

NRE-19: SC19 Network Research Exhibition: Caltech Booth 543 Demonstrations

Hosting NRE-13, NRE-19, NRE-20, NRE-22, NRE-23, NRE-24, NRE-35

Global Petascale to Exascale Workflows for Data Intensive Science Accelerated
by Next Generation Programmable SDN Architectures and Machine Learning Applications

Submitted on behalf of the teams by: Harvey Newman, Caltech, newman@hep.caltech.edu:

Abstract

We will demonstrate several of the latest major advances in software defined and Terabit/sec networks, intelligent global operations and monitoring systems, workflow optimization methodologies with real-time analytics, and state of the art long distance data transfer methods and tools and server designs, to meet the challenges faced by leading edge data intensive experimental programs in high energy physics, astrophysics, climate and other fields of data intensive science. The key challenges being addressed include: (1) global data distribution, processing, access and analysis, (2) the coordinated use of massive but still limited computing, storage and network resources, and (3) coordinated operation and collaboration within global scientific enterprises each encompassing hundreds to thousands of scientists.

The major programs being highlighted include the Large Hadron Collider (LHC), the Laser Interferometer Gravitational Observatory (LIGO), the Large Synoptic Space Telescope (LSST), the Event Horizon Telescope (EHT) that recently released the first black hole image, and others.

Several of the SC19 demonstrations will include a fundamentally new concept of “consistent network operations,” where stable load balanced high throughput workflows crossing optimally chosen network paths, up to preset *high water marks* to accommodate other traffic, provided by autonomous site-resident services dynamically interacting with network-resident services, in response to demands from the science programs’ principal data distribution and management systems.

Some of the cornerstone system concepts and components to be demonstrated include:

- Integrated operations and orchestrated management of resources: Absorbing and advancing the site (DTN-RM) and network resource managers (Network-RM) developed in the SENSE [9] program, edge SDN control, NDN routing and caching, transfer tools control, packet re-routing at network level and real-time resource control over site facilities and instruments.

- Fine-grained end-to-end monitoring and data collection, with a focus on the edges and end sites, enabling data analytics-assisted intelligent and automatic decisions driven by applications supported by optimized path selection and load balancing mechanisms driven by machine learning.
- An ontological model-driven framework with integration of an analytics engine, API and workflow orchestrator extending work in the SENSE project, enhanced by efficient multi-domain resource state abstractions and discovery mechanisms.
- Adapting NDN for data intensive sciences including advanced cache design and algorithms and parallel code development and methods for fast and efficient access over a global testbed, leveraging the experience in the SDN Assisted NDN for Data Intensive Experiments (SANDIE; NSF CC*) project.
- A paragon network at several our partners’ sites composed of P4 programmable devices, including Tofino-based switches and Xilinx FPGA-based smart network interfaces providing packet-by-packet inspection, agile state tracking, real-time decisions and rapid reaction as needed.
- High throughput platform demonstrations in support of workflows for the science programs mentioned. This will include reference designs of NVMeOF server systems to match a 400G network core, and comparative studies of servers with multi-GPUs and programmable smart NICs with FPGAs.
- Integration of edge-focused extreme telemetry data (from P4 switches and end hosts) and end facility /application caching stats and other metrics data to facilitate automated decision-making process.
- Development of dynamic regional caches or “data lakes” that treat nearby as a unified data resource, building on the successful petabyte cache currently in operation between Caltech and UCSD based on the XRootD federated access protocol; extension of the cache concept to more remote sites such as KISTI and KASI in Korea, and TIFR (Mumbai). Applications of the caches to support the LSST science use case and the use of PRP/TNRP

distributed GPU clusters for machine learning and related applications.

