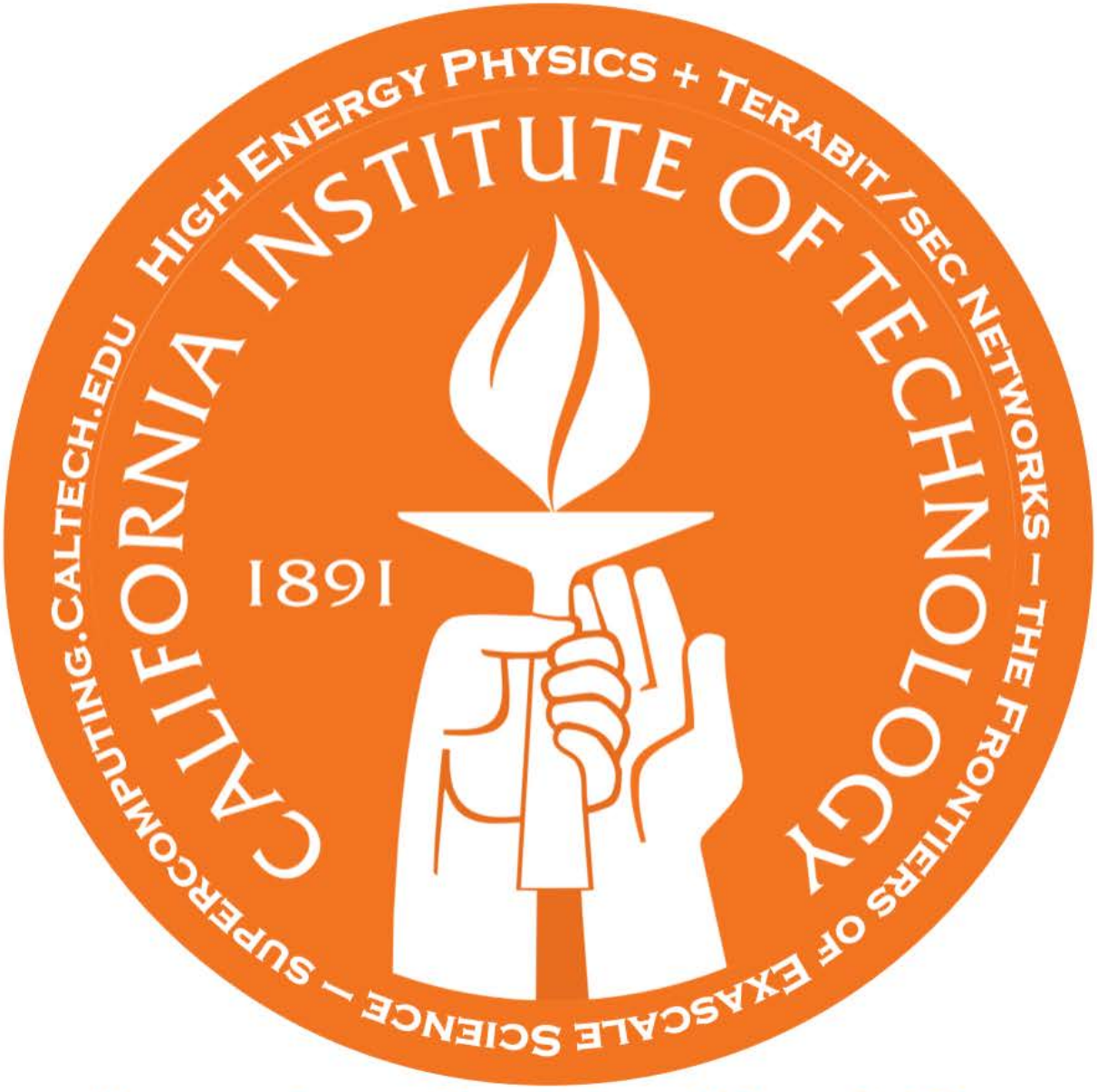
SC18

Global Petascale to Exascale Science Workflows



Accelerated by Next Gen SDN Architectures and Applications

SC18 Network Research Exhibition: Demonstration NRE-016

Global Petascale to Exascale Science Workflows

Accelerated by Next Generation SDN Architectures and Applications

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

Abstract

We will demonstrate several 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 science including the Large Hadron Collider (LHC), the Large Synoptic Space Telescope (LSST), the Linac Coherent Light Source (LCLS II), the Earth System Grid Federation and others.

Several of the SC18 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.

