Final Project Report

Chase Qishi Wu (PI)
The University of Memphis
chase.wu@memphis.edu

Abstract

A number of Department of Energy (DOE) science applications, involving exascale computing systems and large experimental facilities, are expected to generate large volumes of data, in the range of petabytes to exabytes, which will be transported over wide-area networks for the purpose of storage, visualization, and analysis. To support such capabilities, significant progress has been made in various components including the deployment of 100 Gbps networks with future 1 Tbps bandwidth, increases in end-host capabilities with multiple cores and buses, capacity improvements in large disk arrays, and deployment of parallel file systems such as Lustre and GPFS. High-performance source-to-sink data flows must be composed of these component systems, which requires significant optimizations of the storage-to-host data and execution paths to match the edge and long-haul network connections. In particular, end systems are currently supported by 10-40 Gbps Network Interface Cards (NIC) and 8-32 Gbps storage Host Channel Adapters (HCAs), which carry the individual flows that collectively must reach network speeds of 100 Gbps and higher. Indeed, such data flows must be synthesized using multicore, multibus hosts connected to high-performance storage systems on one side and to the network on the other side. Current experimental results show that the constituent flows must be optimally composed and preserved from storage systems, across the hosts and the networks with minimal interference. Furthermore, such a capability must be made available transparently to the science users without placing undue demands on them to account for the details of underlying systems and networks. And, this task is expected to become even more complex in the future due to the increasing sophistication of hosts, storage systems, and networks that constitute the high-performance flows.

The objectives of this proposal are to (1) develop and test the component technologies and their synthesis methods to achieve source-to-sink high-performance flows, and (2) develop tools that provide these capabilities through simple interfaces to users and applications. In terms of the former, we propose to develop (1) optimization methods that align and transition multiple storage flows to multiple network flows on multicore, multibus hosts; and (2) edge and long-haul network path realization and maintenance using advanced provisioning methods including OSCARS and OpenFlow. We also propose synthesis methods that combine these individual technologies to compose high-performance flows using a collection of constituent storage-network flows, and realize them across the storage and local network connections as well as long-haul connections. We propose to develop automated user tools that profile the hosts, storage systems, and network connections; compose the source-to-sink complex flows; and set up and maintain the needed network connections. These solutions will be tested using (1) 100 Gbps connection(s) between Oak Ridge National Laboratory (ORNL) and Argonne National Laboratory (ANL) with storage systems supported by Lustre and GPFS file systems with an asymmetric connection to University of Memphis (UM); (2) ORNL testbed with multicore and multibus hosts, switches with OpenFlow capabilities, and network emulators; and (3) 100 Gbps connections from ESnet and their Openflow testbed, and other experimental connections.

This proposal brings together the expertise and facilities of the two national laboratories, ORNL and ANL, and UM. It also represents a collaboration between DOE and the Department of Defense (DOD) projects at ORNL by sharing technical expertise and personnel costs, and leveraging the existing DOD Extreme Scale Systems Center (ESSC) facilities at ORNL.

1. Award #: DE-SC0010641, The University of Memphis

2. Project Title: Composition and Realization of Source-to-Sink High-Performance Flows: File Systems, Storage, Hosts, LAN and WAN

   PI: Chase Qishi Wu, in Collaboration with Oak Ridge National Laboratory and Argonne National Laboratory

3. Date of the Report: 12/01/2015
   Period Covered by the Report: 09/01/2013 – 09/02/2015

4. Project Accomplishments
   1) Overview
      This is a collaborative project with Oak Ridge National Laboratory (ORNL) and Argonne National Laboratory (ANL). The entire project team had the kickoff meeting at ORNL on November 5, 2013.

      The University of Memphis (UM) participants include the PI, Prof. Chase Qishi Wu, two Ph.D. students, Mr. Daqing Yun and Miss Poonam Dharam, and one undergraduate student, Mr. Mark Berry. We attended weekly teleconferences coordinated by ORNL since the beginning of the project.

      The main task of UM in this project is to design, develop, and test a Transport Profile Generator (TPG), which characterizes and enhances the end-to-end throughput performance of existing transport protocols in high-speed networks. TPG provides end users with a lightweight and easy-to-use toolkit for transport performance profiling and optimization to support big data transfer in data- and network-intensive scientific applications within DOE.

      In this project, we designed and implemented TPG, and conducted extensive tests on a local testbed at UM and on the wide-area testbeds at ORNL and ANL.

   2) Transport Profile Generator (TPG)
      TPG consists of a pair of sender and receiver: the sender (client or source node) transfers a certain amount of test data to the receiver (server or destination node) via a specific transport protocol to establish its corresponding performance profile by strategically varying the values of tunable system and protocol parameters. As shown in Fig. 1, TPG uses one TCP-based channel for profiling control and multiple protocol-specific channels for data transfer. The TPG control flow chart is provided in Fig. 2.

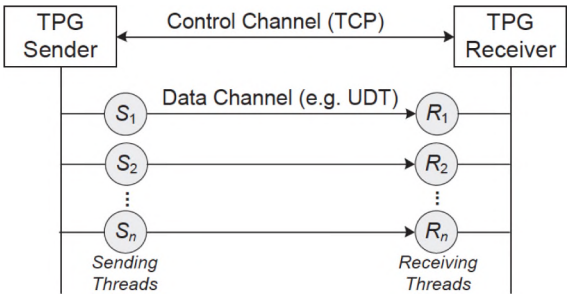

Fig. 1. TPG control and data channels.

