Interaction between Applications and the Network

Malathi Veeraraghavan and Zhenzhen Yan
University of Virginia
Charles L. Brown Dept. of Electrical & Computer Engineering
University of Virginia, POB 400743
Charlottesville, VA 22904-4743
{mvee, zy4d}@virginia.edu

Abstract: Optical circuit-switched networks deployed in core (backbone) networks are used primarily to offer leased-line ("static") services to interconnect IP routers. Most end-user applications are implemented to use the TCP/IP protocol stack of the Internet. Recently, a number of eScience applications have emerged that could rightly be characterized as "heavy-hitters" in that they require a disproportionately larger allocation of rate-hop-duration product when compared to most Internet flows. This has led to the deployment of dynamic circuit services within core networks. This paper describes applications and their interaction with IP-routed networks and optical dynamic circuit switched networks.

OCIS codes: (060.4230) Multiplexing; (060.4250) Networks; (060.4253) Networks, circuit-switched

1. Problem statement

The problem considered in this paper is how applications share network bandwidth. The role of a network is to move data across a path consisting of one or more links between source and destination, and for economic reasons, multiple flows/applications have to be accommodated simultaneously. There are two dimensions to data transfers: rate and duration. For example, a block of size 1MB can be moved within 1 second with a rate allocation of 8 Mbps, or in 8 seconds with a rate allocation of 1 Mbps. Add to this the third dimension of the number of links (hops) on the end-to-end path to define *allocation* as a 3-tuple vector: *{rate, hop, duration}* [1].

What allocations do different applications require, and how do different types of networks meet these needs? These questions are addressed in this paper.

2. Applications on IP routed networks

The end-to-end principle on which the Internet architecture is based has led to loose ties between applications and the network. IP routers within the network have the simple role of forwarding datagrams toward their destinations. Routers store no state information and therefore cannot relate datagrams from any single flow, let alone an application[1]. If the network has no way of knowing about flows or applications, the best it can do is to offer an application resources sufficient to transmit a single datagram on a single link. For example, with Ethernet's Maximum Transmission Unit (MTU) of 1500 bytes, in today's Internet, which is dominated by Ethernet, regardless of the application or its overall needs, each allocation is just 1500B on a single link. That same datagram needs to wait in a buffer at the next router for another 1500B allocation before it can transit the next link on the end-to-end path.

No limits are placed on how many datagrams a particular application can feed into a network simultaneously through multiple interfaces, or back-to-back sequentially through a single interface. With TCP's congestion control algorithm, a TCP sender can gradually increase its sending rate until it is consuming the lion share of a link's bandwidth. Viewed from this context of flows, one sees that large-sized flows (referred to as "elephant" flows in [2]) enjoy a much higher rate-hop-duration allocation than smaller-sized flows ("mice" flows [2]). This approach of bandwidth sharing has both advantages and disadvantages. The advantages are that network bandwidth is well utilized, and large files enjoy lower transfer delays. The disadvantage is that it increases delay variance for inelastic flows, e.g., RTP audio-video flows. These are inelastic flows in that their rates are dictated by the source coding rates, in contrast to file transfers that are "elastic" in the sense that TCP can dynamically adjust sending rates.

Evidence of this disadvantage is seen even in lightly utilized networks. Research-and-Education Networks (RENs) such as Internet2 have a policy of operating their network at light loads (25-30%) to allow these networks to absorb surges in traffic caused by eScience users who often move terabyte-to-petabyte sized datasets [3]. Fig. 1

[1] An application may consist of one or more TCP flows or Real-time Transport Protocol (RTP) flows.

shows delays observed across a lightly loaded (less than 20%) ESnet path. ESnet [5] is a core (backbone) REN that connects national laboratories across the US.