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 virtualization software stacks developed in the SENSE, PRP, AmLight, Kytos 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.

- **Consistent Operations Paradigm:** In association with the use cases cited, flows will be negotiated between the network-resident and site-resident resource and flow orchestration services, up to near-full capacity over a series of load-balanced paths, leading to a "consistent outcome".

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
- **LSST:** Real time 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 NCSA; operational and data quality traffic to SLAC, Tucson and other sites; LSST annual multi-petabyte Data Release emulation from NCSA to La Serena at rates consistent with those required for LSST operations.
- **AmLight Express and Protect (AmLight-ExP)** will support the LSST and LHC-related use cases 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.
- **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 SC18 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 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
- **SDN Federated Network Protocol (SFP):** Yale, Tongji, IBM, ARL and Caltech will demonstrate SFP, the first generic framework supporting fine-grained interdomain routing to address the fundamental mismatch between fine-grained SDN control and the coarse-grained BGP routing in inter-domain networks. Smart filtering and on-demand routing information will be used to address the scalability of fine-grained routing, in a collaborative network composed of both exhibitor booths and U.S. campus science networks.

- **Multi-domain Network State Abstraction (MNSA):** The groups involved in SFP development will demonstrate MNSA, a simple, novel, highly efficient multi-domain network resource discovery and representation system, to provide fine-grained, global network resource information through both SFP and the Application-Layer Traffic Optimization (ALTO) protocol, to support the high throughput needs of the afore-mentioned collaborative data intensive science programs. This demonstration will include (1) efficient discovery of available network resources with extreme low latency, (2) fairer allocations of networking resources in this collaborative network, (3) preservation of the privacy of information among the member networks, and (4) scaling to collaborative networks with hundreds of members.
- **High-Level, Unified SDN Programming (Magellan and Trident):** The groups involved in SFP and MNSA will further demonstrate a high-level, unified SDN and NFV programming framework consisting of two key components: (1) Magellan, which allows systematic, automatic compilation of high-level SDN programs into low-level SDN datapaths; and (2) Trident, which introduces key techniques including live variables, routes algebra, and 3-value logic programming to allow high-level, systematic integration of SDN and NFV, to achieve automatic updates.
- **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.
- **High Throughput and multi-GPU Clusters with Kubernetes:** UCSD and the Pacific Research Platform (PRP) will lead a group of collaborators demonstrating Kubernetes-based virtualized cluster management, and the design and applications of high throughput FIONA servers. The latest multi-GPU server designs being deployed at UCSD and other sites participating in the Cognitive Hardware And Software Ecosystem Community Infrastructure (CHASE-CI) project, which is building a cloud of hundreds of affordable Graphics Processing Units (GPUs), networked together with a variety of neural network machines to facilitate development of next generation cognitive computing, also will be presented.
- **400GE First Data Network:** USC together with Caltech, Starlight/NRL, Arista, SCinet/XNET, Mellanox and 2CRSI are demonstrating the first fully functional 400GE network, as illustrated in the Big Data SCinet diagram of the Caltech, USC and Starlight booths and their wide area connections below. Flows between Supermicro servers equipped with Mellanox ConnectX-5 VPI network interface cardsin the USC and Caltech booths achieved sustained flows close to 800 Gbps in the first round of tests on November 11, prior to SC18, as shown in the InMon traffic monitoring graph below.

Resources

The partners will use two dozen 100G and other wide area links coming into SC18, and several dozen . An inner 400GE core on the showfloor will be composed of a triangle linking the Caltech and USC booths and SCinet, extended to the Starlight booth, in addition to several other booths each connected with 100G links. Waveserver and other data center interconnects and DWDM to SCinet. The network layout highlighting the Caltech, USC and Starlight booths can be seen here:

http://tinyurl.com/SC18-JBDT a snapshot of which is given below.

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, SURFnet, and their science and network partner teams, with support from Ciena, Intel, Dell, 2CRSI, Premio/Echostreams, Arista, Mellanox, Color Chip, IBM and Rackspace.

