What Is Latency?

by Chris E. Gepp

More Data and More Data Movement

These days, data seems to be only getting bigger and storage only cheaper. We take twenty-two megapixel pictures, watch super high-definition video, and increasingly want to store it all in the cloud. In turn, more companies utilize cloud-based services, either simply for backup purposes or as a total replacement for their physical environments. The result is huge amounts of data constantly being pushed across the Internet, but a cyber utopia of perfect data movement has one great enemy: latency. What is "latency" and why is it such a problem?

Brought on-line in the late 1960s, the Internet, originally created by the Department of Defense, most likely was never intended for such wide-scale and heavy use. Since then the Internet has almost inarguably become an essential part of our modern lives. Because of the Internet's architecture and how it is used, though, complete data gridlock is almost always a potential reality.

The Internet is essentially a road system -- a devilishly long tangle of cables and disparate devices interconnected. Into this web a data packet flies, and somehow at the other end it manages to reach a certain specific computer among the countless number possible. We probably often discount this byzantine process of sending data packets as pure magic; not caring in the least about how the data gets there, just that it gets there.

Like cars on a road, though, all the data must actually make a physical journey to reach its destination. And just as with cars, sometimes not every data packet arrives on time and sometimes not at all. The route may be invalid ("a dead-end") or the connection may time out ("an empty gas tank") before the packet can arrive. Slow, inefficient connections serve to create bottlenecks ("traffic jams"), but astonishingly most of the time the data packets *do* make it.

"Latency" is the time spent because of these bottlenecks. Measured in milliseconds, latency is how long it takes for a data packet to travel from its origin to its destination. A quick journey might take 85 ms, whereas 300 ms would be a lengthy, drumming-fingers-on-the-keyboard kind of wait.

Let's take a closer look now at how latency *may* result. (And we say "may" because of the sheer complexity of what lies between Point A and Point B.)

Why So Much Latency?

To answer this question, it would help to look at the roads, or data pathways, themselves. They are not all paved in gold.

Just as when driving somewhere, different kinds of roads take different amounts of time. In the networking world, the cable medium that the data packet uses to reach its destination plays an important role in the data packet's travel time. For example, the ubiquitous phone-cord-like cable plugged into the back of most desktop computers is known as "UTP" (unshielded, twisted pair) cable. While being very "low-latency" -- the data doesn't have to travel far to reach its destination -- this cable type is also "short-distance" and cannot go very far either.

In order to even leave the building let alone travel around the planet, fiber optic cabling becomes the primary choice. Fiber optics work at the speed of light, making for another low-latency solution, in which light pulses rather than electrical signals are utilized. These two cable types are used extensively in building networks and represent a vast chunk of what forms the Internet.

If the above cabling is non-existent, though, how ever does the data get to where it is going? Wireless connections through technologies such as microwave, Wi-Fi, satellite, and cellular, may then become involved. Going wireless, though, almost guaranteed will result in a highly variable, "high-latency" connection. (Think of how random surfing the Web is when using a mobile phone.)

Finally, if only small data must be sent infrequently, the old copper "POTS" ("plain old telephone service") lines can be used through a dial-up connection. (Definitely not a big data solution.)

After knowing all this, one may wonder then, with generally so much low-latency cable media in existence, how there can still be so much latency. The answer: Math.

By adding up the travel time, or the cost, of a data packet using a network segment (a section of physical cable), a real cost starts to becomes apparent. In order for the data packet to reach its destination it usually must traverse multiple network segments. Regardless of media type, a cost is still incurred; like tolls, the more "roads" a data packet must travel along, the higher the cost.

Moving between network segments implies that the data packet is crossing out of one network and into another. Composed of millions of small networks, the Internet uses intermediary devices to interconnect among what would otherwise be disparate networks. Routers primarily serve this function, directing traffic to or away from their host networks. At this point the packet is held up by this traffic cop, who will either point it in the right direction or "drop" the packet, meaning the packet will not be going anywhere else.

Routers, being mechanical devices, move much more slowly than electricity or light. Sometimes these devices do not know the best way to get to the packet's destination. Or they may not know at all. The result is increased wait times or communication failures.

Beyond that, there are still other devices that further delay data packets, such as firewalls and WAN "accelerators." (There is often great irony on the Internet.)

Ways Forward

Hope is not lost, though. The best minds are on the job, and technology only continues to improve. Routers are becoming ever faster. Companies, if they can afford to, may replace their routers and perceive some increases in speed.

Faster still, though, would be getting rid of the device altogether. Doing away with a router implies establishing a more direct path. Essentially, a new bypass is constructed. Vendors such as Netflix now deliver their services much more reliably by reducing the distance between their servers and customer base, similar to Amazon, who have set up warehouses closer to their customers in order to offer same-day shipping.

If reducing the number of devices, or "hops," is not an option, software can also be optimized. By making more efficient use of TCP/IP, the protocol stack on which the Internet runs, vendors can see performance gains as well. Speed increases can come about by altering the size of the data packet or increasing the packet's TTL (its time-out period)

Finally, managing the relative chattiness of TCP/IP is an option. A "connection-oriented" protocol, TCP/IP relies on a constant stream of acknowledgements resulting in additional data being transmitted, adding significant overhead to the entire data transmission. When a file is in the billions of bytes already, this added payload can be quite costly indeed. More advanced protocols that practically eliminate latency have been developed by b2b (business to business) software companies like Signiant for very large file transfers. At the moment, such technology is exclusively used in data-intensive industries such as film and televisions production. But we may see such “next-generation” protocols being more widely adopted, as everyone’s data expands.

Conclusion

Latency will likely continue to be a challenge, as it is already baked into the Internet's design. And the world shows no loss in appetite for more and more big data. Hope exists, though, in the form of various performance enhancements, be it adding more direct paths, improving technology, or optimizing software. We are all stuck in the same traffic jam, but don't worry we'll all get there…eventually.