Final Project Report

Chase Qishi Wu (PI)
Department of Computer Science
New Jersey Institute of Technology
chase.wu@njit.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.

This proposal brings together the expertise and facilities of Oak Ridge National Laboratory (ORNL), Argonne National Laboratory (ANL), and New Jersey Institute of Technology (NJIT). 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-SC0015892, New Jersey Institute of Technology (NJIT)

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: 09/06/2017
   Period Covered by the Report: 09/02/2015 – 09/06/2017

4. Project Accomplishments
   1) Overview
      This project was transferred from University of Memphis in September 2015. 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.

      NJIT participants include the PI, Prof. Chase Qishi Wu, one Ph.D. student, Mr. Daqing Yun, and one undergraduate student, Mr. Mark Berry. We attended weekly teleconferences coordinated by ORNL since the beginning of the project. Mr. Daqing graduated in fall 2016 and joined the faculty at Harrisburg University.

      The main task of NJIT in this project is to design, develop, and test i) Transport Profile Generator (TPG), which characterizes and enhances the end-to-end throughput performance of existing transport protocols in high-speed networks; and ii) PRofiling Optimization Based DAta Transfer Advisor (ProbData). 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. ProbData is intended to help users determine the most effective data transfer method with the most appropriate control parameter values to achieve the best data transfer performance.

      In this project, we designed and implemented TPG and ProbData, and conducted extensive tests on a local testbed at NJIT 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.

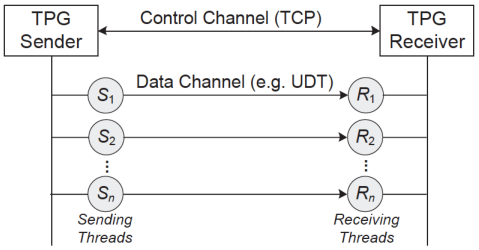


Fig. 1. TPG control and data channels.