- Blending the above innovations with CMS petabyte regional caches and real-time joint-sky-survey analysis data services with a new level of end-to-end performance. This will also help define the near-future workflows, software systems and methods for these science programs.
- System and application optimizations using the latest graphical representations and deep learning methods

This will be empowered by end-to-end SDN methods extending all the way to autoconfigured Data Transfer Nodes (DTNs), including intent-based networking APIs combined with transfer applications such as Caltech's open source TCP based FDT which have been shown to match 100G long distance paths at wire speed in production networks. During the demos, the data flows will be steered across regional, continental and transoceanic wide area networks through the orchestration software and controllers, and automated through orchestration software and controllers such as the Automated GOLE (AutoGOLE) controlled through NSI and its MEICAN frontend, and automated virtualization software stacks developed in the SENSE, PRP/TNRP and Chase-CI, AmLight, and other collaborating projects. The DTNs employed will use the latest high throughput SSDs and flow control methods at the edges such as FireQoS and/or Open vSwitch, complemented by NVMe over fabric installations in some locations.

Elements and Goals of the Demonstrations

- **LHC:** End to end workflows for large scale data distribution and analysis in support of the CMS experiment's LHC workflow among Caltech, UCSD, LBL, Fermilab and GridUNESP (Sao Paulo) including automated flow steering, negotiation and DTN autoconfiguration; bursting of some of these workflows to the NERSC HPC facility and the cloud; use of unified caches to increase data access and processing efficiency.
- **AmLight Express and Protect (AmLight-Exp)** in support of the LSST and LHC-related use cases will be shown, in association with high-throughput low latency experiments, and demonstrations of auto-recovery from network events, using optical spectrum on the new Monet submarine cable, and its 100G ring network that interconnects the research and education communities in the U.S. and South America. For the LSST use case, real time representative low latency transfers for scientific processing of multi-GByte images from the LSST/AURA site in La Serena, Chile, flowing over the REUNA Chilean as well as ANSP and RNP Brazilian national circuits and the AmLight Atlantic and Pacific Ring and Starlight to the conference site are planned, using 300G of capacity between Miami and Sao Paulo, and 200G between Miami and the SC19 exhibit floor.
- **SENSE** The Software-defined network for End-to-end Networked Science at Exascale (SENSE) research project is building smart network services to accelerate

scientific discovery in the era of 'big data' driven by Exascale, cloud computing, machine learning and AI. The SENSE SC19 demonstration showcases a comprehensive approach to request and provision end-to-end network services across domains that combines deployment of infrastructure across multiple labs/campuses, SC booths and WAN with a focus on usability, performance and resilience through:

- Intent-based, interactive, real time application interfaces providing intuitive access to intelligent SDN services for Virtual Organization (VO) services and managers;
 - Policy-guided end-to-end orchestration of network resources, coordinated with the science programs' systems, to enable real time orchestration of computing and storage resources.
 - Auto-provisioning of network devices and Data Transfer Nodes (DTNs);
 - Real time network measurement, analytics and feedback to provide the foundation for full lifecycle status, problem resolution, resilience and coordination between the SENSE intelligent network services, and the science programs' system services.
 - Priority QoS for SENSE enabled flows
 - Multi-point and point-to-point services
- **Multi-Domain, Joint Path and Resource Representation and Orchestration (Mercator-NG):** Fine-grained interdomain routing system (e.g., SFP) and network resource discovery systems (e.g., Mercator) were designed to discover network path and resource information individually in collaborative science networks. Integrating such information is crucial for optimal science workflow orchestration, but a non-trivial task due to the exponential number of possible path-resource combinations even in a single network. The Yale, IBM, ESNet and Caltech team will demonstrate Mercator-NG, the first multi-domain, joint path and resource discovery and representation system. Compared with the original Mercator system published and demonstrated in SC'18, Mercator-NG provides two key novel features, including (1) a fine-grained, compact linear algebra abstraction to jointly represent network path and resource information without the need of enumerating the exponential number of paths in the network; and (2) an efficient science workflow orchestrator to optimize science workflow with the collected network path and resource information. This demonstration will include: (1) efficient discovery of available network path and resource information in a multi-domain wide-area collaborative science network, including Los Angeles, Denver and New York, with extreme low latency, (2) optimal, online science workflow orchestration in this wide-area network, and (3) scaling to collaborative networks with hundreds of members.
 - **Control Plane Composition Framework for Inter-domain Experiment Networks:** This team will demonstrate Carbide, a novel control plane (CP) composition framework for inter-domain LHC experimental network deployment to achieve

