Evaluation of Rate Based Transport Protocols for Lambda-Grids[1]

Ryan X. Wu and Andrew A. Chien
Department of Computer Science and Engineering
University of California, San Diego
{xwu, achien}@ucsd.edu

Abstract *- The notion of lambda-Grids posits collections of plentiful computing and storage resources richly interconnected by dedicated dense wavelength division multiplexing (DWDM) optical paths. Many researchers have proposed new transport protocols to address the opportunities and challenges of this radically different network. We evaluate several promising rate-based transport protocols [1-3], using a range of performance metrics. In lambda-Grids, the DWDM links form a network with plentiful bandwidth, pushing contention and sharing bottlenecks to the end systems (or their network links), inspiring our new Group Transport Protocol (GTP). Our studies for single flows show that several rate-based protocols achieve dramatically better throughput than TCP. As multiple parallel flows are introduced, RBUDP and SABUL incur significant packet loss rates. For converging flows, packet loss rates as high as 50% are observed. Our new group transport protocol (GTP) which focuses on managing end system contention achieves both high throughput and much lower loss rates.*

I. INTRODUCTION

Geometric increases in semiconductor chip capacity predicted by Moore's Law have produced a revolution in computing over the past 40 years. Even more rapid advances in optical networking are producing increases in bandwidth which are even greater. The OptIPuter Project [4] and other efforts such as CANARIE [5] are exploring the new "lambda-grid" environments (low-cost, plentiful wide area bandwidth, plentiful storage and computing), that this revolution enables.

Circuit-switched lambda's can provide transparent end-to-end optical light paths – available at low-cost and delivering huge dedicated bandwidth. Networks of such connections form a lambda-grid (sometimes called a distributed virtual computer) in which the geographically distributed elements can be tightly-coupled. Compared to shared, packet-switched IP networks, the key distinguishing characteristics of these lambda-grids are:

- No internal network congestion and significant end system congestion;
- Small numbers of endpoints (e.g. 10^3, not 10^8);
- Very high speed links (1Gig, 10Gig, etc.) with as much as a terabit across multiple lambdas;
- Coordinated communication across a number of endpoints in a group (e.g. fetching different quantities of data from ten distinct servers to feed a single local computation) .

Delivering communication performance in high bandwidth-delay product networks is a major research challenge for just point to point [6-8, 2]. Here we consider the challenge of achieving high performance more complex configurations of network sharing and communication pattern.

Traditional TCP and its variants (e.g. [9], [10]) were developed for shared networks on which the bandwidth on internal links is a critical and limited resource. As such, the congestion control techniques manage internal network contention, providing a reasonable balance of non-aggressive competition and end-to-end performance. As a result, slow start causes TCP to take a long time to reach full bandwidth when RTT is large and a long time to recover from packet loss because of its AIMD control law. To provide good transport performance for networks with high bandwidth-delay product optical network links, researchers have proposed many TCP variations [7, 6, 11]. Recently, the lambda-grid community has proposed a range of rate-based reliable transport protocols (e.g. [1-3, 12]) based on UDP. They explicitly measure packet loss, and adjust transmission rates in response. We evaluate three of these protocols: RBUDP [2], SABUL [1], and GTP [3], using a point to point single flow, parallel point to point flows, and finally multipoint convergent flows. Our studies provide the following results:

- For a single point-to-point flow, all three rate-based protocols (RBUDP, SABUL, GTP) achieve dramatically higher performance than TCP, while achieving low loss rates.
- For the parallel point-to-point flows, all three rate-based protocols (RBUDP, SABUL, GTP) achieve high bandwidth, but RBUDP aggressiveness causes much higher packet loss rates (20x) than SABUL or GTP.
- For convergent flows, again all three rate-based protocols (RBUDP, SABUL, GTP) achieve high bandwidth, but RBUDP and SABUL flows with various RTT do not converge to a stable rate, and have loss rates 1000x and 100x higher than GTP.

In summary, rate-based protocols can deliver high bandwidth in high bandwidth-delay product networks, but because of their aggressiveness, are subject to high rates of packet loss when multiple flows exist. While only an initial evaluation, our results show that our group transport protocol (GTP) is promising, achieving both high delivered bandwidth and low loss rates for two simple sharing scenarios. We plan continued development and more rigorous evaluation of GTP as a new transport protocol for lambda grids in future work.

II. THE PROTOCOLS

Reliable Blast UDP (RBUDP) [2] targets fast, reliable data transfer on dedicated or QoS-enabled high speed links. It

[1] This work was supported by the National Science Foundation under Cooperative Agreement ANI-0225642 to the University of California, San Diego for "The OptIPuter".

assumes users have explicit knowledge about the link capacity, requiring the sender to specify an initial rate, start, and maintain its transfer at that rate. SABUL (Simple available bandwidth utilization library) [1] is designed for data-intensive applications in high bandwidth-delay product networks. SABUL starts senders at a fixed high initial rate, adjusting based on experienced loss and retransmission. The newest version of SABUL (UDT) [8] uses delay-based rate adaptation to reduce packet loss caused by its aggressiveness. GTP (Group Transfer Protocol) [3] is a receiver–driven transport protocol, exploiting information across multiple flows to manage receiver contention and fairness. We summarize the characteristics of these rate-based protocols in Table 1.

	RBUDP	SABUL	GTP
Initial Rate	Specified by User	A fixed rate. (slow start in UDT)	Negotiated by sender and receiver
Multi-point to Point	No.	No.	Yes.
Rate Adaptation	Optional and Limited	Rate based with delay compensation	Rate estimation and delay compensation
Fairness among flows	Not Considered	To some extent	Max-min fairness among flows at receiver side

TABLE 1: COMPARISON OF RATE BASED PROTOCOLS.