Group Leads and Participants, by Team

- **Caltech HEP + LIGO:** Harvey Newman (newman@hep.caltech.edu), Justas Balcas, Shashwitha Putaswamy, Jean-Roch Vlimant, Joosep Pata, Catalin Iordache, Stuart Anderson
- **Caltech IMSS:** Jin Chang (jin.chang@caltech.edu), Dawn Boyd, Larry Watanabe, Don S. Williams
- **USC:** Celeste Anderson (celestea@usc.edu), Azher Mughal
- **LSST:** Jeff Kantor JKantor@lsst.org , Matt Kollross
- **AmLight/FIU:** Julio Ibarra (Julio@fiu.edu), Jeronimo Bezerra, Vinicius Arcanjo, Adil Zahir
- **Amlight/ISI:** Heidi Morgan (hlmorgan@isi.edu)
- **Yale/Tongi/IBM/ARL:** Richard Yang (yry@cs.yale.edu), Qiao Xiang, Jensen Zhang, X. Tony Wang, Dong Guo, Dennis Yu, May Wang, Christopher Leet, Shenshen Chen, Franck Le, Yeon-sup Lim, Yuki de Pourbaix, Vinod Mishra
- **Maryland:** Tom Lehman (tlehman@umd.edu), Xi Yang
- **UCSD/SDSC/PRP:** Tom deFanti (tdefanti@ucsd.edu), Larry Smarr, John Graham, Tom Hutton, Frank Wuerthwein, Phil Papadopoulos
- **ESnet:** Inder Monga (imonga@es.net), Chin Guok, John MacAuley
- **LBNL:** Alex Sim (asim@lbl.gov)
- **LBNL/NERSC:** Damian Hazen (dhazen@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)
- **SURFnet:** 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
- **Colorado State:** Christos Papadopoulos (christos@cs.colostate.edu), Susmit Shannigrahi
- **UCLA:** Lixia Zhang (lixia@cs.ucla.edu)
- **TIFR Mumbai:** Brij Jashal (brij.jashal@gmail.edu), Kajari Mazumdar
- **CENIC/Pacific Wave:** Dave Reese (dave@cenic.org), 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: 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 will demonstrate 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 SC 2018, we will demonstrate PB data transfers from NCSA to AURA in La Serena, Chile at rates consistent with those required for LSST operations.

(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) Interdomain Routing for SDN federation networks (Y.R. Yang, Q. Xiang, and Franck Le): There are multiple important settings where multiple networks interconnect to form collaborative networks. The de facto interconnection protocol used by these networks is BGP. The deployment of SDN in these BGP-interconnected networks, however, reveals a fundamental mismatch between the fine-grained control by SDN and the coarse-grained routing by BGP. Such a mismatch can lead to serious issues including unnecessary blackholes, unnecessary reduced reachability, and permanent forwarding loops that existing BGP policy routing frameworks do not address. We design SFP, the first fine-grained interdomain routing protocol and system that extends BGP with fine-grained routing, eliminating the aforementioned mismatch. In SFP, we develop a set of novel techniques, including smart filtering and on-demand routing information, to address the scalability of fine-grained routing. At SC18, we will demonstrate the full capacity of SFP in a collaborative network composed of both exhibitor booths and campus science networks across the United States, for supporting efficient messaging and processing of fine-grained interdomain routing information.

(b) Multi-domain Network Resource Discovery (Y.R. Yang, Q. Xiang, and Franck Le):

Many of today's premier science experiments, such as the Large Hadron Collider (LHC), rely on finely-tuned
workflows that coordinate geographically distributed resources (e.g. instrument, compute, storage) to

enable scientific discoveries. The key to supporting these distributed resource workflows is the ability to reserve and guarantee bandwidth across multiple network domains to facilitate predictable end-to-end network connectivity. In the last few years, a number of multi-domain network resource information, e.g., Network Service Interface (NSI), and reservation systems, e.g., the On-Demand Secure Circuits and Advance Reservation Systems (OSCARS) have been developed and deployed, driven by the demand and substantial benefits of providing predictable network resources. Such systems, however, are limited by the fact that they are based on coarse-grained or localized information, resulting in inefficiencies. To fill this gap, we design Explorer, a simple, novel, highly efficient multi-domain network resource discovery system to provide fine-grained, global network resource information, to support high-performance, collaborative data sciences. At SC18, we will demonstrate that Explorer can (1) efficiently discover available networking resources in a collaborative network, whose member networks are from both exhibitor booths and campus science networks across the United States, with extreme low latency, (2) allow fairer allocations of networking resources in this collaborative network, (3) preserve the private information of member networks, and (4) scales to collaborative networks with hundreds of members.

