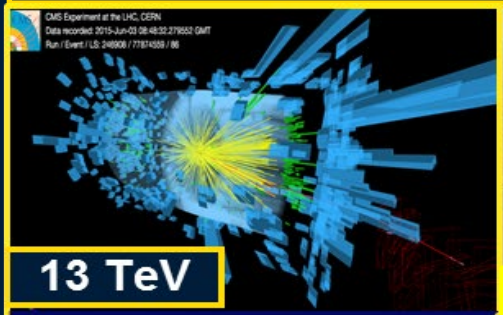
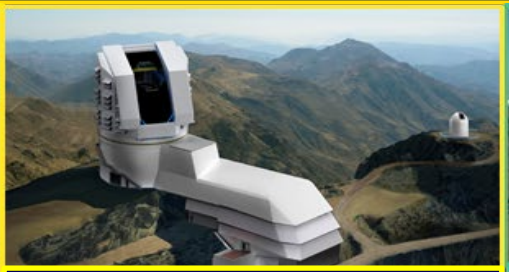
Networking Outlook, R&D and System Development

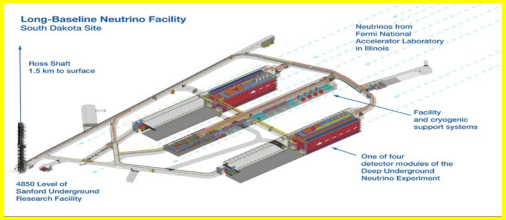
Towards a New Computing Model for the HL-LHC Era

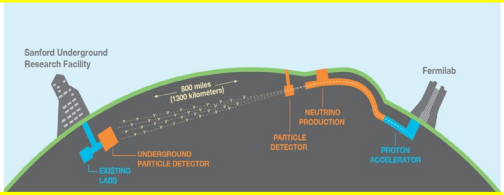


LHC



LSST





LBNF/DUNE

SKA

LHC Run3
and HL-LHC

DUNE

LSST SKA

BioInformatics

Earth
Observation

Gateways to a New Era

Harvey Newman, Caltech

OSG All Hands Network Session

March 5, 2021

Developing the Next Generation Computing Model
A comprehensive R&D program for the HL-LHC era

- **Top Line Message**

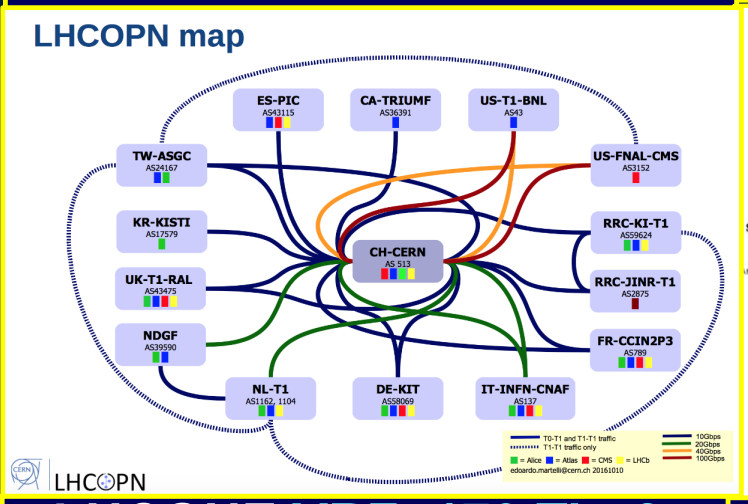
  A comprehensive R&D program to develop the architecture, design, prototyping, scaling and optimization of the HL-LHC Computing Model is required

  - **A new system coordinating worldwide networks as a first class resource along with computing and storage**
  - **Including innovative approaches in several areas**
  - **Leveraging and advancing several key developments: from regional caches/data lakes to networks with "intelligent" control planes and data planes [E.g SENSE, AutoGOLE, NOTED]**
  - **Leveraging regional network developments to form a worldwide fabric supporting OSG/HEP workflow**

- **The OSG, LHC experiments and the R&E Network community should jointly decide how such an effort should be organized and executed, to accomplish the paradigm shift by ~2027**

