Global Petascale to Exascale Workflows
Next Generation Network-Integrated System for Data Intensive Sciences

Booth 2820

SC22 Network Research Exhibition
NRE-19 and Partner NREs: Booth 2820

See https://www.dropbox.com/s/1opcg4vjlhjk6g5/NextGenDiSSystems_hbn111222.pptx?dl=0
Global Petascale to Exascale Workflows for Data Intensive Sciences Accelerated by Next Generation Programmable Network Architectures and Machine Learning Applications


- **A Vast Partnership** of Science and Computer Science Teams, R&E Networks and R&D Projects; Convened by the GNA-G DIS WG; with GRP, AmRP, NRP
- **Mission:** Demonstrate the road ahead
  - To meet the challenges faced by leading-edge data intensive programs in high energy physics, astrophysics, genomics and other fields of data intensive science; *Compatible with other use*
  - Clearing the path to the next round of discoveries
- **Demonstrating a wide range of latest advances in:**
  - Software defined and Terabit/sec networks
  - Intelligent global operations and monitoring systems
  - Workflow optimization methodologies with real time analytics
  - State of the art long distance data transfer methods and tools, local and metro optical networks and server designs
  - Emerging technologies and concepts in programmable networks and global-scale distributed systems
- **Hallmarks:** Progressive multidomain integration; compatibility internal + external; A comprehensive systems-level approach
GNA-G DIS WG: Worldwide Partnerships at SC22

Partners: Group Leads and Participants, by Team

Caltech HEP: Harvey Newman (newman@hep.caltech.edu), Justas Balcas (jbalcas@caltech.edu), Raimondas Svirninkas (raimis.sirvys@gmail.com), Catalin Iordache, Preeti Bhat, Andres Moya, Sravya Uppalapati

Caltech IMSS: Jun Chang (jun.chang@caltech.edu), Acher Mughal (acher@caltech.edu), Dawn Boyd, Larry Watanabe, Don S. Williams

UCSD/SDSC/NRP: Frank Wurthevin (fkwe888@gmail.com), Tom deFanti (tdefanti@eng.ucsd.edu), Larry Smarr, John Graham, Tom Hutton (tutton@ucsd.edu), Dima Moshin, Jonathan Tung, Diego Davila, Igor Silivial, Aashish Arora

Yale: Richard Yang (ryang@cs.yale.edu), Jensen Zhang

Northeastern University: Edmund Yeh (eyeh@ece.neu.edu), Yuanhao Wu, Volkam Mustu, Youzhou Liu

Tennessee Tech: Susmit Shankar (sshankar@dntech.edu), Sanalpa Timalsina

UCLA: Xia Zhang (xia@cs.ucla.edu), Jason Cong (jasonCong@cs.ucla.edu), Michael Lo, Sichen Song

Formlab: Oliver Gustie (gustie@fnal.gov), Phil Demar (demar@fnal.gov)

E3net: Inder Monga (imonga@es.net), Chin Guoq (chin@es.net), Tom Lehman (lehman@es.net), John MacAuley, Xi Yang, Justas Balcas, Mariam Kiar

LBNL/LIERS: Alex Sim (asim@lbl.gov)

Nebraska/UNL: Garth Attenbury (garth.attenbury@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@qualcomm.com), Sriuthi Yellamraju

UFES: Magnus Martinello, Moises R.N. Ribeiro (moises@ufes.br), Christina Domininici (christina.domininici@unesp.br), Everson Borges (everson@ufes.br), Rafael Guimaraes

RNP: Marcos Schwarz (marcos.schwarz@rnp.br), Leandro Ciuffo (leandro.ciuffo@rnp.br)

RENATER/GEANT/RARE: Frederic LOU (frederic.lou@renater.fr)

UNESP (SPRACE NCC UNESP): Sergio Novais (sergio.novais@cern.ch), Rogerio Iope (rogerio.iope@unesp.br)

Redesp: Antonio J F Francisco, Ney Lemke (neylemke@unesp.br), Carlos Antonio Ruggiero (usp) (toto@ifsc.usp.br), Jorge Marcos de Almeida (usp) (jorge@usp.br)

UERJ: Alberto Santoro (alberto.santoro@uern.br)

George Mason/BRIDGES: Bijan Jabbari (bjabbari@gmu.edu), Jerry Sobieski, Liang Zhang

Xiamen: Qiao Xiang (xiangq27@gmail.com), Chenguang Huang, Riji Wen, Xuejin Wang, Jiuwu Shu

Colorado State: Chengyu Fan (chengyu.fan@gmail.com)

CENIC: Louis Fox (fox@cenic.org), Sana Bellamine (sblamine@cenic.org), Tony Nguyen

Pacific Wave/USC: Celeste Anderson (celeste@usc.edu)

Starlight/MRN/CAIR: 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, Vasila Cherghova

AmLight/ISI: Heidi Morgan (hmmorgan@isi.edu)

