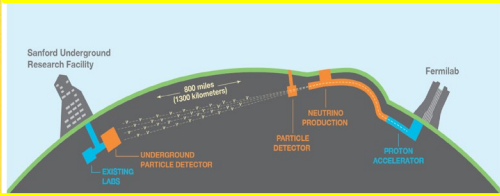
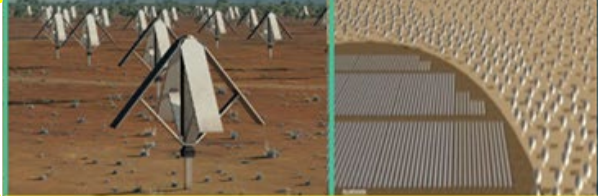
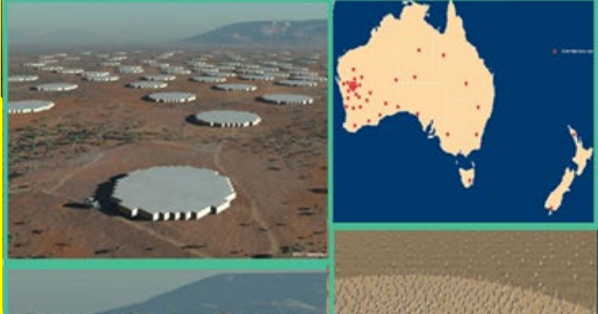
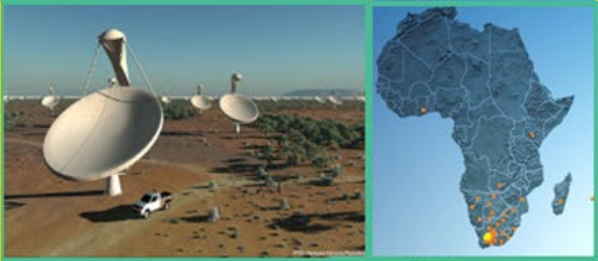
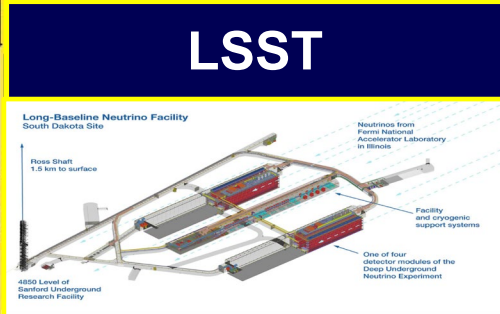
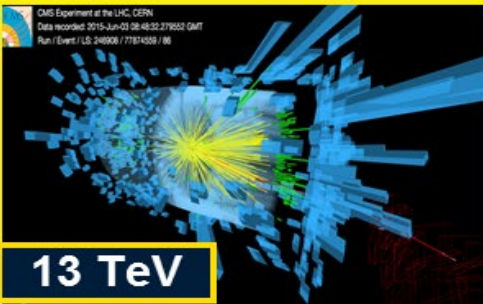


The GNA-G Data Intensive Science Working Group

Next Generation Networks for Global Science Programs



LHC Run3
and HL-LHC

DUNE

VRO SKAO

BioInformatics

Earth
Observation

LHC

LBNF/DUNE

SKA

Gateways
to a New Era



Harvey Newman, Caltech

SC20 XNet Workshop

November 13 2020



<https://www.gna-g.net/>

- **Challenge of Scale: HL-LHC (2028) vs LHC example:**
30X in Storage, 16X in Compute, 16X in Networking by 2028
Terabit/sec Transactions: Cannot be accommodated through technology evolution alone within a ~fixed budget
- **HL-LHC is Not Alone: SKA will Also generate Tbps flows**
- **Challenge of Complexity:**
 - Thousands of scientists and students
 - Tens to hundreds of sites, hundreds of science teams
 - Dozens of Vos; Dozens of network domains
 - Multiple policy and priority frameworks
 - Global reach; DIS programs sharing with the community
- **Conceptual Challenge: Workflow**
VOs have learned to deal effectively with **Computing & Storage:** for distributed data processing, access and analysis; but **View the network as an opaque infrastructure of limitless capacity**
- **Bringing Managed Networks into the picture;** is a necessary step to meet the challenges, This requires **a paradigm shift; and a community-wide effort**

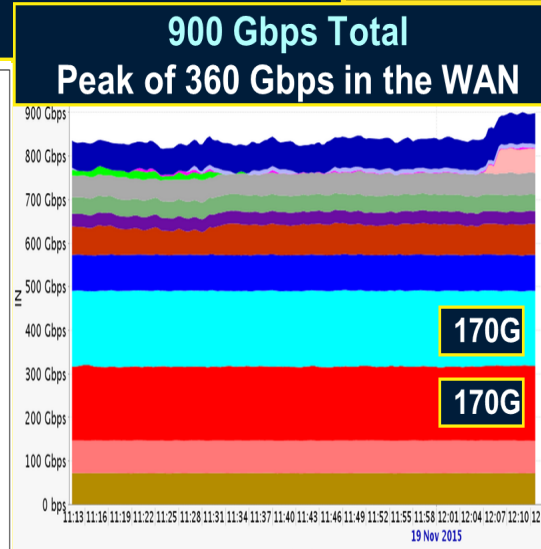
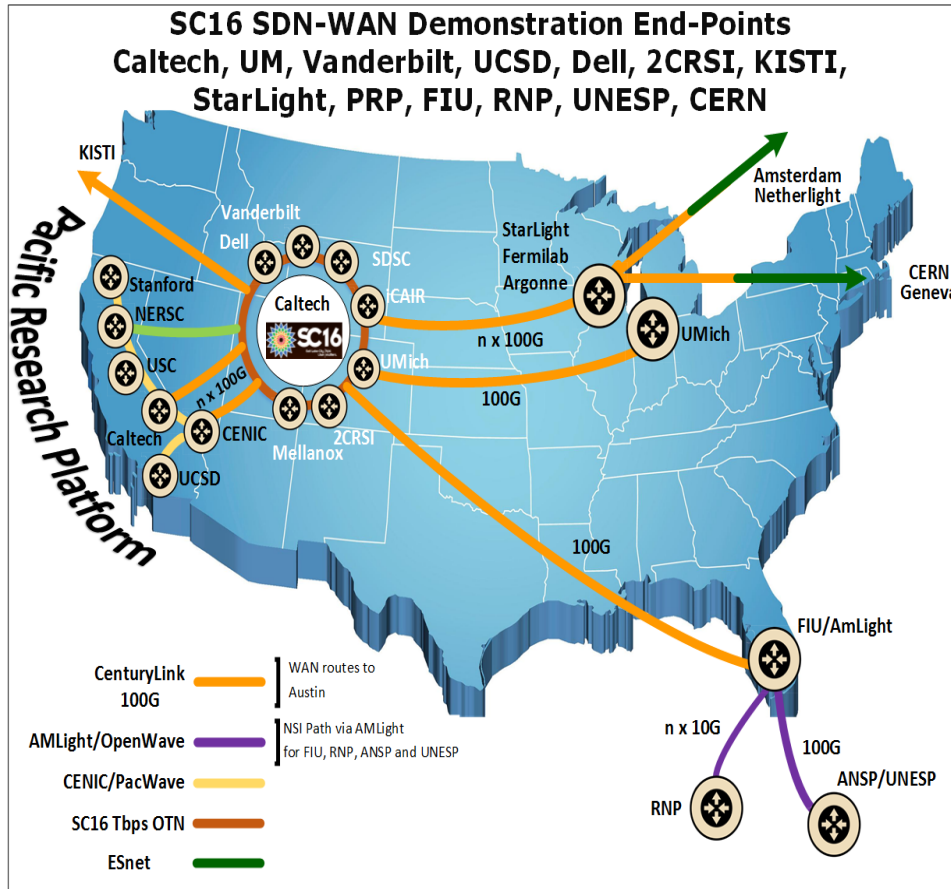
SC15-19: SDN Next Generation Terabit/sec Ecosystem for Exascale Science



SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

SC16+: Consistent Operations with Agile Feedback Major Science Flow Classes Up to High Water Marks

Preview PetaByte Transfers to/from Site Edges of Exascale Facilities With 100G -1600G DTNs

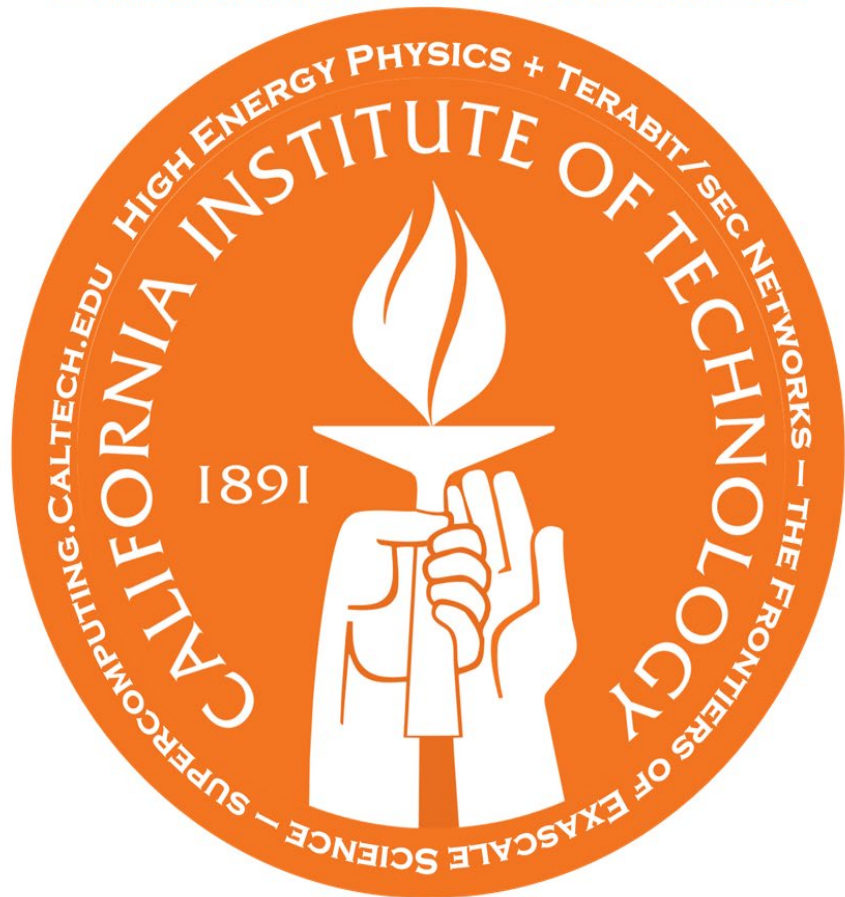


LHC at SC15: Asynchronous Stageout (ASO) with Caltech's SDN Controller

29 100G NICs; Two 4 X 100G and Two 3 X 100G DTNs; 1.5 Tbps Capability in one Rack; 9 32 X100G Switches

Tbps Rings for SC18-19: Caltech, Ciena, Scinet, OCC/StarLight + Many HEP, Network, Vendor Partners

Global Petascale to Exascale Workflows for Data Intensive Sciences




**Accelerated by Next Generation
Programmable SDN Architectures
and Machine Learning Ai Applications**





Demonstrations at Caltech Booth 543

New Approaches to Meet the Challenges

- ***NRE-019** – **Global Petascale to Exascale Workflows for Data Intensive Science Accelerated by Next Generation Programmable SDN Architectures and Machine Learning Applications** 
- NRE-019b** **FPGA-Accelerated Machine Learning [Caltech and 2CRSI]**
Inference for Trigger and Computing at LHC
- NRE-013** – **SENSE: Intelligent Network Services for Science Workflows**
Layer2/3 Services, Full Lifecycle, Multi-Domain, Multi-Resource, Interactive, End-to-End
- NRE-020** – **LHC Multi-Resource, Multi-Domain Orchestration**
via **AutoGOLE** and **SENSE: Inter-Regional Integration**
- NRE-022** – **Toward Unified Resource Discovery and Programming**
in Multi-Domain Networks
- NRE-023** – **International Data Transfer over AmLight Express**
and **Protect (Exp) [Supporting LSST]**
- NRE-024** – **7 X 400GE Ring (Triangle): Caltech-SCinet-Starlight/NRL**
with WAN Extensions to Starlight/iCAIR; PCIe 4.0, Tbps Servers
- NRE-035** – **SANDIE: SDN-Assisted NDN for Data Intensive Experiments (NDN Across AL2S Paths; Persistent Testbed)**