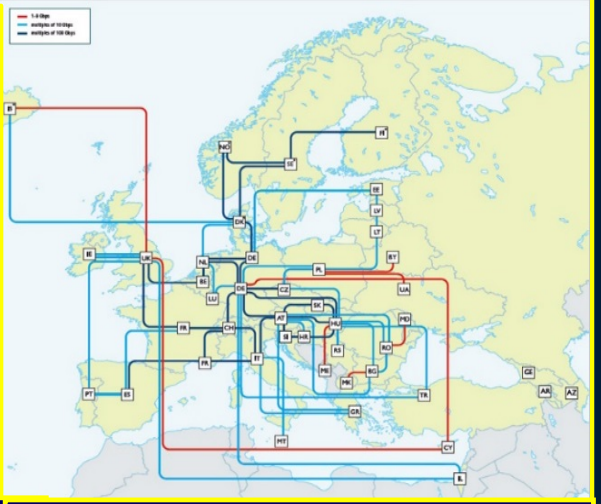
Core of LHC Networking LHCOPN, LHCONE, GEANT, ESnet, Internet2, CENIC…

LHCOPN: Simple & Reliable Tier0+1 Ops

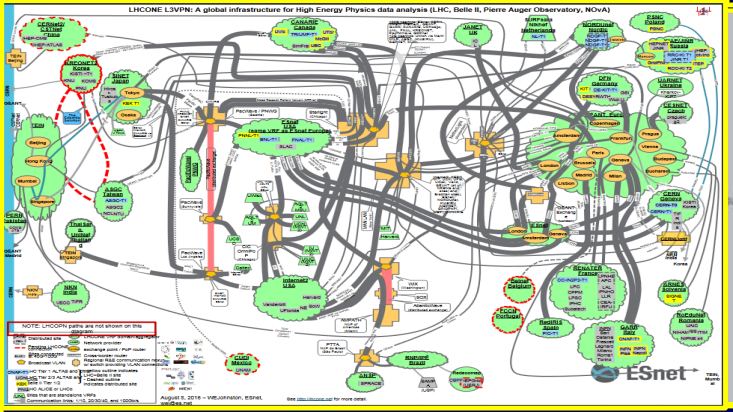


Internet2

GEANT

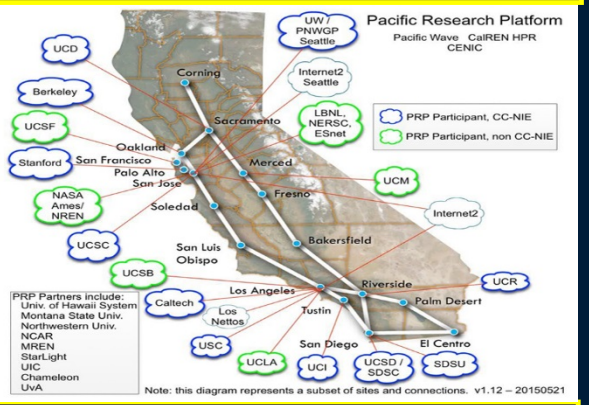


LHCONE VRF: 170 Tier2s



ESnet (with EEX)