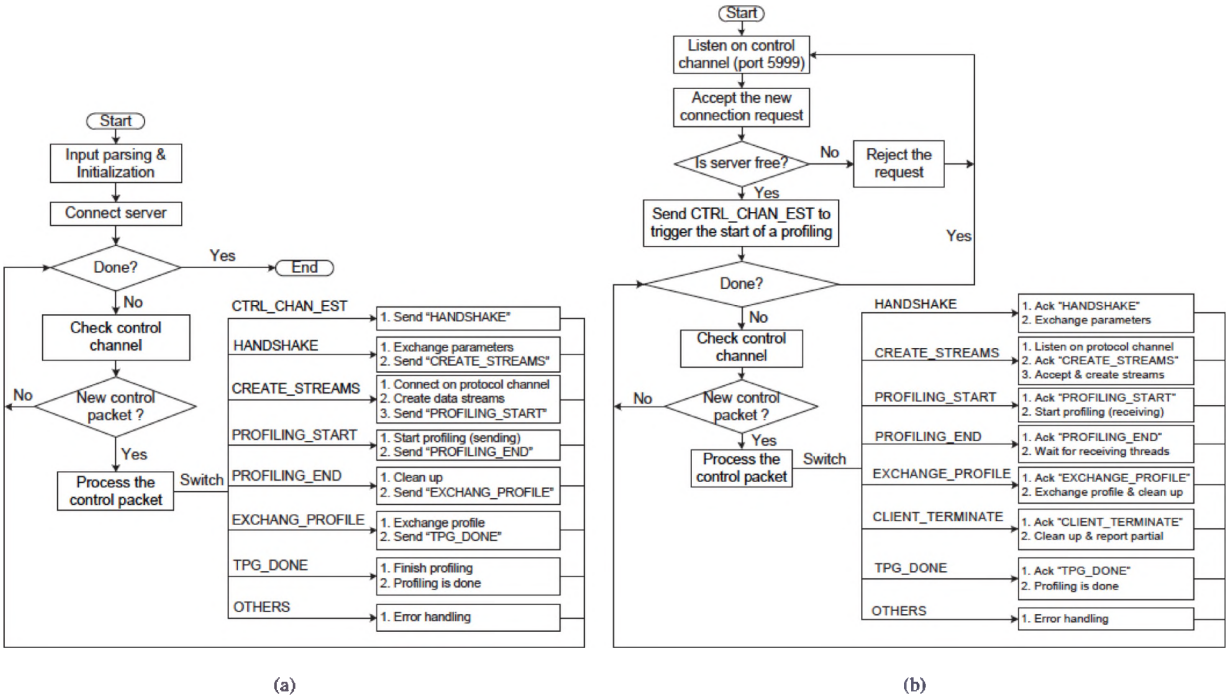

Fig. 2. TPG control flow chart: (a) client, (b) server.

In TPG, we added functionalities to support multiple data streams and multiple NIC-to-NIC connections. Also, we refined TPG with a flexible structure for an easy extension to new protocols, which are defined by their callback functions with a set of tunable control parameters. Therefore, to extend TPG with a new transport protocol, the user only needs to implement a protocol-specific callback function, and (optionally) add an option parameter to TPG for the protocol. For transport profiling, the user can specify the desired transport protocol either with a command-line option or in the profiling strategy function.

We conducted extensive experiments of TPG on a local testbed at UM and the wide-area testbeds at ORNL and ANL using UDT as an example. The experimental results collected over the past several months indicate that TPG-tuned UDT significantly outperforms the default UDT, TCP CUBIC, and Scalable TCP in complex settings with different loss rates and RTTs. In particular, we have made the following key observations, which are critical to the transport performance improvement:

- A jumbo frame generally improves the performance.
- A larger block generally leads to a better performance if there is sufficient buffer space.
- A larger receive buffer generally leads to a better performance, but a larger send buffer may not be always helpful, and hence it is necessary to decide an appropriate send buffer size to achieve a good performance.
- A sufficiently large UDP buffer is required to achieve a good performance.

The details of the TPG design, implementation, analysis, and profiling experiments are provided in several publications listed below.

3) Project-related Publications

- D. Yun, C.Q. Wu, N.S.V. Rao, B. Settlemyer, J. Lothian, R. Kettimuthu, and V. Vishwanath. Profiling Transport Performance for Big Data Transfer over Dedicated Channels. Submitted to *IEEE/ACM Transactions on Networking*, 2015.
- P. Dharam, C.Q. Wu, and N.S.V. Rao. Advance Bandwidth Scheduling in Software-Defined Networks. Submitted to *Globecom*, 2015.
- P. Dharam, L. Wang, and C.Q. Wu. Advance Bandwidth Scheduling for Maximizing Resource Utilization in High-performance Networks. Submitted to *Journal of Communications and Networks*, 2015.
- D. Yun, C.Q. Wu, N.S.V. Rao, B.W. Settlemyer, J. Lothian, R. Kettimuthu, and V. Vishwanath. Profiling Transport Performance for Big Data Transfer over Dedicated Channels. In *Proceedings of the IEEE International Conference on Computing, Networking and Communications, Optical and Grid Networking Symposium*, California, USA, February 16-19, 2015.
- P. Dharam, C.Q. Wu, and Y. Wang. Advance Bandwidth Reservation with Deadline Constraint in High-performance Networks. In *Proceedings of the International Conference on Computing, Networking and Communications* (ICNC), CNC Workshop, Honolulu, Hawaii, USA, February 3-6, 2014.