Restricted Slow-Start for TCP

William Allcock[1], Sanjay Hegde[2] and Rajkumar Kettimuthu[1]

[1]Argonne National Laboratory
Argonne, IL 60439, USA
{allcock, kettimut}@mcs.anl.gov

[2]California Institute of Technology
Pasadena, CA 91125, USA
hegdesan@caltech.edu

Abstract

In network protocol research a common goal is optimal bandwidth utilization, while still being network friendly. The drawback of TCP in networks with large bandwidth-delay products due to its AIMD based congestion control mechanism is well known. The congestion control algorithm of TCP has two phases namely slow-start phase and congestion-avoidance phase. Many researchers have focused on modifying the congestion avoidance phase of the algorithm. In this work, we propose a modification to the slow-start phase of the algorithm to achieve better performance. Restricted slow-start algorithm is a simple sender side alteration to the TCP congestion window update algorithm.

1. Introduction

TCP was originally defined in RFC 793 [1], and several enhancements have been proposed to TCP since then [2-4]. The congestion control algorithm of TCP has two phases namely slow-start phase and congestion-avoidance phase. With slow start, the sender window begins at one segment and is incremented by one segment every time an acknowledgment is received. This opens the window exponentially: send one segment, then two, then four, and so on. With congestion avoidance, the sender window is incremented at most one segment each round-trip time, regardless of how many acknowledgments are received in that round-trip time. The congestion control algorithm starts with the slow-start phase. Whenever congestion is detected, it reduces the sender window to half of its value and enters congestion avoidance. This multiplicative decrease per congestion event is too drastic and linear increase by one packet per round-trip time in the congestion avoidance phase is too slow for networks with large bandwidth-delay products. Recently, researchers have formulated numerous approaches to address the limitations of the AIMD (Additive Increase Multiplicative Decrease) based TCP's congestion control algorithm [5] in long-fat networks (networks with large bandwidth and long delay). These include both loss-based solutions [6,7] and delay-based solutions [8-16].

The current slow-start procedure can result in increasing the sender window by thousands of segments in a single round-trip time for networks with large bandwidth-delay products. Such an increase can easily result in thousands of packets being dropped in one round-trip time. This is often counter-productive for the TCP flow itself, and is

also hard on the rest of the traffic sharing the congested link. In this work, we propose a modification to the slow-start procedure to solve this problem and improve the network utilization.

2. Background and motivation

Congestion occurs when the traffic offered to a communication network exceeds its available transmission capacity. But congestion events are not just pertained to congestion in the network. In some operating systems (for example: Linux), congestion events (send-stalls) are generated due to the saturation of several soft network components such as buffers and queues in the host. Though these are resource constraints at the sending host and are not in any way indicate of congestion in the network, Linux TCP treats these events in the same way as it would treat the network congestion. The impact of these send-stall events was reflected in the demo that we conducted at IGrid2002 [17]. Further analysis revealed that these congestion events (send-stalls) are generated in the slow-start phase rather in the congestion avoidance phase. Motivated by the previous works [18-20], we propose a control theory approach that appropriately paces the TCP sender during the slow-start phase to avoid the saturation of soft component such as device queue. Even though there have been proposals to increase the size of these soft components to overcome the problem, deployment of these solutions revealed that still a considerable amount of available bandwidth goes unutilized. Also, increasing the size of the soft components increases the memory usage. We aim at improving the end-to-end bandwidth utilization without increasing the memory usage at the host.

3. Approach

We use a PID control algorithm [21] to determine the rate of increase during the slow-start phase. In the PID control approach, the gain is calculated using a first order differential equation. The controller gains are configurable. The 90% of the maximum value of the interface queue (IFQ) size is used as the set point and the current value of the IFQ is used as the process variable in the controller. The controller compares the process variable (current IFQ) to its set point (max IFQ) and calculates the error. Based on the error (E), a few adjustable settings and its internal structure, the controller calculates an output that determines the new value of the sender window. The PID transfer function used is

$$K_p * (E) + 1/T_i \int_0^t (E)\, dt + T_d * {}^{d(E)}/_{dt})$$

We use Ziegler Nichols Tuning Method [22] to calculate the PID parameters (K_p, T_i and T_d). A brief description of the method is as follows:

1. Select proportional control alone
2. Increase the value of the proportional gain until the point of instability is reached (sustained oscillations), the critical value of gain, K_c, is reached.
3. Measure the period of oscillation to obtain the critical time constant, T_c.

