Wide-Area experiments with LambdaStream over dedicated high-bandwidth networks

Venkatram Vishwanath, Jason Leigh, Eric He, Maxine D. Brown,
Lance Long, Luc Renambot, Alan Verlo, Xi Wang, Thomas A. DeFanti
Electronic Visualization Laboratory (EVL),
Department of Computer Science, University of Illinois at Chicago.
Chicago, U.S.A.
venkat@evl.uic.edu

Abstract— **LambdaGrid applications can generate hundreds to thousands of parallel flows. These flows emanate from network interfaces in the end-systems (i.e. the compute clusters) to communicate with other end-systems over multiple lightpaths. LambdaStream is an application-level UDP based transport protocol for LambdaGrids. We present wide-area results with LambdaStream over dedicated 10Gbps networks using the TeraGrid and CaveWave. We present results wherein we achieve close to 10Gbps unidirectional and over 18Gbps bidirectional over these high-speed networks.**

Keywords: High-performance Transport Protocols, Optical Networks, Wide-Area Networks.

I. Introduction

Interactive exploration of multi-terabyte datasets has been identified as a critical enabler for scientists to glean new insights in a variety of disciplines, such as biomedical imaging, geoscience and high-energy physics [1] [2] [3]. Practically, these large-scale datasets must flow among a Grid of instruments, physical storage devices, visualization displays, and computational clusters. These applications have a real and insatiable need for tens to hundreds of gigabits-per-second of bandwidth that are best satisfied by interconnecting Grid resources with dedicated networks dynamically created by concatenating optical lightpaths (lambdas). This is called a LambdaGrid.

LambdaGrid applications can generate hundreds to thousands of parallel flows. These flows emanate from network interfaces in the end-systems (i.e. the compute clusters) to communicate with other end-systems over multiple lightpaths. More complex flows include multiple parallel end-systems, inter-communicating over known, but arbitrary, physical network topologies. Efficient management of the myriad parallel data flows among and within the end-systems; and, treating these parallel communication channels as a single problem rather than as entirely uncoordinated flows is extremely critical for LambdaGrid applications. Additionally, as the exponential growth of bandwidth now far exceeds storage and computing, a significant impedance mismatch exists between these high-capacity lambda-based networks and the end-systems that must absorb the bandwidth, resulting in inadequately performing applications. At the recent LCA06 (Linux Conference Australia 2006) conference, keynote speaker Van Jacobson resonated a similar sentiment “The end of the wire isn’t the end of the net,” – that is, the future challenges of high-performance networking reside at the edges [4].

We are currently designing LambdaStream[5], an application-level UDP based, network and end system aware data transport protocol, to address the needs LambdaGrid applications. We present performance results with LambdaStream [5], over high bandwidth dedicated networks between Chicago and San Diego over 10Gbps links.

II. Background

LambdaStream is an application-level UDP based transport protocol, for LambdaGrids. LambdaStream builds on our prior work done in Reliable Blast UPD(RBUDP) [6], a UDP-based transport protocol for data transfers over dedicated networks. The goals of LambdaStream include support for high bandwidth streaming for applications in LambdaGrids. Its key characteristics include a combination loss recovery and a unique rate control. It is also configurable by a user to support reliable and unreliable delivery of data. LambdaStream is being designed to support multipoint-to-multipoint communication that is characteristic of LambdaGrids applications. The current implementation supports point-to-point streaming and multipoint support is currently being implemented.

III. Experiments and Results

The typical usage patterns of application scientists in LambdaGrids, to facilitate scientific insight and discovery, can be categorized as flows.

1. Reliable disk-to-disk transfers. This is typically used to transfer data to a central storage repository or copy data to a local storage repository for analysis.
2. Reliable disk-to-memory transfers. This is typically used to analyze remote datasets without copying them over.
3. Reliable memory-to-memory transfer. Scientists typically would use this model to transfer computed data in memory to remote visualization cluster to display the results or the next stage in the computation pipeline.

Unreliable versions can be considered to be subsets of the above.