(c) High-Level, Unified SDN Programming (Y.R. Yang, Chris Leet, Kai Gao, Jensen Zhang):

The Magellan and Trident teams will demonstrate a high-level, unified SDN and NFV programming framework to support collaborative data sciences. Although much progress has been made in SDN programming and NFV in general settings and in scientific computing in particular, the existing programming frameworks are lacking. For example, there is no automatic way to utilize flexible SDN datapath; the programming of SDN and NFV is largely separated. In this demo, the team will show two key high-level programming techniques: (1) Magellan, which allows systematic, automatic compilation of high-level SDN programs into low-level SDN datapaths; and (2) Trident, which introduces key techniques including live variables, routes algebra, and 3-value logic programming to allow high-level, systematic integration of SDN and NFV, to achieve automatic updates. The Magellan and Trident demo will be integrated with the SFP and MNSA demos to form a single, coherent demo showing the possibility of automated, high-level networking for collaborative data sciences.

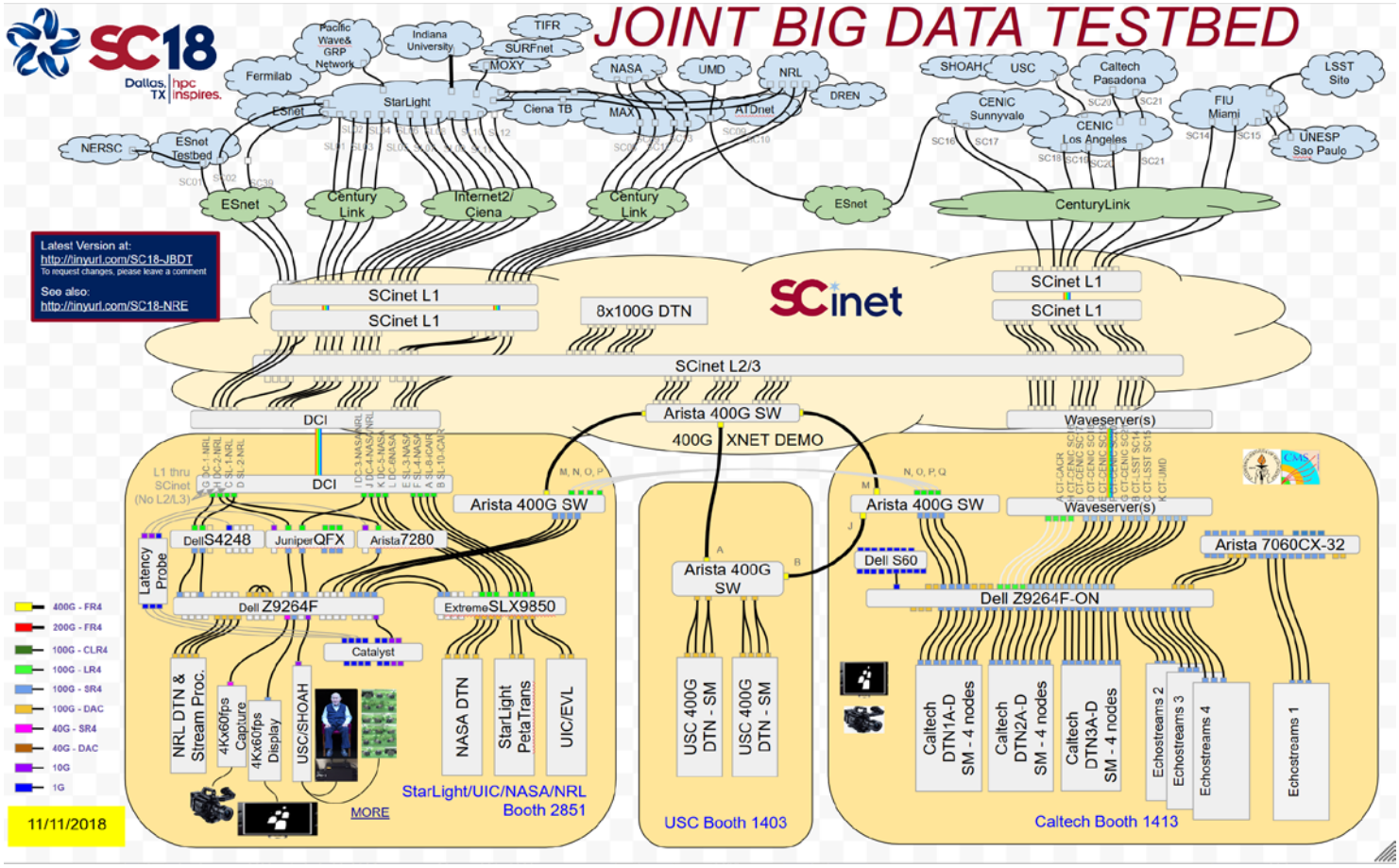
A paper accepted for the SC18 Technical Program on this body of work is available here:

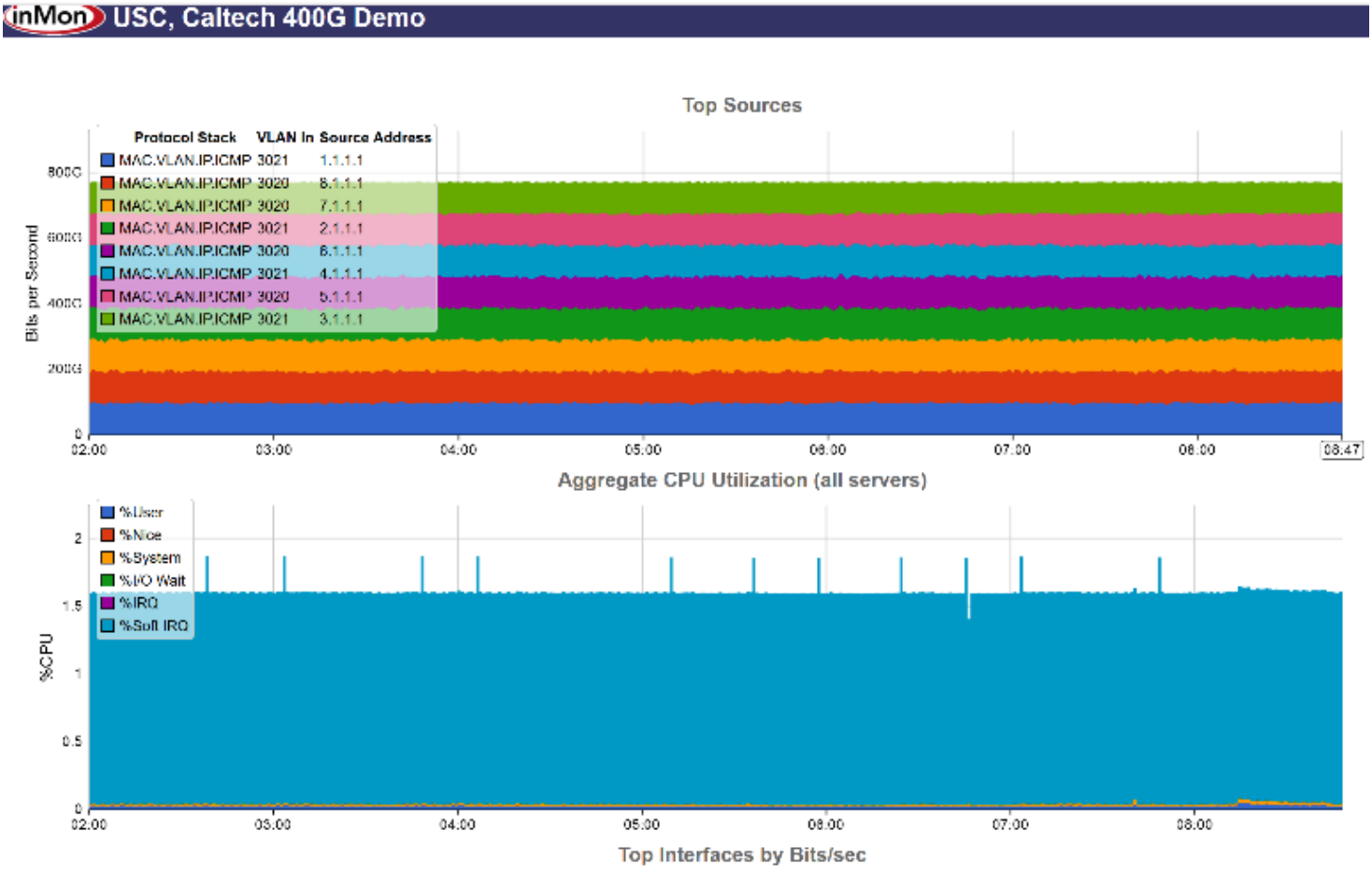
https://www.dropbox.com/sh/1mg4jfgcxuahrwj/AAAprKgHS78Oh5dzL4j-dOg7a?dl=0&preview=Explorer-long.pdf

(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 recently developed routing and caching algorithms of NDN combined with SDN-based path allocations and end-to-end provisioning across the SC18 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, J. Chiu, L. Mercer): 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 network paths with round trip times of more than 120 milliseconds. Low impact on the end-system resources will be maintained, including offloading of the transfer processes to on-board network processors where available. Data transfer applications will use low latency protocols such as NVMe over Fabric (NVMeoF), combined with either iWARP or 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.


