SC22 Network Research Exhibition:
Caltech Booth 2820 Demonstrations
Hosting a Wide Range of NREs

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

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

Abstract

We will demonstrate a wide range 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, and emerging technologies and concepts in programmable networks which are designed to meet the challenges faced by leading edge data intensive experimental programs in high energy physics, astrophysics, genomics and other fields of data intensive science to clear the path to the next round of discoveries.

These advances, being developed within the Global Network Advancement Group (GNA-G) framework and its SENSE/AutoGOLE and Data Intensive Sciences (DIS) working groups, the Global Research Platform, and its many science, computer science and R&E network teams aim to address the key challenges including: (1) global data distribution, processing, access and analysis, (2) the coordinated use of massive but still limited computing, storage and network resources, (3) coordinated operation and collaboration within global scientific enterprises each encompassing hundreds to thousands of scientists, and (4) enabling the science programs to make efficient use of the available network and site infrastructures, while simultaneously accommodating and in concert with the network operations required to supporting the worldwide academic and research community across national, regional and transoceanic boundaries.

The major programs being highlighted include the Large Hadron Collider (LHC), BioGenome and the human genome as well as astrophysics projects such as the SKA and the Vera Rubin Observatory and others. We will also highlight the latest developments in bottleneck structures and analysis for real-time congestion resolution others well as state of the art 5G "holodeck" applications running over a segment-routed transcontinental links between the UCSD and NYU campuses. The 5G applications also will help pave the way for other real-time highly differentiated services and applications including extended reality (XR) and vehicle to anything (V2X) as well as the future transition to 6G wireless applications.

Several of our teams' demonstrations include state of the art programmable open source network operating systems, layer 2 multi-domain virtual circuit overlays coupled to extensive structured use of IPv6 address spaces, segment routing approaches to efficient flow distribution, steering and load balancing, and stateful traffic engineering and workflow acceleration taking network segment and site state, policy and priority and real-time SLAs into account.

The teams' several branches of advanced development, driven by a diverse set of challenging use cases within the GNA-G framework and associated R&D projects, and enabled by the National Research Platform and Global Research Platform, are embedded and interoperate within an emerging "composable architecture" of subsystems, components and interfaces organized into several areas:

- **Visibility:** monitoring and information tracking and management including IETF ALTO/OpenALTO, BGP-LS, sFlow/NetFlow, Perfsonar, Traceroute, Reservoir Labs/G2 congestion information, Kubernetes statistics, LibreNMS, P4/Inband telemetry
- **Intelligence:** NetPredict, Hecate, RL-G2, Yale Bilevel optimization, stateful decisions using composable metrics (policy, priority, network- and site-state, SLA constraints, responses to 'events' at sites and in the networks, ...), Coral, Elastiflow/Elastic Stack
- **Controllability:** SENSE/OpenNSA/AutoGOLE, P4/PINS, segment routing (SRv6, PolKA), BGP/PCEP
- **Network OSes and Tools:** SONIC, GEANT RARE/freeRtr, Calico VPP, Bstruct-Mininet environment, ...
- **Orchestration**: SENSE, Kubernetes, dedicated code and APIs for interoperation and progressive integration

An overarching 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.

The cornerstone system concepts, components and developments demonstrated will include:

- Integrated operations and orchestrated management of resources: interworking with and advancing the site (Site-RM) and network resource managers (Network-RM) developed in the SENSE program.
- 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.

- The integration of Qualcomm Technology's GradientGraph (G2) with 5G / OpenALTO / SRv6 will be used to demonstrate how applications including XR, auto/V2X, and IoT can benefit from the intelligent routing, rate limiting, and service placement decisions computed by G2. We will also demonstrate how the integration of G2 with science networks through the integration of the LHC Rucio / FTS data management system, AutoGOLE / SENSE virtual circuit and orchestration services / and OpenALTO comprehensive monitoring may be used towards optimizing large-scale data transfers to share data from the Large Hadron Collider at CERN and scientists across the globe.
- 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 NDN for Data Intensive Science Experiments (N-DISE)
- A paragon network at several our partners' sites Composed of P4 programmable devices, including Tofino and Tofino2-based switches, Smart NICs and Xilinx FPGA-based 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 NVMe server systems to match a 400G network core, as well as 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 and in ESNet based on the XRootD federated access protocol; extension of the cache concept to more remote sites such as Fermilab, Nebraska and Vanderbilt.
- Creation of an overlay network with PolKA tunnels forming virtual circuits to validate the data-intensive transfer over 10G and 100G+ as a proof-of-principle of PolKA source routing mechanisms. The overlay will be built by integrating persistent resources from GNA-G AutoGOLE/SENSE and GEANT RARE testbeds. Underlay congestion can be detected by tunnel monitoring and signalized to the overlay so that the traffic is steered from congested tunnels to other paths. Comparisons between segment routing and PolKA regarding controllability and performance metrics are also planned. From the application's point of view, PolKA full deployment enables us to meet the extreme traffic engineering demanded by data-intensive sciences by offering a new range of network functionalities such as: multipath routing, in-network telemetry and proof-of-transit with path attributes to support higher level stateful traffic engineering decisions.
- Network traffic prediction and engineering optimizations using the latest graph neural network and other emerging deep learning methods, developed by ESnet's Hecate /DeepRoute project.

