Report from the "DOE Workshop on Ultra High-Speed Transport Protocols and Dynamic Provisioning for Large-Scale Science Applications"

(Held: April 9-10, 2003)

Wu Feng, Los Alamos National Laboratory & The Ohio State University

Program Co-Chairs:
Nageswara S. Rao and William R. Wing

DOE HQ:
Thomas Ndousse, George Seweryniak, Walter Polansky

LA-UR-03-3940

"We engineered the Internet, and it works fine for e-mail and the web; but to do *world-class* scientific research, we need to develop a science of networking that delivers usable performance to DOE scientific applications."

- Allyn Romanow
Cisco Systems

Participants

- Representation
  Balanced participation from universities, industry and national laboratories.

  - Total: 32
    - *National Laboratories: 10*
      - ANL: 2, ESnet: 1, LANL: 2, ORNL: 3, PNNL: 1, SLAC: 1
    - *Universities: 11*
      - Caltech: 1, Georgia Tech: 2, PSC: 1, UC-Davis: 1, UMass: 1, UVa: 1, UIC: 2, Indiana: 1, UTenn: 1
    - *Industry: 8*
      - Celion, Cienna, Cisco, Juniper, Level(3), LightSand, MCNC R&D Institute, Qwest
    - *DOE HQ: 3*

Workshop Goal

- Address the research, design, development, testing and deployment aspects of transport protocols and network provisioning as well as the application-level capability needed to build operational ultra-speed networks to support emerging DOE distributed large-scale science applications over the next 10 years.

Workshop Focus

Given the results from the "High Performance Network Planning Workshop" in August 2002, formulate an R&D roadmap in three specific (and critical) areas over a 1- to 10-year horizon:

- Ultra High-Speed Transport Protocols
  - Session Co-Leads
    - Wu Feng, Los Alamos National Laboratory
    - Don Towsley, Univ. of Massachusetts
- Dynamic Network Provisioning
  - Session Co-Leads
    - Biswanath Mukherjee, UC-Davis
    - William Wing, Oak Ridge National Laboratory
- Network Testbeds
  - Joint effort between the above two groups.

Why these areas? The next generation of DOE scientific breakthroughs will depend primarily on these research areas as well as a core set of services needed from ESnet.

Plann...
roadmap in thr... areas over a 1- to
10-year horizo...

- Ultra High-Speed Transport Protocols
  - Session Co-Leads
    - Wu Feng, Los Alamos National Laboratory
    - Don Towsley, Univ. of Massachusetts
- Dynamic Network Provisioning
  - Session Co-Leads
    - Biswanath Mukherjee, UC-Davis
    - William Wing, Oak Ridge National Laboratory
- Network Testbeds
  - Joint effort between the above two groups.

Applications Perspective (Single Stream)

Application	Now	5 Years Out	10 Years Out
Climate	Authenticated data streams thru firewalls	Robust (reliable) access via multiple sites/paths Petabyte transfers.	Robust access with BW & latency for remote analysis and visualization
SNS	(Facility comes on-line in 2006.)	50-80 Mb/s sustained 320 Mb/s peak	1 Gb/s sustained
MMC	100 Mb/s sustained 200 Mb/s peak	200 Mb/s sustained 400 Mb/s peak	2 Gb/s sustained 4 Gb/s peak
HEP	1 Gb/s & end-to-end QoS	100 Gb/s over lambda and network monitoring	1 Tb/s
FES	Authenticated data streams thru firewalls at 30 Mb/s sustained.	100 Mb/s sustained 500 Mb/s peak (for 20 sec out of 15 min, i.e., QoS)	QoS for network latency and reliability to support real-time remote experiments
Chem Sci	Robust (reliable) access w/ security for *long* times.	10+ Gb/s sustained (collab. viz & data mining)	100+ Gb/s sustained (distributed simulations)
Bioinfo	?	?	?

Some applications actually want bandwidths that are order(s) of magnitude *higher*. Above → "realistic" expectations.

Collaborative work environments need QoS.

- Only 250 kb/s sustained bandwidth over Access Grid at workshop!

Application Requirements

- Usage Scenarios
  - Bulk data transfer / replication, remote visualization, computational steering, data exploration and mining, instrument control
- Issues
  - Service Classification / Quality of Service
    - Guaranteed vs. best effort
    - Unicast vs. multicast vs. broadcast
    - High bandwidth
    - Stable bandwidth (low burstiness)
    - Low latency
    - Low resource (CPU, memory, etc.) utilization
    - Fairness
    - Timeliness (real-time vs. non-real-time)
    - Robustness/Reliability/Error Rate and Patterns (bit-level/packet-level)
  - Performance Gaps
    - Out-of-box vs. network wizard vs. what apps want
  - Byte vs. Block Orientation in Transferring Data
  - Security: Authentication & authorization (Session Layer)
    Privacy & integrity (Presentation Layer)