III. EVALUATION

A. Methodology

We compare RBUDP, SABUL, GTP, and standard un-tuned TCP. Throughout our experiments, we use the latest available version (RBUDP v0.2, SABUL/UDT 1.0, and GTP prototype). Our experimental test bed is the TeraGrid [13] at SDSC (San Diego Supercomputer Center), and NCSA/UIUC (National Center for Supercomputing Applications), and ANL (Argonne National Lab). The achievable bandwidth between SDSC and NCSA on each connection is 1Gbps (NIC speed limit).

In the full paper, we will report results from a broader range of network structures, explored by simulation with DummyNet [14] and other TeraGrid sites as listed below.

B. Performance Metrics

1. *Instantaneous transmission and loss rate,*
2. *Sustained throughput on a 100GB transfer (Point to point and multi-point to point),*
3. *Total throughput in the first 50 RTT, and*
4. *Fairness for multi-point to point transmission.*

C. Scenarios

1. *Point to point with multiple connections*
2. *Multi-point convergent transmission with varied delay and bandwidth for each link*
3. *Flow adaptation to dynamic changes:* Including the dynamics when flows join or leave, and changing workload on end systems;
4. *Transmission from fast (slow) to slow (fast) machines.*

D. Preliminary Results

Scenario 1: Point-to-point: Transfer 10GB data from SDSC to NCSA (1Gbps link with 58ms RTT).

	TCP	RBUDP	SABUL	GTP
Time (s)	1639	9.08	8.90	8.92
Avg. Rate	4.88Mbps	881Mbps	898 Mbps	896Mbps
Loss Rate	unknown[2]	0.07%	0.01%	0.02%

TABLE 2: POINT TO POINT, FROM SDSC TO NCSA

Scenario 2: Point-to-point, parallel flows: Transfer 10GB data from SDSC to NCSA on the same 1Gbps link with three parallel connections.

	TCP	RBUDP[3]	SABUL	GTP
Aggregate Rate (Mbps)	14.5	931	912	904
Avg. Loss	unknown	2.1%	0.1%	0.03%
System stability	Yes	Yes	Yes	Yes
Fairness	Fair	Fair	Fair	Fair

TABLE 3: PARALLEL FLOWS: SDSC TO NCSA

Scenario 3: Multi point, convergent flows: Transfer 10GB data; one receiver at SDSC, and three senders with two from NCSA, one from SDSC. Each sender-receiver connection is with 1Gbps dedicated link.

	TCP	RBUDP	SABUL	GTP
Aggregate Rate[4] (Mbps)	677	443	811	865
Avg. Loss	unknown	53.3%	8.7%	0.06%
System stability	Yes	No	No	Yes
Fairness	No	No	No	Yes

TABLE 4: MULTI-POINT, CONVERGENT FLOWS

Our results show: For single, point-to-point high bandwidth delay product links, three rate based protocols achieve much higher throughput than traditional TCP while maintaining low loss rate. And all three protocols perform well when there are parallel flows between sender and receiver, but RBUDP has a much higher loss rate. When multiple senders are connected with receiver with different RTT, all three rate based protocols achieve high bandwidth, but loss rates vary over a range of 1000x, with GTP having lowest loss rate by a large margin.

A large number of challenges remain, including how to manage end contention at receiver and how to achieve a stable fair rate allocation in equilibrium among flows. In the full paper, by utilizing DummyNet, we will extend our evaluation to address some of these issues.

The major challenges remaining for rate-based protocols include rate convergence, rate allocation, fairness among flows with different RTT, auto-scaling, fast rate adaptation to dynamic flow changes (when new flow join or some flows complete), and TCP-friendliness.

[2] We are not able to measure instant TCP loss rate, due to the lack of root privilege on TeraGrid.

[3] We assume each flow has no knowledge about others, and starts with the rate of full bandwidth.

[4] Aggregate rate and loss rate vary for RBUDP and SABUL, and numbers listed are the average values of several measurement.

V. REFERENCES

1. H. Sivakumar, R. Grossman, M. Mazzucco, Y. Pan, and Q. Zhang, *Simple Available Bandwidth Utilization Library for High-Speed Wide Area Networks.* to appear in Journal of Supercomputing, 2003.
2. E. He, J. Leigh, O. Yu, and T. DeFanti, *Reliable Blast UDP: Predictable High Performance Bulk Data Transfer.* IEEE Cluster Computing, 2002: p. 317.
3. X. Wu and A. Chien, *GTP: A Group Transport Protocol.* To submit for publication, available online soon., 2003.
4. *The OptIPuter Project.* www.optiputer.net.
5. *CANARIE.* www.canarie.ca.
6. *FAST TCP.* http://netlab.caltech.edu/FAST/.
7. S. Floyd, *HighSpeed TCP for Large Congestion Windows.* Internet draft.
8. Y. Gu, X. Hong, M. Mazzucco, and R.L. Grossman, *SABUL: A High Performance Data Transfer Protocol.* Submitted for publication.
9. L. Brakmo and L. Peterson, *TCP Vegas: End to End Congestion Avoidance on a Global Internet.* IEEE Journal of Selected Areas in Communications,. **13**(8): p. 1465-1480.
10. M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow., *TCP Selective Acknowledgement Options.* RFC2018, Internet Engineering Task Force (IETF), October 1996.
11. D. Katabi, M. Handley, and C. Rohrs. *Congestion Control for High Bandwidth-Delay Product Networks*. in *Proceedings on ACM Sigcomm*. 2002.
12. *Tsunami.* http://www.indiana.edu/~anml/anmlresearch.html.
13. *Teragrid.* www.teragrid.org.
14. *DummyNet.* http://info.iet.unipi.it/~luigi/ip_dummynet/.