SC18
Dallas, TX
hpc inspires.
JOINT BIG DATA TESTBED
Pacific Wave& GRP Network
Indiana University
TIFR
SURFnet
MOXY
Fermilab
ESnet
StarLight
NASA
UMD
NRL
DREN
Ciena TB
MAX
ATDnet
SHOAH
USC
Caltech Pasadena
LSST Site
CENIC Sunnyvale
CENIC Los Angeles
FIU Miami
UNESP Sao Paulo
NERSC
ESnet Testbed
ESnet
Century Link
Internet2/ Ciena
Century Link
ESnet
CenturyLink
Latest Version at:
http://tinyurl.com/SC18-JBDT
To request changes, please leave a comment
See also:
http://tinyurl.com/SC18-NRE
SCinet L1
SCinet L1
8x100G DTN
SCinet
SCinet L1
SCinet L1
SCinet L2/3
DCI
Arista 400G SW
400G XNET DEMO
Waveserver(s)
L1 thru SCinet (No L2/L3)
DCI
Arista 400G SW
Arista 400G SW
Waveserver(s)
Latency Probe
DellS4248
JuniperQFX
Arista7280
Arista 7060CX-32
Dell Z9264F
ExtremeSLX9850
Arista 400G SW
Dell S60
Dell Z9264F-ON
Catalyst
NRL DTN & Stream Proc.
4Kx60fps Capture
4Kx60fps Display
USC/SHOAH
NASA DTN
StarLight PetaTrans
UIC/EVL
USC 400G DTN - SM
USC 400G DTN - SM
Caltech DTN1A-D SM - 4 nodes
Caltech DTN2A-D SM - 4 nodes
Caltech DTN3A-D SM - 4 nodes
Echostreams 2
Echostreams 3
Echostreams 4
Echostreams 1
400G - FR4
200G - FR4
100G - CLR4
100G - LR4
100G - SR4
100G - DAC
40G - SR4
40G - DAC
10G
1G
MORE
StarLight/UIC/NASA/NRL Booth 2851
USC Booth 1403
Caltech Booth 1413
11/11/2018


inMon
USC, Caltech 400G Demo
Top Sources
Protocol Stack VLAN In Source Address
MAC.VLAN.IP.ICMP 3021 1.1.1.1
MAC.VLAN.IP.ICMP 3020 8.1.1.1
MAC.VLAN.IP.ICMP 3020 7.1.1.1
MAC.VLAN.IP.ICMP 3021 2.1.1.1
MAC.VLAN.IP.ICMP 3020 6.1.1.1
MAC.VLAN.IP.ICMP 3021 4.1.1.1
MAC.VLAN.IP.ICMP 3020 5.1.1.1
MAC.VLAN.IP.ICMP 3021 3.1.1.1
Bits per Second
800G
600G
400G
200G
0
02:00
03:00
04:00
05:00
06:00
07:00
08:00
08:47
Aggregate CPU Utilization (all servers)
%User
%Nice
%System
%I/O Wait
%IRQ
%Soft IRQ
%CPU
2
1.5
1
0.5
0
02:00
03:00
04:00
05:00
06:00
07:00
08:00
Top Interfaces by Bits/sec


CALIFORNIA INSTITUTE OF TECHNOLOGY
1891
CALIFORNIA INSTITUTE OF TECHNOLOGY
1891
US Large Hadron Collider
US LUA
Users Association
ESnet
Energy Sciences Network
US CMS
ATLAS
LHC
Fermilab
CERN
INTERNET 2
AMERICA'S LIGHTPATHS
AMLIGHT
RESEARCH · EDUCATION
COLLABORATION
LUX ET VERITAS
YALE
YALE
UNIVERSITY OF MARYLAND
18 56
NORTHEASTERN UNIVERSITY
LVX VERITAS VIRTVS
FOUNDED A.D. 1898
BOSTON, MASSACHUSETTS
UNIVERSITY OF SOUTHERN CALIFORNIA
1880
STARLIGHT
The Optical STAR TAP℠
SURF NET
Argonne
NATIONAL LABORATORY
PRP
PACIFIC RESEARCH PLATFORM
BERKELEY LAB
Lawrence Berkeley National Laboratory
tifr
KiSTi
www.kisti.re.kr
CENIC
PACIFIC WAVE
PACIFIC NORTHWEST GIGAPOP
DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA
NSF
+ INDUSTRY