collaborative network with both scientific and campus/domain-specific network. The demonstration will include three key features of the framework. 1) High Security: It composes different layers of CPs, each of which is associated with a real-time, distributed verification model to guarantee the desired traffic policy is not violated. 2) High Reliability: When the LHC CP is crushed or link failure, the framework can use underlay CP as a backup, without affecting any policies. 3) Flexibility: It allows a) partially specified CP for LHC, and b) modularity of existing CP/network, so the LHC CP can be deployed in a plug-in and incremental manner. The virtual LHC CP is specified by the users and includes both intradomain and interdomain. LHC CP can coexist with any existing/instantiated CP underlay.

- **NDN Assisted by SDN:** Northeastern, Colorado State and Caltech will demonstrate Named Data Networking (NDN) based workflows, accelerated caching and analysis in support of the LHC and climate science programs, in association with the SANDIE (SDN Assisted NDN for Data Intensive Experiments) project. Specifically, we will demonstrate (1) increased throughput over (the high speed DPDK-based) NDN network using our OSS NDN based XRootD plugin and the NDN producer, (2) a new implementation of caching which enables multiple types of storage devices, (3) an extended testbed topology with additional node at Northeastern University, and (4) an adaptive optimized joint caching and forwarding algorithm over the SANDIE testbed.
- **FPGA-accelerated Machine Learning Inference for LHC Trigger and Computing:** UCSD and MIT will lead a group of collaborators demonstrating real-time FPGA-accelerated machine learning inference. Machine learning is used in many facets of LHC data processing including the reconstruction of energy deposited by particles in the detector. The training of a neural network for this purpose with real LHC data will be demonstrated. The model deployment and acceleration on a Xilinx Alveo card using a custom compiler called hls4ml will also be shown. An equivalent setup utilizing an NVIDIA GPU will also be presented allowing for direct comparison. This demonstration will serve to illustrate a first prototype of a new approach to the real-time triggering and event selection with LHC data aimed at meeting the challenges of the second phase of the LHC program, the High Luminosity LHC, which is planned to run from 2026-2037, following the development of further prototypes following this approach during the upcoming LHC data taking runs in 2021-2023.
- **400GE First Data Networks:** Caltech, Starlight/NRL, USC, SCinet/XNET, Ciena, Mellanox, Arista, Dell, 2CRSI, Echostreams, DDN and Pavilion Data, as well as other supporting optical, switch and server vendor partners will demonstrate the first fully functional 3 X 400GE local ring network as well as 400GE wide area network ring, linking the Starlight and Caltech booths and Starlight in Chicago. This network will integrate storage using NVMe over Fabric, the latest high throughput methods, in-depth monitoring and realtime flow steering. As part of these demonstrations, we will make use of the latest DWDM, Waveserver Ai, and 400GE as well as 200GE switch and network interfaces

from Arista, Dell, Mellanox and Juniper as part of this core set of demonstrations.

Resources

The partners will use approximately 15 100G and other wide area links coming into SC19, and the available on-floor and DCI links to the Caltech and partner booths. An inner 1.2 Tbps (3 X 400GE) core network on the showfloor will be composed linking the Caltech, SCinet, Starlight and potentially other partner booths, in addition to several other booths each connected with 100G links. Waveserver Ai and other data center interconnects and DWDM to SCinet. For example, the network layout highlighting the Caltech and Starlight booths, SCinet, and the many wide area network links to partners' lab and university home sites can be seen here: <http://tinyurl.com/SC19-JBDT>

The SC19 optical DWDM installations in the Caltech booth and SCinet will build on this progress and incorporate the latest advances.

