Understanding Latency – A Root Cost and Mitigation Approach

Abstract: Latency has always been a key part of network performance, but recent increases in network bandwidth, router forwarding speed, and end-system computational resources have elevated it to a primary focus for electronic traders, search engines, name-servers, data centers, and home Internet users. Even so, latency has typically been challenging to both measure and to mitigate, largely because previous approaches have addressed symptoms resulting from complex combinations of root causes and because end-to-end interaction delays are a multidimensional combination of these effects. This tutorial presents a comprehensive exploration of latency by focusing on the independent root causes and their associated costs, and exploring mitigations directly focused on those causes. We begin with an analysis of the transaction latency budget and its relation to the communicating parties - whether human or computer, pairwise or group. We explore boundaries of that budget, whether binary or gradated, and the complex ways in which latency costs can be usefully expressed. The root causes of *generation*, *transmission*, *processing*, *multiplexing*, and *grouping* are discussed in depth as well as corresponding mitigations of *relocation*, *speed-up*, *resource dedication*, and *avoidance*. We address these causes and mitigations in the context of examples including 'bufferbloat', Internet and big-data search, and protocols specifically aimed at reducing or tolerating latency for a variety of environments including home Internet access, data center operation, high-speed trading, and interplanetary communication. This tutorial also focuses on emerging opportunities for latency reduction that resulted from resource trade-off changes, including reducing component message sizes and a variety of anticipation techniques. Finally, we discuss how to apply these mitigations and new opportunities in both current and future network architectures, including wireless and optical physical layers; packet and circuit systems; location, location-independent, and name-based services; and network management.

Presenter: Joe Touch, University of Southern California

Joe Touch (IEEE Senior Member '02) is the Postel Center Director at the University of Southern California's Information Sciences Institute (ISI) and a Research Associate Professor in USC's Computer Science and EE/Systems Departments. He received a B.S. in both biophysics and CS from the Univ. of Scranton in 1985, an M.S. in CS from Cornell Univ. in 1987, and a Ph.D. in CS from the Univ. of Pennsylvania in 1992. He joined ISI in 1992 and his research involves three distinct areas: virtual networks, digital optical processing, and network security simplification. He developed network virtualization for concurrent overlays (the X-Bone) supporting spread-spectrum DDOS defense and satellite constellations, which led to the Recursive Network Architecture (RNA) - a first-principles model for the current and future Internet. He is implementing all-optical Internet router functions within the NSF CIAN ERC and exploring digital optical processing compatible with optical transmission in the NSF INSPIRE Optical Turing Machine. He is also designing zero-identity, high-speed Internet security. His other interests include Internet protocols, network services, and network device design. He holds 5 US patents and has published over 100 papers in conferences and journals. Joe is a member of Sigma Xi and an ACM Distinguished Scientist. Within the IEEE Communications Society he received the TCCC Outstanding Service Award and is a Distinguished Lecturer. He participates actively in the IETF Transport, Internet, and Security Areas and chairs IANA's Transport Port Expert Review team. He serves on numerous conference committees and the editorial boards of *IEEE Network* and *Elsevier's Journal of*

Computer and System Sciences. Joe has over 25 years of experience analyzing communication latency and his latency-focused publications include his Ph.D. dissertation - see: http://www.latency.org/