# The Best Post on Latency You Will Ever Read

**Lori MacVittie, 2008-25-08**

No, it's not this one. It's not even mine. It's this one on High Scalability written by Todd Hoff. Not only does he explain latency and its sources, but its costs. Then he goes on to offer a plethora of ways to reduce latency.

A couple of suggestions he offers are:

> **Use a TCP Offload Engine (TOE)**. TOE tech offloads the TCP/IP stack from the main CPU and puts it on the network controller. This means network adapters can respond faster which means faster end-to-end communication. Network adapters respond faster because bus wait time is reduced as the number of transactions across the system I/O bus and memory bus are reduced.
>
> **Make TCP Faster**. FastTCP, for example, tweaks TCP to provide smoother and faster data delivery.

I'd also suggest combining his suggestions of load-balancing and caching with TCP optimizations and connection management optimization by deploying an application delivery controller instead of a legacy load-balancer.

Todd also mentions "**use Ajax to minimize perceived latency to the user.** Clever UI design can make a site feel faster than it really is", but be careful with AJAX. It often needs to be tuned itself and there are some libraries that are better at this than others, and you can also employ more advanced features in an application delivery controller to help you out there.

I could go on about how much I love this post and how well it syncs up with the benefits of an integrated application delivery controller but I won't. Read Todd's post. It's a must read if you're building or thinking about building a scalable web application.