Wu-chun Feng
feng@lanl.gov

Ultra High-Speed Transport

Networking Perspective: Today

Network Environments

- LAN, e.g., Ethernet + IP + TCP + ftp
  - Throughput: 4-5 Gb/s. Latency: 20 μs. (LANL)
- SAN, e.g., Quadrics + OS-bypass + src routing + MPI
  - Throughput: 6-7 Gb/s. Latency: 5 μs. (LANL, OSU)
- WAN, e.g., Ethernet/DWDM + IP + TCP + ftp
  - Throughput: 2-3 Gb/s. Latency: 90 ms transoceanic. (Internet2 Land Speed Record: Caltech, CERN, LANL, SLAC, Feb. 2003)
- Hybrid network environments, e.g., SAN + WAN
  - MicroGrid or distributed cluster. Interaction of different transport protocols.

Current Problems

1. *Not* numbers that applications actually see. Achieving the above numbers required *many network wizards* and *many months* to accomplish.
2. Deploying R&D onto a production network is not a good idea. There exists a strong need for network testbeds.
3. QoS not being addressed, e.g., Access Grid is a tremendous capability, but it is fraught with QoS issues.
4. Deployment of hardware & software infrastructure. Is OC-768 (40 Gb/s) by 2008 enough? No, need 2-20 Tb/s (Bill Wing, ORNL).

Vision for Short Term (now – 5 yrs): Problems with Current Transport Protocols

- Algorithmic
  - Byte- vs. block-oriented → SCTP/RDMA/R-UDP/Tsunami
  - Sequence numbers → PAWS & "Fat" TCP (RFC 1263)
  - Checksums / CRC → SCTP/RDMA
  - Slow start. Too "slow" for interactive? → Packet probes / tomography / history
  - Congestion control (as it relates to stability & convergence rate) → FAST TCP, scalable TCP, HS-TCP, XCP, stochastic approximation (SA), control theoretic
  - Congestion loss vs. non-congested loss → Not addressable in TCP. All losses viewed as congestion losses.
    - Self-clocking ACKs
  - Fairness (not needed all the time)
    - Ability to turn off congestion ctrl. → Not implemented but could be.
  - QoS guarantees → Not addressable in TCP (nor in ESnet at the present time)
  - Assumes shared, packet-switched network → Fundamental assumption of TCP.
  - Striping / parallel streams → SCTP, RDMA, R-UDP?
- Implementation
  - Flow control with respect to advertised window → DRS / Web100 / Net100.
  - MTU size → TCP/IP → device driver (virtualize MSS/MTU).
  - "Excessive" CPU & memory utilization → TCP off-load & RDMA over TCP/IP.
- In user space or kernel?

Wu-chun Feng
feng@lanl.gov

Los Alamos NATIONAL LABORATORY

Vision for Short Term (now – 5 yrs): Problems with Current Transport Protocols

- Algorithmic
  - Byte- vs. block-oriented → SCTP/RDMA/R-UDP/T
  - Sequence numbers → PAWS & "Fat" TCP (RFC
  - Checksums / CRC → SCTP/RD
  - Slow start … history
  - Congest… scalab…
  - C… …wed
  - Fa…
  - Q… …t time)
  - Assumes s… …ion of TCP.
  - Striping / paral…

Problem:

So many requirements …

So many issues …

Solution?

Short-Term: Address them as noted.

Long-Term: (Next slide)

- Implementation
  - Flow control with respect to advertised window → DRS / Web100 / Net100.
  - MTU size → TCP/IP → device driver (virtualize MSS/MTU).
  - "Excessive" CPU & memory utilization → TCP off-load & RDMA over TCP/IP.
- In user space or kernel?

Vision for Long Term (5 – 10 yrs): Problems with Current Transport Protocols

- Composable ("Lego") Transport Protocols
  - Configuration time vs. run-time loadable modules.
  - User hints, network-passed information.
  - Definition of functions that need to be composed:
    - Parallelism: parallel streams, network striping, multipath, data aggregation
    - Unit of "Bookkeeping": Byte- vs. Block- vs. Stream- vs. File-Oriented.
    - Error Control: Strong CRC and FEC.
    - QoS: Best Effort vs. Reservation. Best Effort vs. Guarantees (Soft & Hard).
    - Information Export to Application or Network Manager.
    - Environment: LAN vs. SAN vs. WAN vs. mixed.
    - (Control: RTT-dependent or not …)
- Analytical Design Based on Control and Statistics.
  - Effects of composing certain transport features with others.
- "Smart" Transport
  - Knowing where data is headed ahead of time may influence how transport protocols are composed.
  - Interface with I/O issues.
- Compatibility & Future Issues
  - Legacy Problem / Interoperability / Bridging: Packet & circuit switched.