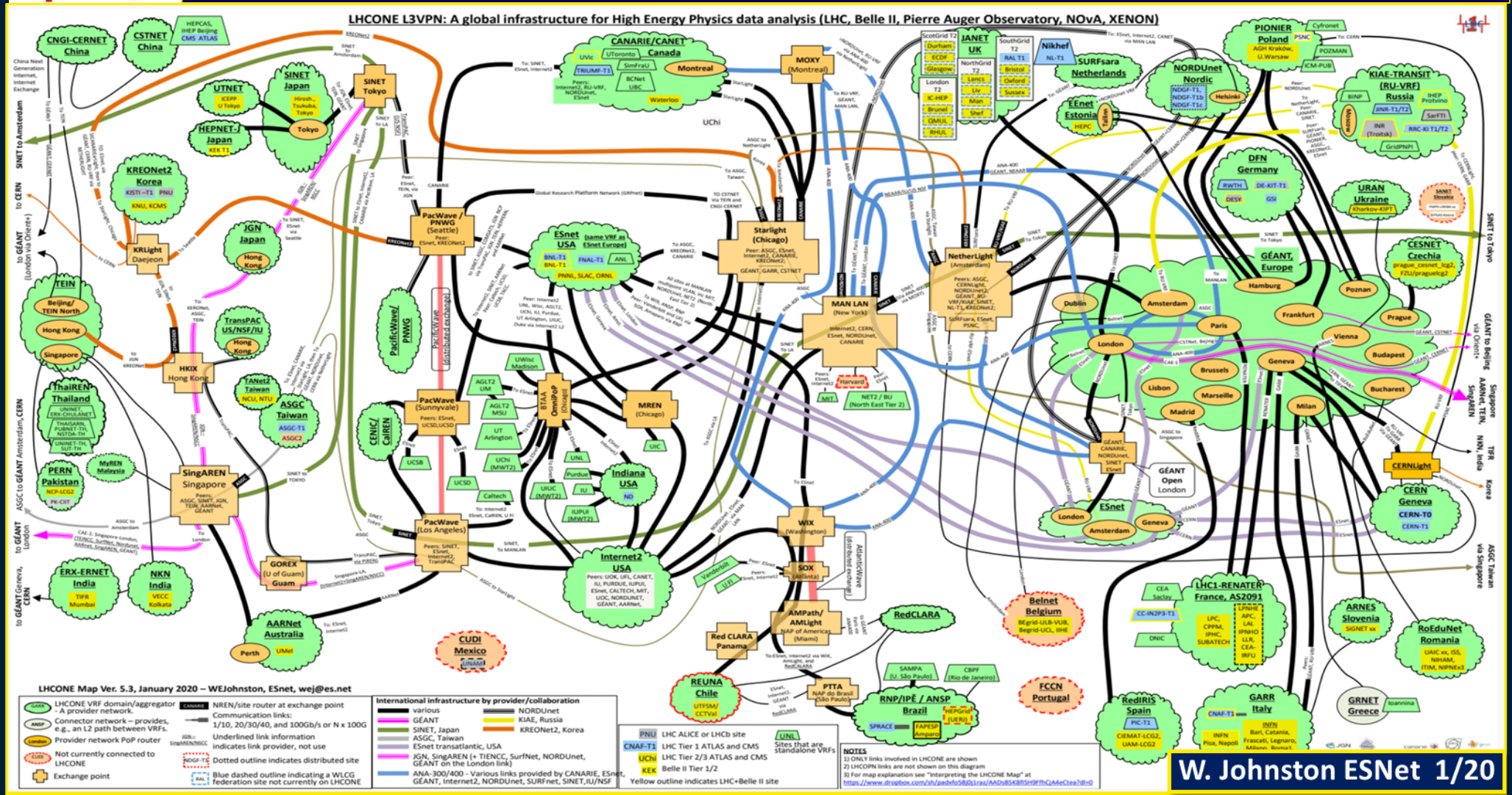
CENIC and PRP



+ NRENs in Europe, Asia, Latin America, Au/NZ; *US State Networks*

LHCONE VRF: The Challenge of Complexity and Global Reach

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



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 ?

Towards a Computing Model for the HL LHC Era.
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
  - This is projected to outstrip the affordable capacity
- At the January 2020 LHCONE/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, during and after the upcoming LHC Run3 (2022-24) and Beyond
- These needs are further specified in "blueprint" Requirements documents by US CMS and US ATLAS, submitted to the ESnet Requirements Review in August, and under continued discussion/development for a 2021 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

HL-LHC Network Needs and Data Challenges
Current Understanding: 3/2021

- **Export of Raw Data from CERN to the Tier1s (350 Pbytes/Year):**
  - **400 Gbps *Flat* each for ATLAS and CMS; +100G each for other data formats; +100 G each for ALICE, LHCb**
- **"Minimal" Scenario [*]: Network Infrastructure from CERN to Tier1s Required**
  - **4.8 Tbps Aggregate: Includes 1.2 Tbps Flat (24 X 7 X 365) from the above, x2 to Accommodate Bursts, and x2 for overprovisioning, for operational headroom: including both non-LHC use, and other LHC use.**
  - **This includes *1.4 Tbps Across the Atlantic for ATLAS and CMS alone***
- **Note that the above Minimal scenario is where the network is treated as a scarce resource, unlike LHC Run1 and Run2 experience in 2009-18.**
- **In a "Flexible Scenario" [**]: *9.6 Tbps, including 2.7 Tbps Across the Atlantic* Leveraging the Network to obtain more flexibility in workload scheduling, increase efficiency, improve turnaround time for production & analysis**
  - **In this scenario: Links to Larger Tier1s in the US and Europe: ~ 1 Tbps (some more); Links to Other Tier1s: ~500 Gbps**
- **Tier2 provisioning: 400Gbps bursts, 100G Yearly Avg: ~Petabyte Import in a shift**
  - **Need to work with campuses to accommodate this: it may take years**

[*] NOTE: Matches numbers presented at ESnet Requirements Review (Summer 2020)
[**] NOTE: Matches numbers presented at the January 2020 LHCONE/LHCOPN Meeting