We performed three different LambdaStream tests over the TeraGrid[8] and CAVEwave networks between EVL, Chicago and CalIT2, UCSD, San-Diego: moving data reliably from memory to memory, from disk to memory, and from disk to disk; the three ways scientists typically use networks to access data. The aim of the experiments was to measure the performance of LambdaStream over wide-area networks and the performance of the Teragrid and the CAVEWave Networks.

The wide-area cluster-to-cluster experiments were conducted between a 30-node OptIPuter cluster at EVL, Chicago and a 28-node OptIPuter cluster at Calit2, San Diego. The two clusters were connected using two distinct networks: namely the CaveWave and the Teragrid. The Teragrid Network used for the experiments was a 10Gb SONET routed network using MPLS between Chicago and San Diego. The Force10 at Starlight, Chicago connects to the TeraGrid Juniper T640, also located at StarLight; its circuit connects Chicago to Los Angeles to the UCSD San Diego Supercomputer Center where, for this test, it was directly connected to the OptIPuter cluster at Calit2. The rtt between the clusters at Chicago and San Diego over the Teragrid was 56.3ms. The Force10 at StarLight also connects directly to CAVEwave, which is a 10Gb switched LAN PHY wave on the National Lambda Rail (NLR); the circuit connects Chicago to San Diego through Seattle, where it connects to a 28-node OptIPuter cluster at the Calit2 building on the University of California, San Diego (UCSD) campus. The rtt between the cluster at Chicago and San Diego over the CaveWave is 78.3 ms.

The 30-node dual opteron OptIPuter cluster at EVL, with 1Gb NICs per node, connects to an in-house switch that aggregates traffic and sends it to StarLight over a 10Gb link.

The 28-node OptIPuter cluster at CalIT2 is also a dual opteron cluster with 1Gb NICs per node. The dataset used was a 0.3-meter high-resolution map of 5,000 square miles of the city of Chicago provided by the U.S. Geological Survey (USGS) National Center for Earth Resources Observation and Science (EROS). The map consists of 3,000 files of tiled images that are 75MB each, for a total of 220GB of information.

MRTG (Multi Router Traffic Grapher) was used to monitor traffic through StarLight's Force10 switch, and a similar MRTG-like traffic utilization tool called Cricket was used to monitor traffic through the TeraGrid's Juniper T640 router, also located at StarLight. The MRTG-measured bandwidth speeds were verified by LambdaStream, which is instrumented to monitor the amount of information it sends and receives. Moreover, MRTG measurements are five-minute averages, but LambdaStream measurements are continuous, and therefore more accurate.

The Results of the experiments summarized in Table I show that we were able to achieve 1:1 Memory to Memory Bisection bandwidth, 1.6:1 Disk to memory bandwidth and 2:1 Disk to Disk bandwidth over the 10Gbps Wide-Area Networks. Additionally, CAVEwave achieved 18.19Gbps and TeraGrid achieved 18.06Gbps doing bi-directional reliable memory-to-memory transfers. The optimizations performed at the protocol level are beyond the scope of this paper.

Table I: LambdaStream Throughput Over 10Gbps Links

	Iperf: Unreliable UDP (Gbps) 10 streams per direction[1]	LambdaStream- Reliable Memory-to-Memory (Gbps) 10 streams per direction[2]	LambdaStream- Reliable Disk-to-Memory (Gbps) 16 streams per direction[3]	LambdaStream- Reliable Disk-to-Disk (Gbps) 20 streams per direction[3]
CAVEwave Chicago to San Diego	9.73	9.19	9.08	9.30
CAVEwave San Diego to Chicago	9.75	9.23	9.21	9.01
TeraGrid Chicago to San Diego	9.43	9.14	9.03	9.22
TeraGrid San Diego to Chicago	9.45	9.16	9.15	9.02

1: Using iperf, CAVEwave moved 10 UDP streams with 0% packet loss and effective throughputs, as shown.Using iperf, TeraGrid moved 10 UDP streams with 1% packet loss and effective throughputs, as shown.

2: Bidirectionally, CAVEwave achieved 18.19Gbps and TeraWave achieved 18.06Gbps doing memory-to-memory transfers.

3: Disk bandwidth is 500Mb maximum, so more nodes were used to saturate the 10Gb link.

