On the Treatment of Application-Limited Streams

Andreas Petlund[†], Anna Brunstrom[‡], Jonas Markussen[†], Markus Fuchs[‡*]
[†]Simula Research Laboratory, [‡]Karlstad University, *University of Kaiserslautern

Introduction

For streams that probe actively for bandwidth, a lot of work has been done to define how to behave fairly, mainly by dividing the resource of bottleneck throughput between the competing streams over time. Although this work targets greedy traffic it has a tendency to steer our thinking of what is fair for all types of traffic.

For streams that are application-limited, latency is more important than throughput. When reliable transport is required, the latency induced by the need to retransmit lost packets can cause problems. Redundancy and more aggressive retransmissions may improve latency for such streams. The use of aggression and redundancy has also been explored as a means of reducing retransmission latency [1, 2, 3]. However, to which degree such aggression should be allowed, and how to weigh the need for throughput against the need for latency, has only been superficially treated so far. Proposals that advocate more aggressive behaviours for application-limited streams are often met with scepticism and arguments that such behaviour will not be fair to competing traffic.

In this position paper, we present experimental results showing how application-limited streams lose against greedy streams in the "traditional", throughput-based fairness regime. Even when more aggressive retransmission mechanisms are applied, the application-limited streams still lose the battle against the greedy streams. Our results suggest that it is defendable to use aggressive retransmissions to reduce latency in many cases. We invite a discussion on how to define the level of aggression that can be applied without clogging the tubes and how to explore and formulate guidelines that help constantly application-limited streams recover in a timely fashion.

Sharing behaviour of application-limited streams

Figure 1 shows results from a set of experiments where we send an increasing number of streams over a 1Mbps bottleneck. The dotted line shows the expected bandwidth consumption for the thin streams if given their "fair share". We here define the fair share for the thin streams as the aggregate bandwidth needed for the thin streams as long as that number is less than half the bottleneck capacity. We also ran this experiment on thin streams using two mechanisms that increase aggressiveness for retransmissions: one where the sender performs a "fast retransmit" on the first dupACK it receives (mFR), thus reacting on the first indication that loss has happened; and one where six retransmissions using the base RTO are performed before the RTO is exponentially increased (LT), thus increasing the chance of recovering the segment without extreme delays. The two mechanisms were applied both individually and in combination. The goal of this experiment was to see if being slightly more aggressive will skew the throughput fairness in a severely congested scenario. The results show no significant difference between achieved throughput for competing greedy streams when aggressiveness for thin-stream retransmissions is raised. We can see in Figure 1 that, as competition gets tougher, the thin streams consistently lose the struggle for throughput-resources.

We focus in this position paper on the sharing characteristics of application limited streams, as arguments for why we believe more aggressive behaviours for such streams are well justified. The positive benefits on latency of the two more aggressive retransmission mechanisms used in the experiments were demonstrated in laboratory experiments in [2] and in a "live" game server evaluation in [4]

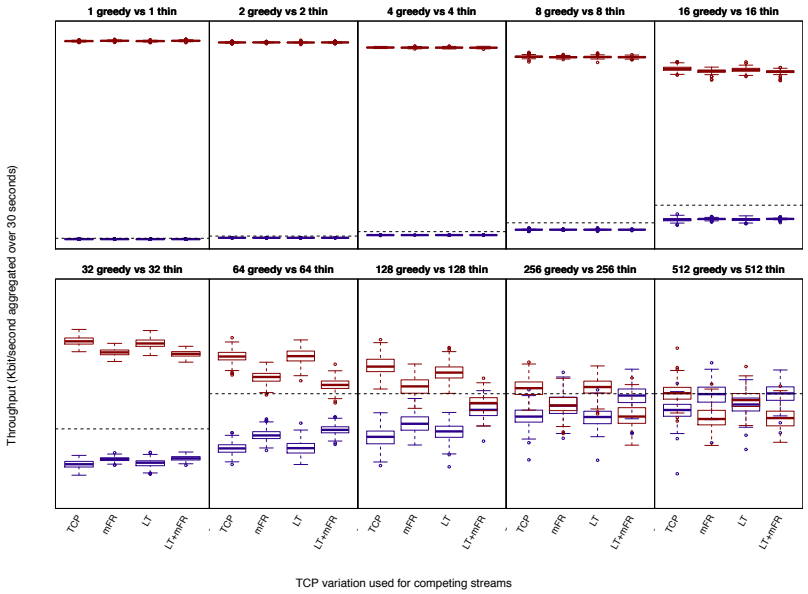

Figure 1: Aggregated throughput of thin streams competing with greedy TCP streams. The plot shows aggregate throughput of all greedy (red) streams and all thin streams (blue) normalised over 30 second intervals. Thin-stream properties: Packet-size: 160 B, Intertransmission-time: 150 ms. Link properties: Bottleneck bandwidth: 1Mbps, RTT: 100ms, queue size: 1 BDP (9 packets * 1500B)

Discussion

For congestion-controlled streams, fairness is deeply investigated, modelled and researched. Most commonly used when measuring fairness is Jain's fairness index [5], which is a general metric. Although it allows to assess fairness in different dimensions like latency or loss rate, throughput is still the predominant choice. For application-limited streams that are not aggressively probing for bandwidth, however, there is no clear consensus on how to limit aggressiveness. There has been a common practise of condemning spurious retransmissions in order to conserve bandwidth resources for goodput. Our view is that this argument should be reviewed when retransmission latency is important. Our experimental results show that application-limited streams are currently at a disadvantage when sharing resources with greedy streams, suggesting that more aggressive behaviour is appropriate.

Furthermore, we believe that improved mental models, as well as operational performance metrics, for how streams of different types should share network resources are required. It is unclear by which standards one should guide the evaluation of resource allocation in order to justly evaluate schemes based on redundancy and more aggressive retransmissions.

Acknowledgement

The authors are funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project(ICT-317700). The views expressed are solely those of the authors.

References

[1] M. Allman, K. Avrachenkov, U. Ayesta, J. Blanton, and P. Hurtig, "Early Retransmit for TCP and Stream Control Transmission Protocol (SCTP)," RFC 5827 (Experimental), Internet Engineering Task Force, May 2010. [Online]. Available: http://www.ietf.org/rfc/rfc5827.txt

[2] A. Petlund, "Improving latency for interactive, thin-stream applications over reliable transport," Ph.D. dissertation, Simula Research Laboratory / University of Oslo, Unipub, Kristian Ottosens hus, Pb. 33 Blindern, 0313 Oslo, December 2009.

[3] T. Flach, N. Dukkipati, A. Terzis, B. Raghavan, N. Cardwell, Y. Cheng, A. Jain, S. Hao, E. Katz-Bassett, and R. Govindan, "Reducing Web Latency: the Virtue of Gentle Aggression," in *Proceedings of the ACM Conference of the Special Interest Group on Data Communication (SIGCOMM '13)*, 2013.

[4] A. Brunstrom, A. Petlund, and M. Rajiullah, "Reducing Internet Transport Latency for Thin Streams and Short Flows," in *Proceedings of Future Network and MobileSummit (poster paper)*, 2013.

[5] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems," DEC-TR-301, Digital Equipment Corporation, Tech. Rep., Sep. 1984. [Online]. Available: http://arxiv.org/abs/cs.NI/9809099