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric; + LHCOPN

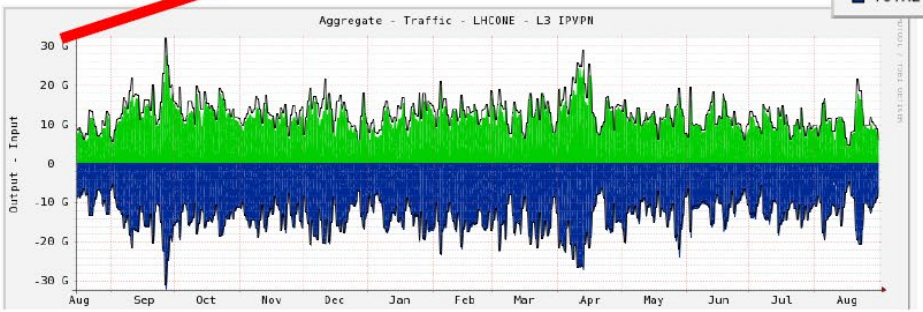
Global infrastructure for *HEP (LHC, Belle II, NOvA, Auger, Xenon)* data flows

Where were we? LHCONE

LHCONE in Europe
GEANT

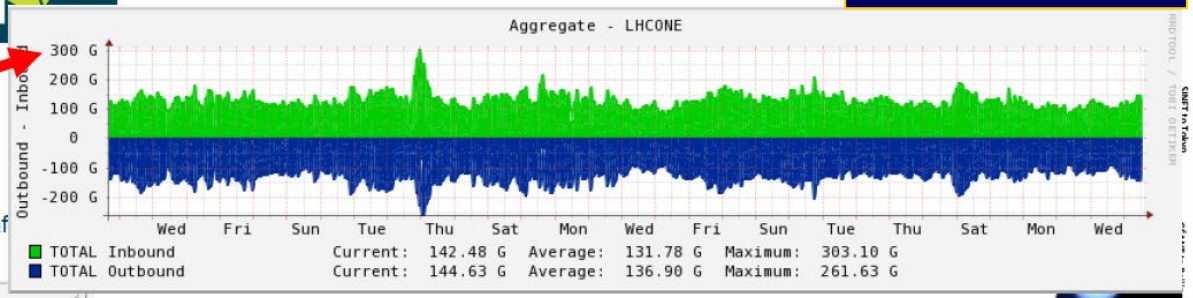
- Aggregate LHCONE traffic from all the NRENs and Peers
  - Average traffic ~25Gbps
  - Sustained Peaks ~35Gbps
  - Trans-Atlantic Traffic ~ 20Gbps (Peak)
- Graphs shows 1 day average traffic over last 12 months the peak traf is much highe

