DOE Workshop on

Ultra High-Speed Transport Protocols and Dynamic Provisioning for Large-Scale Science

April 10-11, 2003
Argonne National Laboratory, Argonne, IL

Background

The next generation ultra-scale supercomputers proposed for large-scale science computations promise speeds approaching 120 teraflops within the next few years and petaflops further along. DOE's facilities such as the spallation neutron source, represent some of the most valuable resources for leading edge basic and applied scientific experiments. These computational and experimental resources are vital to the success of a number of DOE's large-science applications (DLSA) from fields as diverse as earth science, high energy and nuclear physics (HENP), astrophysics, fusion energy science, molecular dynamics, nanoscience, and genomics. This data generated at the ultra-scale computing and experimental facilities must be transferred, visualized and steered by geographically distributed teams of scientists. Such networking capabilities add a whole new dimension to the access to these supercomputers, high-performance storage systems and experimental facilities, and thereby eliminate the "access bottlenecks" that currently plague these valuable resources.

The network capabilities required to support this scale of data volumes far exceeds today's leading-edge high-speed network technologies. Large-science applications demand networks capable of delivering hundreds of Gigabps throughputs to the users in near future and several Terabps within the next decade. While such speeds are currently possible only at the link level - based on *dense wavelength division multiplexing* (DWDM) technologies - several architectural and design factors of the transport stacks, I/O systems, network interface cards, and related software currently limit the throughputs to a mere fraction of 1 Gbps and offer very little throughput stability. Experts in the field now agree that sustaining multi-Gbps throughputs at the application level will not be achieved by simply replacing existing low speed links with ultra fast ones. For instance (to give a somewhat dated example), when OC3 (150Mbps) backbone was replaced by OC12 (600Mbps), the application throughput improved only marginally (25-50%) instead of fourfold as expected. Indeed, it took several years of protocol tuning and enhancements to reach 300Mbps of sustained throughput but only under low traffic levels. And similar fate awaits the simple-minded approach of just replacing the current links with faster links in the backbone networks. In a nutshell, to meet the demands of DLSA, highly focused efforts are needed to: (a) suitably provision the networks, (b) optimize the transport processes, and (c) effectively interface with the applications.

Workshop Goals and Objectives

The objective of this workshop is to identify the provisioning and transport technologies as well as application-level issues needed to build operational ultra-speed networks within the timeframe of next few years. Our task comprises of identifying the relevant existing technologies, together with their strengths and weaknesses, and the needed future technologies. The workshop goals include:

- Assess the viability of TCP and non-TCP methods for ultra high-speed operations required in the large-scale science computations and experiments;
- Explore innovative and efficient strategies for provisioning ultra high-speed capacity links to support ultra high-speed network operations; and

?? Identify and assess the generic network functionalities needed for the wide-class of DOE large-science computations and experiments.

Workshop Format
This is a "working" workshop with active discussions and participation from the experts in various topics within the focused groups as well as all together. The participants will be provided with documents summarizing the network requirements of DOE large-science applications prior to the workshop. This advanced knowledge of the "unique" network requirements will help participants to identify research, development, and deployment issues of ultra high-speed networks. The workshop participants will be organized into three groups corresponding to the main objectives of the workshop outlined above. The output of the workshop is a report of the discussions, findings, and research and development directions of the groups.

Sponsor: Office of Science, U. S. Department of Energy

Program Co-Chairs: Nagi Rao and Bill Wing, Oak Ridge National Laboratory, phone:(865) 574-7517; email; {raons,wingwr}@ornl.gov

Workshop website: www.csm.ornl.gov/net/wk2003.html

Local arrangements: The workshop will take place at Argonne National Laboratory.

?? Participants need to register ahead of time to get badges to enter ANL. Please register as soon as possible at the workshop website. It takes extra time for non-US citizens, particularly citizens of sensitive countries.
?? A block of rooms is booked at the Argonne Guesthouse at the rate of $65/night for single and $75/night for double occupancy.

Contact Cheryl Zidel, zidel@mcs.anl.gov for any questions about local arrangements.

Travel Expenses and Reimbursements: For participants whose expenses are paid by DOE, the easiest route is to submit all the bills after the workshop to be reimbursed. For more details about this please contact Tina Lanning, lanningt@ornl.gov.