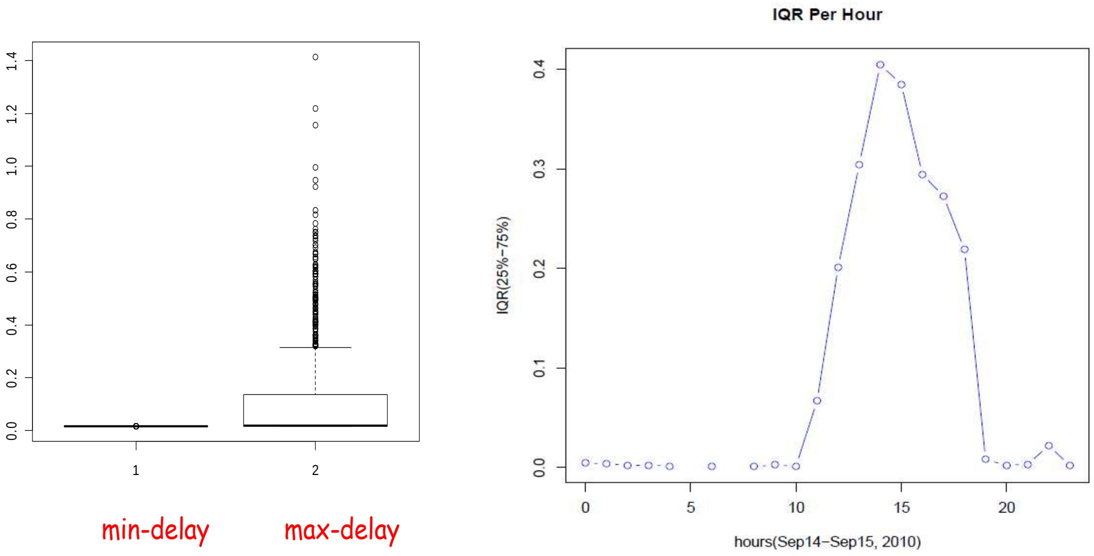

Fig. 1 OWAMP (one-way ping) [4] delays between servers directly connected to ESnet's El Paso router and Boise router collected over a 24 hour period from Sep. 14, 23:00 to Sep. 15, 2010. Approximately 600 packets are sent every minute, and the minimum delay and maximum delay in each minute is reported. The boxplot shows that the spread in minimum delay over the approximately 1440 data points obtained for the day is small, while the spread in maximum delay is large. The minimum delay is 15.4 ms, while the 75% quartile of maximum delay is 137 ms, and the absolute maximum delay is 1.4 s. The right-hand side plot shows the Inter-Quartile Range (IQR) for maximum delay on a per-hour basis, with IQR increasing during business hours.

3. Applications on optical circuit-switched networks

With a hypothesis that elephant flows are causing the type of delay variance (jitter) seen in Fig. 1, ESnet has deployed a virtual circuit network called the Science Data Network (SDN) to complement its IP routed network. It uses Multi-Protocol Label Switching (MPLS). Circuits are provisioned for the elephant flows created by eScience users across ESnet, and datagrams from these flows are redirected to the circuits at provider-edge routers owned by ESnet but located at their customer sites. This brings us to the question of how applications interact with such networks.

The key difference between datagram networks and circuit/virtual circuit networks is that the latter support admission control. By definition, admission control requires applications to specify characteristics of their traffic before transmission begins. Clearly, this can be constraining and is a negative when compared to datagram networks where no such specification of expected traffic is required. For example, how does an application such as a Web client anticipate the characteristics of the traffic exchanged between itself and a Web server. This is highly dependent on what the user of the Web client does. Nevertheless there are applications where such an a priori specification is feasible. For example, consider plain old telephone service (POTS). When a user dials digits, signaling messages are sent from switch-to-switch asking for an allocation of 64 kbps for an unspecified duration of time. With this information about the expected traffic on that "flow," the network is able to make an allocation for the flow on a multi-hop basis. Consider the 3-tuple mentioned in Section 1. The "rate" is 64 kbps, "hops" consist of all the links on the end-to-end path of the circuit, and "duration" is left open. By making such a guaranteed allocation, the delay and delay variance for data sent on this circuit is kept shielded from other traffic. ATM switched virtual circuit (SVC) services were designed to operate in a similar manner in that applications could communicate their traffic specifications to networks and obtain multi-hop rate-hop allocations with open-ended durations.

In a 2006 paper [6], we showed that such open-ended allocations are feasible even when the desired utilization is high, i.e., call blocking rate is acceptably low, if the rate allocated per flow is a small fraction of link capacity. On the other hand, if the rate allocation is a significant fraction of link capacity, e.g., one-tenth (1 Gbps is allocated for a single flow on a 10 Gbps link), then a scheduled dynamic circuit service (SDCS) [7] is required to allow for high utilization operation at acceptable performance levels. Such high rate allocations are primarily useful for large file transfers.