Caltech and Partners at SC19



- ❑ **LHC/HEP, LSST/Astrophysics; AmLight Express+Protect, SENSE, SANDIE(NDN), SDN NGenIA, Mercator, Carbide Multicontroller SDN Projects**
- ❑ **Ai Presentations:** CMS Trigger w/Fast Training and Interference, Higgs Bosons and Interaction Networks, Quantum ML, Inline Monitoring + Decisions
- ❑ **Ciena DWDM+Waveserver Ais in the Caltech booth: 400G waves, 16 100G clients**
- ❑ **“Caltech-Starlight-SCinet” Triangle: 400GE Arista, Dell + Mellanox 200GE Switches ~8 Tbps Server Capacity at the Caltech Booth in 1/2 Rack; to 1 Tbps per rack unit**
 - ❑ **2CRSI: 4 AMD Rome (PCIe Gen4) +1 Intel Server; Echostreams Servers; Pavilion IO NVMeoF; 28 processors, 28 200GE + ~40 100GE, ~160 SSDs**
 - ❑ **QSFP56 DD 400GE + 200GE DAC; 400G to 2 X 200G Splitters [Bleeding edge]**
- ❑ **Network, Server, Storage Partners:** SCinet, Ciena, Arista, Mellanox, Dell, 2CRSI, Intel, Echostreams, Pavilion IO, NVIDIA, XiLinx
- ❑ **Science+Network Partners:** USC, AmLight, Starlight, CENIC, PWNWG, KIT, SURFnet, UMD/MAX, MIT, NUE, CSU, FNAL,USCD, UERJ, UNESP, KISTI/KASI, UMich, TIFR
- ❑ **WAN Sites:** Caltech, FIU, Maryland, Starlight, UCSD, MIT, LBL, CENIC, FNAL, NEU, CSU, LSST (Chile), GridUNESP, UERJ (Rio), SURFnet, KISTI/KASI, CERN, TIFR
- ❑ **Caltech Booth to WAN: 400G to Caltech + USC campuses; 400G to PRP/TNRP via CENIC (UCSD, LBNL, UCSC, et al); 300G to Brazil+Chile via AmLight Express (200G Scinet to FIU); 200G to ESnet via Sunnyvale**
- ❑ **Caltech campus/CENIC LA Waveserver Ai 2 X 200G+10X10G upgrade Persists**

★ **Creating the Future of SCinet and of Networks for Science**



A New Generation Data Intensive SDN Facility and *Persistent 400G WS Ai Super-DMZ*

Campus Connection: Caltech/SCinet/Caltech Booth

**SC19 Rack: 5 400GE, 2 200GE,
2 100GE Switches, AMD Rome**

**Caltech HEP
+SDN Lab + iBanks GPU**

Caltech Tier2

SC19 Rack Layout

A	Dell Z9332F 400G Switch	42
B	Arista 7060 PX4-32 400G Switch	40
B2	Arista 7060 DX4-32 400G Switch	39
C	Arista 7060 DX4-32 400G Switch	38
D	Top of Rack Switch Dell S60	37
		36
		35
		34
E	Echostreams SANDIE5	33
F	Echostreams SANDIE6	32
G	2CRSI 4 Node H262-Z62 Rome1 (2U)	31
G2	2CRSI 4 Node H262-Z62 Rome 2 (2U)	29
G3	2CRSI 4 Node H262-Z62 Rome 3 (2U)	28
		27
H	Mellanox SN3700 200G Switch	25
I	Mellanox SN3700 200G Switch	24
J	Mellanox SN2700 100G Switch	23
K	Dell Z9100 100G Switch	22
	Console	21
		20
		19
		18
		17
		16
		15
L	Supermicro 6027 4 Node (2U)	15
M	Dell Z9264F 100G 64 Port Switch (2U)	14
		13
		12
		11
P	Pavilion IO Storage Unit (4U)	10
		9
		8
		7
N,O	Additional Servers (3U)	6


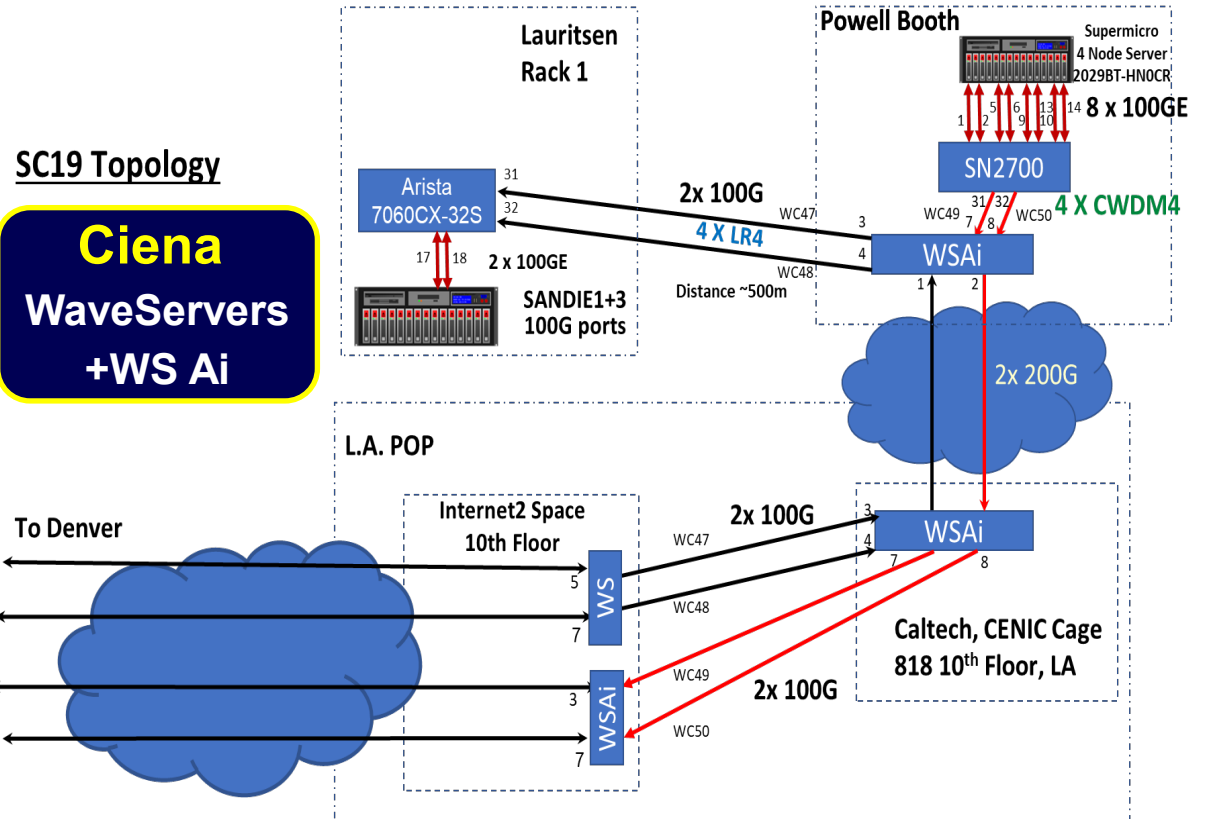
Total 42 Rack Units

V6 11/10/19

On Hand

Reserved

PreCommissioned at Caltech

**2CRSI, Arista, Mellanox, Dell,
Pavilion IO**

To SCinet Denver

Internet2 LA PoP

CENIC LA PoP

★ Creating the Next Generation of Data and Network CyberSystems



Next Generation Computing and Networking

~ 7 Tbps Rack at Booth 543, + ~1 Tbps Caltech and Partner Sites

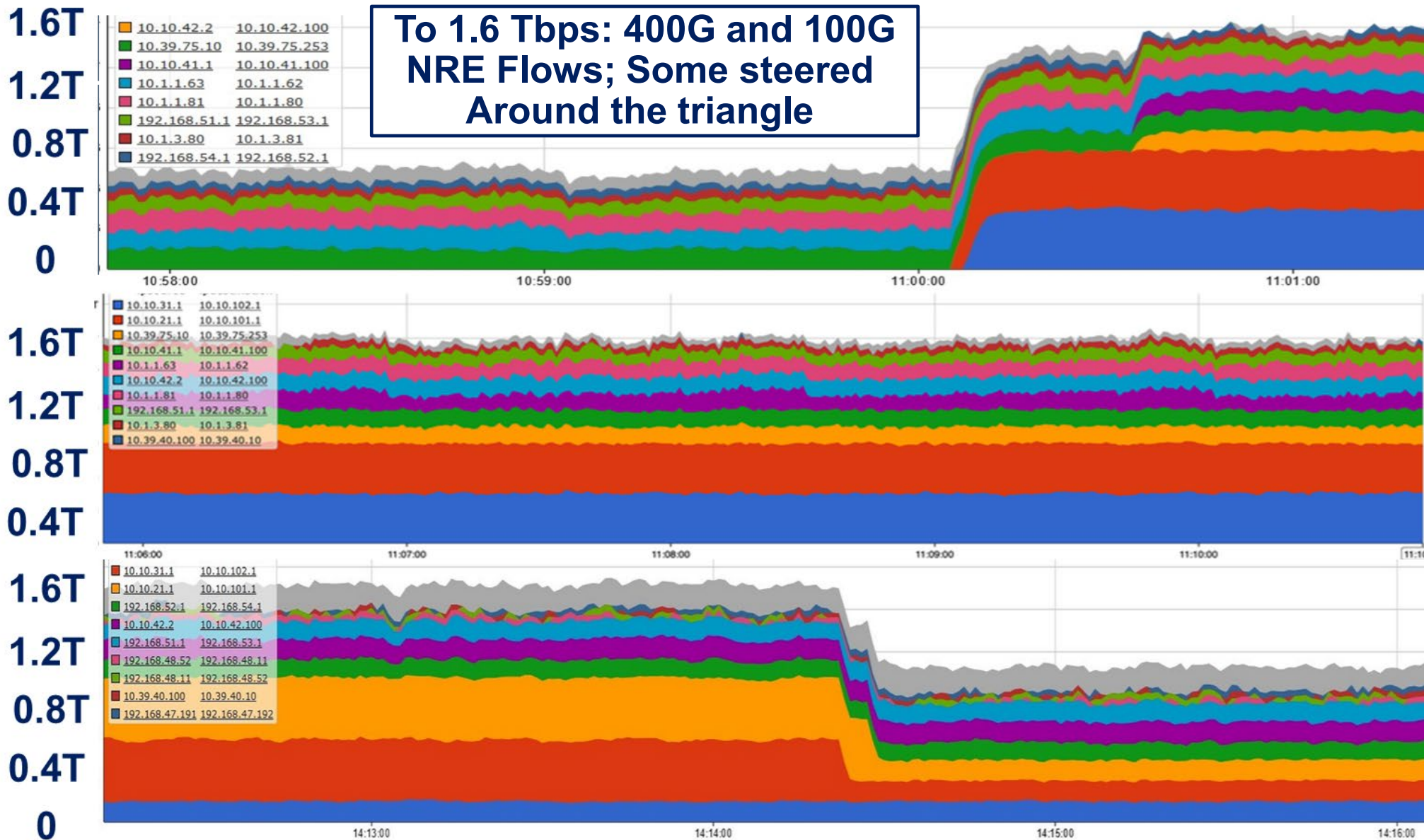
- **Three 2CRSI Servers Gigabyte H262-X62 PCIe 4.0 4-Node, 8 CPU Hyperconverged Servers: to 6 Tbps in 6 RU [Projects to 42 Tbps per rack]**
- **With Mellanox NICs: 24 ConnectX-6 200GE and 12 ConnectX-5 100GE; + Many ConnectX-5s in 2CRSI and Echostream Servers at SC, Caltech, CERN**
- **3 Arista 7060 DX-4 (PX-4) & 1 Dell Z9332F-ON 32 X 400G Switches; Mellanox SN3700 32 X 200G Switches; QSFP56-DD (OSFP) Standards**
- **Brand New 400GE Transceivers: Arista & Dell FR4 (2km); + Arista DR4 (500m)**
- **DAC Cables Beyond 100G: Arista and Dell 400GE; Arista & Mellanox 400G to 2X200G Splitter Cables; Mellanox 200G Switches**
- **100GE Switches: Dell 9264F-ON (64 port), Dell Z9100 and Mellanox 2700 (32 port) Switches**
- **Echostreams Supermicro servers: 4 X 100GE**
- **See <http://tinyurl.com/sc19-jbdt>**

Precommissioned at Caltech

	42
Dell Z9332F 400G Switch	41
Arista 7060 PX4-32 400G Switch	40
Arista 7060 DX4-32 400G Switch	39
Arista 7060 DX4-32 400G Switch	38
Top of Rack Switch Dell S60	37
	36
	35
	34
Echostreams SANDIE5	33
Echostreams SANDIE6	32
2CRSI 4 Node H262-Z62 Rome1 (2U)	31
2CRSI 4 Node H262-Z62 Rome 2 (2U)	30
2CRSI 4 Node H262-Z62 Rome 3 (2U)	29
	28
	27
	26
Mellanox SN3700 200G Switch	25
Mellanox SN3700 200G Switch	24
Mellanox SN2700 100G Switch	23
Dell Z9100 100G Switch	22
	21
Console	20
	19
	18
	17
	16
Supermicro 6027 4 Node (2U)	15
Dell Z9264F 100G 64 Port Switch (2U)	14
	13
	12
	11
	10
Pavilion IO Storage Unit (4U)	9
	8
	7
	6