10x

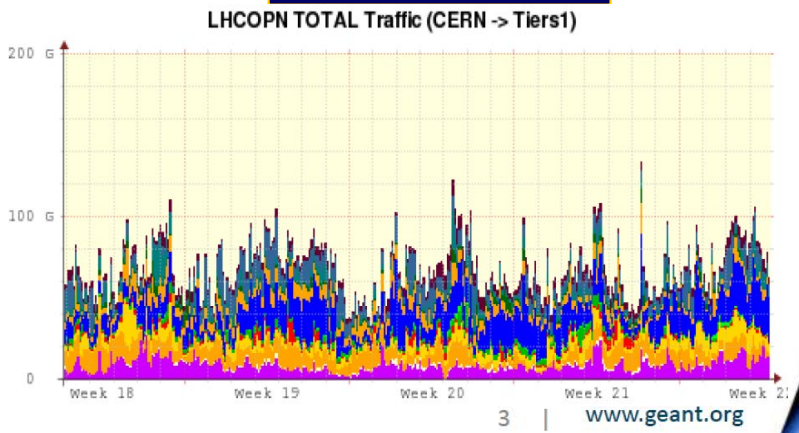


Connect | Communicate | Collaborate 17

LHCONE



+ LHCOPN



3 | www.geant.org

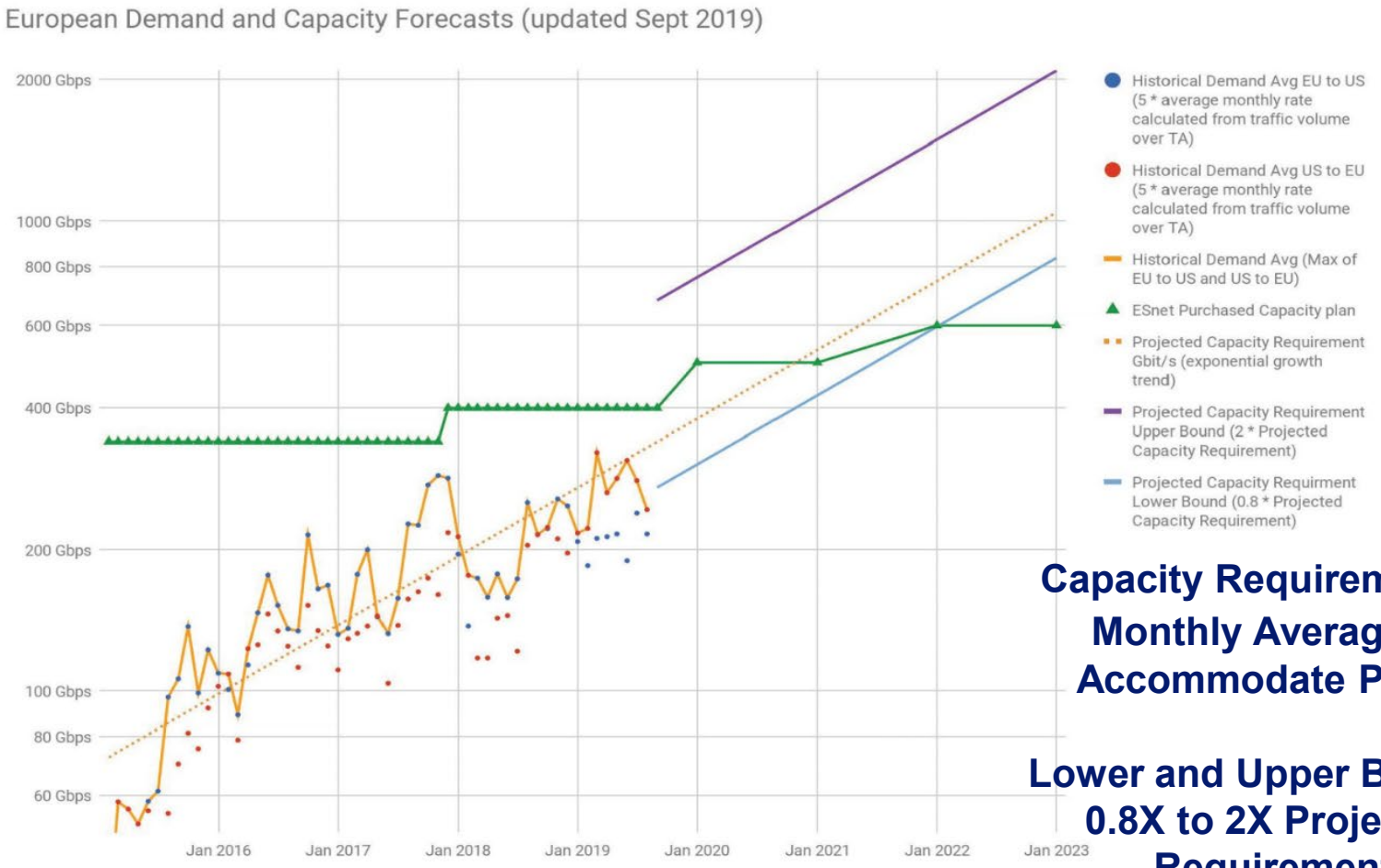
GÉANT

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

A complex system with limited scaling properties. So: Multi-ONE ? New Mode of Sharing ?

LHCONE traffic growing by 60%/Yr: a challenge already in LHC Run3 (2022-4)

Esnet Transatlantic Network Traffic Projections



Capacity Requirement = 5x Monthly Average to Accommodate Peaks

Lower and Upper Bounds: 0.8X to 2X Projected Requirement

- 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.

https://www.dropbox.com/s/yi9b1gc8v5q8jke/DeMar-US-CMS-BluePrint_3-17-20.pdf?dl=0

Capacity Requirements Analysis, Using ESnet Transatlantic Network Traffic Projections

GNA-G
Global Network Advancement Group

- Requirements based on recent traffic: 0.35 – 0.85 Tbps
  [based on 0.8 to 2X the 2016-19 traffic projection]
- Long Term Growth Rate 1.4X per year, or 2X every two years on average
- Hence *16X capacity requirement from 2020 - 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
- Long-term capacity per unit cost growth 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: 28-68 200G links across the Atlantic
  (for example: 7 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 jointly develop a new system that
  manages and coordinates the use of limited network resources

R&D on Network Capabilites

Issues: Bandwidth requests can (over)match capacity on some routes/segments; Aggregate of requests can overwhelm the available capacity and impede or block other use of the shared network infrastructures (e.g. across oceans, on campuses)

Approach: Develop a stateful network management system to address the issues, and reduce some of the need for over-provisioning

- **Prerequisites:** Such a system requires detailed monitoring information along network paths and at sites, to track progress/times to completion, and evaluate the impact of further significant allocations.

