Final Scientific Report

DOE Award No. DE-FG02-02ER25528
PI Vishal Misra
Columbia University

1 Executive Summary

This project studied transport protocols in the context of scientific applications. There were two main objectives of the project

- Study the viability and reliability of the Transport Control Protocol (TCP), widely used on the Internet, for delay sensitive scientific applications.
- Design high speed versions of TCP, appropriate for high data volume scientific applications.

Substantial progress was made on both fronts. We demonstrated that contrary to the widely held belief at the time, TCP did not have “chaotic” or self-similar behavior. We investigated the effect of different protocol parameters on the behavior of TCP and how it impacted the perceived long range dependence. More details are provided later in the document. In terms of designing protocols suitable for high data volume applications, we proposed a modification to TCP that enabled it to scale to high bandwidth, high delay links. The project also proposed to implement the high speed protocols on a DOE testbed. Initial discussions of plans occurred with Nageswara Rao at Oak Ridge National Laboratories, however the project funding was not renewed before that could be successfully accomplished. In terms

2 Comparison of actual accomplishments with the goals and objectives of the project

We had proposed to study two kinds of problems in the project. Both problems were extensively studied, analyzed and code was released for simulators. In terms of theoretical goals, they were met and in some cases exceeded beyond the stated ones, as detailed in the following sections. However, one part of the project was deployment of running code on production networks, and that could not be achieved. The original grant was for 3 years, however funding was not renewed for the third year by the time the PI had a graduate student well trained in system building aspects of the project.

3 Project Activities

In the following sections, we detail the technical activities undertaken as part of the project. Two of the activities, primarily done at Columbia University (with some collaborators elsewhere) successfully accomplished the goals of the project. The third part, regrettably, remained unfinished.
3.1 High Speed TCP

The PI was invited to attend the First International Workshop on Protocols for Fast Long-Distance Networks (PFLDNET 2003), held at CERN, Geneva in February, 2003. We presented two papers, directly related to work proposed under the grant. The publications were

1. Abhinav Kamra, Vishal Misra and Don Towsley


In the first paper, we proposed simple modifications to the TCP protocol that in conjunction with AQM and ECN mechanisms enables flows to better use high bandwidth pipes over long round trip delays. Our work made more effective use of feedback and we implemented our scheme in the ns-2 simulator, and released the code. Simulation results indicated that our protocol performed much better than other proposed schemes.

In the second paper we studied the behavior of some high speed TCP variants and their relationship to the notion of “TCP fairness”, i.e., how closely their throughput tracks TCP throughput formulas. Our study indicated that there is a significant stochastic bias in protocols based on increased-decrease mechanisms and in particular High Speed TCP (proposed by Floyd et. al.) is unfair to regular TCP for almost all loss probability levels.

We identified a crucial drawback in other schemes that have been proposed recently for high throughput applications, namely HSTCP and ScalableTCP. Both protocols change the behavior of TCP when the window size is high, adjusting the response function to be more aggressive with respect to loss. Although the steady state behavior of the protocols yields a unique operating point at a fixed loss rate, our experiments indicate that the protocols are susceptible to a bi-stable behavior. If a flow starts early and the window size reaches beyond the threshold where the response function becomes more aggressive, then the flow can “lock out” later arriving flows from gaining a fair share of the bandwidth. This creates a temporal unfairness where the late arriving flows are prevented from reaching a window size high enough to change into the more aggressive mode.

3.2 Delay sensitive applications and TCP

In another piece of work proposed in the project, studied self-similarity and chaos in network traffic.