Ciena: Scott Kohlert (skohlert@ciena.com), Rod Wilson

KIST/KREONET: Byungung Cho (bcho@kisti.re.kr), Mazahir Hussain, Tergel Munkhbat

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 (cane@hawaii.edu)

SURFNet: Hans Trompert (hans.trompert@surfnet.nl)

CESNET: Michal Hazlinsky (mazlinsky@cesnet.cz)

Clemson: Cole McKnight (clemson@clemson.edu)

NCHC/TAWREN: Li-Chi Ku, liku@nch.org.tw

GNA-G/ANET: David Wilde (David.Wilde@anet.edu.au)

GNA-G AutoGOLE / SENSE WG Members: https://www.gna-g.net/join-working-group/auto-gole-sense

GNA-G Data Intensive Science WG Members: https://www.gna-g.net/join-working-group/data-intensive-science

STORDIS: Waldemar.Scheck@stordis.com

SC22 and Beyond: Persistent Development on a Global Testbed; Trajectory to Production
Mission: **To Support global research and education using the technology, infrastructures and investments of its participants**

The GNA-G exists to bring together researchers, National Research and Education Networks (NRENs), Global eXchange Point (GXP) operators, regionals and other R&E providers, in developing a common global infrastructure to support the needs.
Global Petascale to Exascale Workflows for Data Intensive Sciences

- Advances Embedded and Interoperate within a ‘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, Qualcomm Gradient Graph congestion information, Kubernetes statistics, LibreNMS, P4/Inband telemetry
  - **Intelligence**: Stateful decisions using composable metrics (policy, priority, network- and site-state, SLA constraints, responses to ‘events’ at sites and in the networks, ...), using NetPredict, Hecate, G2, Yale Bilevel optimization, Coral, Elastiflow/Elastic Stack
  - **Controllability**: SENSE/OpenNSA/AutoGOLE, P4/PINS, segment routing with SRv6 and/or PoIKA, BGP/PCEP
  - **Network OSes and Tools**: GEANT RARE/freeRtr, SONIC, Calico VPP, Bstruct-Mininet environment, ...
  - **Orchestration**: SENSE, Kubernetes (+k8s namespace), dedicated code and APIs for interoperation and progressive integration
Next Generation Network-Integrated System

- Top Line Message: In order to address the challenges and meet the needs, we need a new dynamic and adaptive software-driven system, which
  - Coordinates worldwide networks as a first class resource along with computing and storage, across multiple domains
  - Simultaneously supports the LHC experiments, other major DIS programs and the larger worldwide academic and research community
  - Systems design approach: A virtualized global dynamic fabric that flexibly allocates, balances and conserves the available network resources
  - Negotiating with site systems that aim to accelerate workflow; Use of ML
  - Builds on ongoing R&D projects: from regional caches/data lakes to intelligent control and data planes to ML-based optimization
    [E.g. SENSE/AutoGOLE, NOTED, ESNet HT, GEANT/RARE, AmLight, Fabric, Bridges; NetPredict, DeepRoute, Hecate, ALTO, PolKA ...]
  - A key milestone: integration of SENSE + network services with FTS & Rucio
  - We are also leveraging the worldwide move towards a fully programmable ecosystem of networks and end-systems (P4, PINS; SRv6; PolKA), plus operations platforms (OSG, NRP; global SENSE Testbed; BRIDGES)
  - The LHC experiments together with the WLCG, the GNA-G and its Working Groups, and the worldwide R&E network community, are the key players
  - Directions also taken up by other programs: LBNF/DUNE, VRO, SKA
<table>
<thead>
<tr>
<th>NRE</th>
<th>Name</th>
<th>Email</th>
<th>Demonstration</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRE-001</td>
<td>Edmund Yeh</td>
<td><a href="mailto:eyeh@ece.neu.edu">eyeh@ece.neu.edu</a></td>
<td>N-DISE: NDN for Data Intensive Science Experiments</td>
</tr>
<tr>
<td>NRE-004</td>
<td>Joe Mambretti</td>
<td><a href="mailto:j-mambretti@northwestern.edu">j-mambretti@northwestern.edu</a></td>
<td>1.2 Tbps Services WAN Services: Architecture, Technology and Control Systems</td>
</tr>
<tr>
<td>NRE-005</td>
<td>Joe Mambretti</td>
<td><a href="mailto:j-mambretti@northwestern.edu">j-mambretti@northwestern.edu</a></td>
<td>400 Gbps E2E WAN Services: Architecture, Technology and Control Systems</td>
</tr>
<tr>
<td>NRE-007</td>
<td>Edoardo Martelli</td>
<td><a href="mailto:edoardo.martelli@cern.ch">edoardo.martelli@cern.ch</a></td>
<td>LHC Networking And NOTED</td>
</tr>
<tr>
<td>NRE-008</td>
<td>Joe Mambretti</td>
<td><a href="mailto:j-mambretti@northwestern.edu">j-mambretti@northwestern.edu</a></td>
<td>IRNC Software Defined Exchange (SDX) Multi-Services for Petascale Science</td>
</tr>
<tr>
<td>NRE-009</td>
<td>Jim Chen</td>
<td><a href="mailto:jim-chen@northwestern.edu">jim-chen@northwestern.edu</a></td>
<td>High Speed Network with International P4 Experimental Networks for The Global Research Platform and Other Research Platforms</td>
</tr>
<tr>
<td>NRE-010</td>
<td>Magnos Martinello</td>
<td><a href="mailto:magnos.martinello@ufes.br">magnos.martinello@ufes.br</a></td>
<td>Demonstrating PolKA Routing Approach to Support Traffic Engineering for Data-intensive Science</td>
</tr>
<tr>
<td>NRE-011</td>
<td>Qiao Xiang</td>
<td><a href="mailto:xiangq27@gmail.com">xiangq27@gmail.com</a></td>
<td>Coral: Fast Data Plane Verification for Large-Scale Science Networks via Distributed, On-Device Verification</td>
</tr>
<tr>
<td>NRE-013</td>
<td>Tom Lehman</td>
<td><a href="mailto:tlehman@es.net">tlehman@es.net</a></td>
<td>AutoGOLE/SENSE: End-to-End Network Services and Workflow Integration</td>
</tr>
<tr>
<td>NRE-015</td>
<td>Tom Lehman</td>
<td><a href="mailto:tlehman@es.net">tlehman@es.net</a></td>
<td>SENSE and Rucio/FTS/XRootD Interoperation</td>
</tr>
<tr>
<td>NRE-016</td>
<td>Marcos Schwarz</td>
<td><a href="mailto:marcos.schwarz@rnp.br">marcos.schwarz@rnp.br</a></td>
<td>Programmable Networking with P4, GEANT RARE/freeRtr and SONIC/PINS</td>
</tr>
</tbody>
</table>
Global Petascale to Exascale Workflows for Data Intensive Sciences