Key System Features include: ⇨ **R&D**

- Handling multiple requests taking policy and priority into account; (according to a new paradigm "to be defined")
- Giving weight to: performance/throughput, load balancing, good use of site resources, organizational and geographical preferences in assigning paths;
  - Eventually: a hierarchy of objectives + constraints in a multi-objective optimization strategy
- Identification, diversion and assignment to alternate, additional, or privileged paths when available, OR otherwise constraining the allocations not to impede others' existing best effort traffic on shared routes
- Deciding how to deal with the constraints as real-time requests keep coming in, via: Queueing and/or real-time adjustments of allocations, with notifications to and from the client workflow/data-management system
- Setting break-points on taking back capacity when the application does not well-use the allocation(s) it has been given

Steps to Arrive at a Fully Functional System by 2027
the Data Challenge Perspective (with thanks to Fkw)

- Three Types of Challenges
  1. Functionality Challenge : Where we establish the functionality we want in our software stack, and do so incrementally over time
  2. Software Scalability Challenge: Where we take the products that passed the previous challenge, and exercise them at full scale but not on the final hardware infrastructure
     - E.g. Use the cloud in 2021/22 and then FABRIC in 2023
  3. End-to-end Systems Challenge: On the actual hardware; can only be done once the actual hardware systems are in place.
- In US CMS: Targets are Q4 2022, 2023 (or 2024, 2025 if not all components are ready earlier) for 1 & 2; Q4 of 2026 for 3
- Remark: it's conceivable, maybe even likely that it takes multiple attempts to achieve sustained performance at scale with all of the new software we need, with the functionality we want.
- + Scaling Challenges: Demonstrate capability to fill ~50% full bandwidth required in the minimal scenario with production-like traffic: Storage to storage, using third party copy protocols and data management services used in production: *2021: 10%; 2023: 30%; 2025: 60%; 2027: 100%*

SDN Enabled Networks for Science at the Exascale

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

Creates Virtual Circuit Overlays. Orchestrator, Site and Network RMs

Model-based Site and Network Resource Managers

Designed to Adapt to Available SDN Systems

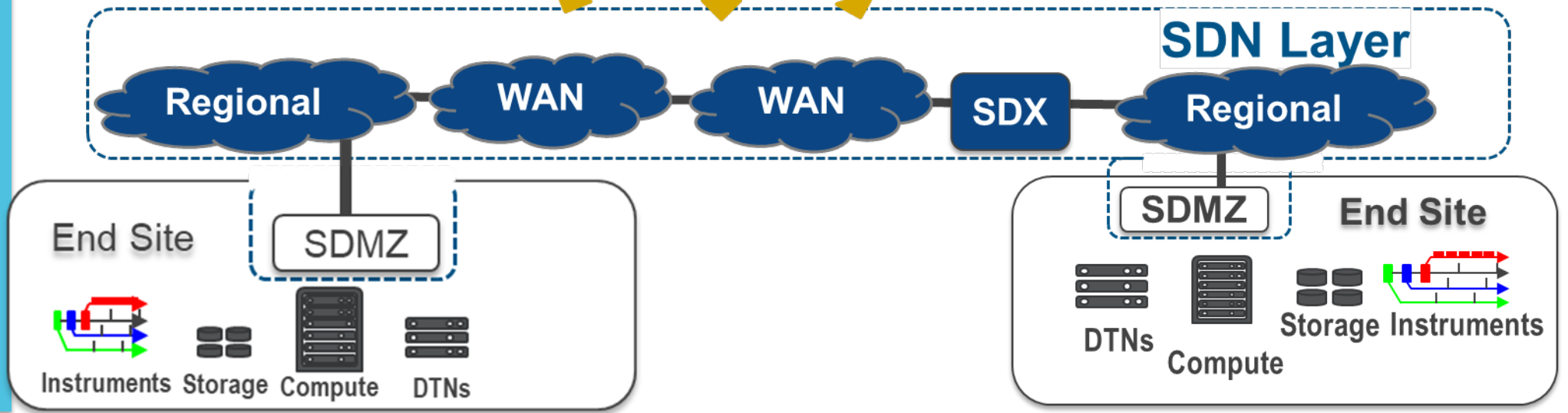
SENSE Native RMs are Available if no current automation layer