Partners

Physicists, network scientists and engineers from Caltech, Pacific Research Platform, Fermilab, FIU, UNESP, Yale, Tongji, UCSD, UMaryland, LBL/NERSC, Argonne, KISTI, Michigan, USC, Northeastern, Colorado State, UCLA, TIFR (Mumbai), SCinet, ESNet, Internet2, StarLight, ICAIR/ Northwestern, CENIC, Pacific Wave, Pacific Northwest GigaPop, AmLight, ANSP, RNP, REUNA, NetherLight, SURF, and their science and network partner teams, with support from Ciena, Intel, Dell, 2CRSI, Premio/Echostreams, Arista, Mellanox, and Pavilion Data

Group Leads and Participants, by Team

- **Caltech HEP + LIGO + LSST:** Harvey Newman (newman@hep.caltech.edu), Justas Balcas, Raimondas Sirvinskas, Catalin Iordache, Joseph Chiu, Stuart Anderson, Juan Barayoga
- **Caltech IMSS:** Jin Chang (jin.chang@caltech.edu), Dawn Boyd, Larry Watanabe, Don S. Williams
- **USC:** Azher Mughal (azheramin@gmail.com)
- **LSST:** Jeff Kantor (JKantor@lsst.org), Matt Kollross
- **AmLight/FIU:** Julio Ibarra (Julio@fiu.edu), Jeronimo Bezerra, Adil Zahir
- **Amlight/ISI:** Heidi Morgan (hlmorgan@isi.edu)
- **Yale/Tongi/IBM/ARL:** Richard Yang (vry@cs.yale.edu), Qiao Xiang, Jensen Zhang, X. Tony Wang, Dong Guo, Dennis Yu, May Wang, Christoher Leet, Shenshen Chen, Franck Le, Yeon-sup Lim, Yuki de Pourbaix, Vinod Mishra
- **Maryland:** Xi Yang (maxyang@umd.edu)
- **Virnao:** Tom Lehman (tom.w.lehman@gmail.com)
- **UCSD/SDSC/PRP:** Javier Duarte (jduarte@ucsd.edu), Tom deFanti (tdefanti@ucsd.edu), Larry Smarr, John Graham, Tom Hutton, Frank Wuerthwein, Phil Papadopoulos
- **MIT:** Phil Harris
- **ESnet:** Inder Monga (imonga@es.net), Chin Guok, John MacAuley
- **LBL:** Alex Sim (asim@lbl.gov)

- **LBL/NERSC:** Damian Hazen (dhazen@lbl.gov)
Alex Sim (asim@lbl.gov)
- **UNESP:** Sergio Novaes (Sergio.Novaes@cern.ch),
Rogerio Iope, Beraldo Leal, Marco Gomes,
Artur Beruchi
- **Starlight:** Joe Mambretti
(j-mambretti@northwestern.edu), Jim Chen
- **Johns Hopkins:** Alex Szalay (szalay@jhu.edu)
- **SURF:** Gerben van Malenstein
(gerben.vanmalenstein@surfnet.nl)
- **Fermilab:** Phil Demar (demar@fnal.gov)
- **Argonne:** Linda Winkler (winkler@mcs.anl.gov)
- **Michigan:** Shawn McKee (smckee@umich.edu)
- **Northeastern University:** Edmund Yeh
(eyeh@ece.neu.edu), RanLiu, Yuanhao Wu
- **Colorado State:** Christos Papadopoulos
(christos@cs.colostate.edu),
Chengyu Fan
- **Tennessee State:** Susmit Shannigrahi
(susmit.shannigrahi@gmail.com)
- **UCLA:** Lixia Zhang (lixia@cs.ucla.edu)
- **TIFR Mumbai:** TIFR Mumbai: Brij Jashal
(brij.jashal@tifr.res.in),
Kajari Mazumdar (mazumdar@tifr.res.in)
- **CENIC/Pacific Wave:** John Hess, Louis Fox
- **ANSP (Brazil):** Luis Lopez
- **RNP (Brazil):** Michael Stanton (michael@rnp.br),
Alex Moura
- **REUNA (Chile):** Sandra Jaque (sjaque@reuna.cl), Albert
Astudillo (aastudil@reuna.cl)
- **Ciena:** Marc Lyonnais (mlyonnai@ciena.com),
Rod Wilson, Nick Wilby, Lance Williford

Additional Information Submitted by Partners

(1) LSST (J. Kantor, Matt Kollross et al.):