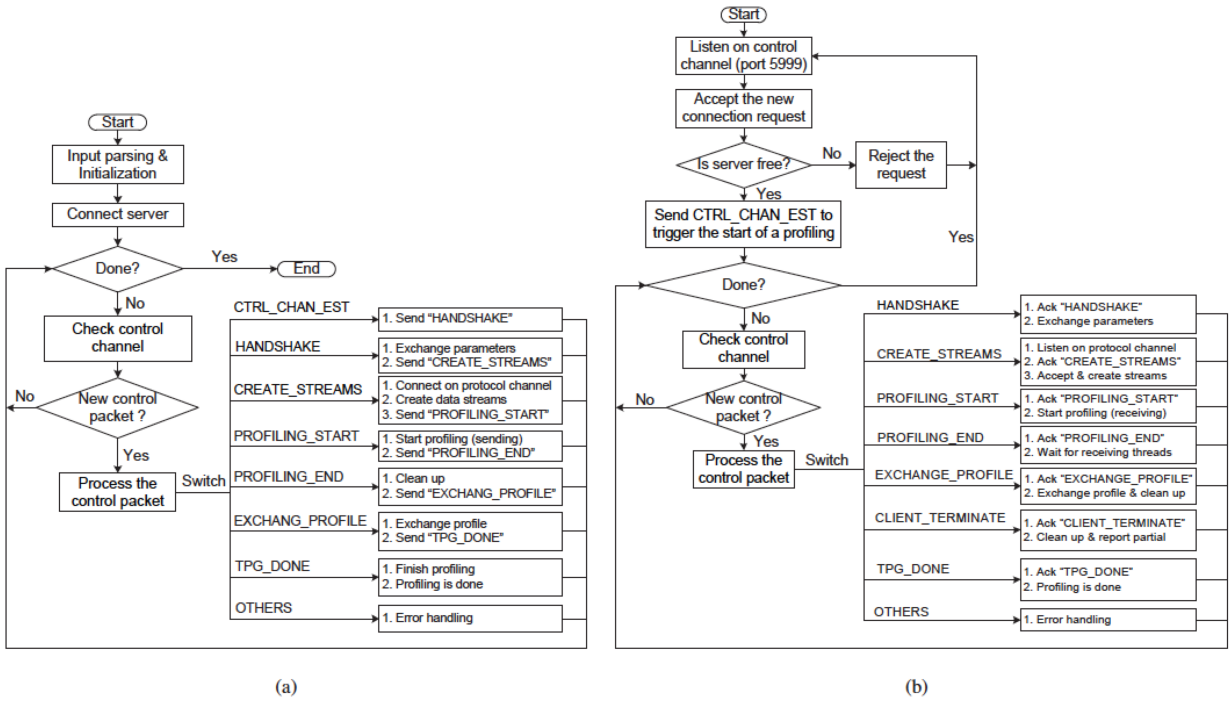


Fig. 2. TPG control flow chart: (a) client, (b) server.

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.

- 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 at the end of this report.

3) PRofiling Optimization Based DAta Transfer Advisor (ProbData)

ProbData is implemented with 19,000+ lines of C/C++ code in Linux to support memory-to-memory data transfer profiling for TCP and UDT. As shown in Fig. 3, ProbData uses iperf3 to perform profiling for TCP and uses TPG to perform profiling for UDT. Also, an SPSA-based stochastic optimization approach is employed to improve profiling speed.

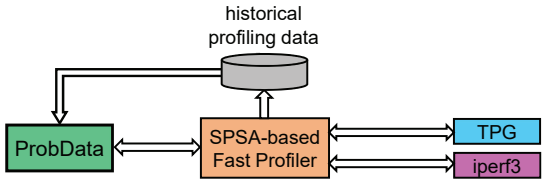
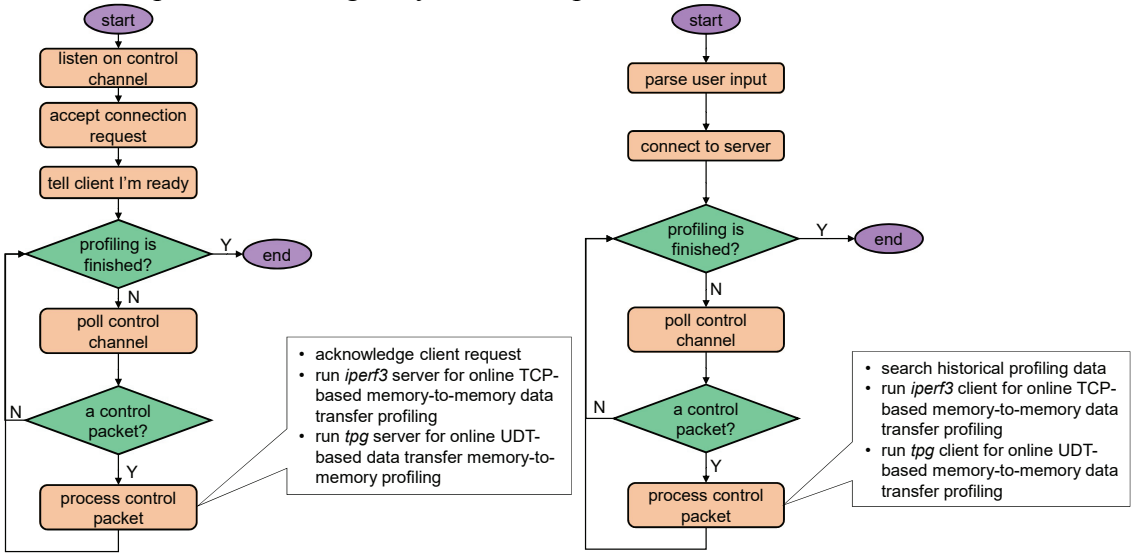


Fig. 3. Design of ProbData

ProbData consists of a pair of sender and receiver, which communicate with each other to exchange control parameters of ProbData and move the profiling process forward via a TCP-based control channel. The client and server are also responsible for running the clients and servers of TPG and iperf3 to conduct data transfer profiling. The main steps and control flow charts of the client and server of ProbData are shown in Figs. 4(a) and (b), respectively. The entire profiling process is mainly driven by the client, in which each step is acknowledged by the server prior to the actual execution.