IV. Discussion and Future Work

The results of this experiment, summarized in Table I, prove that data-intensive science can effectively use hybrid networks (routed and switched) to move large files effectively using LambdaStream. Routed networks are typically shared networks, as the cost of routers makes it prohibitive to permanently dedicate links among specific sites. Switched networks are less costly, so can be dedicated, and provide applications with known and deterministic characteristics, such as guaranteed bandwidth (for data movement), guaranteed latency (for visualization, collaboration and data analysis) and guaranteed scheduling (for remote instruments).

We are currently extending LambdaStream to a multi-point-to-multi-point protocol to efficiently manage the myriad parallel data flows among and within the end-systems; treating these parallel communication channels as a single problem rather than as entirely uncoordinated flows. This enables better utilization and a more predictable performance for multi-point-to-multi-point communication typically used for cluster-to-cluster and multi-cluster data transfer, streaming of visualization, video-streams, etc. We are also co-developing MAGNET [7], a Monitoring apparatus for General Kernel Event Tracing, to provide timely systemic feedback to LambdaStream to adapt to end-system conditions that are one of the main sources for packet losses in LambdaGrids. Our goal is to design an end system and Network aware synergistic protocol to enable next generation cluster to cluster high performance applications.

V. References

[1] http://wwwiepm.slac.stanford.edu/monitoring/bulk/sc2005/hiperf.html

[2] J. Leigh, L. Renambot et al, “The Global Lambda Visualization Facility: An International Ultra-High-Definition Wide-Area Visualization Collaboratory”, Journal of Future Generation Computer Systems (FGCS), in review.

[3] http://www.igrid2005.org

[4] V. Jacobson and B. Felderman, “A modest proposal to help speed up & scale up the linux networking stack.” Proceedings of Linux Conference Australia, (LCA 2006), Dunedin, New Zealand, Jan 23-28, 2006,

[5] C. Xiong, Leigh, J., He, E., Vishwanath, V., Murata, T., Renambot, L., DeFanti, T., “LambdaStream – a Data Transport Protocol for Streaming Network-intensive Applications over Photonic Networks,” Proceedings of The Third International Workshop on Protocols for Fast Long-Distance Networks, Lyon, France 02/02/2005 - 02/03/2005

[6] E. He, J. Leigh, O. Yu, T. DeFanti, “Reliable Blast UDP: Predictable High Performance Bulk Data Transfer,” IEEE Cluster Computing 2002, Chicago, IL, September 1, 2002.

[7] Mark K. Gardner, Wu-chun Feng, Michael Broxton, Adam Engelhart, Gus Hurwitz, “MAGNET: A Tool for Debugging, Analysis and Adaptation in Computing Systems,” 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2003), Tokyo, Japan, May 2003.

[8] www.teragrid.org

VI. Acknowledgements

We would like to thank TeraGrid for providing us access to their Networks. We are extremely grateful to Charlie Catlett, Linda Winkler, Larry Smarr, Phil Papadopoulos, Tom Hutton, Greg Hidley, Praveen K.J., John Ratcliffe, Qian Liu for all their support. Linda Winkler was instrumental in setting up the TeraGrid for the tests. The Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago specializes in the design and development of high-resolution visualization and virtual-reality display systems, collaboration software for use on multi-gigabit networks, and advanced networking infrastructure. These projects are made possible by major funding from the National Science Foundation (NSF), awards CNS-0115809, CNS-0224306, CNS-0420477, SCI-9980480, SCI-0229642, SCI-9730202, SCI-0123399, ANI-0129527 and EAR-0218918, as well as the NSF Information Technology Research (ITR) cooperative agreement (SCI-0225642) to the University of California San Diego (UCSD) for ”The OptIPuter” and the NSF Partnerships for Advanced Computational Infrastructure (PACI) cooperative agreement (SCI-9619019) to the National Computational Science Alliance. EVL also receives funding from the State of Illinois, General Motors Research, the Office of Naval Research on behalf of the Technology Research, Education, and Commercialization Center (TRECC), and Pacific Interface Inc. on behalf of NTT Optical Network Systems Laboratory in Japan.