4 32X400GE + 2 32X200GE Switches

SC19 Results on the 400G Triangle



The GNA-G Data Intensive Sciences WG

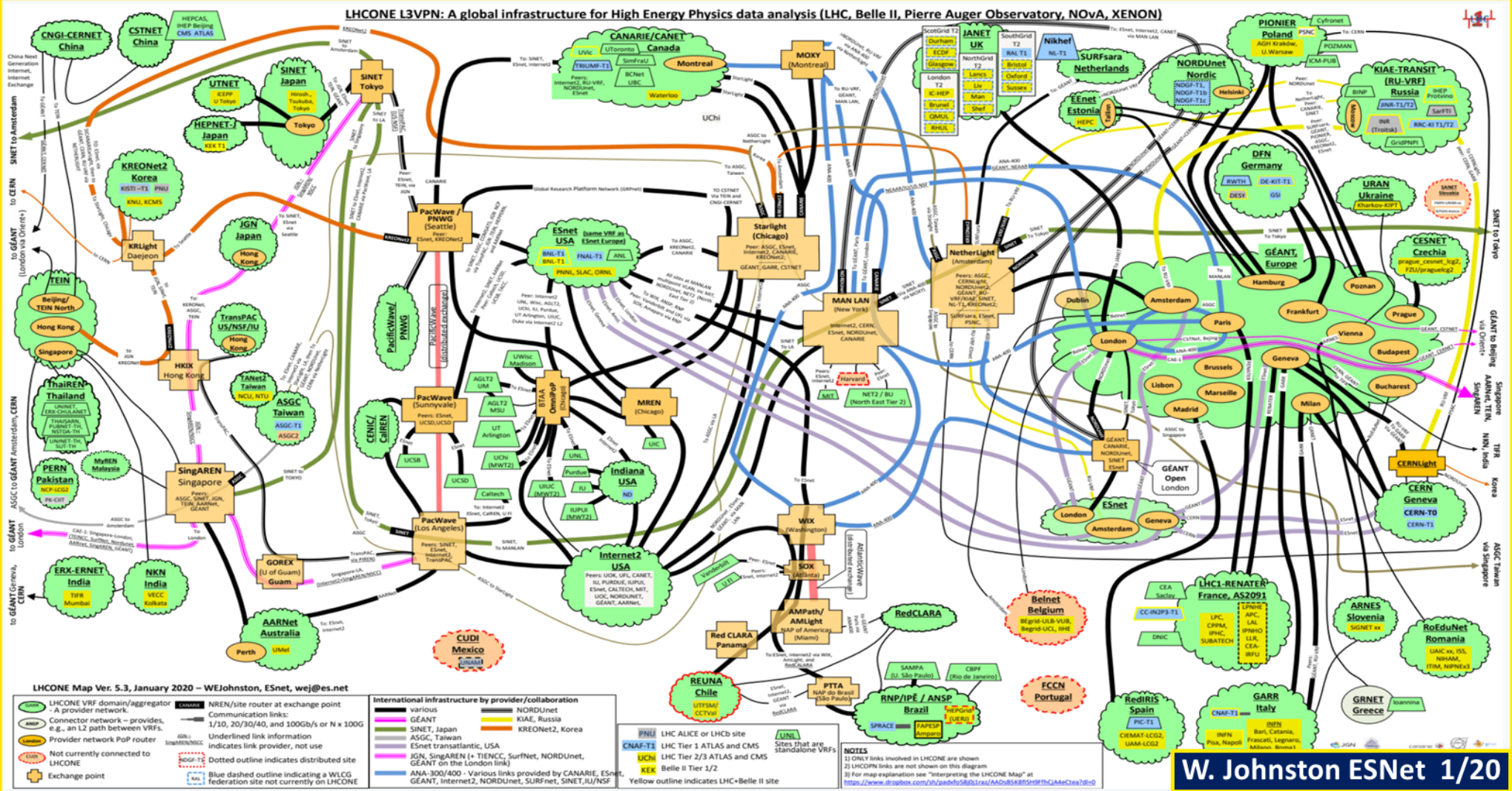
Challenges: Capacity in the Core and at the Edges

- Programs such as the LHC have experienced rapid exponential traffic growth, at the level of 40-60% per year, projected to outstrip the affordable capacity
 - At the January 2020 LHCON/LHCOPN meeting at CERN, CMS and ATLAS expressed the need for Terabit/sec links on major routes by the start of the HL-LHC in 2028
 - This is to be preceded by data & network 1-10 Petabyte/day “challenges” before and during the upcoming LHC Run3 (2021-24)
 - These needs were further specified in “blueprint” Requirements documents by US CMS and US ATLAS, submitted to the ESnet Requirements Review in August, and under continued discussion for a 2/21 DOE Review
 - Three areas of capacity-concern by 2028 were identified:
 - (1) Exceeding the capacity across oceans, notably the Atlantic, served by ANA
 - (2) Tier2 centers at universities requiring 100G annual average with sustained 400G bursts, and
 - (3) Terabit/sec links to labs and HPC centers (and edge systems) to support multi-petabyte transactions in hours rather than days
 - Analysis of the transatlantic shortfall follows, as an example



LHCONE VRF: The Challenge of Complexity and Global Reach

Global infrastructure for HEP (LHC, Belle II, NOvA, Auger, Xenon) data flows



W. Johnston ESNet 1/20

Good News: The Major R&E Networks Have Mobilized on behalf of HEP

Challenge: A complex system with limited scaling properties.

Response: New Mode of Sharing? Multi-One?

Hierarchical Storage via Data Lakes

Regional Caches



- Store most data on “active archive” on inexpensive, high latency media (e.g. Tape).
- Keep a “golden copy” on redundant high availability disk [fewer copies].
 - This defines the working set allowed to be accessed.
 - Jobs requesting data not in working set will queue up until data is recalled from archive
- Regional Caches at processing centers (e.g. Tier1s & 2s; ~1 petabyte)
 - Size of region determined by latency tolerance of application
 - Cost trade-off: between cache size vs network use
- Useful distance metric: 10% IO penalty among merged caches
- EU example: ~500-1000 km
- Advanced protocol, caching methods: could extend distance



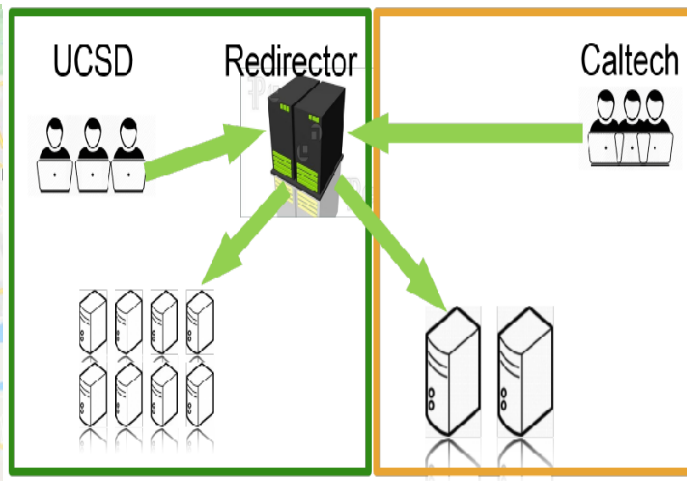
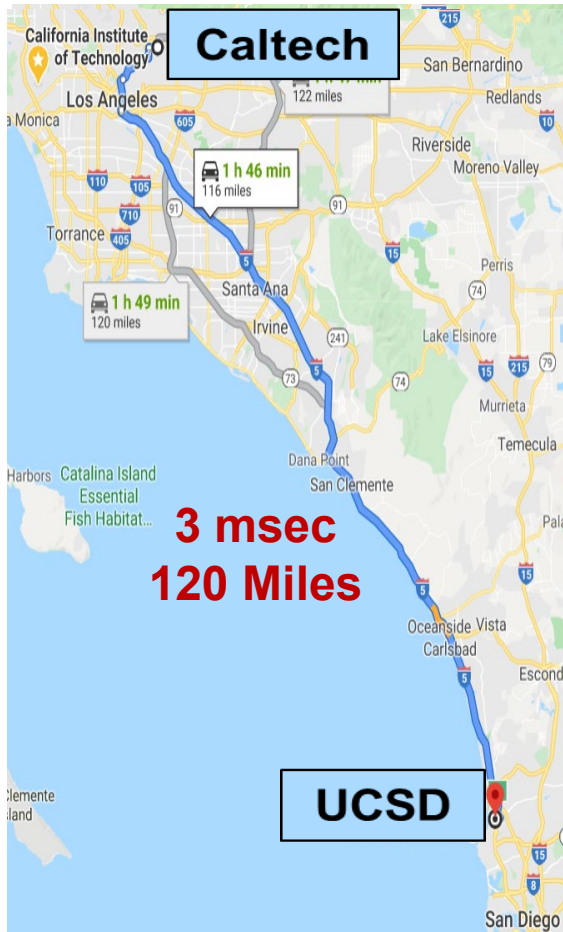
Examples in Production:

“SoCal” (UCSD + Caltech); INFN

F. Wuerthwein (UCSD) et al

(Southern) California ((So)Cal) Cache

(Roughly 20,000 cores across Caltech & UCSD ... half typically used for analysis)

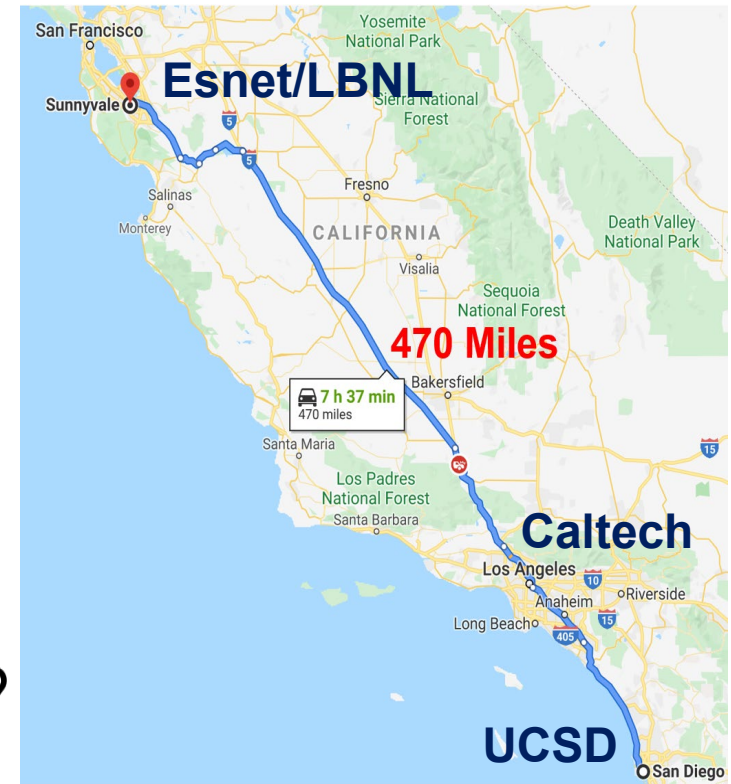


CPU in both places can access storage in both places.

How much disk space is enough?

Cache MINI and measure working set accessed:

0.45 Petabytes in October 2019



In early May, we added a cache at the ESNet POP in Sunnyvale to the SoCal cache.

Global Network Advancement Group (GNA-G) Leadership Team: Since September 2019

leadershipteam@lists.gna-g.net



Erik-Jan Bos
NorduNet



Buseung Cho
KISTI



Dale Finkelson
Internet2



Gerben van
Malenstein SURFnet



Harvey Newman
Caltech



David Wilde
Aarnet

- The GNA-G is an open volunteer group devoted to developing the blueprint to make using the Global R&E networks both simpler and more effective, operating under GNA-G.
- Its primary mission is to support global research and education using the technology, infrastructures and investments of its participants.
- ★ The GNA-G needs to be a data intensive research & science engager that facilitates and accelerates global-scale projects by
 - (1) enabling high-performance data transfer, and
 - ★ (2) acting as a partner in developing next generation intelligent network systems that support the workflow of data intensive programs

Charter: https://www.dropbox.com/s/4my5mjl8xd8a3y9/GNA-G_DataIntensiveSciencesWGCharter.docx?dl=0

▪ Principal aims of the GNA-G DIS WG:

(1) To meet the needs and address the challenges faced by major data intensive science programs

- Coexisting with support for the needs of individuals and smaller groups

(2) To provide a forum for discussion, a framework and shared tools for short and longer term developments meeting the program and group needs

- To develop a persistent global persistent testbed as a platform, to foster ongoing developments among the science and network partners

- While sharing and advancing the **(new)** concepts, tools & systems needed
- Members of the WG will partner in joint deployments and/or developments of generally useful tools and systems that help operate and manage R&E networks with limited resources across national and regional boundaries
- A special focus of the group is to address the growing demand for
 - Network-integrated workflows
 - Comprehensive cross-institution data management
 - Automation, and
 - Federated infrastructures encompassing networking, compute, and storage
- Working Closely with the AutoGOLE/SENSE WG on the **Global persistent testbed**

The GNA-G Data Intensive Sciences WG

Charter: https://www.dropbox.com/s/4my5mjl8xd8a3y9/GNA-G_DataIntensiveSciencesWGCharter.docx?dl=0

