haven't been successful.

100VG-AnyLAN

From 1991 to 1994, Hewlett-Packard developed the new 100VG-AnyLAN 100Mbps shared-media technology. The 100VG incorporated a new MAC called *demand priority*, which allowed time-critical applications to transmit ahead of other noncritical frames. 100VG uses a shared-media token-passing bus architecture without collisions to achieve very high throughput rates in heavily loaded shared-media networks. For backward compatibility, 100VG could use either Ethernet or Token Ring frame formats. Hewlett-Packard positioned 100VG as technically superior to 100BASE-T, whose supporters chose to maintain the existing CSMA/CD MAC. In the end, the IEEE decided that the CSMA/CD MAC would be maintained for the 100Mbps version of Ethernet. The IEEE did standardize 100VG as a new technology under the 802.12 umbrella. We also discuss 100VG in more detail in Chapters 1 and 2.

BLAM/IEEE 802.3w

Ethernet uses the Binary Exponential Backoff (BEB) algorithm to deal with overloading on the network. When a station senses a collision, it waits for a certain backoff time before it retries. If a collision occurs again, the backoff time is doubled from the previous attempt. If a condition of overloading occurs, the BEB ensures that different nodes wait longer and longer before being able to transmit their data. This causes the short-term load to adjust to a level that the network can support, providing for an overload control mechanism.

An individual node will attempt a retransmission up to 16 times. If transmission is still unsuccessful at that point, the backoff timer expires and the transmission is aborted. The frame is lost, and an error occurs. (The protocol stack will always ensure retransmission.)

The BEB mechanism does not treat all nodes equally. Assume that a node has been attempting to transmit for some time but has been unsuccessful. Its BEB counter has been escalated to ten, which is a significant lapse of time. If a new node now tries to attempt a transmission, its counter is still set to zero. This means the newer node may complete a transmission before the node that has been waiting. If a network is very busy, a node that has been waiting for some time may then suddenly gain access to the network and be allowed to send a series of frames in a row. This is known as the *channel capture effect* because it allows a node to capture the channel for quite some time, which is an undesirable condition especially in busy networks.

In 1995, the IEEE started the 802.3w working group to investigate an improved CSMA/CD backoff algorithm. The binary logarithmic arbitration (BLAM) was supposed to improve the existing CSMA/CD BEB algorithm. This would improve the maximum utilization levels that shared Ethernet can operate (discussed in more detail later). The BEB algorithm would also fix the Ethernet channel capture effect, which can allocate bandwidth in a preferential and unfair manner to certain nodes.

Unfortunately, BLAM was too little, too late. While the IEEE 802.3w engineers were investigating BLAM, other IEEE groups were already working on full-duplex and Gigabit Ethernet. Full-duplex Ethernet, in effect, turns off the Ethernet MAC, so BLAM is not applicable. The 802.3w efforts were abandoned in 1997.

802.3z/Gigabit Ethernet — can be CSMA/CD

With Gigabit Ethernet, the familiar CSMA/CD MAC has been modified for the first time. The IEEE 802.3z specification defines both half- and full-duplex operation. Half-duplex Gigabit Ethernet specifies a minimum slot time of 512 bytes, as opposed to 64 bytes for 10 and 100Mbps Ethernet. When frames of less than 512 bytes are transmitted, the CSMA/CD MAC adds carrier extension (CE) bits to meet the minimum slot time requirement. This makes Gigabit Ethernet very inefficient for small frames because large amounts of useless carrier extension bits are transmitted.

For example, let's assume that we wanted to transmit only 64 byte minimum size frames. The carrier extension would add 438 bytes of carrier extension to meet the spec of a 512 byte slot time. To calculate the overall efficiency, we still need to add the IFG overhead of 12 bytes: 64/(512+12) = 12% efficiency, or 122 Mbps. This is only marginally better than 100BASE-T!

Small frames are quite common, so this inefficiency for small frame sizes needed to be addressed. The IEEE 802.3z Gigabit MAC also includes a feature called *burst mode*. In this case, a station may continuously transmit multiple smaller frames, up to a maximum of 8,192 bytes of data. The interframe intervals will also be filled with carrier extension so that the wire never appears free to any other stations during the burst cycle. Note that all these modifications apply to half-duplex shared operation only. So far, no half-duplex 802.3z equipment is available. Please refer to Chapter 3 for further details on the modified Gigabit Ethernet MAC.

Jumbo Frames

Another proposal is called *jumbo frames*, as in very large frames. Jumbo frames are not a CSMA/CD modification; in fact, they only work in a full-duplex environment. Jumbo frames are, however, one of those attempts to touch the "Holy Grail" of Ethernet, namely frame size and the CSMA/CD MAC.

