Performance Evaluation

8.1 Introduction

Historically, performance evaluation was initially concerned with computer systems. During the 1970's and 1980's, computer system performance evaluation emerged as an essential component of Computer Engineering due to rapid and concurrent advancements in computer hardware and computer operating systems. The resultant increased complexity of modern computer systems made understanding and evaluating computer systems more difficult.

Performance evaluation is the application of the scientific method to the study of computer systems. Viewed as distinct from computer system design, the goal of performance evaluation is to determine the effectiveness and fairness of a computer system that is assumed to work correctly. Performance evaluation techniques have been developed to accurately measure the effectiveness with which computer system resources are managed while striving to provide service that is fair to all customer classes.

Up to this point, we have focused primarily on the functional aspects of networks. Like any computer system, however, computer networks are also expected to perform well, since the effectiveness of computations distributed over the network often depends directly on the efficiency with which the network delivers the computation's data. While the old programming adage "First get it right and then make it fast" is valid in many settings, in networking it is usually necessary to "design for performance." It is therefore important to understand the various factors that impact network performance.

8.2 Computer Network Performance Metrics

A Metric is basically a descriptor used to represent some aspect of a computer network's performance. The goal is obtain objective performance indices. For computer networks, metrics discussed in the following sections can capture performance at multiple layers of the protocol stack, e.g., UDP throughput, IP

packet round trip time, MAC layer channel utilization. Performance metrics can be positive and negative. e.g., good throughput, packet loss rate, MAC layer retries.

8.2.1 Bandwidth and Throughput

Bandwidth and throughput are two of the most confusing terms used in networking. While we could try to give you a precise definition of each term, it is important that you know how other people might use them and for you to be aware that they are often used interchangeably.

First of all, bandwidth is literally a measure of the width of a frequency band. For example, a voice-grade telephone line supports a frequency band ranging from 300 to 3300 Hz; it is said to have a bandwidth of 3300 Hz − 300 Hz = 3000 Hz. If you see the word "bandwidth" used in a situation in which it is being measured in hertz, then it probably refers to the range of signals that can be accommodated.

When we talk about the bandwidth of a communication link, we normally refer to the number of bits per second that can be transmitted on the link. We might say that the bandwidth of an Ethernet is 10 Mbps. For example, a network might have a bandwidth of 10 million bits/second (Mbps), meaning that it is able to deliver 10 million bits every second. It is sometimes useful to think of bandwidth in terms of how long it takes to transmit each bit of data. On a 10-Mbps network, for example, it takes 0.1 microseconds (μs) to transmit each bit.

A useful distinction might be made, however, between the bandwidth that is available on the link and the number of bits per second that we can actually transmit over the link in practice. We tend to use the word "throughput" to refer to the measured performance of a system. Thus, because of various inefficiencies of implementation, a pair of nodes connected by a link with a bandwidth of 10 Mbps might achieve a throughput of only 2 Mbps. This would mean that an application on one host could send data to the other host at 2 Mbps.

Throughput or ***network throughput*** is the average rate of successful message delivery over a communication channel. The throughput is usually measured in bits per second (bit/s or bps), and sometimes in data packets per second or data packets per time slot. The ***system throughput*** or ***aggregate throughput*** is the sum of the data rates that are delivered to all terminals in a network. Users of

telecommunications devices, systems designers, and researchers into communication theory are often interested in knowing the expected performance of a system. From a user perspective, this is often phrased as either "which device will get my data there most effectively for my needs?", or "which device will deliver the most data per unit cost?" Systems designers are often interested in selecting the most effective architecture or design constraints for a system, which drive its final performance. In most cases, the benchmark of what a system is capable of or its 'maximum performance' is what the user or designer is interested in. When examining throughput, the term 'Maximum Throughput' is frequently used.

Finally, we often talk about the bandwidth requirements of an application—the number of bits per second that it needs to transmit over the network to perform acceptably. For some applications, this might be "whatever I can get"; for others, it might be some fixed number (preferably no more than the available link bandwidth); and for others, it might be a number that varies with time.

8.2.2 Bandwidth and Latency

Network performance is measured in two fundamental ways: bandwidth (also called throughput) and latency (also called delay). Latency corresponds to how long it takes a message to travel from one end of a network to the other. (As with bandwidth, we could be focused on the latency of a single link or an end-to-end channel.) Latency is measured strictly in terms of time. For example, a transcontinental network might have a latency of 24 milliseconds (ms); that is, it takes a message 24 ms to travel from point A to Point B. There are many situations in which it is more important to know how long it takes to send a message from one end of a network to the other and back, rather than the one-way latency. We call this the round-trip time (RTT) of the network.

We often think of latency as having three components. First, there is the speed-of-light propagation delay. This delay occurs because nothing, including a bit on a wire, can travel faster than the speed of light. If you know the distance between two points, you can calculate the speed-of-light latency, although you have to be careful because light travels across different mediums at different speeds: It travels at 3.0×10^8 m/s in a vacuum, 2.3×10^8 m/s in a cable, and 2.0×10^8 m/s in a fiber.

Second, there is the amount of time it takes to transmit a unit of data. This is a function of the network bandwidth and the size of the packet in which the data is carried. Third, there may be queuing delays inside the network, since packet switches generally need to store packets for some time before forwarding them on an outbound link. So, we could define the total latency as

Latency = Propagation + Transmit + Queue

Propagation = Distance/Speed-Of-Light

Transmit = Size/Bandwidth

where Distance is the length of the wire over which the data will travel, Speed-Of-Light is the effective speed of light over that wire, Size is the size of the packet, and Bandwidth is the bandwidth at which the packet is transmitted. Note that if the message contains only one bit and we are talking about a single link (as opposed to a whole network), then the Transmit and Queue terms are not relevant, and latency corresponds to the propagation delay only.