- **Mission: Meet the challenges of globally distributed data and computation faced by the major science programs**
- **Mission: Coordinate provisioning the feasible capacity across a global footprint, and enable best use of the infrastructure:**
 - **While meeting the needs of the participating groups, large and small**
 - **In a manner Compatible and Consistent with other use**
- **Members:**
- **Alberto Santoro, Azher Mughal, Bijan Jabbari, Buseung Cho, Caio Costa, Carlyn Ann-Lee, Chin Guok, Ciprian Popoviciu, Dale Carder, Dale Finkelson, David Lange, David Wilde, Edoardo Martelli, Eduardo Revoredo, Eli Dart, Frank Wuerthwein, Frederic Loui, Gerben van Malenstein, Harvey Newman, Heidi Morgan, Iara Machado, Inder Monga, Jeferson Souza, Jensen Zhang, Jeonghoon Moon, Jeronimo Bezerra, Jerry Sobieski, Joe Mambretti, John Graham, John Hess, John Macauley, Julio Ibarra, Justas Balcas, Kai Gao, Karl Newell, Kaushik De, Kevin Sale, Lars Fischer, Marcos Schwarz, Matt Zekauskas, Michael Stanton, Mike Hildreth, Mike Simpson, Ney Lemke, Phil Demar, Raimondas Sirvinskas, Richard Hughes-Jones, Rogerio Iope, Sergio Novaes, Shawn McKee, Siju Mammen, Susanne Naegele-Jackson, Tom de Fanti, Tom Hutton, Tom Lehman, William Johnston, Xi Yang, Y. Richard Yang**
- **Participating Organizations/Projects:**
- **ESnet, Nordunet, SURFnet, AARNet, AmLight, KISTI, SANReN, GEANT, RNP, CERN, Internet2, CENIC/Pacific Wave, StarLight, NetherLight, Southern Light, Pacific Research Platform, FABRIC, RENATER, ATLAS, CMS, VRO, SKAO, OSG, Caltech, UCSD, Yale, FIU, UERJ, GridUNESP, Fermilab, Michigan, UT Arlington, George Mason, East Carolina, KAUST**
- **Meets Weekly or Bi-weekly; all are welcome to join.**

Capacity Requirements Analysis, Using ESnet Transatlantic Network Traffic Projections

- Requirements based on recent traffic: 0.35 – 0.85 Tbps [based on 0.8 to 2X the 2016-19 traffic projection]
- Growth Rate 1.4X per year, or 2X every two years on average
- Hence 16X capacity requirement in 2028 = 5.6 to 13.6 Tbps; Since this is an ESnet only, and not a global projection, the upper limit may be the better requirements metric
- Traditional long-term capacity per unit cost rate: +15-20 % per year; Hence 3.1 to 4.3 times affordable capacity by 2028 (source: Telegeography)
- Implied Shortfall: 3.7 to 5.2X
- Naïve Implementation Outlook by 2028: 52-68 200G links across the Atlantic (for example: 13 to 17 200G links on each of 4 disjoint paths); compare the ANA consortium today: 9 100G links at present
- Ways to bring down the costs: Acquire spectrum IRUs on undersea cables; Move towards co-ownership on undersea cables if and where possible
- Outlook: These can get us part of the way there (within a factor of 2?)
- Bottom Line: Need to develop a new system that comprehensively monitors, tracks, manages and controls use, coordinated with compute and storage use

Beyond Capacity Alone: the Challenges of Complexity and Global Reach

- Working to adopt, extend, and/or interface highly capable toolsets and best practices across a global footprint, via:
- Common adoption, or interfacing via APIs, or mediation/impedance-matching code
 - ★ Leverage developments underway in projects such as SENSE, AutoGOLE, AmLight, PRP, NOTED and SANDIE. Testbeds: ESnet, FABRIC and BRIDGES
 - ★ Ongoing discussions should continue to define what the new services and classes of work required entail
 - ★ Solutions will vary by region and by network
- ★ A change in paradigm to a system of end-to-end services will be required involving coordinated operation and responses among sites and networks
 - A real-time orchestration system that responds to Constraints: resource allocation and operational decisions become network-state, site-state, policy and priority dependent, and potentially complex
- ★ An important part of this is the persistent testbed being deployed by the AutoGOLE/SENSE WG in collaboration with AutoGOLE and other projects.
- ★ This is proceeding: starting with the current SENSE testbed sites, plus extensions to CERN, Starlight in Chicago, SURFnet in Amsterdam, KISTI, UCSD, and other sites in the US, Europe, Latin America and Asia



SDN Enabled Networks for Science at the Exascale

SENSE: <https://arxiv.org/abs/2004.05953>

Model-based Site and Network Resource Managers

Designed to Adapt to Available SDN Systems

SENSE Native RMs are Available if no current automation layer

Application Workflow Agents

SENSE

SENSE operates between the **SDN Layer** controlling the individual networks/end-sites, and science workflow agents/middleware

Intent-Based APIs with Resource Discovery, Negotiation, Service Lifecycle Monitoring/Troubleshooting

Regional

WAN

WAN

SDX

SDN Layer

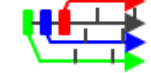
Regional

End Site

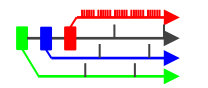
SDMZ

End Site

SDMZ



Instruments Storage Compute DTNs



DTNs

Compute

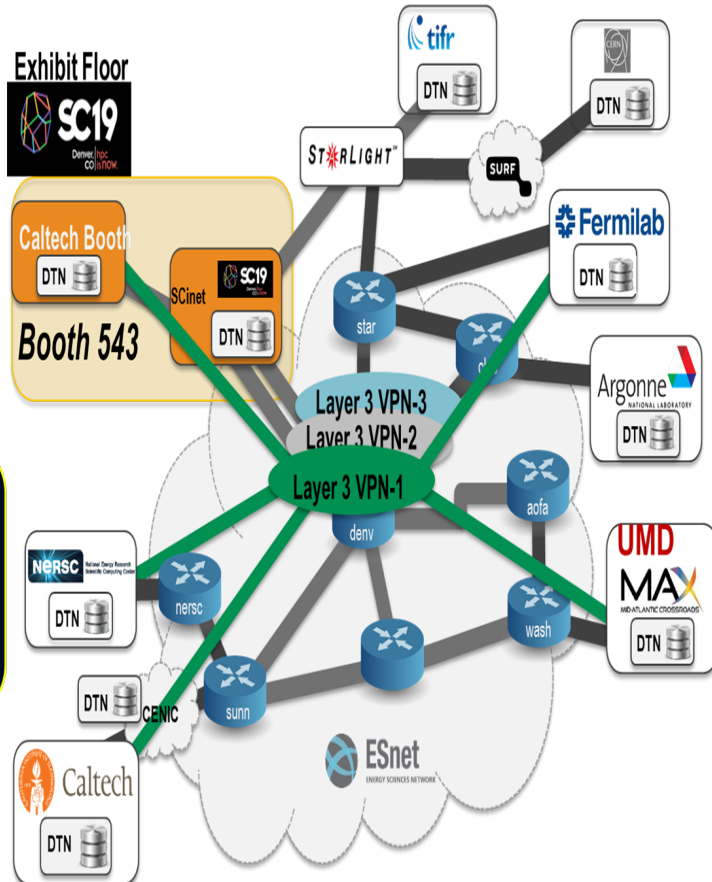
Storage Instruments

SENSE SC19 Demonstration Topology

SENSE Testbed and L3 VPN Service

SENSE enabled resources at DOE Labs, Universities, Research Facilities, + SC19

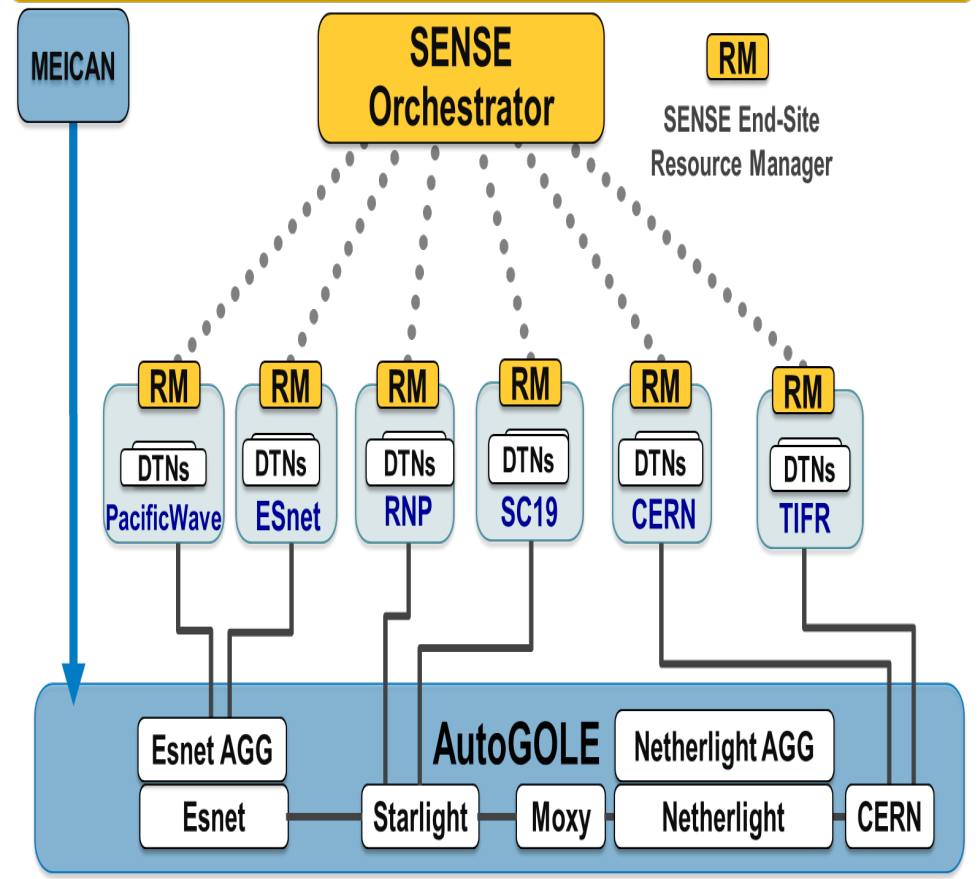
Dynamic attachment of End Site resources to L3VPNs advertised by ESnet



Provisioning SENSE

AutoGOLE Topology

SC19-NRE-020 Intercontinental Demonstration Multi-Resource Orchestration via AutoGOLE and SENSE



SENSE - AutoGOLE Joint Interworking Demo

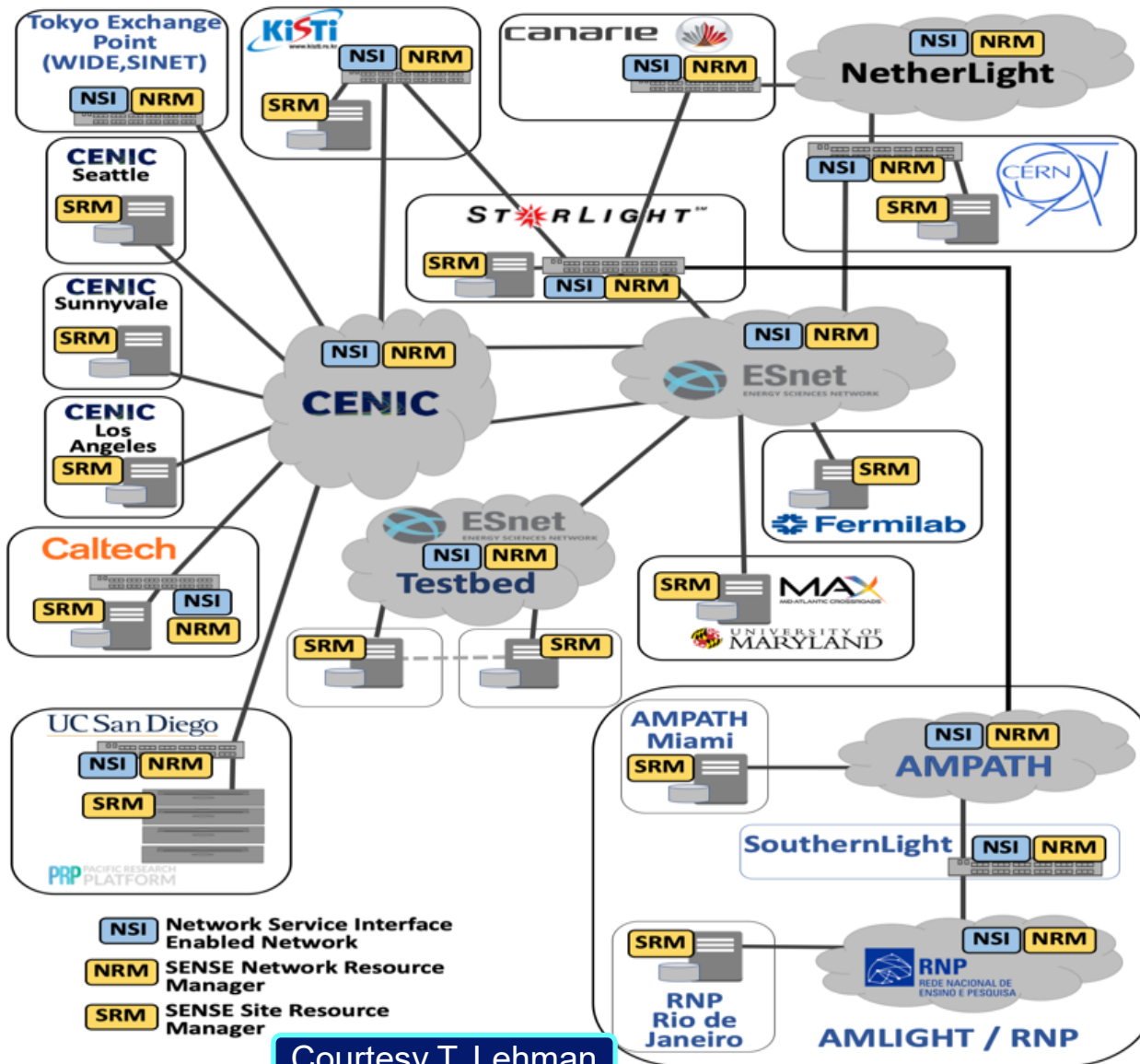
Candidate Inter-regional Mediation Layer for Global Workflows (as discussed in GNA-G)