(a) Client (b) Server

Fig. 4. Control flow charts of ProbData.

The details of the ProbData design, implementation, analysis, and advising experiments are provided in several publications listed at the end of this report.

4) Project-related Publications

- D. Yun, C.Q. Wu, N.S.V. Rao, Q. Liu, R. Kettimuthu, E.S. Jung. Data Transfer Advisor with Transport Profiling Optimization. In *Proceedings of the 42nd IEEE Conference on Local Computer Networks*, Singapore, October 9-12, 2017 (LCN17).
- N.S.V. Rao, Q. Liu, S. Sen, J. Hanley, I. Foster, R. Kettimuthu, C.Q. Wu, D. Yun, D. Towsley, and G. Vardoyan. Experiments and Analyses of Data Transfers Over Wide-Area Dedicated Connections. In *Proceedings of the 26th International Conference on Computer Communications and Networks*, Vancouver, Canada, July 31-August 3, 2017 (Invited paper, ICCCN17).
- C.Q. Wu. Bandwidth Scheduling in Overlay Networks with Linear Capacity Constraints. In *Proceedings of the IEEE International Conference on Computer Communications*, Atlanta, GA, USA, May 1-4, 2017 (INFOCOM17, acceptance rate: 20.93%).
- N.S.V. Rao, Q. Liu, S. Sen, G. Hinkel, N. Imam, I. Foster, R. Kettimuthu, B. Settlemyer, C.Q. Wu, and D. Yun. Experimental Analysis of File Transfer Rates Over Wide-Area Dedicated Connections. In *Proceedings of the 18th International Conferences on High Performance Computing and Communications*, Sydney, Australia, December 12-14, 2016 (HPCC16, Best Paper Award).
- Q. Liu, N.S.V. Rao, C.Q. Wu, D. Yun, R. Kettimuthu, and I.T. Foster. Measurement-Based Analysis of Performance Profiles and Dynamics of UDP Transport Protocols. In *Proceedings of the 24th IEEE International Conference on Network Protocols*, Singapore, November 8 – 11, 2016 (ICNP16).
- D. Yun, C.Q. Wu, N. S.V. Rao, Q. Liu, R. Kettimuthu, and E.-S. Jung. Profiling Optimization for Big Data Transfer Over Dedicated Channels. In *Proceedings of the 25th International Conference on Computer Communication and Networks*, Waikoloa, Hawaii, USA, August 1-4, 2016 (ICCCN16).
- C.Q. Wu and R. Kettimuthu. Distance-Agnostic, Application- and Resource-Aware Transport for Next-Generation Networks. In *Proceedings of DOE Network 2025 Challenges Workshop*, Washington DC, USA, February 1-2, 2016 (DOENET2025).
- P. Dharam, C.Q. Wu, and N.S.V. Rao. Advance Bandwidth Scheduling in Software-Defined Networks. In *Proceedings of the IEEE Globecom*, San Diego, CA, USA, December 6-10, 2015 (Globecom15).
- D. Yun and C.Q. Wu. An Integrated Transport Solution to Big Data Movement in High-performance Networks. In *Proceedings of the 23rd IEEE International Conference on Network Protocols*, PhD Forum, San Francisco, CA, USA, November 10-13, 2015 (ICNP15).
- D. Yun and C.Q. Wu. An Integrated Transport Solution to Big Data Movement in High-performance Networks. A poster presentation at *the NSF Data Science Workshop*, Seattle, WA, USA, August 5-7, 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.