Application Workflow Agents

SENSE

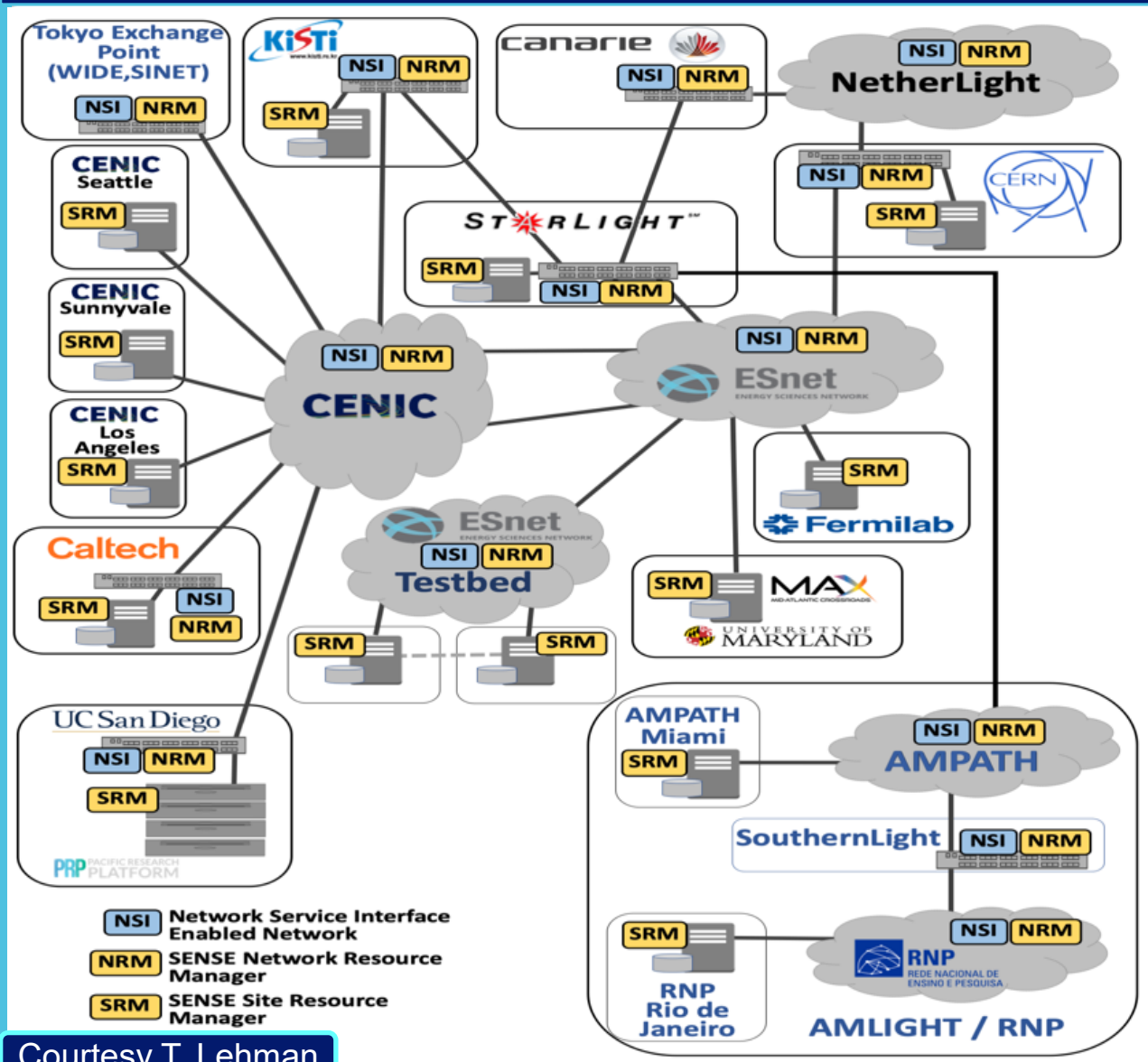
SENSE operates between the SDN Layer controlling the individual networks/end-sites, and science workflow agents/middleware