For a global fabric, including Australia and Africa we will include genomics, AMLight/VRO, SKAO, and others in the overall concept along with HEP

- **Computing:** Technology evolution + **Code improvements**
 - + **Hybrid architectures** (GPU, FPGA)
 - + Greater use of **HPC exascale + pre-exascale systems**
 - + **Cloud resources an option** for peak needs
- **Storage:** **Data Lakes as Regional Caches; including streaming access** [**Compact Event Forms + Caching Strategies + Improved Architectures**]
- **Networking:** Tuned end systems + QoS via virtual circuits,
 - + allocated resources with prioritization, policy;Interworking with **LHCONE and the major R&E networks**
- **Common Services Framework Foundation:**
 - [**Networks**] SENSE/AutoGOLE: Integration, Adaptation, Mediation
 - [**VO Workflow Interface**] Rucio/FTS/XRootD: **Serving > 30 VOs, Many PIs**
- **Developed on a Persistent, Global Federated Testbed:** Now being deployed
- **Vision: A Stateful, Adaptive Real-Time System**
Full lifecycle services overseeing task completion
- **Network management-enabling VO workflow:** a bigger picture
- **Interactions: VO Orchestrators with Network Orchestrators**
Sites Resource Managers with Network Resource Managers

SC20 AutoGOLE/SENSE Persistent Testbed:

ESnet, SURFnet, Internet2, StarLight, CENIC, Pacific Wave, AmLight, RNP, KISTI, Tokyo, Caltech, UCSD, PRP/TNRP, FIU, CERN, Fermilab, Umd, DE-KIT



Caltech/
UCSD/
Sunnyvale
Moving to
400G/
2 X 200G
with CENIC

Courtesy T. Lehman

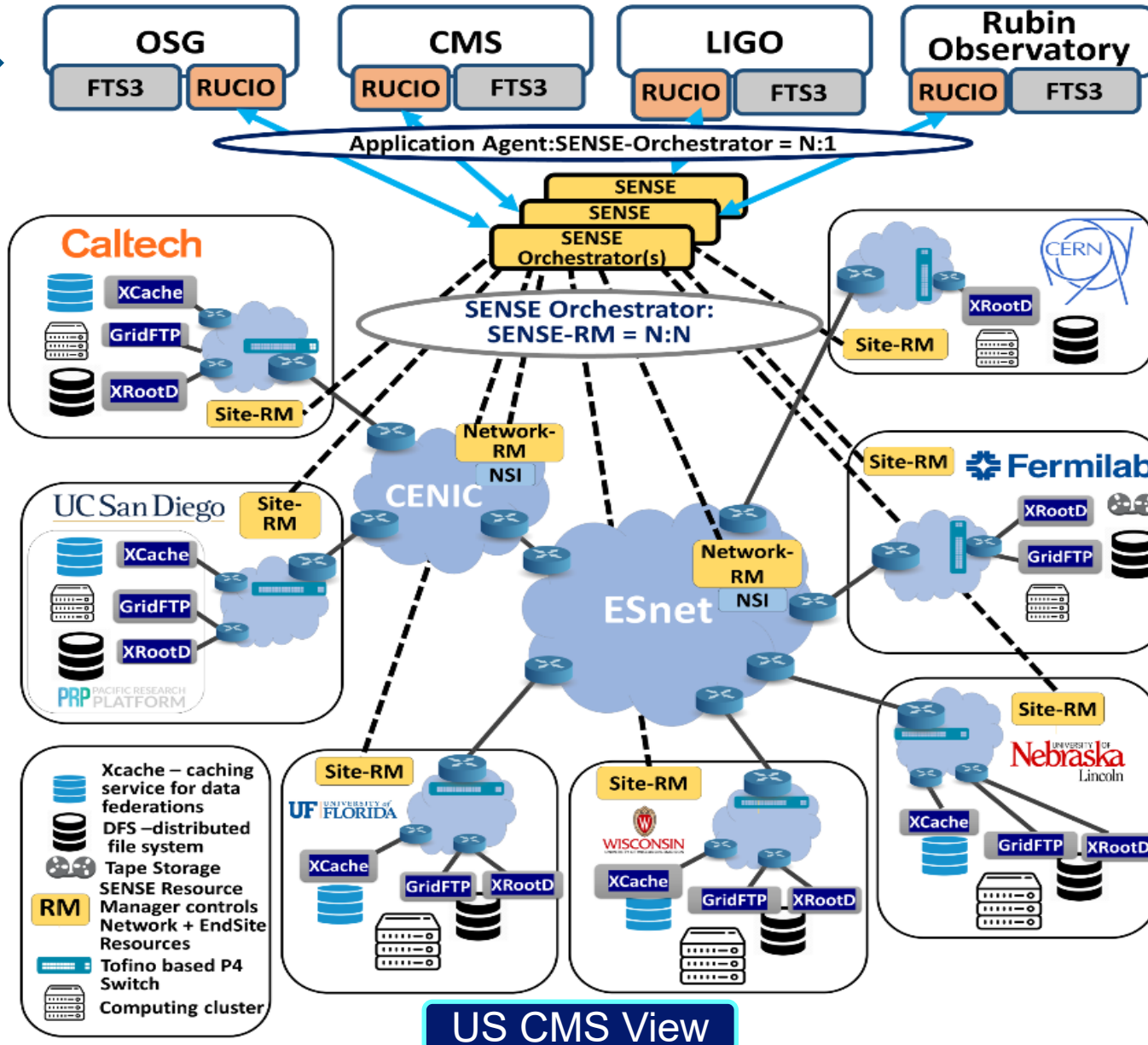
2021
ESnet6/
High Touch
FABRIC
BRIDGES

US CMS Tier2s
UERJ
UNESP
KAUST
SKAO
AarNet
TIFR et al

Federation with
the StarLight &
GEANT/RARE
P4 Testbeds

Interfacing to Multiple VOs With FTS/Rucio/XRootD

SENSE Orchestrator, Site and Network RMs



Courtesy
T. Lehman

Interfacing to Multiple VO's With FTS/Rucio/XRootD

LHC, Dark Matter, ν , Heavy Ions, VRO, SKAO, LIGO/Virgo/Kagra; Bioinformatics

OSG Data Federation

- Cache at institution
- Cache in the backbone
- Future Deployments



More than a dozen caches deployed across 3 continents

Collaboration	Working Set	Data Read	Reread Multiplier
DUNE	25GB	131TB	5.4k
LIGO (private)	41.4TB	3.8PB	95
LIGO (public)	4.3TB	1.5PB	318
MINERVA	351GB	116TB	340
DES	268GB	17TB	66
NOVA	268GB	308TB	1.2k
RPI_Brown	67GB	541TB	8.3k

7 most popular data areas



European Science Data Center



Vera Rubin Observatory



SKAO Key Science Drivers

- Testing General Relativity
(Strong Regime, Gravitational Waves)

Cosmic Dawn - EOR
(First Stars and Galaxies)

Cradle of Life
(Planets, Molecules, SETI)

Galaxy Evolution
(Normal Galaxies $z \sim 2-3$)

Cosmology
(Dark Matter, Large Scale Structure)

Cosmic Magnetism
(Origin, Evolution)

Exploration of the Unknown

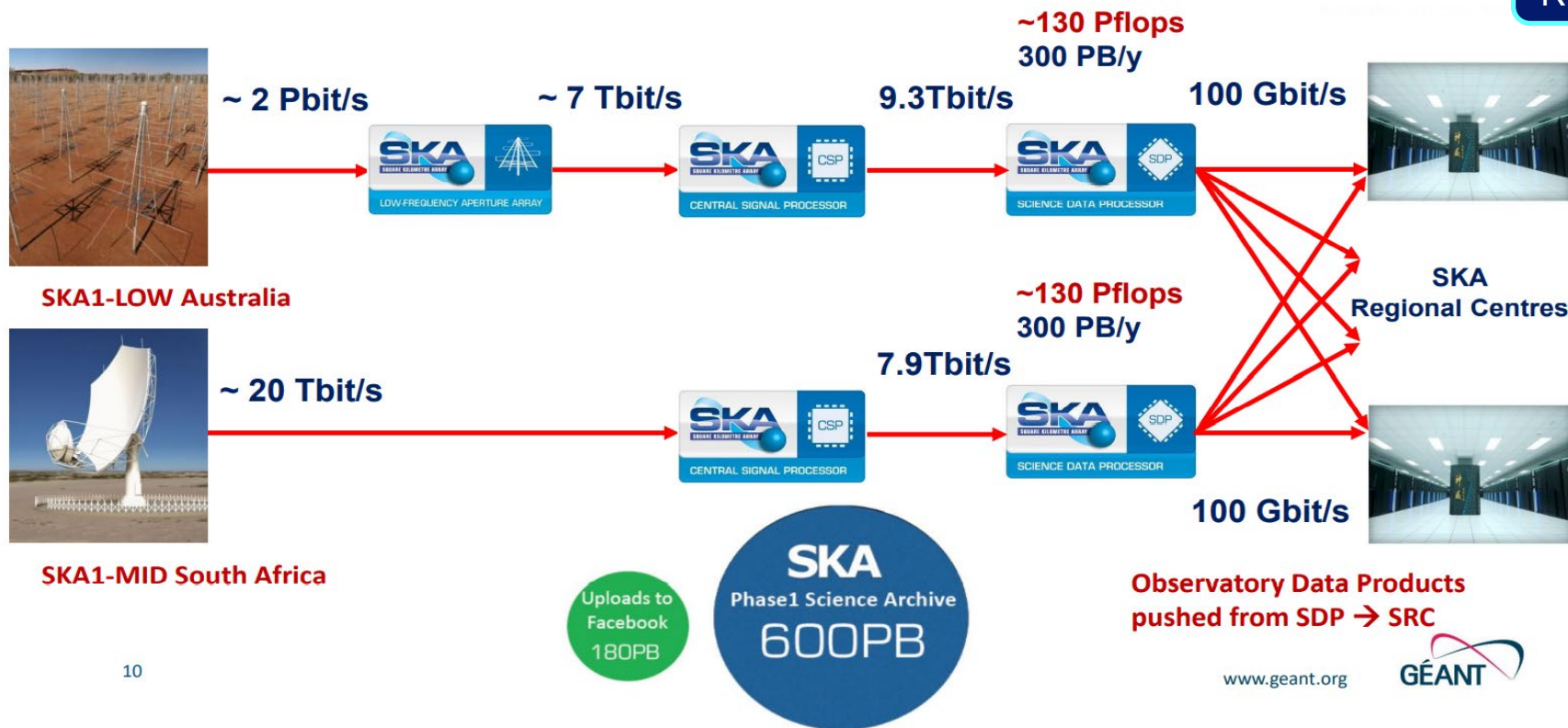


Courtesy
R. Hughes-Jones

SKAO Phase1 Data Flows: Telescope Arrays to Central Signal Processors to Science Data Processors to Science Regional Centers

SKA Phase1 Data Flows

Courtesy
R. Hughes-Jones



CSP – SDP Network

- Long-haul: 8.1 Tbit/s over 820 km SKA1-Low 9.5 Tbit/s over 912 km SKA1-Mid

Exabyte Archive; ~10 Tbps Flows;
1 to 80 X 100G Bursts

Traffic Pattern:

Visibility, Transients 80* 100 Gigabit Bursts

VLBI 100 Gigabit continuous

Pulsar Search 740 * 1 Gig = 8 * 100 Gigabit Bursts

Pulsar Timing 1 * 100 Gigabit Bursts

Protocol:

