

# Why Latency Is Important

## Why Latency Is Important

Traditionally, the metric of focus in networking has been *bandwidth*. As more and more parts of the Internet have their capacity upgraded, bandwidth is often not the main problem anymore. However, network-induced *latency* as measured in [One-Way Delay \(OWD\)](#) or [Round-Trip Time](#) often has noticeable impact on performance.

### It's not just for gamers...

The one group of Internet users today that is most aware of the importance of latency are *online gamers*. It is intuitively obvious that in real-time multiplayer games over the network, players don't want to be put at a disadvantage because their actions take longer to reach the game server than their opponents'.

However, latency impacts the other users of the Internet as well, probably much more than they are aware of. At a given bottleneck bandwidth, a connection with lower latency will reach its achievable rate faster than one with higher latency. The effect of this should not be underestimated, since most connections on the Internet are short - in particular connections associated with Web browsing.

But even for long connections, latency often has a big impact on performance (throughput), because many of those long connections have their throughput limited by the [window size](#) that is available for TCP. And when the window size is the bottleneck, throughput is inversely proportional to round-trip time. Furthermore, for a given (small) loss rate, RTT places an upper limit on the achievable throughput of a TCP connection, as shown by the [Mathis Equation](#)

### But I thought latency was only important for multimedia!?

Common wisdom is that latency (and also jitter) is important for *audio/video* ("multimedia") applications. This is only partly true: Many applications of audio/video involve "on-demand" unidirectional transmission. In those applications, the real-time concerns can often be mitigated by clever buffering or transmission ahead of time. For *conversational* audio/video, such as "Voice over IP" or videoconferencing, the latency issue is very real. The principal sources of latency in these applications are not backbone-related, but related to compression/sampling rates (see [packetization delay](#)) and to transcoding devices such as H.323 MCUs (Multi-Channel Units).

## References

- [It's the Latency, Stupid](#) by Stuart Cheshire, May 1996 (periodically revised)
- [Why the Long Wait?](#) with contributions from David P. Reed, in FastCompany.com, September 2000
- [It's Latency](#) by Geoff Huston, in the ISOC's *ISP Column*, January 2004
- [Fighting Physics: A Tough Battle](#) by Jonathan M. Smith, ACM Queue, April 2009