In this work we analyzed the traces of an Infocom 2000 paper “The Chaotic Nature of TCP Traffic” and concluded that erroneous analysis had led the authors to the conclusion of chaos in TCP traffic. The traces themselves did not present evidence of chaos or long range dependence. We further explained the behavior of TCP traffic by a detailed Markovian model we derived for TCP.
We also analyzed the question whether Self-Similarity really existed on the Internet, regardless of TCP contributing to it. Our work re-examined the evidence, causes and implications of Self-Similarity, questioning many of the widely held beliefs at the time. Our work appeared in the following publication:


After the grant funding ended, we continued work on analyzing the delay sensitivity of TCP. We constructed a detailed Markovian model of TCP and analyzed its behavior with respect to latency and loss. All prior studies of TCP had focused primarily on throughput performance. The following publication resulted from this work:

- Eli Brosh, Salman A. Baset, Dan Rubenstein and Henning Schulzrinne, **The Delay-Friendliness of TCP**, *SIGMETRICS ’08*, Annapolis, Maryland, USA, June, 2008.

Note that the PI is not listed as an author in this publication. The paper was published at the Sigmetrics 2008 conference, and the PI was the PC co-chair of the conference, and hence took his name out as an author to avoid any appearance of impropriety. In the journal version of the paper that has been accepted in the IEEE/ACM Transactions on Networking, the PI is the primary faculty author.

### 3.3 Collaboration with a DOE lab: Differential service over dedicated links

In addition, the PI initiated a collaborative effort with a project run by Nageswara Rao at the Oak Ridge National Laboratories. Due to Visa restrictions, the PI was not able to pay a site visit to ORNL, however work was initiated on a particular problem that Nageswara Rao was investigating. The problem is guaranteeing throughput in a dedicated link to a small number of flows. The applications include high bandwidth data transfers, as well as control of processes over wide-area networks, for tasks that include computational steering, remote interactive visualization and instrument controls. Each of these flows have different bandwidth requirements, and it is difficult for “vanilla” TCP to share bandwidth in the link according to different demands. To initiate work on this project, we first replicated the setup at ORNL in our lab at Columbia as closely as possible. We instrumented Linux based PCs with gigabit NICs, and connected them to a CISCO catalyst switch that has gigabit interfaces. We implemented a number of algorithms, including TCP-LP and an algorithm described next to achieve the desired goal of sharing of bandwidth over a dedicated link.

One of the problems with TCP is that the performance is completely determined by the round trip times and loss rates experienced by flows in their paths\(^1\). Thus, if flows share a common path, TCP divides throughput equally amongst the flows. To achieve an unequal distribution, as dictated by application demands, flows have to be prioritized. This can be achieved by the DiffServ architecture, in which differential service can be provided to flows in an application transparent manner, which is a very important operational requirement. We have extensively studied this problem, and have derived a number of important theoretical results related to it. We have shown under what conditions it is possible to guarantee minimum rates to different flows over a single shared link, and have developed simple algorithms that achieve the same. We performed extensive validation of the scheme via simulations, and a paper describing the results was accepted in the IEEE/ACM Transactions of Networking.

\(^1\)the Advertised window can also affect (restrict) throughput

4 Final year of the grant

We had proposed a number of activities for the final year of the grant.

• Continue with study of high speed transport protocols. Compliment simulation study with theoretical analysis of unfairness in high speed protocols

• Continue collaboration with Nageswara Rao on the bandwidth sharing problem. Implement and test algorithms on the experimental testbed at Columbia University.

• Release code for the DiffServ algorithm as well as high speed protocols developed at Columbia University.

• Implement protocols on the DOE UltraScience Net experimental testbed.

However, due to many reasons, including the PIs status at that time as a “temporary worker” with an H1-B visa, formal relationship with ORNL could not be developed. Although both Nageswara Rao and the PI had finalized arrangements for site visits, the program manager decided that the collaborative aspects with a DOE lab of this project was not satisfactory at this point. The program manager, in July of 2004 decided not to renew the funding for the grant for the third year. All the work that was accomplished was done in the first two years of the original grant.

5 Products developed under the award

The primary products developed under the Award have been the publications that have been listed in the preceding sections. Some code for the ns-2 simulator was also released and was used by other researchers studying high speed TCP protocols. We also identified a bug in the ns-2 simulator that showed spurious bi-stable behavior and that fix was released to the simulation community as well.