SC26 Network Research Exhibition: Demonstration Abstract

Prototype Multi-Tbps WAN Services: Architecture, Technology and Control Systems

Joe Mambretti, Jim Chen, Fei Yeh, International Center for Advanced Internet Research - Northwestern University

j-mambretti, jim-chen, fyeh@northwestern.edu, Linden Mercer, linden.mercer.ctr@nrl.navy.mil, Bill Fink, billfink@mindspring.com, Tom Lehman, tlehman@es.net, Rod Wilson (Ciena), rwilson@ciena.com, Marc Lyonnais (Ciena), mlyonnai@ciena.com, Scott Kohlert (Ciena), skohlert@ciena.com, Gaurauvdeep Shami (Ciena), gshami@ciena.com

Abstract

Large scale data production within and among science research collaborations and sites continues to increase, a long term trend that continues to accelerate, especially driven by the deployment of new science facilities, including high luminosity research instrumentation and next generation high performance computation science centers (many AI centric). Consequently, the science networking community has begun to prepare infrastructure that can support high-capacity services, with a key thematic focus on services beyond 400 Gbps LANs and WANs, including those approaching 800 Gbps, 1.2 Tbps, 1.6 Tbps, and multi-Tbps WAN and LAN services. ***NB: A companion NRE describes demonstrations of the multi-Tbps infrastructure required for a large scale multi-Tbps services.*** This NRE describes prototype multi-Tbps WAN services that can be supported by the infrastructure described in that NRE. Also, Ref NRI 103.

Overview

When 400 Gbps WAN/LAN services were initially being anticipated, the science networking community began to develop techniques to effectively utilize that level of capacity. Standard implementations and configurations as commonly used were not able to rise to the level of available network capacity. Similarly, today, the requirements and implications of WAN and LAN services beyond 400 Gbps are being explored, including technologies that can support 800 Gbps, 1.2 Tbps, 1.6 Tbps WAN services and multi-Tbps services E2E. (NB: LHC Tier 1 site and several Tier 2 sites are planning for minimal 1 Tbps connections.) These SC26 demonstrations will showcase capabilities and technologies that will assist in providing such services by demonstrating prototype multi-Tbps WAN transport services from the StarLight International/National Communications Exchange Facility in Chicago to the SC26 venue, multi-Tbps from the JBDT Facility in McLean Va. to the StarLight Facility, including 3*1.2 Tbps and 3*800 Gbps between the StarLight Exchange (including the JBDT facility at that site) and the JBDT Facility at McLean, Virginia, and 1.2 Tbps (3*400 Gbps) between the LHC testbed at the University of Victoria in Canada, over CANARIE to PacWave, to the PNWGP, to 3*400 Gbps NA REX paths to the StarLight Exchange, to the SC26 venue.

Goals

With its research partners including the SCinet WAN group, the Ciena research group, NASA, the NRL, and ESnet, the International Center for Advanced Internet Research (iCAIR) at Northwestern University is designing a multi-Tbps WAN service among the sites noted. Multiple issues are being investigated to enable the utility of multi-Tbps WAN services E2E.

1. One area of investigation is the potential to implement these services across E2E lightpaths established on open line systems across multidomain networks, avoiding as much as possible (in some cases completely) L3 devices and/or services, including by using segmented/isolated slices across shared infrastructure.

2. At the edges of the WAN paths, this project is implementing multiple optimized Gen5 400 Gbps Data Transfer Nodes (DTNs, with a

minimal of three DTNs at each site. Gen6 DTNs are also currently being evaluated.

3. Each DTN has been implemented with 400 Gbps smart NICs configured with high performance optical transport connectors. 800 Gbps smart NICs are also being evaluated.

4. These demonstrations leverage experimental research into the optimal design, configuration, components, and integration technologies for DTNs, including 100, 200, 400, and 800 Gbps NICs, NVMe over FABRIC, and techniques for kernel bypass using zero-copy for memory and disk copy to avoid bottlenecks in large scale data transfer over multi-Tbps WANs and optimal affinity bindings for NUMA architecture for higher resource utilization.

5. These demonstrations also showcase middleware required to orchestrate (including by using SENSE/AutoGOLE) highly distributed infrastructure resources for reliable, optimized high-speed multi-domain network services supporting high capacity, high performance data transfers over thousands of miles.

6. Technologies include Software Defined Networking (SDN), RDMA, RoCE, NVMe over fabric (e.g., enabling less than 1 minute for a one Terabyte NVMe-to-NVMe transfer over thousands of miles), GPU acceleration, and federated P4 and SDN controllers managing multi-domain paths.

7. These demonstrations also demonstrate measurement techniques for real-time monitoring, benchmarking and evaluation including those based on AI/ML toolsets.

8. These techniques include analytics based on real-time in-network (in-band) processing to avoid sending data to processes external to the network, using programmable hardware such as FPGA within accelerated frameworks.

9. At Tbps and above, conventional monitoring tools (e.g., standard SNMP or simple flow exporting processes) cannot provide required granularity and overly rely on CPUs. The approach here uses In-Network Telemetry (INT) and hardware-accelerated processing to achieve hi-fidelity per-packet visibility, using telemetry logic embedded directly into the data plane of high-performance hardware, e.g., enabling the network to include telemetry information (e.g., switch ID, port residency time, and queue utilization) directly into the packet headers as they pass through nodes. Hardware used includes: a) Tofino2 programmable switches used for line-rate packet inspection, e.g., using P4 programmable logic to edit payloads, extracting metadata-rich headers for analysis from general scientific traffic for analysis, b) FPGA-Based SmartNICs (e.g., Xilinx Alveo U55C), processing header-only data streams, using P4 and c) NVIDIA BlueField-2 DPUs for real-time in-network machine learning, running P4 pipelines via DOCA, enabling sophisticated telemetry not traditionally addressed by P4. AI/ML integration uses fixed-point inference within the P4 data plane to allow the telemetry process to perform real-time anomaly detection at line rate.