LSST Science Use Case 1 (at SC18): Prompt processing

LSST acquires 3.2 Gigapixel (6.4 GB uncompressed) images approximately every 15 seconds and must transfer those images from AURA in La Serena, Chile to NCSA in Urbana-Champaign, Illinois in 5 seconds. This is in order to perform “prompt processing” to detect astronomical transient events, such as supernovae explosions, and send out alerts to the scientific community within 60 seconds of image readout from the instrument. Approximately 2000 full focal plane images per night are generated (in pairs of exposures over a single telescope pointing called a “visit”). Each image is composed of 21 files, with each file containing the image data from 1 LSST Camera Raft (an array of 3 x 3 CCDs, each 4k x 4k pixels). At SC 2018, we demonstrated low latency transfers simulated or pre-cursor images from AURA in La Serena Chile to the Chicago Starlight point, and from there to NCSA and/or to the SC venue.

LSST Science Use Case 2: Data Release

At NCSA in Illinois and a satellite processing center at CC-IN2P3 in Lyon, France, LSST reprocesses all of the accumulated survey images every year, to produce deep, co-added images and astronomical object catalogs with extremely precise measurements of very faint objects up to 13B light years distance from Earth. The output of this annual processing is a Data Release, and the size of each Data Release increases each year, from approximately 6 PB in year 1 up to 60 PB in year 10. On completion and quality assessment, the entire Data Release is transferred to our Data Access Centers located at NCSA and at AURA in La Serena, Chile. The transfer from NCSA to La Serena is accomplished over the network, over a period of months. At SC19, we will demonstrate PB data transfers from NCSA to AURA in La Serena, Chile at rates consistent with those required for LSST operations, working with AmLight.

(2) **The AmLight Express and Protect (AmLight-Exp) Project** (J. Ibarra, J. Bezerra, H. Morgan et al.): AmLight-Exp plans to support high-throughput, low latency experiments using optical spectrum on the new Monet submarine cable, and its 100G ring network that interconnects the research and education communities in the U.S. and South America including Chile and Brazil. Use cases for LSST, requiring high throughput image transfers, low latency, and rapid recovery from network events will be tested.

(3) (a) **Multi-Domain, Joint Path and Resource Representation and Orchestration (Qiao Xiang, Francke Le, Y. Richard Yang):** Fine-grained interdomain routing system (e.g., SFP) and network resource discovery systems (e.g., Mercator) were designed to discover network path and resource information in collaborative science networks, driven by the demand and substantial benefits of providing predictable network resources for distributed science workflows. However, a major lack of existing systems is that they are designed to discover different types of information, e.g., network path and network resource, individually, leading to substantial inefficiency when such information is used to optimize science workflows. Integrating the discovery and representation of such information, however, is a non-trivial task even in a single network, due to the exponential number of possible path-resource combinations. Toward addressing this challenge, the Yale, IBM, ESNet and Caltech team will demonstrate Mercator-NG, the first multi-domain, joint path and resource discovery and representation system. Compared with the original Mercator system published and demonstrated in SC'18, Mercator-NG provides two key novel features, including (1) a fine-grained, compact linear algebra abstraction to jointly represent network path and resource information without the need of enumerating the exponential number of paths in the network; and (2) an efficient science workflow orchestrator to optimize science workflow with the collected network path and resource information. This demonstration will include: (1) efficient discovery of

available network path and resource information in a multi-domain wide-area collaborative science network, including Los Angeles, Denver and New York, with extreme low latency, (2) optimal, online science workflow orchestration in this wide-area network, and (3) scaling to collaborative networks with hundreds of members.