Once the values for K_c and T_c are obtained, the PID parameters are calculated as follows: $K_p = 0.33\ K_c$; $T_i = 0.5\ T_c$; and $T_d = 0.33\ T_c$.

4. Experimental Results

Our scheme is implemented in a 2.4.19 linux kernel and the performance is evaluated through experiments conducted over a 100 mbps link between Argonne National Laboratory and Lawrence Berkeley National Laboratory, a round-trip time of 60 ms. We use web100 [23] to get detailed statistics of the TCP state information. Preliminary results show that our scheme is able to achieve 40% improvement in throughput compared to the standard TCP. Figure 1 compares the cumulative send-stall signals over time in modified TCP with that of the standard Linux TCP.

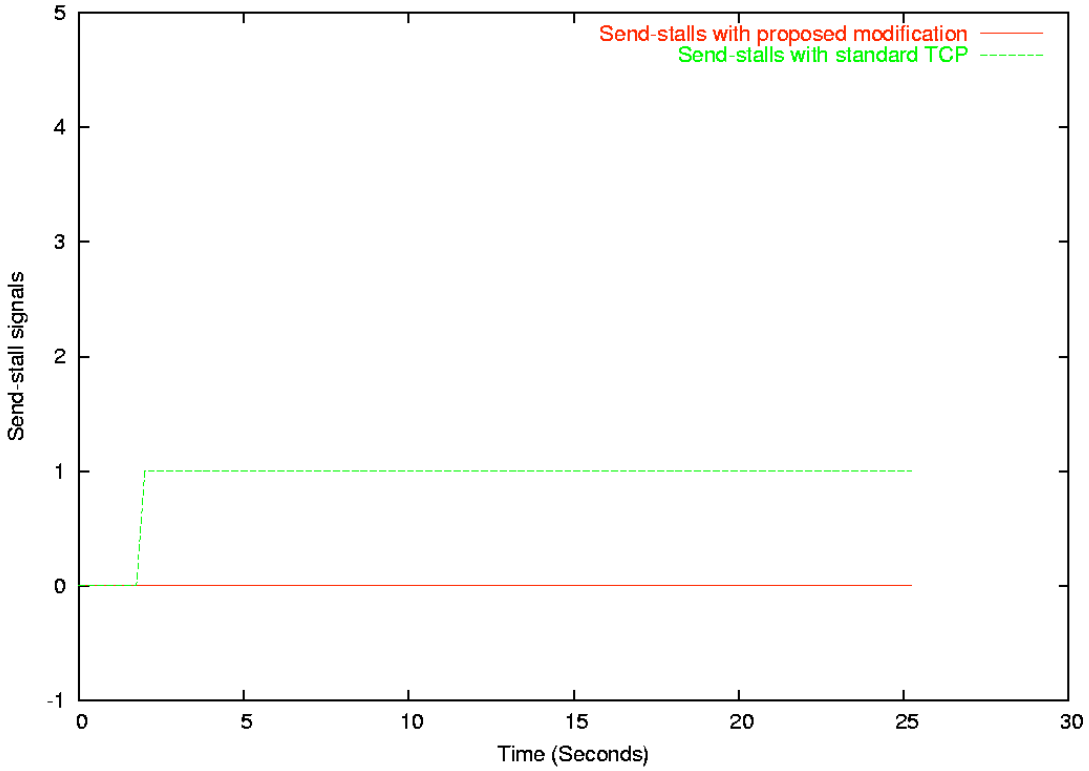

Figure 1: Comparison of send-stall signals in the standard Linux TCP and the modified TCP

References

[1] J. Postel, "Transmission Control Protocol," RFC-793, September 1981

[2] V. Jacobson, R. Braden, and D. Borman. RFC 1323: TCP Extensions for High Performance, 1992.