UDP/IP

UDP/IP

TCP/IP

TCP/IP

Design for peak rates

GNA-G Data Intensive WG: Activities Towards the Goals

- **Identify open source tools and services, and those of the partners, that can be used to build, grow and operate the new infrastructure**
- **Design and establish a development framework: 3 pathways**
 - **Consider which tools can be used (as is, or nearly) for common use**
 - **Design APIs to enable interworking among tools/subsystems where needed**
 - **Design and develop “mediation code” at the interfaces where needed.**
- **Follow open source software, technology and pricing roadmaps to identify and exploit opportunities, to address the challenges**
- **Follow and/or influence the develop paths of our WG member projects to address or mitigate the challenges, and help set our design paths**
- **Look for opportunities to engage our partner projects, and/or build new ones (including new funding sources) to develop the development and operational manpower needed**
 - **Develop funding agency and industry relationships**
- **Coordinate with the Telemetry, Virtualization, and Anomaly Detection WGs in the above**
- **Establish metrics of success, from simple to complex (See backup slides)**
 - **Engage with CS/EE, SDN and optimization experts as needed**

- **Leveraging, Coalescing, Integrating: the communities' tools & services**
- **How much of the infrastructure is devoted to major science programs**
 - **Sharing and funding models**
 - **New modes of operation with real-time in depth information; trends in industry**
 - **New controlled modes of use; both the managed and "unmanaged" parts**
- **VOs need to develop top to bottom operational models; accounting for classes of work, resource usage by class**
- **Adaptive and Predictive: Data transactions with times to completion**
- **Network and Site Engineers and Scientists: learning to work together, with a global real time system**
- **Capacity versus complexity: what are the tradeoffs between capacity cost versus complexity and development costs ?**
- **Human capital: developing a new generation of engineers and scientists able to develop, operate and/or optimize the new class of systems**
- ★ **The wide-ranging societal value of developing such a workforce & system**

Charter: https://www.dropbox.com/s/4my5mjl8xd8a3y9/GNA-G_DataIntensiveSciencesWGCharter.docx?dl=0

- 1. Set up a group of data management and development POCs among the partner science programs and network organizations**
- 2. Consider or else help develop roadmaps for the estimated requirements of the science programs, and a complementary roadmap of the affordable capacity along the routes that interconnect the partner's sites.**

This implies engagement through the POCs to understand the requirements resulting from each program's workflow, and technology tracking, projections and operational scenarios to match the affordable capacity to the requirements.
- 3. Work with the AutoGOLE/SENSE WG to define and evolve a common set of services, and the interfaces to the data management software system/stacks of the partner projects and the services needed to support their workflow.**
- 4. Coordinate this WG's efforts with the NSF IRNC, PRP, FABRIC, AutoGOLE/SENSE, Bridges and other testbeds to create an at-scale network testbed infrastructure for prototyping and development.**
- 5. Develop an Architecture and Proof of Concept(s) software and demonstrations to help develop and validate the operational aspects and required parameters and performance of the common services and interfaces to the various science programs' workflows.**
- 6. Work with the Telemetry WG, and partners including PRP and AmLight to define and evolve the network monitoring services needed to support the partner organizations' workflow.**

- 7. Build a software infrastructure to interface with partner organizations & projects**
- 8. Define interfaces/APIs to work with each of a starting list of partners' data management systems, and the tools used for production dataset processing and distribution for analysis**
 - **Define and develop tools that allow partner organizations to allocate bandwidth along defined paths, within available limits, coexisting with best effort services.**
- 9. Define and develop mechanisms and tools that allow flows to be identified and associated with a series of "priority" activities of the major partners.**
 - **Under constrained conditions provide functions that allow each partner to prioritize their allocations**
- 10. Define and develop mechanisms and tools that allow fair sharing among multiple partners using the shared global testbed.**
- 11. Develop metrics, algorithms and services that seek to optimize operation of the testbed according to the metrics**
- 12. Work with the partners to setup a process by which the methods and tools developed on the testbed are integrated into preproduction services supporting the workflows of the partners**
- 13. Work to scale the prototypical and pre-production services to production, on an agreed upon timescale, set by the major milestones of the partner programs.**

The GNA-G and a Next Generation Networking System for Data Intensive Sciences

- **Mission: To meet the challenges of globally distributed Exascale data and computation faced by the major science programs**
- **Coordinate provisioning the feasible capacity across a global footprint, and enable best use of that infrastructure**
- **Beyond capacity alone, enable the science within constraints. Approach:**
 - **Stable, resilient high throughput flows**
 - **Controls at the network edges, and in the core**
 - **Dynamic, adaptive operations among the sites and networks; Increasing negotiation, adaptation, with built-in intelligence**
 - **Real-time coordination among the VO and Network Orchestrators**
 - **A new “Consistent Operations” paradigm: goal-oriented, policy-driven**
- ★ **Bringing Exascale, pre-Exascale HPC and Cloud facilities, into the data intensive ecosystems of global science programs**
 - ★ **Petabyte transactions and caching using state of the art + emerging network and server technology generations; Tbit/sec demonstrators**
- ★ **We require a comprehensive, forward looking global R&D program**
- ★ **The GNA-G and its DIS WG, have key roles in this essential endeavor**



Extra Slides

Follow

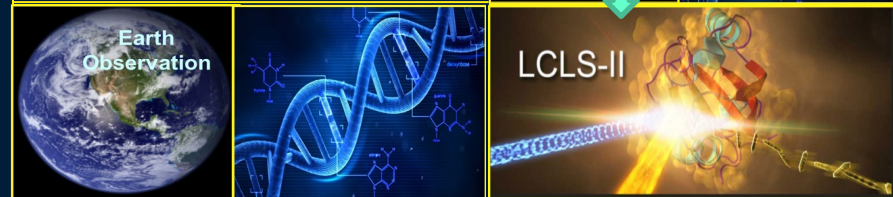
A New Era of Challenges: Global Exabyte Data Distribution, Processing, Access and Analysis



- **Exascale Data for the LHC Experiments**
 - ~1 Exabyte Stored by 2019;
to ~ 10-50 EB during HL LHC Era
- **Network Flow: 45-60 Gbytes/sec**
 - ~1.5 Exabyte flowed over WLCG in 2019
- **Emergence of 400-800G in Hyper-Data Centers, 100-200G on Terrestrial WANs**
 - 400G in Wide Area by 2022 ?
- **Network Dilemma: Per technology generation (~10 years)**
 - Capacity at same unit cost: 4X
 - Bandwidth growth: 35-70X in Internet2, GEANT, ESnet
- **LHC Run3: likely reach a network limit**
- **Unlike the past: Optical and switch advances are evolutionary**
Physics Limits by ~HL LHC Start

New Levels of Challenge

- **Global data distribution, processing, access and analysis**
- Coordinated use of massive but still limited *diverse* compute, storage and network resources
- **Coordinated operation and collaboration *within and among* scientific enterprises**

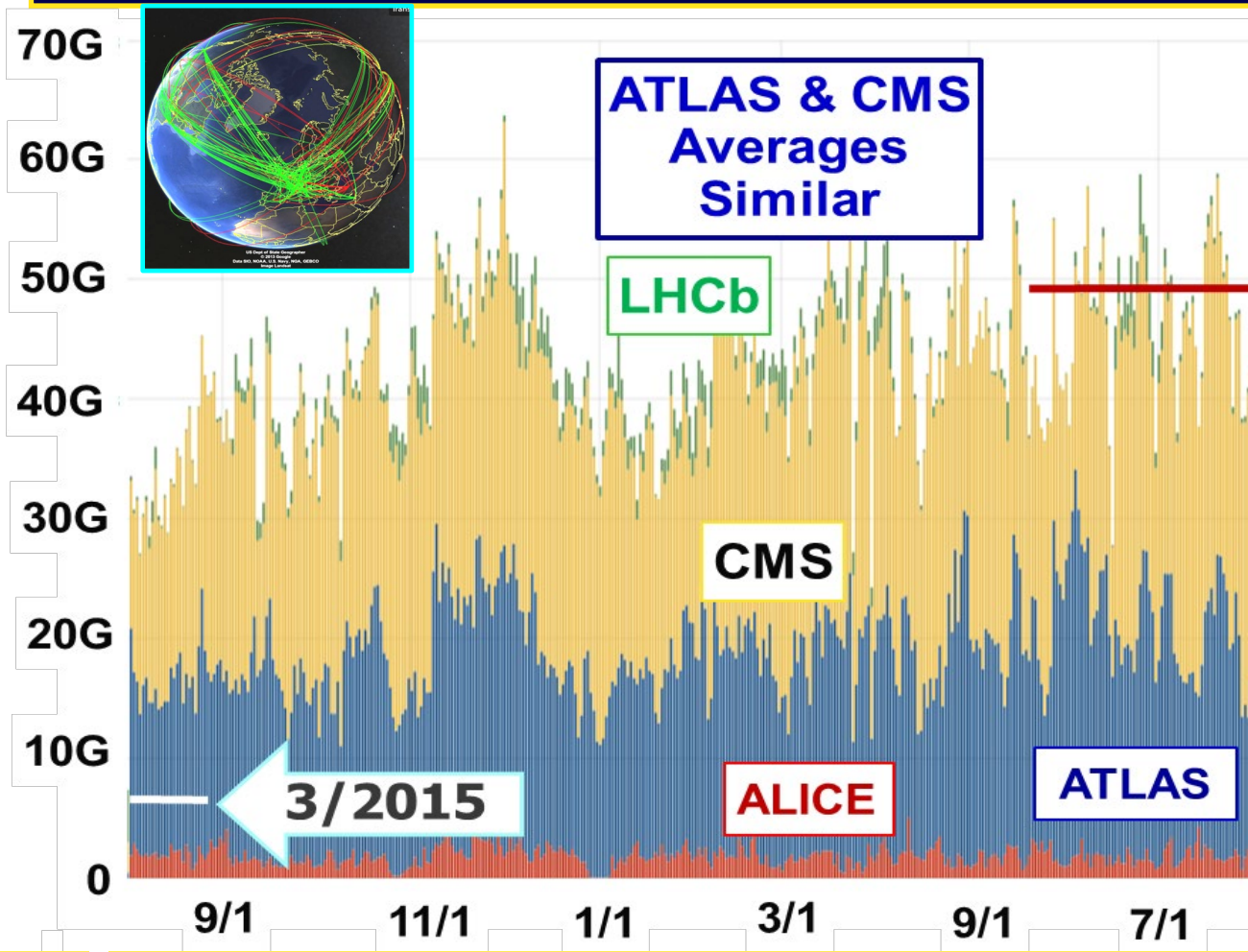


- **HEP will experience increasing Competition from other data intensive programs**
 - **Sky Surveys: LSST, SKA**
 - **Next Gen Light Sources**
 - **Earth Observation**
 - **Genomics**

LHC Data Flows Have *Increased* in **Scale and Complexity** since the start of LHC Run2 in 2015



WLCG Transfers Dashboard: Throughput Aug. 2018 – Aug. 2019



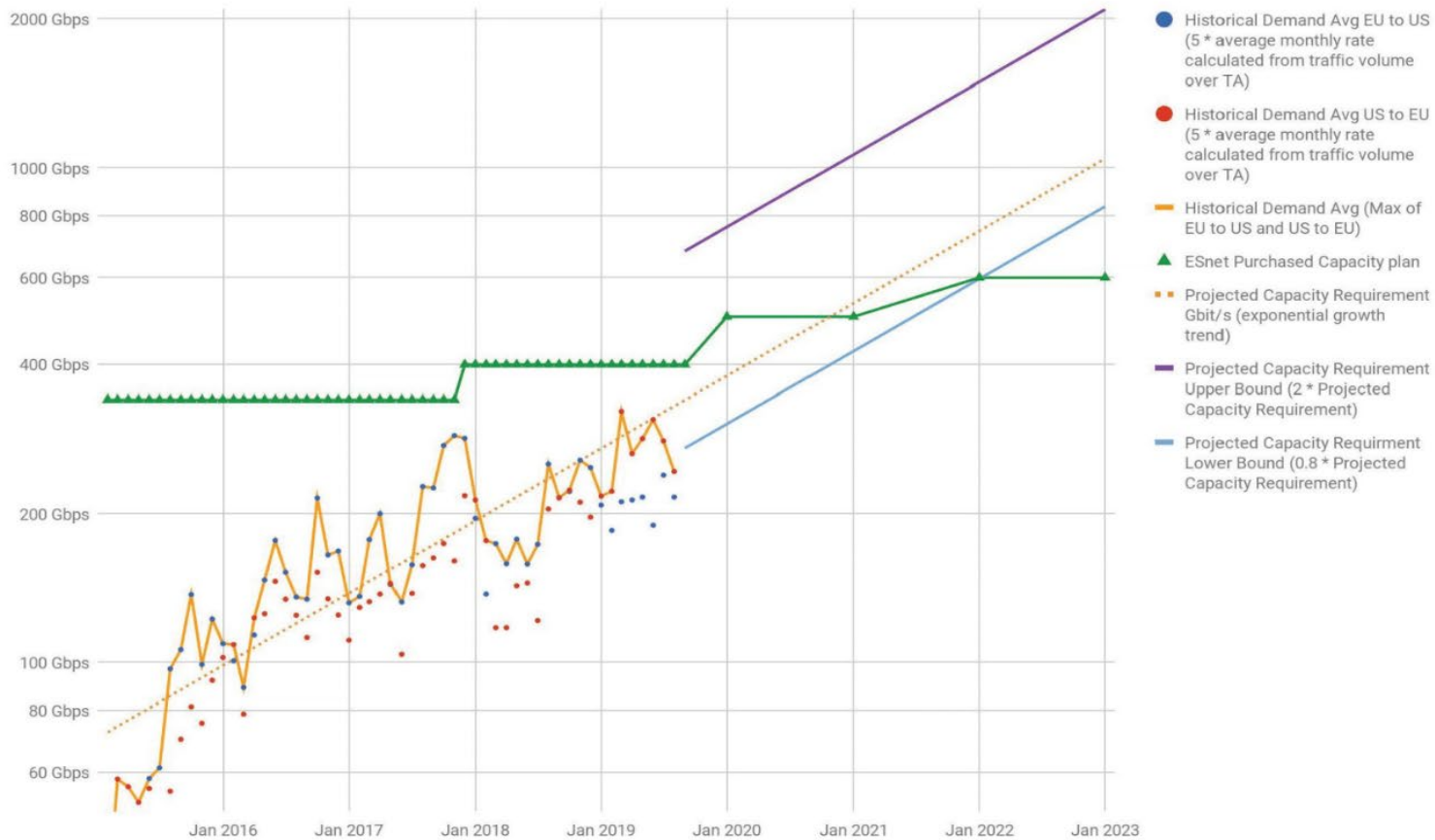
45-50 *GBytes/s Sustained*
60+ *GBytes/s Peaks*