10. Another technique is in-network classification, which identifies whether packet payloads that contain new information relative to references to earlier data to enhance distributed scientific workflows (e.g., detecting key astronomical or physics events in real time to allow for dynamic responses.

11. This WAN service approach also includes examining implications for integrating multi-Tbps WAN services with existing production operation processes.

Impacts

1) These research projects are directed at developing resources to meet demands for increasingly large capacity WAN transport for large scale (global) data intensive science. This NRE is focused on muti-Tbps WAN services (vs underlying infrastructure). (ref related NRE focused on multi-Tbps WAN infrastructure).
2) SCinet supports capabilities that enable innovative multi-Tbps WAN services and techniques to be demonstrated at an international scale.
3) Next steps include migration to network software stacks based on Gen6 servers, 800 G smart NICs, extending capabilities to additional international sites, integrating new protocols, developing additional monitoring and analytics using INT AI techniques, including DTNs based on AI models, etc.

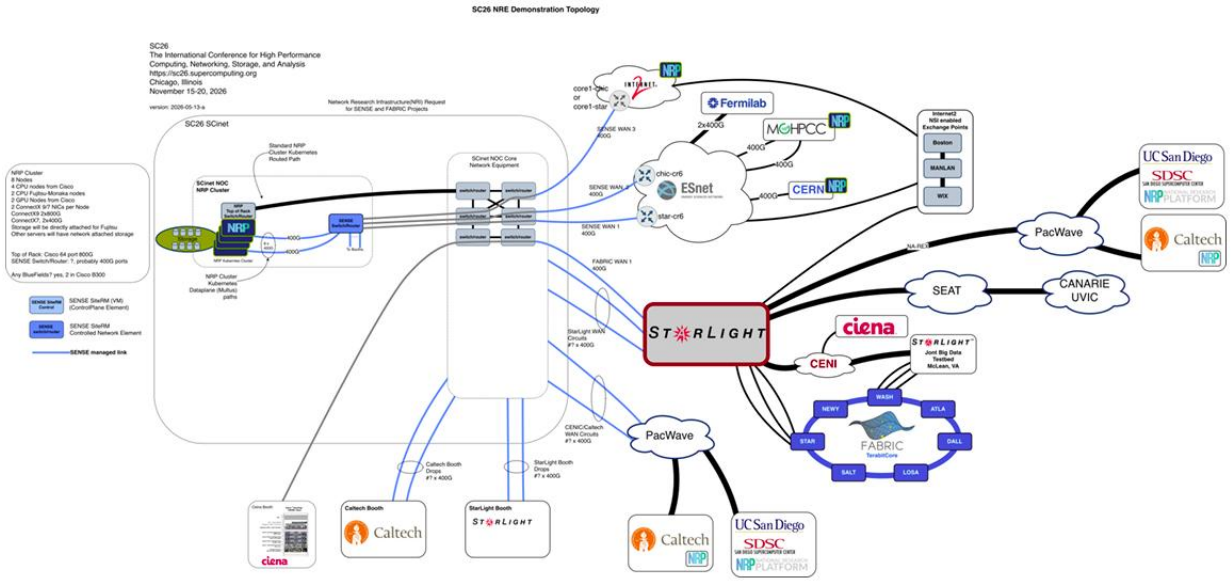
Resources

Required resources from SCinet WAN include 3*1.2 Tbps and 3*800 Gpbs WAN transport services among the StarLight International/National Communications Exchange in Chicago, the Joint Big Data Testbed Facility in McLean Virginia, and the SC26 venue (NB: existing testbeds, e.g., 3*400 Gbps on FABRIC and 2*1.2 Gbps on the CENI testbed). Also, to develop LHC networking capabilities (e.g., scitags), a 1.2 Tbps testbed has been implemented from the University of Victoria in Canada to PacWave in Seattle, to the PNWGP, across NA REX to the StarLight Facility. The primary requests from SCinet are the optical fiber paths from the StarLight Facility to the StarLight booth at the SC26 venue.

Tb

Involved Parties

- Joe Mambretti, iCAIR, jmambretti@northwestern.edu
- Jim Chen, iCAIR, jim-chen@northwestern.edu
- Fei Yeh, iCAIR,fyeh@northwestern.edu
- Linden Mercer, linden.mercer.ctr@nrl.navy.mil
- Bill Fink billfink@mindspring.com
- Tom Lehman, tlehman@es.net
- Rod Wilson, Ciena, rwilson@ciena.com
- Marc Lyonnais, Ciena, mlyonnai@ciena.com
- Scott Kohlert, Ciena, skohlert@ciena.com
- Gaurauvdeep Shami, Ciena, gshami@ciena.com
- StarLight International/National Communications Exchange Facility and Consortium
- Metropolitan Research and Education Network (MREN)
- SC26 SCinet



Initial SCinet topology – Tom Lehman – This topology does not include all paths, e.g., the complete NA REX infrastructure and connections between StarLight and the SC26 venue