TIMELY: RTT-based congestion control for the datacenter – Public Review

Mohammad Alizadeh
MIT Computer Science and Artificial Intelligence Laboratory
alizadeh@csail.mit.edu

New technology sometimes presents an opportunity to revisit how we design systems by changing our basic underlying assumptions. This is the story of this paper, which challenges some conventional wisdom about delay not being an accurate congestion signal in low latency networks by using recent advances in NIC hardware.

The context is datacenter congestion control. Traditional TCP transport stacks fare poorly in this environment, which has led to considerable interest in recent years in developing specialized transports that aim to deliver high bandwidth utilization at extremely low, microsecond-level packet latency. This is important for demanding datacenter applications such as cloud storage and near-realtime web services, where potentially 1000s of coordinated backend servers need to communicate to answer a single user query.

Existing proposals, ranging from enhancements to TCP to completely redesigned network fabrics, have largely ignored a classic indicator of congestion: delay in the form of round-trip time (RTT) measurements. The thinking has been that RTT measurements are too noisy to be useful in low latency networks because microsecond timescale increases in queueing delay can be drowned out in measurement noise at the servers.

In this paper, Mittal et al. show that recent NIC hardware features such as precision timestamping of packet events make accurate RTT measurement possible, thus largely bypassing the limitations of software. These high-precision RTT measurements provide a very reliable estimate of network queuing latency. There is a figure in the paper that illustrates this beautifully. It compares measured RTTs with direct queue length measurements at the switch — the match is uncanny.

The authors proceed to develop TIMELY, a new congestion control algorithm that uses high precision RTT measurements to great effect. For example, it lowers 99th percentile tail latency by an order of magnitude in a kernel-bypass, RDMA messaging stack. TIMELY's control algorithm, based on an interesting use of the RTT *gradient* for feedback stabilization, overcomes several important practical challenges such as infrequent RTT samples and burstiness due to hardware offloads.

TIMELY's evaluation is solid and persuasive, especially, the methodical way in which the RTT measurement framework is benchmarked. The paper has several excellent examples of how to design experiments to gain insight into system behavior. In one experiment, for instance, random noise of varying magnitude is added to the RTT measurements to investigate TIMELY's sensitivity and prove that the NIC hardware support is key.

The paper also leaves several avenues open for future work. While TIMELY's gains with respect to existing baselines are impressive, it still incurs over 100 microseconds of latency at the tail under high utilization. For RDMA applications, which TIMELY's implementation is based on, this latency can be quite significant, as single-digit microsecond latencies for small reads and writes are readily achievable today in the absence congestion. This raises important questions about how to get closer to the limits of the hardware. How should delay-based congestion control evolve at higher link speeds? What are the fundamental barriers due to traffic burstiness caused by NIC offloads? What role can hardware pacing or other hardware mechanisms play? The paper does not answer these questions, but provides useful data points towards their resolution.

Another area for further research is a more formal theoretical analysis of TIMELY's control loop. The algorithm has several tunable parameters begging for a stability analysis. Finally, TIMELY's implementation benefits from kernel bypass, but its comparison against other datacenter transports is limited to conventional DCTCP. Comparing TIMELY against user-space or otherwise optimized implementations of non RTT-based datacenter transports (e.g., in a dataplane OS like the recently proposed IX) would be interesting.

In summary, this paper takes an old idea — delay-based congestion control — and pushes it to new heights with an interesting use of modern hardware and wonderfully meticulous system design. But it won't be the last word in this space. With new use cases like RDMA emerging, we can expect more work towards bringing the full capabilities of modern hardware to bear on demanding datacenter applications.