On Reducing Latencies below the perceptible

*We should strive towards imperceptible latencies in
all human-computer interactions.*

Every moment of our lives is precious, and every moment spent waiting on a computer, wasted. We optimize our lives, almost subconsciously, in a quest for lower latency between what we want and how fast we get it. We move next to good schools when we have children (and away when we don't). We look for jobs in our general area, go to "convenience" stores, drive rather than walk, build wider highways, take planes rather than trains, skip through commercials, and take extraordinary steps throughout our lives to optimize for whatever we enjoy most.

In the last 200 years we've shrunk the world from where it took months to send a newspaper coast to coast to where a multitude can arrive in milliseconds.

In the last 30 years we've shrunk the Internet, too. Packets are now only rarely bounced off satellites in geosynchronous orbit, even attempts at low earth orbit failed - most traffic is now carried by an ever-expanding network of cables beneath the earth and sea. Various organizations are building straighter cables or even converting to wireless to shave nanoseconds off transit times. Data centers compressmachines into ever smaller rack spaces, and on every chip we squeeze more transistors into ever smaller spaces in the quest for faster, faster, faster response times to ever bigger questions.

In the last 10 years we've taken something of a step backwards – the SR-71 and the Concorde fly no more - but that is no reason to accept similar declines in performance on our networks and interconnects.

Now we have an always-on Internet, a reliable, if sometimes unpredictable friend, that can outstrip the speed of sound and touch closely at the speed of light, but doesn't. We almost have enough bandwidth, worldwide, but the bandwidth increases of the last decade have not been met by corresponding improvements in latency.

The behavior of interactive traffic (DNS, X11/RDP, ssh, voice and videoconferencing) are no better than 1999, and in many ways worse.

In 1991, I dreamed of a Jamaphone - a device that could connect a group of musicians and performers together over 1/3 of a continental distance. The speed of sound across an orchestra is roughly 40ms, and thus an inter networked orchestra could cover half of Europe!

The early experiments were promising but the raw capability is there no more in many cases.

I'd also had a dream - where we - as humans - wouldn't connect to the Internet, but to each other. The vast repositories of data that we carry around would be shared at the shortest distance and highest bandwidth - multiple orders of magnitude better than what can be achieved by looping everything through the interconnected InterNet, a *network of networks*.

I'd thought then that the edge of the network would retain the intelligence and that the core of the Internet would just route it. Never in my wildest dreams did I imagine what actually happened.

The InterNet stopped being a "network of networks" and became the "Internet", with a capital I, and you connected to "it" rather than "through it". Formally essential local services like email moved from inside of the local network to outside of it, with a corresponding increase in latency and jitter.

The same technologies that drove cross-continental latencies down to nearly the speed of of light in fiber, can also apply to the edges of the network.

The same technologies that have shrunk the world, have also driven local communities further apart. Why should we be better connected to a data center than to our next door neighbors or to our internal networks? We've girdled the world with cables but why can't we achieve latencies in the last miles measured as good as the original, 1981, Ethernet?

Google has a project: they want to reduce RTT latencies below 100ms to the user - in contrast, I'd like to see *cross-town* latencies drop below 2ms! I'd like to see protocols and services use the local-est network wherever possible, too, and wifi and wireless networks that also share the less than 2ms characteristic. And more data should move as close to the user as physically possible.

With latencies that low, instant tele-presence, severely time dependent services like earthquake alarms, and interactively participating in a band or performance by transporting high quality audio and video, all become feasible.

A skype call forces all your traffic, even if it is from upstairs to downstairs, through the internet, rather than your most local connection. The latest webrtc code and codec can carry samples of multi-channel audio on a 2.7ms period, over the closest connections it can find. That sounds about right, I can feel that much latency -

But after I sat down to write this I realized that to achieve this effective latency on encrypted https traffic - which require 6 RTTs just to start today - would require 330 microseconds as the baseline goal instead - And then I thought : "why stop there?"

Why not always aim for – in our networks, protocols, and applications - the lowest latencies possible under the laws of physics? At the very least: Why not always aim for latencies below that of human perception?

Lest you think that goals like this are impossible, the speed of light across San Francisco (6 miles) is 32 microseconds. There's plenty of margin left for overhead! Your neighbor is vastly closer (as is your child upstairs)

The structure of that InterNet would be very different from today's "hub and spoke" model - it would have fiber spread to every home – and interconnects in every neighborhood - and low power devices could handle much of what's outsourced to virtual machines in the data center today.

In everyone's hands today are machines vastly more powerful than any that existed 10 years ago on desktops - which have vast amounts of storage, yet eat milliwatts. There are terabits of potential local bandwidth in the air, unused most of the time, at frequencies with very short ranges.

The original InterNet was distributed, resilient, error, loss and recovery transparent, and while not as pervasive as today's Internet, was a more survivable and balanced medium.

"Progress" in wifi, has led to a pair of standards that consume bandwidth at the expense of latency and interoperability with your neighbor, where I'd rather have 5Mbits of low latency bandwidth in a crowded urban environment than 500Mbits in a Faraday cage.

With less men in the middle, with well designed protocols, we can reduce the latencies on all our networks and make them more sharable. We can replace the phone system, upgrading it along the way to have full videoconferencing ability, and one day have inter-networks so fast there would be essentially no difference in perception between being in the reality and the media that brought you there.

In improving mankind's interpersonal latency, I'd like us to always aim at the speed of light, and ever closer to the speed of thought.[1]

Dave Täht
CEO, TekLibre

1) http://the-edge.blogspot.com/2003/08/inner-workings-of-internet-mind.html