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, Nebraska, Vanderbilt, 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 both edge and in-network caches to increase data access and processing efficiency.
- **SENSE/AutoGOLE:** The GNA-G SENSE/AutoGOLE WG demonstration will present key technologies, methods and a system of dynamic Layer 2 and Layer 3 virtual circuit services to meet the challenges and address the requirements of the largest data intensive science programs, including the Large Hadron Collider (LHC) the Vera Rubin Observatory and programs challenges and programs in many other disciplines. The services are designed to support multiple petabyte transactions across a global footprint, represented by a persistent testbed spanning the US, Europe, Asia Pacific and Latin American regions. Two key components that will be demonstrated are AutoGOLE services and SENSE (see below).
- **Global Ring and KAUST:**
  These demonstrations will also showcase the power of collaboration in the global research and education network (REN) community. The demo will have KAUST as a collaborator in the Asia-Pacific Global Ring (AER [*]), closing the global ring by interconnecting Amsterdam to Singapore, and then onto Los Angeles with support from partners including SingaREN, JGN-X and APOnet.

  [*] See Fast and stable Network Ring between Asia-Pacific and Europe | AARNet
- **400 Gbps Next Generation Wide Area Networks:** Science research collaborations continuing to prepare for increased network capacity, e.g., transitioning from 100 Gbps paths to 400 Gbps, 800 Gbps and beyond. An international collaboration led by the StarLight/ Caltech/ UCSD/ NRP/GRP consortium, with many will demonstrate the architecture, services and techniques to enable efficient management of high volume science data streams within multi-hundred Gbps channels.

- **Global Research Platform (GRP:** An international collaboration will demonstrate services and capabilities of the prototype Global Research Platform, an international Science DMZ, optimized for large scale world-wide science projects, especially those based on instruments producing high volumes of data.
- **Software Defined Exchanges (SDXs):** Several international open exchanges that are developing capabilities for Software Defined Exchanges (SDXs) will showcase the capabilities of these facilities for providing highly programmable capabilities for international science data transport.
- **AmLight Express and Protect (AmLight-ExP)** in support of the SENSE/AutoGOLE Demonstrations 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 Monet submarine cable, and its 100G ring network that interconnects the research and education communities in the U.S. and South America.
- **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 SC22 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
- **Integration of OpenALTO, SRv6 and Qualcomm Technologies' GradientGraph**