Bandwidth and latency combine to define the performance characteristics of a given link or channel. Their relative importance, however, depends on the application. For some applications, latency dominates bandwidth. For example, a client that sends a 1-byte message to a server and receives a 1-byte message in return is latency bound. Assuming that no serious computation is involved in preparing the response, the application will perform much differently on a transcontinental channel with a 100-ms RTT than it will on an across-the-room channel with a 1-ms RTT. Whether the channel is 1 Mbps or 100 Mbps is relatively insignificant, however, since the former implies that the time to transmit a byte (Transmit) is 8 μs and the latter implies Transmit = 0.08 μs.

In contrast, consider a digital library program that is being asked to fetch a 25-megabyte (MB) image—the more bandwidth that is available, the faster it will be able to return the image to the user. Here, the bandwidth of the channel dominates performance. To see this, suppose that the channel has a bandwidth of 10 Mbps. It will take 20 seconds to transmit the image, making it relatively unimportant if the image is on the other side of a 1-ms channel or a 100-ms channel; the difference

between a 20.001-second response time and a 20.1-second response time is negligible.

8.2.3 Delay × Bandwidth Product

It is also useful to talk about the product of these two metrics, often called the delay × bandwidth product. Intuitively, if we think of a channel between a pair of processes as a hollow pipe (see Figure 1.22), where the latency corresponds to the length of the pipe and the bandwidth gives the diameter of the pipe, then the delay × bandwidth product gives the volume of the pipe—the number of bits it holds. Said another way, if latency (measured in time) corresponds to the length of the pipe, then given the width of each bit (also measured in time), you can calculate how many bits fit in the pipe. For example, a transcontinental channel with a one-way latency of 50 ms and a bandwidth of 45 Mbps is able to hold

$$50 \times 10^{-3} \text{ seconds} \times 45 \times 10^{6} \text{ bits/second}$$

$$= 2.25 \times 10^{6} \text{ bits}$$

or approximately 280 KB of data. In other words, this example channel (pipe) holds as many bytes as the memory of a personal computer from the early 1980s could hold.

The delay × bandwidth product is important to know when constructing high performance networks because it corresponds to how many bits the sender must transmit before the first bit arrives at the receiver. If the sender is expecting the receiver to somehow signal that bits are starting to arrive, and it takes another channel latency for this signal to propagate back to the sender (i.e., we are interested in the channel's RTT rather than just its one-way latency), then the sender can send up to two delay × bandwidth's worth of data before hearing from the receiver that all is well. The bits in the pipe are said to be "in flight," which means that if the receiver tells the sender to stop transmitting; it might receive up to a delay × bandwidth's worth of data before the sender manages to respond.

In our example above, that amount corresponds to 5.5×10^{6} bits (671 KB) of data. On the other hand, if the sender does not fill the pipe—send a whole delay ×

bandwidth product's worth of data before it stops to wait for a signal—the sender will not fully utilize the network.

Note that most of the time we are interested in the RTT scenario, which we simply refer to as the delay × bandwidth product, without explicitly saying that this product is multiplied by two. Again, whether the "delay" in "delay × bandwidth" means one-way latency or RTT is made clear by the context.

8.3 Private and public networks

A ***private network*** is one in which all devices on the network and all links between those devices are used and administratively controlled by a single organization. Prior to the Internet, most large corporate networks were private networks. In a private network, when a network is spread out among multiple geographic locations in a WAN, the organization arranges for one or more telephone-company-provided leased lines, to connect each pair of sites, which will communicate with each other.

The major advantage of private networks is privacy. No one but the companies connected at either end of the leased line can access the data traveling across them, providing a high degree of security in terms of data privacy and ensuring that any users accessing machines on the network are authorized to do so.

The primary downside to private networks is cost. If you need to connect machines in location A and Location B, it's necessary to pay the phone company a tariff-based fee (usually based mostly on distance) for the cross-country leased line. If you add offices in Location C and Location D, you have to pay more tariff-based fees for thousands of miles of leased lines to add these sites to your private network.

A ***public network*** is one where network connectivity and resources are shared by many different administrative units. Typically no one company using the network has control over every piece of the network. One often shares the links between sites on a public network or more entities, with multiple organizations' data intermixed on the same network lines. The most well known public network is the Internet.

Public networks enable organizations to take advantage of economies of scale, as it's often the case that Wide Area Network links that handle a large amount of traffic aren't much more expensive than slower WAN links. With the cost of a faster link split among multiple organizations, users can enjoy faster network speeds at lower cost than would be possible on a private network.

Another advantage of public networks is that they allow organizations to basically time-share connectivity, and not pay for a line when they're not using it. Not every organization uses network capacity at the same time. If you transmit large amounts of data from Location A to Location B between 3am and 6am, but don't otherwise need connectivity between those two locations, why pay for the capability to network between those two points, the other 21 hours of the day?

The primary disadvantage of public networks is a reduced amount of control over data and host security. Since your data is traveling along paths that other organizations' data are also using, it is possible

Another disadvantage is that public networks often provide less control over bandwidth than do private networks. Unless you pay your public network service provider a substantial premium for guaranteed bandwidth (specified as a CIR, or committed information rate), odds are that when many other people are using the network, the speed of the network will seem to slow down. In some situations, this may not matter. But if you have a business-critical need to move data within a specific time window, guaranteed bandwidth may be an important criterion when selecting a network service provider.