To be able to schedule an allocation, durations have to be specified in one form or another without which the network scheduler cannot know when existing calls will terminate to allow for the scheduling of a new request. For elastic applications such as file transfers, file size is a proxy for duration because the network scheduler can determine duration based on the allocated rate. Users could be allowed to optionally specify deadlines for file transfers, which just adds another constraint to the scheduling task. There are many examples of inelastic applications where users are perfectly willing to specify both rate and duration. These include video-conferencing, remote visualization, remote instrument control, distance learning, video-on-demand, cloud computing, etc. Users of these applications often desire advance reservation capability (contrast this to a POTS call where no mechanism exists to process such requests). This has led researchers to refer to this as “book-ahead” [8] or “advance-reservation” services, but we prefer the term “scheduled” because the key parameter is duration rather than the desired start time. We propose the terminology, Specified Start Time (SST) and Earliest Start Time (EST), as sub-classes of scheduled dynamic circuit services. The SST sub-class allows applications to specify a set of optional start times, and is thus suitable for applications such as distance learning where other resources need to be co-reserved. The EST sub-class supports file transfer applications that typically want an immediate start time but can accept an earliest start time allocation as determined by the network scheduler. An Inter-Domain Controller Protocol (IDCP) [9] has been developed to support such scheduled dynamic circuit services across multiple RENs.

4. Applications in hybrid networks

Comparable to the problems encountered with extending ATM to the desktop, service providers are facing difficulties in extending their core dynamic circuit services to the desktop. While core RENs, such as ESnet and Internet2 in the US, GEANT2 in Europe, and JGN2Plus in Japan, and commercial providers, such as AT&T and Verizon, are offering such high-rate optical dynamic circuit services, regional and access providers, and enterprise networks, have not yet done so. This has led to solutions such as Lambdastation [10] in which applications are modified to communicate with servers that signal core networks before initiating elephant flows allowing core providers to set up circuits and create policy based routing entries at edge IP routers to redirect these flows to the circuits.

For commercial applications, such dynamic circuit services would be useful on congested access links, such as residential access links. Even as passive optical networks (PONs) emerge in this market, the load during busy (evening) hours will no doubt be high as large file downloads take an unfair share of link bandwidth.

5. Summary

A number of eScience and commercial applications require a larger allocation of rate-hop-duration product than most Internet flows. This paper describes how such applications interact with IP-routed networks, and optical dynamic circuit switched networks.

6. Acknowledgment

This work was supported by the DOE grant DE-SC002350. The authors thank Zhiyang Yu, University of Virginia, for his programming support, and Mark Karol, Telcordia, for his insights.

7. References

[1] M. Veeraraghavan, M. Karol, G. Clapp, “Optical dynamic circuit services,” IEEE Communications Magazine, 48, 11, 109-117 (November 2010)

[2] K-c. Lan and J, Heidemann, “A measurement study of correlations of Internet flow characteristics,” Comput. Netw. 50, 1, 46-62 (January 2006).

[3] R. P. Vietzke, “Internet2 Headroom Practice,” Aug. 15, 2008, https://wiki.internet2.edu/confluence/download/attachments/17383/Internet2+Headroom+Practice+8-14-08.pdf?version=1

[4] “OWAMP: One-Way Ping,” http://www.internet2.edu/performance/owamp/

[5] “Energy Sciences Network (ESnet)”, http://www.es.net/

[6] M. Veeraraghavan, X. Fang, X. Zheng, On the suitability of applications for GMPLS networks, in Proc. of IEEE Globecom 2006, San Francisco, Nov. 27 - Dec. 1, 2006.

[7] M. Veeraraghavan and D. Starobinski, “A Routing Architecture for Scheduled Dynamic Circuit Services,” in *Proc. of ACM ReArch Workshop* (held in conjunction with ACM CoNEXT), Philadelphia, PA, Nov. 30, 2010.

[8] A. G. Greenberg, R. Srikant, and W. Whitt, “Resource sharing for book-ahead and instantaneous-request calls,” IEEE/ACM Trans. Netw., 7, 1, 10–22 (1999).

[9] DANTE, Internet2, Canarie and ESNet (DICE), “Inter-Domain Controller (IDC) Protocol Specification,” May 30, 2008, http://hpn.east.isi.edu/dice-idcp/dice-idcp-v1.0/idc-protocol-specification-may302008.doc

[10] “The Lambda Station Project,” http://www.lambdastation.org