Development Trajectory: Parallel developments + mission-driven progressive interfacing and system-level integration

Overarching Concept: Consistent Network Operations:
- Stable load balanced high throughput workflows cross optimally chosen network paths
- Provided by autonomous site-resident services dynamically interacting with network-resident services
- Up to preset or flexible *high water marks* to accommodate other traffic
- Responding to (or negotiating with) site demands from the science programs’ principal data distribution and management systems

Data Center Analogue for Networks
- Classes of “Work” (work = transfers, or overall workflow), defined by task parameters and/or priority and policy
- Adjust rate of progress in each class to respond to network or site state changes, and “events”
- Moderate/balance the rates among the classes to optimize a multivariate objective function with constraints
P4 Tofino + Tofino2 + SONIC
Programmable Global Persistent Testbed

22 Active GNA-G/RARE P4 Testbed Sites/Devices
- Caltech, Pasadena-US: 4 x FreeRtr/P4+SONIC
- CERN, Geneva-CH: FreeRtr/P4
- FIU, Miami-US: FreeRtr/P4
- GEANT, Amsterdam-NL: FreeRtr/P4
- GEANT, Budapest-HU: FreeRtr/P4
- GEANT, Frankfurt-DE: FreeRtr/P4
- GEANT, Paris-FR: FreeRtr/DPDK
- GEANT, Poznan-PL: FreeRtr/P4
- HEAnet, Dublin-IE: FreeRtr/P4
- RENATER, Paris-FR: FreeRtr/P4
- RNP, Rio de Janeiro-BR; FreeRtr/P4
- SouthernLight (FIU/Red Clara/Rednesp/RNP), São Paulo-BR: FreeRtr/P4

- StarLight, Chicago-US: FreeRtr/P4
- SWITCH, Geneva-CH: FreeRtr/P4
- TCD, Dublin-IE: FreeRtr/P4
- Tennessee Tech, Cookeville-US: FreeRtr/P4
- UFES, Vitória-BR: FreeRtr/P4
- UMd, College Park, Maryland-US: FreeRtr/P4

+ 7 Sites in October – November (by SC22):
- JISC, London-UK: FreeRtr/P4
- KAUST, Saudi Arabia: FreeRtr/DPDK
- KISTI, South Korea: SONIC/P4
- RNP, Rio de Janeiro-BR+1 FreeRtr/P4
- SC22 Caltech Booth, Dallas-US: FreeRtr/P4
- UCSD, San Diego-US: SONiC/P4
- UFES, Vitória-BR: +1 FreeRtr/P4
PolKA: Polynomial Key-based Architecture for Source Routing

Creation of an overlay network with PolKA tunnels forming virtual circuits, integrating persistent resources from the GNA-G AutoGOLE/SENSE and GEANT RARE testbeds, validated using 100G+ transfers of science data.

- Underlay congestion will be detected by tunnel monitoring and signaled to the overlay so that the traffic is steered away from congested tunnels to other paths.
- Comparisons between SRv6 segment routing and PolKA regarding controllability and performance metrics.
- PolKA full deployment enables extreme traffic engineering demands of data-intensive sciences to be met, through 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.