(b) Control Plane Composition Framework for Interdomain Experiment Networks (Y. R. Yang, Geng Li, Kerim Gokarslan): The Carbide team will demonstrate a novel control plane (CP) composition framework for inter-domain LHC experimental network deployment to achieve collaborative network with both scientific and campus/domain-specific network. Existing work in network verification relies on centralized computation at the cost of fault tolerance, while other approaches to compose multiple control planes do not provide any guarantees of correctness. Carbide provides both control plane composition and network verification via an online composition layer and a novel real-time distributed verification framework. The demonstration of the Carbide team will include three key features of the framework. 1) High Security: It composes different layers of CPs, each of which is associated with a real-time, distributed verification model to guarantee the desired traffic policy is not violated. 2) High Reliability: When the LHC CP is crushed or link failure, the framework can use underlay CP as a backup, without affecting any policies. 3) Flexibility: It allows a) partially specified CP for LHC, and b) modularity of existing CP/network, so the LHC CP can be deployed in a plug-in and incremental manner. The virtual LHC CP is specified by the users and includes both intradomain and interdomain. LHC CP can coexist with any existing/instantiated CP underlay.

(4) SENSE: SDN for End-to-end Networked Science at the Exascale (I. Monga, J. Balcas, P. Demar, C. Guok, D. Hazen, T. Lehman, H. Newman, L. Winkler, X. Yang) Distributed application workflows with big-data requirements depend on predictable network behavior to work efficiently. The SENSE project vision is to enable National Labs and Universities to request and provision end-to-end intelligent network services for their application workflows, leveraging SDN capabilities. Our approach is to design network abstractions and an operating framework to allow host, Science DMZ / LAN, and WAN auto-configuration across domains, based on infrastructure policy constraints designed to meet end-to-end service requirements.

(5) SANDIE: Named Data Networking (E. Yeh, H. Newman): The SANDIE project teams working on NDN will demonstrate a new highly effective approach to data distribution, processing, gathering and analysis of results to accelerate the workflow for the CMS experiment at the LHC, and to provide a model for the other LHC experiments. This will be accomplished through integration of NDN and SDN systems concepts and algorithms with the mainstream data distribution, processing and management systems of CMS, leveraging the most recent code implementation, storage integration

and routing and caching algorithms of NDN combined with SDN-based path allocations and end-to-end provisioning across the SC19 and wide area network footprint. The goal is to provide more rapid and reliable data delivery, with varying patterns and granularity over complex networks, progressing in scale from the Terabyte to eventually the Petabyte range in support of the LHC physics program.

(6) NVMe Over Fabric High Throughput DTN Server Designs (A. Mughal, C. Anderson; H. Newman): USC working with Caltech and NRL will demonstrate real-time processing of large scale science datasets coupled to transfers across national and international networks using state of the art data transfer solutions. Servers designed to meet the high throughput requirements at 100 Gbps and beyond, over long network paths. Data transfer applications will use low latency protocols such as NVMe over Fabric (NVMeoF), combined with RoCE as an underlay providing remote data memory access (RDMA), in order to achieve the maximum disk read or write throughput while minimizing the CPU load.

Partner Demonstrations:

The NRE demonstrations hosted at the Caltech Booth 534 include:

- 1) SC19-NRE-013 – SENSE: Tom Lehman, Chin Guok, Harvey Newman, Justas Balcas, John MacAuley, Xi Yang
- 2) SC19-NRE-019 - Global Petascale to Exascale Workflows for Data Intensive Science: Harvey Newman, Raimondas Sirvinskas, Joseph Chiu, Azher Mughal, Javier Duarte, Phil Harris
- 3) SC19-NRE-022 - Multi-domain, Joint Path and Resource Representation and Orchestration: Richard Yang, Jensen Zhang, Qiao Zhang
- 4) SC19-NRE-020 - LHC Multi-Resource, Multi-Domain Orchestration via AutoGOLE, SENSE, and TIFR: Gerben van Malenstein, Tom Lehman, Harvey Newman, Brij Jashal, Chin Guok, Justas Balcas, John MacAuley, Xi Yang
- 5) SC19-NRE-023 - LSST and AmLightExpress/Protect: Julio Ibarra, Jeronimo Bezerra, Heidi Morgan
- 6) SC19-NRE-024 - 400GE Ring: Azher Mughal, Harvey Newman
- 7) SC19-NRE-035 – SANDIE: Edmund Yeh, Ran Liu, Harvey Newman, Raimondas Sirvinskas, Catalin Iordache