Complex Workflow

- 700k jobs (threads) simultaneously
- Multi-TByte to Petabyte Transfers;
- 6-17 M File Transfers/Day
- 100ks of remote connections

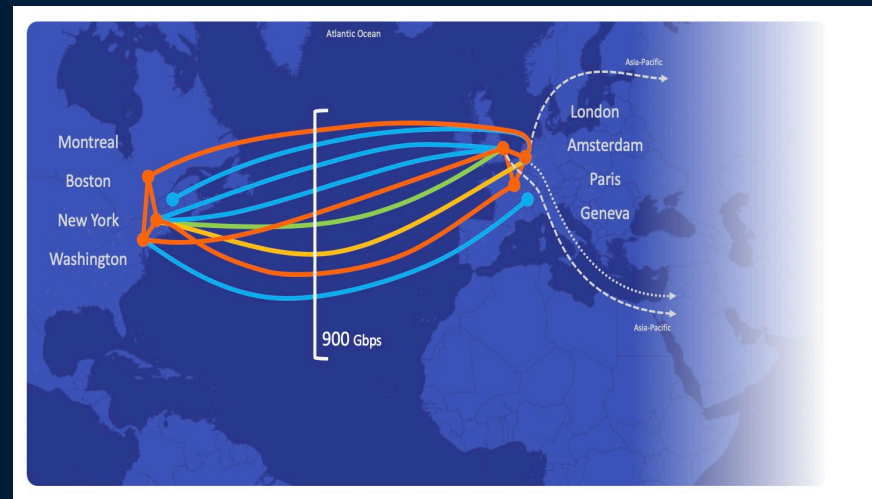
7X Growth in Sustained Throughput in 4.3 Years: +60%/Yr; ~100X per Decade

European Demand and Capacity Forecasts (updated Sept 2019)



- Recommendation from ESnet6 technical review:
ESnet should consider spectrum acquisition as an option for the non-OLS footprint to serve the science community that depends upon capacity growth of this connectivity.

- **Currently: 9x 100 Gbit/s lambdas between GXP**
 - 7: Internet2, NORDUnet, **ESnet**, SURFnet, CANARIE, and GÉANT
 - 1: NSF-funded **NEAAR** Project
 - 1: Japan's NII/SINET
- **Started in 2012**
- **First light in 2013**



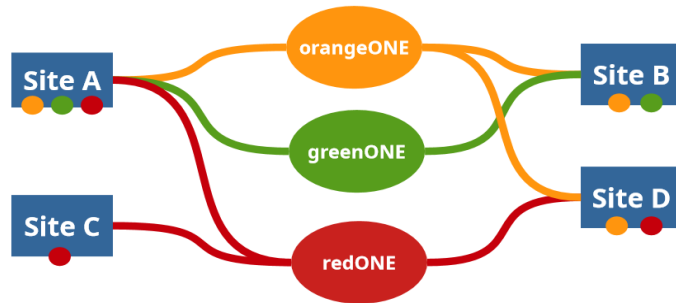
- **Possible Future Directions from Late 2020 or 2021**
 - **ANAv2: Long-term commitments on bandwidth or spectrum**
 - **ANAv3: At the table with new cable builds, anchor tenantry?**

**Aim: Rightsized, upgradable, resilient bandwidth
for less money across the North Atlantic Ocean**

MultiOne and DUNEOne

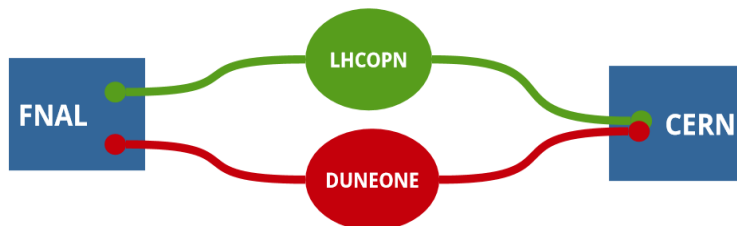
Recap: multiple "ONEs"

- Each site joins only the VPNs it is collaborating with, to reduce the exposure of their data-centre/Science-DMZ
- If doable, each Collaboration funds its own VPN



DUNEONE prototype

ProtoDUNE and DUNE identified as possible use case to build a multiONE prototype

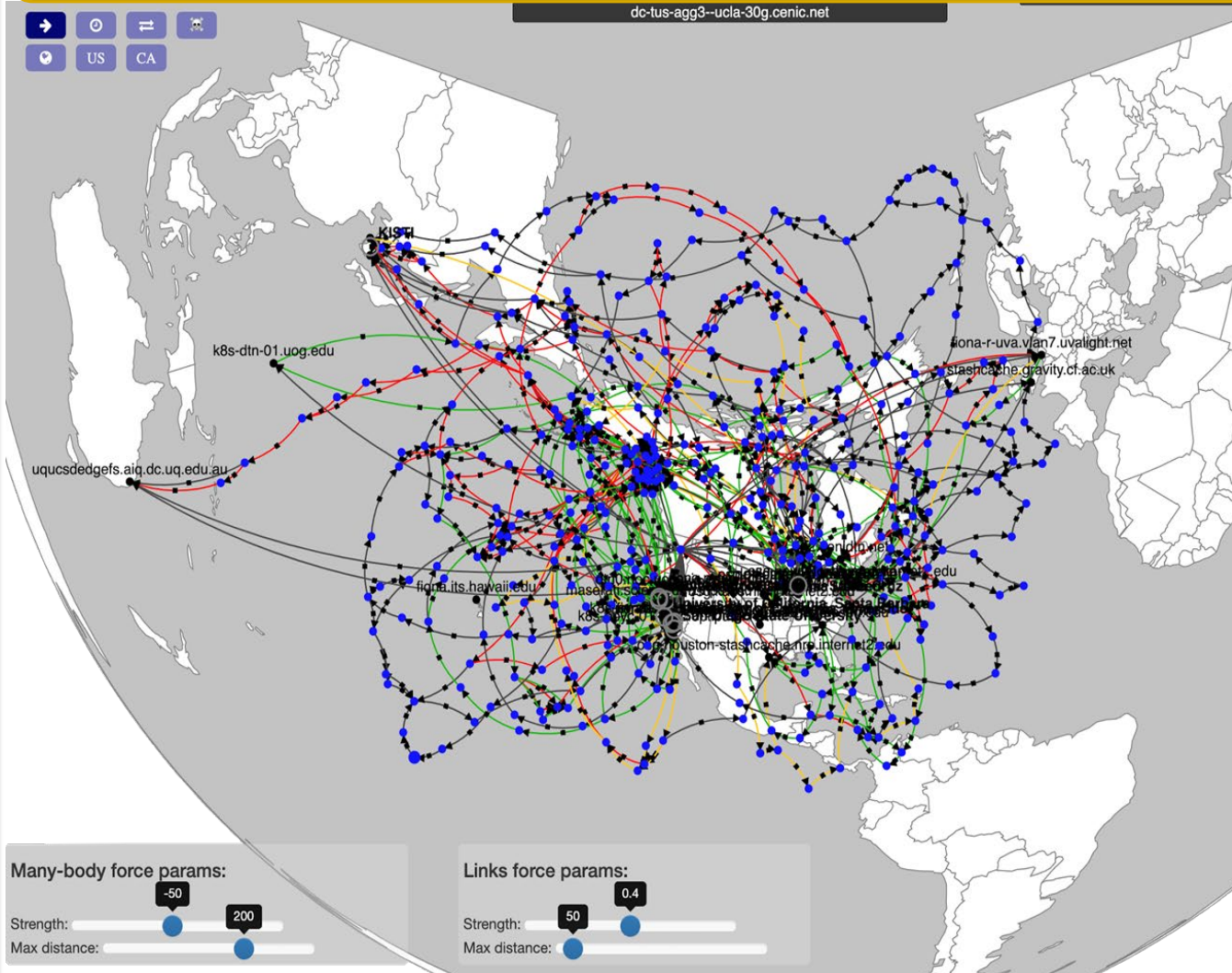


Status

- Not identified a solution to easily separate traffic, yet
- Explored traffic marking for policy routing with router vendor. Not possible with existing network processor, but it may be possible with upcoming ones
- ESnet is ready to implement a L2 circuit between CERN and FNAL. L3VPN will be considered when necessary at a later stage
- Analysing protoDUNE traffic to check if it could be identified by src and destination addresses

Edoardo Martelli at LHCOPN/LHCONE Meeting May 13, 2020

PRP and the Interactive Global Research Observatory Knowledge Base (IGROK)



Distributed Clusters in the Continental US, Netherlands, United Kingdom, Australia, Korea, Hawaii, Guam

Highly Capable “FIONAs”:
Data Transfers
SDN with Smart NICs
Machine Learning
Apps with GPUs

Automated Provisioning, Operations, Monitoring with an extensive toolset:
K8S+, netbox, Prometheus, Thanos etc.

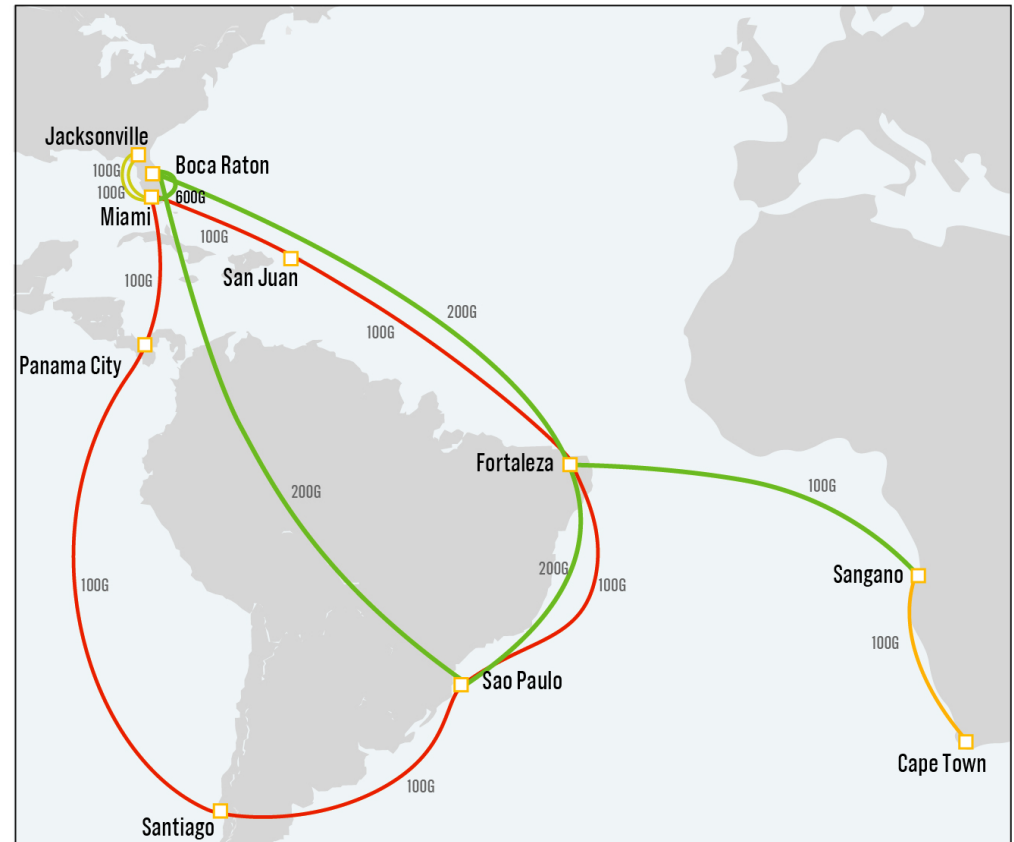
FABRIC Core: <https://fabric-testbed.net/>



https://nsf.gov/awardsearch/showAward?AWD_ID=1935966

Current AmLight Network Infrastructure

- AmLight Express path (green), 600Gbps in service:
 - 200G from Boca Raton to Sao Paulo
 - 200G from Boca Raton to Fortaleza
 - 200G from Sao Paulo to Fortaleza
- AmLight-SACS (green+yellow) extends AmLight Express from Fortaleza to Cape Town at 100Gbps
- 100G AmLight Protect ring Miami-San Juan, San Juan-Fortaleza, Fortaleza-Sao Paulo, Sao Paulo-Santiago, Santiago-Panama, and Panama-Miami (solid red)
- Express and Protect rings are diverse, operating on multiple submarine cables



