Workshop on Reducing Internet Latency

Position paper by Mat Ford, Internet Society

Competitive markets for Internet service provision are characterised by offerings defined along one axis only: the headline bandwidth of the connection. This made some sense when transitioning from dial-up access at speeds of 56kbps to broadband access at speeds 20 or more times that. It makes less sense when comparing different broadband offerings where the bandwidth of the last mile is no longer the bottleneck on the overall performance of the connection for a typical end user.

For a growing number of Internet applications and users, headline speed (above some threshold) is not the key, determining factor of the quality of the Internet experience: latency and latency variation are the most important factors. Even bandwidth intensive applications like streaming high-definition video services require less bandwidth than is commonly available in many broadband markets today.

Content distribution networks and widely distributed caching infrastructures give content providers a means to reduce end-to-end service latency by moving their end of the connection physically closer to their users. But the challenge of ensuring end-to-end latency management across the full range of Internet applications is obviously bigger than that. As the Internet and Internet-enabled services become ever more central to economic and social development, the pressure on all elements of the ecosystem to collectively engineer unnecessary latency out of the system will continue to grow.

As latency management has grown in importance for content providers and end users alike, we have seen and continue to see web, application and browser developers deploy techniques that appear to result in some measure of improved performance. However, these techniques are not necessarily widely tested across a broad range of deployment scenarios before going live, and are not well coordinated with other, equally well-intentioned, actions of other players. Domain sharding, for example, can improve website performance in some scenarios, and opening multiple HTTP connections can improve browser performance in some scenarios, and the combination of the two can be very detrimental to performance and levels of network congestion. It is unlikely that any significant progress will be made in a general sense for Internet performance while individual actors seek ways to 'work around' other components of the system. A holistic approach is the only viable, long-term strategy.

There is therefore a need for a more co-ordinated cross-industry focus on latency management for the Internet. Building as broad a consensus as possible around a roadmap for application and browser developers, stack vendors, chipset vendors, CPE, router and network appliance manufacturers and network operators would be a positive step in the right direction. Better cross industry co-ordination implies a need for awareness-raising activities as well as consensus-building activities to both develop and promulgate promising approaches to the widest set of potential stakeholders.