  We intend to demonstrate the integration of OpenALTO, SRv6 and Qualcomm Technologies' GradientGraph, showing how applications including XR, auto/V2X, and holographic telepresence can benefit from the intelligent routing, rate limiting, and service placement decisions computed by the GradientGraph platform. We also intend to demonstrate the integration of OpenALTO and GradientGraph with science networks, Rucio, FTS, AutoGOLE, SENSE and OpenALTO towards optimizing large-scale data transfers from the Large Hadron Collider at CERN to scientists across the globe.
- **5G/Edge Computing Application Performance Optimization**. UCSD, the Pacific Research Platform, Caltech, Yale and Qualcomm Technologies will attempt to demonstrate the first end-to-end integrated traffic optimization framework based on bottleneck structure analysis for 5G applications. This intended demonstration will focus on showing the integration of Qualcomm technology's GradientGraph with the IETF ALTO open standard, to support the optimization of edge computing applications such as XR, holographic telepresence, and vehicle networks. Holodecks at UCSD and NYU will be interconnected across the CENIC and NYSERNet regional networks via a transcontinental AP-REX link.
- **High-Performance Routing of Science Network Traffic**. LHC/CERN, Caltech, the Pacific Research Platform, ESnet, Yale and Qualcomm Technologies will attempt to demonstrate the first integration of the IETF ALTO open standard with the Rucio, FTS applications and AutoGOLE/SENSE to optimize the steering of large-scale data transfers through global science networks. Based on the open-source project OpenALTO, the IETF standard exposes highly detailed network state information that the application can use to optimize its performance. This intended demonstration will show how Qualcomm Technologies' GradientGraph platform leverages this information to compute optimized application placement and flow routing strategies.
- **N-DISE: Named Data Networking for Data Intensive Science Experiments:**The NDN for Data Intensive Science Experiments (N-DISE) project aims to accelerate the pace of breakthroughs and innovations in data-intensive science fields such as the Large Hadron Collider (LHC) high energy physics program and the BioGenome and human genome projects. Based on Named Data Networking (NDN), a data-centric future Internet architecture, N-DISE will deploy and commission a highly efficient and field-tested petascale data distribution, caching, access and analysis system serving major science programs. The N-DISE project will build on recently developed high-throughput NDN caching and forwarding methods, containerization techniques, leverage the integration of NDN and SDN systems concepts and algorithms with the mainstream data distribution, processing, and management systems of CMS, as well as the integration with Field Programmable Gate Arrays (FPGA) acceleration subsystems, to produce a system capable of delivering LHC and genomic data over a wide area network at throughputs approaching 100 Gbits per second,

while dramatically decreasing download times. N-DISE will leverage existing infrastructure and build an enhanced testbed with high performance NDN data cache servers at participating institutions.

The N-DISE demonstration is designed to exhibit improved performance of the N-DISE system for workflow acceleration within large-scale data-intensive programs such as the LHC high energy physics, BioGenome and human genome programs. To achieve high performance, the demonstration will leverage the following key components: (1) the transparent integration of NDN with the current CMS software stack via an NDN based XRootD Open Storage System plugin, (2) joint caching and multipath forwarding capabilities of the VIP algorithm, (3) integration with FPGA acceleration subsystems, (4) SDN support for NDN through the work of the Global Network Advancement Group (GNA-G) and its AutoGOLE/SENSE and Data Intensive Sciences Working Group.

- **High performance networking with the Bella Link and the Sao Paulo Backbone SP linking 9 universities (Rednesp, RNP, UNESP and USP):**

With the recent availability of an L2 circuit linking the University of São Paulo State (UNESP) to CERN in Geneva, through the Bella Link, Brazil (and São Paulo State) now has three 100 gbps academic links to the USA and to Europe. This new circuit is supposed to present low latency, and that should open new opportunities for collaboration among Brazilian and European academic institutions. All links pass through the SP4 Equinix datacenter in the greater Sao Paulo area. At the SP4 datacenter, rednesp (Research and Education Network at Sao Paulo) has a node called SouthernLight which is part of the GNA Autogole/SENSE testbed.

At the same time, the Rednesp Sao Paulo regional network in finishing the "Backbone SP" connecting 9 important universities in Sao Paulo state. In this demo, we intend to compare properties such as effective bandwidth, latency to different continents and jitter using different paths of the 3 intercontinental links as well as the integration of the "Backbone SP" to international connections. We also expect to show results and metrics of parallel computations using some of the national and international computing facilities connected to this infrastructure.

- **Coral: Scalable Data Plane Checking via Distributed, On-Device Verification**

Data plane verification (DPV) is important for finding network errors, and therefore a fundamental pillar for achieving consistently operating, autonomous-driving, high-performance science networks. Current DPV tools employ a centralized architecture, where a server collects the data planes of all devices and verifies them. Despite substantial efforts on accelerating DPV, this centralized architecture is inherently non-scalable. To tackle the scalability challenge of DPV, the team of Xiamen University, China, circumvents the scalability bottleneck of centralized design and design Coral, a distributed, on-device DPV framework.

The key insight of Coral is that DPV can be transformed into a counting problem on a directed acyclic graph, which can be naturally decomposed into lightweight tasks executed at network devices, enabling scalability. Coral consists of (1) a declarative requirement specification language, (2) a planner that employs a novel data structure DVNet to systematically decompose global verification into on-device counting tasks, and (3) a distributed verification (DV) protocol that specifies how on-device verifiers communicate task results efficiently to collaboratively verify the requirements. During SC'22, we will demonstrate, through both a testbed reconstruction of the Internet2 WAN environment and large-scale simulations of WAN/LAN/DC with real-world datasets, that Coral consistently achieves scalable DPV under various scenarios. A preliminary manuscript of Coral can be found at [1] and a set of small-scale demos can be found at [2].