Wu-chun Feng
feng@lanl.gov

Recommendations from Transport Group

1. Theory of Transport
   - Stochastic control, statistics, optimization, scalability, robustness
2. Algorithm Design
   - "Autonomic", adaptive, modular, composable …
3. Experimentation (Simulation, Testbeds)
4. Instrumentation & Diagnostic Tools
   - e.g., Web100/Net100 (end host) + for network
   - Statistical inferencing techniques. Data collection methods.
5. DOE Deployment, Wider Adoption, and Legacy Integration

Dynamic Network Provisioning

Dynamic Provisioning: Barriers Over the Next Five Years

- Limited deployment of ultra-long haul DWDM links.
- Enhanced support for striped/parallel transport needed.
  - At the core, edge/GMPLS routers, OS, and app levels.
- Lack of high-speed circuit-switched infrastructure. (e.g., a la recent achievements of the Internet2 Land Speed Record)
  - At the core and at the edge/GMPLS routers to support (dynamic) on-demand reconfigurability.
  - Emulated circuit-switched to the end host.
- Lack of well-developed methods and APIs for
  - Connection set-up, scheduling & reservation, and allocation.
- DOE apps do *not* follow commercial scaling model.
- Lack of a security model.
- Lack of a robust multicast solution.

Recommendations from Dynamic Provisioning Group

1. Scalable network technologies to deliver bandwidth and lambdas on demand.
2. New radical technologies to dynamically schedule and manage a multiple-lambda network.
3. Scalable technologies to design and manage *multi-tier*, logical , ultra high-speed network.
4. An application-centric, cross-country, ultra high-speed network infrastructure for R&D.
5. Necessity to engage and involve ESnet first and foremost as well as researchers and developers in middleware, transport, operating systems, and of course, applications.

Wu-chun Feng
feng@lanl.gov

Network Testbeds

Visions & Recommendations for Enabling Network Research

- Now
  - Evaluate granularity of work (i.e., MTU and MTU-like methods).
  - Directions and recommendations for a production network to support experimental *application* research and experimental testbeds to support experimental *network* research, e.g., initial deployment of experimental testbed with QoS and GMPLS, for tomorrow's apps.
- Five-Year Horizon
  - Further deployment of experimental testbed with QoS and GMPLS support.
  - Striping infrastructure for meeting high-throughput needs. (Complementary to deployment of experimental testbeds.)
  - Support for splicing & cascading transport methods from end-to-end.
    - Splicing: Multiple (homogeneous) TCP flows.
    - Cascading: Multiple (heterogeneous) transport flows, e.g., ULNI/InfiniBand → TCP → ULNI/Quadrics
- Ten-Year Horizon
  - Evolution of experimental testbeds, keeping in mind that aggregate network demands are doubling every year.
  - Deployment and support for composable transport methods.

One for All and All for One … A View from 10,000 Feet

- Network Research
  - Focus on end-to-end.
  - Need support from ESnet to enable research in network provisioning and scheduling, transport (traditional TCP/IP to RDMA over WAN to composable protocols), and OS.
  - Even with support, need network testbeds to validate and verify.
- ESnet
  - Focus on the core.
  - Core services to provide to network researchers (ultimately, applications)
    - QoS / MPLS / λ–switching
    - Multi-tiered, ultra high-speed
    - Programmable routers bridging O-E interface.
    - Support for the non-standard, exploratory, researchy: jumboframes, RDMA over WAN.

There is obviously much more to this picture … this is just a starting point …

One for All and All for One …
A Vi… 10,000 Feet

- N…
  - … the core.
  - …ces to provide to …searchers … applications)
  - …PLS / λ–switching
  - …ered, ultra high-speed
  - …ammable routers …g O-E interface.
  - Ev… netwo… and verify.
  - … for the non-standa… exploratory, researchy: jumboframes, RDMA over WAN.

Example

Latest Internet2 Land Speed Record *required* end-to-end support for jumboframes to achieve 2.38 Gb/s between Sunnyvale and Geneva. (Only ~300 Mb/s with standard 1500-byte MTUs.)

Sponsors: DOE, NSF, European Commission

There is obviously much more to this picture … this is just a starting point …

Consequences of Not Doing This Research in DOE

- Why should DOE invest in this research, and hence ESnet? Internet is optimized for the mass market, not for scientific applications.
- U.S. is losing competitive advantage for worldwide scientific resources, e.g., Japanese Earth Simulator, European data grid (LHC: Large Hadron-Collider Computing Grid), which requires superior networking infrastructure.
- Brain Drain
  - People leaving. Inability to retain top scientists.
- "Gray Matter" Tax
  - Scientist spend time doing networking rather than science, e.g., 15-20 scientists and months of planning and tuning to break the Internet2 Land Speed Record.
- Harmful effects to existing DOE-funded projects, e.g., SciDAC projects.