Understanding Latency

Latency refers to the round-trip time for a request—specifically, the time for the initial packet to reach its destination, for the destination machine to reply, and for that reply to reach the requestor. Every network has latency. The total amount of latency that you'll get when connecting to a given remote host can vary widely, depending on network conditions.

Assuming your network connection is not overloaded, the bulk of a connection's latency comes from the laws of physics. The minimum latency between two points on the earth can be calculated by dividing the distance by the speed at which light or electricity moves in a particular medium (which is usually a large fraction of the speed of light in a vacuum, c).

For example, consider a packet traveling round trip from New York to San Francisco (about 2,900 miles, or 4,670 km):

- Over copper wire, the data moves somewhere between .66c and c (depending on the type of wire). Thus, a packet takes at minimum 15–24 ms each way, or about 30–48 ms round trip. This delay is barely noticeable.
- Over an optical fiber, the data moves at about 0.65c. Thus, a packet takes at minimum 24 ms each way, or about 48 ms round trip. This delay is also barely noticeable.
- Over a satellite connection, the packet must go to geostationary orbit (at an altitude of 35,786 km) and back down again. For a round trip, it must do this twice. Thus, at approximately c, the minimum round-trip latency is about 477 milliseconds, or almost half a second. This delay is *painfully* noticeable.

These calculations represent an absolute *lower* bound for the latency of a connection through those media. There are several other factors that can add additional latency on top of the link latency:

- Routing delays. A network packet can be further delayed by buffering at routers along its route. Whenever the rate of data exceeds the capacity of a particular network hop, the packets must be delayed until they can be sent. If your packets must travel through a highly congested network hop, this buffer delay can add considerable latency.

  For example, if a particular network hop is limited to 100 packets per second, a router can send only one packet down that wire every 10 milliseconds. Thus, when a router receives an additional packet to send down that wire, it must buffer the packet until the next free slot. If there are no other packets waiting to be sent, then the packet will be sent in ten milliseconds or less (five, on average). If there are already three packets waiting, then the packet will be delayed by an additional 30 milliseconds. Latency caused by time slots can be particularly noticeable in some types of cellular communications (EDGE, for example).
- Retransmission and exponential backoff. Whenever a host sends a TCP packet, it waits a while for the other end to acknowledge receipt of the packet. If it does not receive that acknowledgement after a period of time, the sender retransmits it. As the number of failures increases, the sender increases the retransmission delay exponentially, under the assumption that the packet loss was probably caused by a saturated link en route to the destination. Thus, a network connection with high packet loss (more than 1–2 percent) can significantly reduce performance.
- Signal propagation delays within hardware that receives, transmits, forwards, or repeats packets. For example, an Ethernet switch must receive an entire packet before it can begin sending the packet to its destination. Although a single switch or repeater adds only a small delay, those delays can add up over long distances. For example, early fiber optic cable systems required repeaters every 10 km (at most). Thus, an older fiber run from New York to San Francisco could easily have nearly 500 repeaters.