Guidelines to workshop participants:

This workshop serves a very important role in helping to chart the networking directions of Mathematics, Information and Computer Science Division of DOE's Office of Science. This area represents one of the most crucial enabling technologies to support several DOE's large-science applications in radically new and effective ways. The objective is to develop a comprehensive understanding of the DLSA Networking Needs (DLSANN) to identify those that are met by (a) current technologies with small-scale incremental efforts, and (b) focused DOE efforts in the near term (around two years) and further along. The outcome of this workshop could provide an important advice in shaping the networking priorities of MICS and DOE for DLSA efforts. Some of the most important questions concern the *deployable solutions* in the near term, that is technologies leading to operational solutions that can be used by non-experts in ESnet and other environments.

There are three important parts to the workshop activities:

1. **Pre-Workshop Activities**: Participants are from federal agencies, national laboratories, universities and industry. Due to their diverse backgrounds and interests, they are requested to look at the DOE networking needs for high-impact science prior to the workshop. A comprehensive report on these matters was produced as a result of a DOE workshop. Please refer to "Background Information" on our workshop website. Participants are requested to identify the networking technologies in their individual areas that can (partially) meet these DOE networking challenges currently and in future if suitable projects are supported. Ideally, participants will prepare *individual lists* of DOE applications and the technologies from their areas before the workshop.

2. **Workshop Activities**: The workshop contains a series of group sessions interspersed with joint sessions. In the earlier group sessions (Day 1 morning session), the participants will make brief presentations of their lists to communicate their views to the other participants (please note the diversity of backgrounds of the participants). Then these lists are consolidated through open discussions to identify the currently available technologies together with their strengths and limitations (Day 1, afternoon sessions 1 and 2). Also, perhaps more importantly, the technologies that are to be developed by DOE to meet the challenges will be identified through open discussions. These results will be presented (in a preliminary form) by the group leaders to all participants in a joint session during the first day. At this time the connections between topics of various groups will be identified and discussed. In the second day, these group lists will be further refined with details and their connections with topics from other groups. These group lists will be combined and eventually converted into a workshop report. Work assignments will be identified to generate the final workshop report.

3. **Post-Workshop Activities:** Based on the workshop activities, a detailed report will be produced after the workshop. Each group will contribute to their individual area as well to the overall report. This report consists of three important parts: (a) explicit transport, provisioning and functional capabilities needed for DOE networks; (b) existing networking technologies that meet some of the needs; and (c) future directions to develop the technologies that are needed to meet the needs within the timeframe of few years.

Answers to the following questions could be useful for the participants in their contributions to this workshop.

General questions to all groups:

1. What are the current technologies in your expertise areas that can contribute to DLSA?
2. What DLSA networking capabilities are not achieved by current network technologies in your area?
3. What networking technologies can be developed and deployed in operational networks for DLSA within next two years or so?
    a. With explicit DOE efforts
    b. With industry and other agency efforts alone
4. What DLSA networking requirements are not likely to be met if we rely on
    a. With explicit DOE efforts
    b. With industry and other agency efforts alone
5. What DLSA requirements are too challenging to be completely met within next two-years in spite of various DOE and other efforts

The following questions for individual groups will be refined and expanded within next few days.

Applications Group A

1. List of DSLA disciplines and areas that require ultra high-speed networks
    a. List of individual projects
2. Refined list of individual projects requiring
    a. high-performance throughput
    b. visualization of large data sets over wide-area networks
    c. computational steering
    d. closed-loop control of remote devices and computations

Transport Group B

Assessment of TCP and non-TCP protocols for

a. high-performance throughput
b. visualization of large data sets over wide-area networks
c. computational steering
d. closed-loop control of remote devices and computations

Infrastructure Group C

What type of network levels provisioning are appropriate for immediate attention for DLSANN?
    Routed IP networks vs.All Optical Networks vs. mixed networks

What is the role in DLSANN of

1. Optical Burst switching
2. Fiber channel over SONET

Please note that we explicitly deal with the networking requirements of DOE's large-science applications and do not target areas that are not relevant to DLSANN. Interesting network research areas that do not address DLSANN are not appropriate for this workshop – for example, simply coming up with a list of interesting topics to be studied over next several years is not very useful to this workshop. Also, our focus extends beyond the research part in that we like the capabilities to be deployed and tested so that they can be provided as production modules to application users.