AmLight is collaboration between FIU, NSF, ANSP, AURA, RNP, REUNA, RedCLARA, TENET/SaNREN

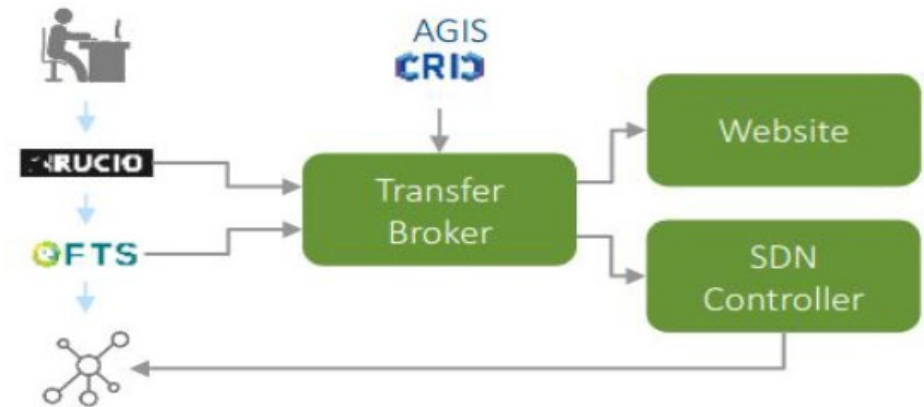
AmLight SDN Tools

- **OpenFlow Sniffer**: Developed to troubleshoot OpenFlow messages exchanged between Controllers and OpenFlow devices
- **SDNTrace** is an inter-domain path tracing tool, useful for discovery and troubleshooting inter-domain SDN networks
- **SDN Looking Glass** consolidates tools for monitoring and troubleshooting SDN networks on AmLight
 - Provides Topology Discovery;
 - Runs Path Traces of the Control Plane and Data Plane;
 - Sends alerts via e-mail and Slack; and,
 - Provides a REST API to be used by external SDN apps, auditing tools, and external NMS.
- **Kytos SDN Platform** is an open source project that aims to develop an SDN framework to facilitate the development of network applications (NApps)
 - Kytos started as an open source project funded by the State University of Sao Paulo to manage the LHC data transfers between LHC Tier 1 and Tier 2s
 - AmLight adopted Kytos to respond to the SLA requirements of the Vera Rubin Observatory science data transfers and transient alerts.
- **Kytos E-Line Napp** is a circuit provisioning application was developed on top of the Kytos platform
 - Service type defined by the Metro Ethernet Forum for connecting exactly two User Network Interfaces (UNI), so they can communicate only with each other
 - The Kytos E-Line application will be used to fully support the Vera Rubin network needs, including bandwidth reservation and prioritization

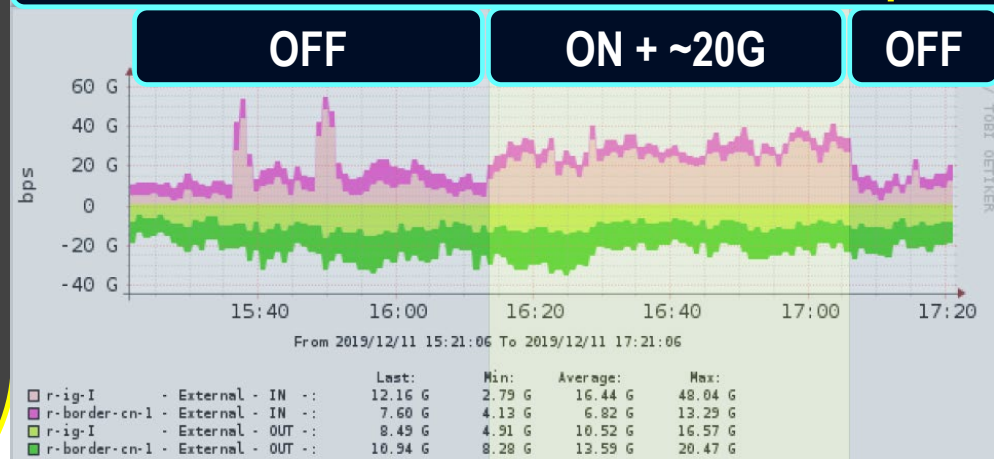
NOTED: Network Optimized Transfer of Experimental Data CERN/IT Project

- NOTED publishes network aware information on on-going massive data transfers, that can be used to provide additional capacity by orchestrating the network behavior (e.g. more effective use of existing network paths; finding alternates; load balancing).
- The advantage of starting with NOTED is that its Transfer Broker, as shown, can already interpret Rucio and FTS queues and translate them into network aware information with the help of the WLCG's database.
- While still in the prototyping stage, NOTED has already demonstrated the full chain with transfers between CERN and the Tier1s in Germany (DE-KIT) and the Netherlands (NLT1).

Transfer Broker Interfaces to Job Queues, SDN Controller, WLCG Database



Switch some traffic to DE-KIT LHCOPN path



Eduardo Martelli et al.

Application-Network Integration for Data-Intensive Science

Y. Yang, J. Zhang, K. Gao et al.: *IETF Standards Based*

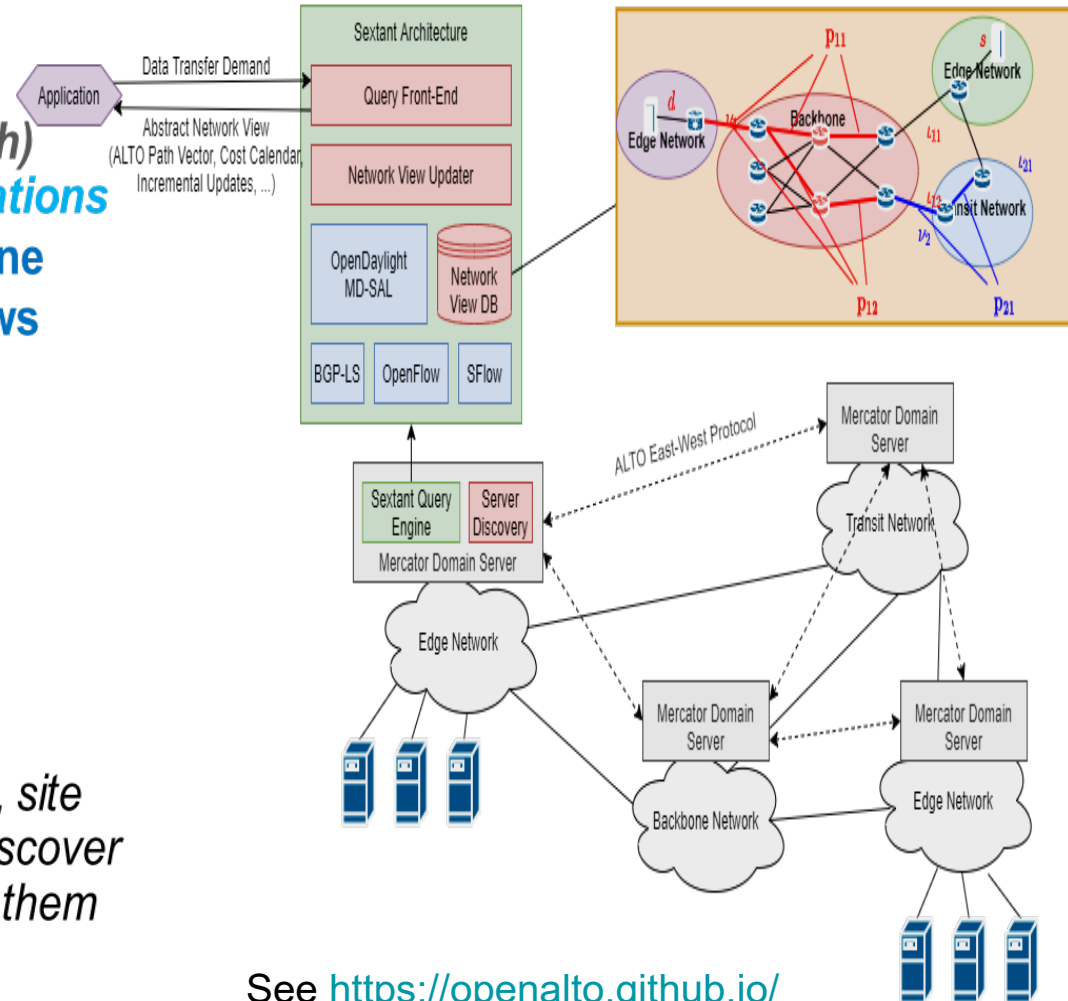
- **Automated Network Information Exposure**
automatically collect low-level network information (e.g., prefixes, routes, bandwidth) and expose abstract network view to applications

- Provide an advanced network query engine
- Expose on-demand abstract network views (e.g., ALTO path-vector, cost calendar, incremental updates, ...)
- Based on OpenDaylight controller
- Listen to multiple Southbound protocols e.g., BGP-LS, OpenFlow, SFlow, ...

- **Multi-domain Network Resource Discovery & Orchestration:**

Coordinate with multiple network domains (e.g., site networks, transit networks, backbones, ...) to discover on-demand network resources and orchestrate them

- Enable fine-grained inter-domain routing discovery
- Provide fine-grained, multi-domain resource discovery using a linear-inequality-based resource abstraction



See <https://openalto.github.io/>
for more details

ALTO IETF RFCs and Documents:

<https://datatracker.ietf.org/wg/alto/documents/>

Application-Network Integration for Data-Intensive Science

Y. Yang, J. Zhang, K. Gao et al.

Network Simulation

GNS3: A graphical network simulator

Mininet: Virtual OpenFlow network simulator

SDN Network Controller Platform

OpenDaylight: Open source SDN controller & platform

Network Management Tools

(1) Sextant: Automated network information collection, abstraction & exposure

❑ **Current features:**

- Information: IP aggregation & network distance
- Northbound: ALTO
- Southbound: BGP & BGP-LS, OpenFlow
- New features are still under active development

(2) Mercator: Multi-domain network resource discovery & orchestration

❑ **Current features:**

- Multi-domain resource queries for multiple flows
- **Flow-level (L4) resource reservation** using OpenFlow

Technical Stack

Network
Management
Services



Network
Controller



Network
Simulator



Mininet

A View of Metrics of Success: from Simple to Complex

- **While the service elements and approaches above provide a useful foundation to begin development, it will be up to the experiments and other client developers to build and test the system that helps each organization manage its workflow.**
- **The metrics of success can start out simple, but as resources become constrained, effective metrics become naturally more complex:**
- **Stage 1 Factors: Time to completion (TC) of a given transaction, percentage of successful transactions; average TC and maximum TC.**
 - **Avoid long tails in the TC distribution.**
- **Stage 2 Factors: Coordination of network resources with the use of computing and storage resources, as reflected in: Efficiency of CPU usage, efficient storage use within limits; minimize queue lengths. Balanced workflow: avoid starving a site**
- **Stage 3 Factors: Apply priority profiles. Define classes of work and queue profiles. Optimize through operational experience according to the above (simpler) metrics**
- **Stage 4: Construct abstract metrics of success based on the above metrics. Include policy-based elements such as preferred use of in-region resources, avoiding bottlenecks and other workflow issues at “system” level.**
 - **Learn through prototypes and pre-production systems which abstract metrics are effective with the right balance among performance, resource-use efficiency, policy and other system level (including common sense) factors.**

- **Stage 5: once effective metrics are developed in Stages 1-4, construct real-time adjustment mechanisms, and a foundation for automated adjustment and control**
- **Stage 6 (potential): once the operational foundation is built launch trials of automated optimization procedures, through reinforcement learning and/or graph neural nets or other techniques.**
- **Stage 7 (potential): Given the shifting requirements of the client virtual organizations, driven by deadlines such as data processing and simulation campaigns over periods of months, and the approach of major conferences, it may also be useful to evaluate long-term as well as short-term fair-sharing concepts.**
 - **Metrics could thus follow an “economic” model, and have terms that take into account the resources used over a quarterly or longer period.**
 - **Such models also can account, if needed, for above-standard priority use that is arranged recognizing the increased impact on other operations, translated to a scaling factor or other penalty applied to such high priority use in the accounting of resource usage.**
- **In the latter stages, the design and use of complex metrics could benefit from experts in (one or more of):**
 - **Control systems; Multi-objective optimization; Game theory**