[1] Qiao Xiang et al., Switch as a Verifier: Toward Scalable Data Plane Checking via Distributed, On-Device Verification, https://arxiv.org/abs/2205.07808

[2] Qiao Xiang et al., Coral System Functionality Demonstration, http://distributeddpvdemo.tech/

Resources

The partners will use approximately 4 400G and other 100G wide area links coming into SC22, and the available on- floor and DCI links to the StarLight and Caltech booths. A 1.2 Tbps wide area network including 400GE switches from Arista, Edgecore , Dell and Juniper, and links among Starlight, McLean, Caltech and the SC22 venue in Dallas, using Waveserver Ai's and DWDM to SCinet. Together with servers with 100, 200 and potentially 400G smart NICs and NVMe storage systems, and programmable switch routers at the sites running SONIC and/or freeRtr as well as fixed function switches in a global network architecture, this will host the wide range of demonstrations summarized in this and its partner NREs as well as others.

As of this writing the GEANT/RARE freeRtr is running in production mode on a 400GE Tofino2 Edgecore switch in the SWITCH Swiss national R&E network. It will be demonstrated over the wide area as soon as a 400GE path connecting Tofino2 switches are available. Server to server transfers between 400G DTNs also will be shown as soon as they are available. If not by SC22, then shortly afterward.

Partners: **Group Leads and Participants, by Team**

- **Caltech HEP:** Harvey Newman (newman@hep.caltech.edu), Justas Balcas (jbalcas@caltech.edu), Raimondas Sirvinskas (raimis.sirvis@gmail.com), Catalin Iordache, Preeti Bhat, Andres Moya, Sravya Uppalapati
- **Caltech IMSS:** Jin Chang (jin.chang@caltech.edu), Azher Mughal (azher@caltech.edu), Dawn Boyd, Larry Watanabe, Don S. Williams
- **UCSD/SDSC/NRP:** Frank Wuerthwein (fkw888@gmail.com), Tom deFanti (tdefanti@eng.ucsd.edu), Larry Smarr, John Graham, Tom Hutton (hutton@ucsd.edu), Dima Mishin, Jonathan Guiang, Diego Davila, Igor Sfiligoi, Aashay Arora,
- **Yale:** Richard Yang (yry@cs.yale.edu), Jensen Zhang
- **Northeastern University:** Edmund Yeh (eyeh@ece.neu.edu), Yuanhao Wu, Volkan Mutlu, Yuezhou Liu
- **Tennessee Tech:** Susmit Shannigrahi (sshannigrahi@tntech.edu), Sankalpa Timilsina
- **UCLA:** Lixia Zhang (lixia@cs.ucla.edu), Jason Cong (cong@cs.ucla.edu), Michael Lo, Sichen Song
- **Fermilab:** Oliver Gutsche (gutsche@fnal.gov), Phil Demar (demar@fnal.gov)
- **ESnet:** Inder Monga (imonga@es.net), Chin Guok (chin@es.net), Tom Lehman (tlehman@es.net), John MacAuley, Xi Yang, Justas Balcas, Mariam Kiran
- **LBNL/NERSC:** Alex Sim (asim@lbl.gov)
- **Nebraska/UNL:** Garhan Attenbury (garhan.attebury@unl.edu)
- **Vanderbilt:** Andrew Melo, andrew.m.melo@accre.vanderbilt.edu
- **CERN:** Edoardo Martelli (edoardo.martelli@cern.ch), Carmen Misa (carmen.misa@cern.ch)
- **Qualcomm Gradient Graph:** Jordi Ros-Giralt (jros@qti.qualcomm.com), Sruthi Yellamraju
- **UFES:** Magnos Martinello, Moises R.N. Ribeiro (moises@ele.ufes.br), Christina Dominicini (cristina.dominicini@ifes.edu.br), Everson Borges (everson@ifes.edu.br), Rafael Guimaraes
- **RNP:** Marcos Schwarz (marcos.schwarz@rnp.br), Leandro Ciuffo (leandro.ciuffo@rnp.br)
- **RENATER/GEANT/RARE:** Frédéric LOUI (frederic.loui@renater.fr)
- **UNESP (SPRACE NCC UNESP):** Sergio Novaes (Sergio.Novaes@cern.ch),Rogerio Iope (rogerio.iope@unesp.br)
- **Rednesp:** Antonio J F Francisco, Ney Lemke **(UNESP)** (ney.lemke@unesp.br), Carlos Antonio Ruggiero **(USP)** (toto@ifsc.usp.br), Jorge Marcos de Almeida (USP) (jorge@usp.br)
- **UERJ:** Alberto Santoro (Alberto.Santoro@cern.ch)
- **George Mason/BRIDGES:** Bijan Jabbari (bjabbari@gmu.edu), Jerry Sobieski, Liang Zhang
- **Xiamen:** Qiao Xiang (xiangq27@gmail.com), Chenyang Huang, Ridi Wen, Yuxin Wang, Jiwu shu
- **Colorado State:** Chengyu Fan (chengy.fan@gmail.com)