[3] S. Floyd and T. Henderson. RFC 2582: The NewReno Modification to TCP's Fast Recovery Algorithm, 1999.

[4] M. Allman, V. Paxson, W. Stevens, "TCP Congestion Control," RFC-2581, April 1999

[5] Dina Katabi, Mark Handley, and Charles Rohrs, "Internet Congestion Control for High Bandwidth-Delay Product Networks," ACM Sigcomm 2002, Pittsburgh, August, 2002

[6] Sally Floyd, “HighSpeed TCP for Large Congestion Windows”, Internet-draft draft-floyd-tcp-highspeed-02.txt, Work in progress, February 2003.
[7] Tom Kelly, “Scalable TCP: Improving Performance in HighSpeed Wide Area Networks,” First International Workshop on Protocols for Fast Long Distance Networks, Geneva, February 2003.
[8] Lawrence S. Brakmo and Larry L. Peterson, “TCP Vegas: end-to-end congestion avoidance on a global Internet,” *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1465–80, October 1995.
[9] E. Weigle and W. Feng, “A case for TCP Vegas in high-performance computational grids,” in *Proceedings of the 9th International Symposium on High Performance Distributed Computing (HPDC’01)*, August 2001.
[10] C. Casetti, M. Gerla, S. Mascolo, M. Sansadidi, and R. Wang, “TCP Westwood: end-to-end congestion control for wired/wireless networks,” *Wireless Networks Journal*, vol. 8, pp. 467–479, 2002.
[11] R. Wang, M. Valla, M. Sanadidi, B. Ng, and M. Gerla, “Using adaptive rate estimation to provide enhanced and robust transport over heterogeneous networks,” in *Proc. of IEEE ICNP*, 2002.
[12] Dina Katabi, Mark Handley and Charlie Rohrs, “Congestion Control for High Bandwidth-Delay Product Networks“, Proceedings on ACM Sigcomm 2002.
[13] Shudong Jin, Liang Guo, Ibrahim Matta, and Azer Bestavros, “A spectrum of TCP-friendly window-based congestion control algorithms,” *IEEE/ACM Transactions on Networking*, vol. 11, no. 3, June 2003.
[14] R. Shorten, D. Leith, J. Foy, and R. Kilduff, “Analysis and design of congestion control in synchronised communication networks,” in *Proc. of 12th Yale Workshop on Adaptive and Learning Systems*, May 2003.
[15] L. Xu, K. Harfoush, and I. Rhee, “Binary increase congestion control for fast long-distance networks,” in *Proc. of IEEE Infocom*, 2004.
[16] A. Kuzmanovic and E. Knightly, “TCP-LP: a distributed algorithm for low priority data transfer,” in *Proc. of IEEE Infocom*, 2003.
[17] William E. Allcock, John Bresnahan, Julian J. Bunn, S. Hegde, Joseph A. Insley, Rajkumar Kettimuthu, Harvey B. Newman, S. Ravot, T. Rimovsky, Conrad Steenberg, L. Winkler, “Grid-enabled particle physics event analysis: experiences using a 10 Gb, high-latency network for a high-energy physics application,” Future Generation Comp. Syst. 19(6): 983-997 (2003).
[18] C. Hollot, V. Misra, D. Towsley, and W. Gong, “On designing improved controllers for AQM routers supporting TCP flows,” In Proceedings of IEEE INFOCOM, Apr. 2001.
[19] S. Kunniyur and R. Srikant, “Analysis and design of an adaptive virtual queue,” In Proceedings of ACM SIGCOMM, 2001.
[20] S. H. Low, F. Paganini, J. Wang, S. Adlakha and J.C. Doyle, “Dynamics of TCP/AQM and a scalable control,” In Proceedings of IEEE INFOCOM, June 2002.
[21] Gerry, J.P, "A Comparison of PID Control Algorithms,” Control Engineering, pp 102-105, March 1987.
[22] Ziegler J.G. and Nichols N.B, “Optimum settings for automatic controllers,” Trans. ASME, pp. 759-768, 1942.
[23] http://www.web100.org/