Jumbo frames would increase the maximum frame size from the current 1,518 bytes to 9,000 bytes. As the overhead is fixed per frame, increasing the frame size will make Ethernet more efficient. For 1,518 byte frames, the overhead and IFG account for about 2.5% efficiency loss. Increasing the frame size to 9,000 bytes would increase the efficiency to over 99%. The second, and more important, reason some people are advocating jumbo frames is to improve server performance. Every individual frame transmitted requires the server CPU (or desktop for that matter) to perform some data processing. Once the frame transmission is underway, the server CPU can leave the task unattended. Increasing the frame size reduces the number of CPU interrupts and therefore improves the computers' utilization and throughput. Preliminary tests show that today's servers are not capable of completely filling a Gigabit Ethernet pipe. With 100% CPU utilization, servers at this point are only capable of delivering a maximum data rate of around 400Mbps. Increasing the frame size by a factor of 5 reduces the number of CPU interrupts considerably, and therefore allows servers to get closer to the 1Gbps mark. (1Gbps still can't be accomplished today due to other performance limitations, primarily the PCI bus.)

Ethernet Capacity Recap

We have no single rule of thumb as to what utilization a shared Ethernet network can accommodate because utilization is a function of numerous variables. In real networks with bimodal frame data flow, an average utilization of 50% and a peak utilization of 80% are good guidelines to determine the limits of your Ethernet LAN.

Delay or response time increases exponentially with the number of nodes. Consequently, we recommend that you stay below 200 users for a shared LAN segment. Collisions are a normal part of the control mechanism of Ethernet. Collisions of up to 20% are quite acceptable.

Switched Ethernet connections have two major advantages in that the available bandwidth does not have to be time-shared with numerous other stations. The connection can also operate at up to 90% utilization, if required. Full-duplex Ethernet provides even more performance by allowing for simultaneous transmission and reception. Under ideal circumstances, 95% utilization on each cable pair can be achieved, which increases the theoretical utilization limit to 190%. Table 7.2 summarizes the real-world throughput capabilities of 10Mbps, 100Mbps shared, Switched, and Switched FDX Ethernet.

TABLE 7.2 A RULE-OF-THUMB GUIDELINE FOR ETHERNET UTILIZATION LIMITS

Connection Type	Wire Speed	Average Utilization Limit
Shared Ethernet	10Mbps	30%
Shared Ethernet used for multimedia traffic	10Mbps	20%
Switched Ethernet	10Mbps	85%
Switched FDX Ethernet	10Mbps	190%
Shared Fast Ethernet	100Mbps	30%
Shared Fast Ethernet used for multimedia traffic	100Mbps	20%
Switched Fast Ethernet	100Mbps	85%
Switched FDX Fast Ethernet	100Mbps	190%
Shared FDX Gigabit Ethernet[1]	1000Mbps	60%
Switched FDX Gigabit Ethernet	1000Mbps	190%

[1] *At the time of writing this book, all Gigabit Ethernet products shipping were full-duplex. No products or even plans exist for half-duplex repeaters. A few vendors are, however, shipping a new kind of device called a full-duplex repeater or buffered distributor. This device doesn't share the bandwidth but essentially acts like a switch that operates permanently in broadcast mode. The utilization limits for FDX repeaters lie somewhere in between those of classic repeaters and switches and depend to a large degree on the internal buffering capabilities of the particular FDX repeater. To play it safe, we recommend the same utilization rates as for regular half-duplex repeaters, times a factor of two for FDX operation.*

Tools to Measure Loading

The nice thing about a shared-media network is that all the traffic is transmitted to all points on the network simultaneously. That means you can "tune in" or listen for traffic at any point on the LAN and find out what the utilization is like. For switched or segmented networks, the procedure becomes much more cumbersome. You may have to measure each segment or connection physically in order to get an accurate reading of the traffic on that segment. (Remote monitoring probes now allow you to measure utilization in different segments from one physical point. We discuss these in more detail in Chapter 11, "Managing Switched and Fast Ethernet Networks.")

Raw Data Throughput Limit	Peak Utilization Limit	Peak Data Throughput Limit
3.0Mbps	80%	8Mbps
2.0Mbps	50%	5Mbps
8.5Mbps	90%	9Mbps
19Mbps	190%	19Mbps
30Mbps	80%	80Mbps
20Mbps	50%	50Mbps
85Mbps	90%	90Mbps
190Mbps	190%	190Mbps
600Mbps	120%	1600Mbps
1900Mbps	190%	1900Mbps

Many kinds of tools are available for analyzing the loading of Ethernet networks. We mention a variety of products here that allow you to measure utilization levels in the hope that you will have at least one of them around to do the job. The most common tools available for network analysis are software-based traffic analyzers. Protocol analyzers will also do the job because they all include some traffic monitoring tools. Hardware-assisted protocol analyzers, the ultimate in networking tools, also allow you to measure traffic levels, although they are a complete overkill for the task of measuring network utilization. Last, a new breed of hand-held diagnostic tools may be the best choice for measuring utilization.