- **CENIC:** Louis Fox (lfox@cenic.org), Sana Bellamine (sbellamine@cenic.org), Tony Nguyen
- **Pacific Wave/USC:** Celeste Anderson (celestea@usc.edu)
- **Starlight/MREN/iCAIR:** Joe Mambretti (j-mambretti@northwestern.edu), Jim Chen, Fei Yeh
- **Internet2:** Christian Todorov, (ctodorov@internet2.edu), Rob Vietzke (rvietzke@internet2.edu)
- **AmLight/FIU:** Julio Ibarra (Julio@fiu.edu), Jeronimo Bezerra, Vasilka Chergarova
- **Amlight/ISI:** Heidi Morgan (hlmorgan@isi.edu)
- **Ciena:** Scott Kohlert (skohlert@ciena.com), Rod Wilson
- **KISTI:** Buseung Cho (bscho@kisti.re.kr). Jeonghoon Moon
- **CANARIE:** Thomas Tam (Thomas.Tam@canarie.ca)
- **KAUST:** Alex Moura (alex.moura@kaust.edu.sa), Kevin Sale
- **DE-KIT:** Bruno Hoeft (bruno.hoeft@kit.edu)
- **JPL:** Lee, Carlyn-Ann (Carlyn-Ann.Lee@jpl.nasa.gov)
- **NIST:** Davide Pasavento (davide.pasavento@nist.gov)
- **Hawaii:** Chris Zane (czane@hawaii.edu)
- **SURFNet:** Hans Trompert (hans.trompert@surfnet.nl)
- **CESNET:** Michal Hažlinský, (hazlinsky@cesnet.cz)
- **Clemson:** Cole McKnight (cbmckni@g.clemson.edu)
- **NCHC/TAWREN:** Li-Chi Ku, lku@narlabs.org.tw
- **GNA-G/AArNet:** David Wilde (David.Wilde@aarnet.edu.au)
- **GNA-G AutoGOLE / SENSE WG Members**: https://www.gna-g.net/join-working-group/autogole-sense
- **GNA-G Data Intense Science WG Members:** https://www.gna-g.net/join-working-group/data-intensive-science/

Partner NRE Demonstrations: To be Added

The NRE demonstrations hosted at or partnering with the Caltech Booth 2820 include:


CALIFORNIA INSTITUTE OF TECHNOLOGY
1891
US Large Hadron Collider
US LUA
Users Association
ESnet
Energy Sciences Network
CENIC
CONNECTING CALIFORNIA 1997-2022
25
US CMS
ATLAS
LHC
Fermilab
CERN
INTERNET 2
NRP
NATIONAL RESEARCH PLATFORM
AMLIGHT
LUX ET VERITAS
YALE
UNIVERSITY OF CALIFORNIA
SAN DIEGO
NORTHEASTERN UNIVERSITY
BOSTON, MASSACHUSETTS
GridUNESP
iCAIR
International Center for Advanced Internet Research
STARLIGHT
The Optical STAR TAP
PACIFIC NORTHWEST GIGAPOP
PACIFIC WAVE
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
DOCETE OMNES GENTES
UFES
RNP
UERJ
UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO
KIT
Karlsruher Institut für Technologie
GEORGE MASON UNIVERSITY
50
1972-2022
GÉANT
Qualcomm
KAUST
GNA-G
Global Network Advancement Group
SCinet
30 years
NSF
+ INDUSTRY