KEEPING THE INTERNET
OPEN • INNOVATIVE • FREE

www.cdt.org

CENTER FOR DEMOCRACY
& TECHNOLOGY

1634 I Street, NW
Suite 1100
Washington, DC 20006

P +1-202-637-9800
F +1-202-637-0968
E info@cdt.org

WORKSHOP ON REDUCING INTERNET LATENCY

CDT POSITION PAPER

June 21, 2013

One key area of ongoing discussion concerning Internet latency relates to technical solutions available to network operators for reducing the latency experienced by real-time interactive applications (voice and video calling, for example). Two classes of complementary solutions that have been suggested for operators to investigate are (1) improved buffer management throughout the network via active queue management or other means, and (2) traffic classification and prioritization of latency-sensitive traffic. These solution tracks present several quandries from a deployment perspective.

Many network operators offer their own real-time services (particularly voice) that make use of segregated bandwidth or Quality of Service mechanisms to ensure low-latency performance. These operators therefore may not have sufficient incentives to invest in technical solutions whose main beneficiaries are over-the-top services that compete with their own real-time services, or to make the solutions in which they invest available to competing applications.

Traffic classification and prioritization requires some scheme for determining which traffic will be assigned to which traffic class and prioritized for delivery. Allowing individual applications to self-declare their needs for priority may be both difficult to manage (since operators cannot anticipate the proportion or schedule of traffic sent and received in each traffic class) and open to abuse by applications that falsley claim priority needs. Conversely, requiring prior negotiation or payment between application providers and operators as a means to select traffic for priority delivery raises scalability concerns for application providers, who would need to separately negotiate with many different network operators to improve performance for users globally. It also creates problematic incentives for improving network performance in the long term, since it creates a revenue stream for operators that relies on prioritized delivery being a necessity for real-time applications to succeed.

These quandries point to a few potentially fruitful directions for technical solutions that require network operator involvement. First, emphasis should be placed on solutions that maximize performance benefits across different kinds of applications. Many operators likely have an interest in reducing the latency of interactive web applications, games, or collaboration software. Some mobile operators use TCP proxies to improve web browsing performance, but have been more reluctant to manage large buffers in the network, perhaps because the perceived benefits of buffer management largely accrue to over-the-top voice

services. Solutions that are seen as benefitting more than voice and video telephony would appear to have a greater chance of being deployed.

Second, within the realm of traffic classification and prioritization, emphasis should be placed on classification schemes based on generic criteria, unrelated to either application self-declaration or prior negotiations between application developers and operators. For example, stochastic queueing schemes, or those that segregate traffic based on specific network- or transport-layer attributes that applications are unlikely to falsley report, could deliver benefits to latency-sensitive applications without the drawbacks described above.

In sum, focus should be directed at solutions that align with operators' immediate deployment incentives, are simple for applications to leverage without being abused, and support long-term network performance improvement. All of these goals may not be concurrently achievable, but they are worthwhile principles for guiding the discussion.