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



[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



Courtesy T. Lehman

2021 Outlook
ESnet6/
High Touch
FABRIC
BRIDGES

US CMS Tier2s
UERJ
Grid UNESP
KAUST
SANReN
SKAO
AarNet
TIFR et al

Federation with the StarLight GEANT/RARE & AmLight P4 Testbeds

400G Link(s) NetherLight-CERN

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

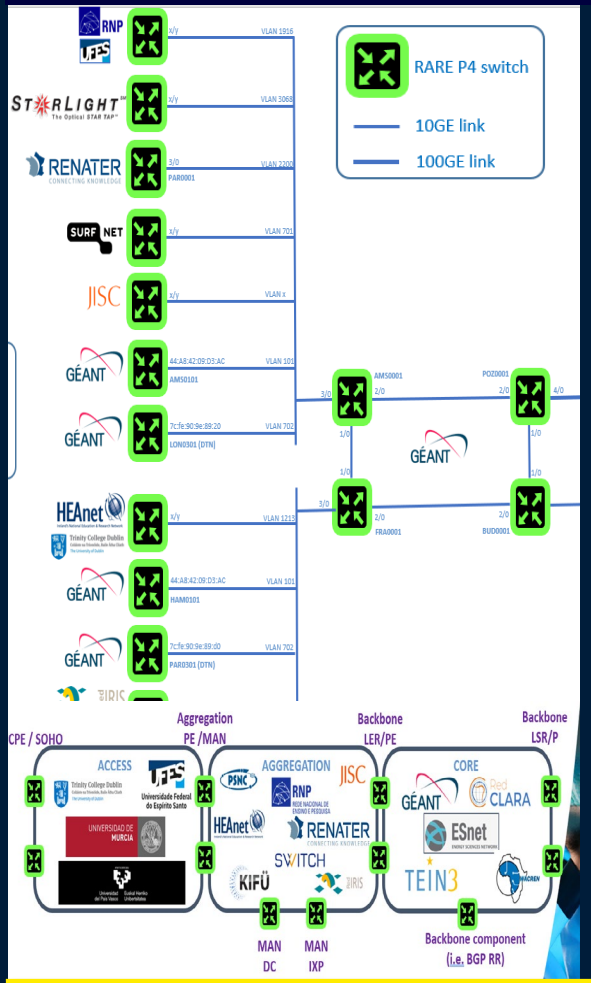
Automation Following Atlantic Wave SDX

R&D on Network Capabilities
Key Technologies

- **Overlay Networks based on Virtual Circuits across multiple domains: SENSE and its Orchestrator, Network & Site RMs**
  - **Adapts to multiple regional overlays, integrates with traditional networks**
  - **Allows emerging paradigms (SENSE, P4 programmable networks, NDN) to co-exist with traditional networks, migrate into production**
- **Packet marking, Traffic Shaping (Shawn's talk)**
- **Network telemetry: precision timestamps, classification of sets of flows, higher level services to handle flows by class (e.g. ESnet6 high touch, Pacific Research Platform)**
- **P4-based production switches;**
- **E.g. RARE Freertr in GEANT: Both production-ready open images in inexpensive switches; and fully programmable images for the academic and research community**
  - **Runs on Tofino-based (Edgecore, STORDIS) & standard (Mellanox) Spectrum2 and -3 Switches**
  - **Key functionality: define packet headers under full user control. With all needed attributes and state information at the edges; and in parts of the core when possible**
- **Also P4 on SmartNICs, Xilinx accelerators (PRP, ESnet)**

RARE
Router for Academia Research & Education

https://wiki.geant.org/display/RARE/Home

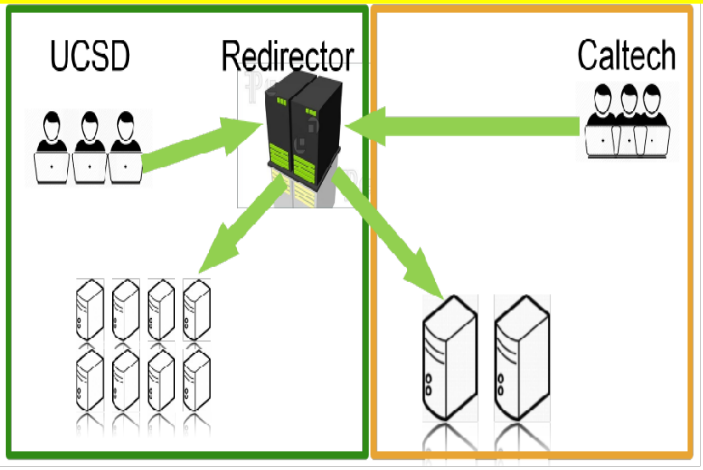


+ UCSD, Caltech, Umd/MAX, Tennessee Tech, Fermilab

(Southern) California ((So)Cal) Cache

Roughly 20,000 cores across Caltech & UCSD … half typically used for analysis

A 1.5 Pbyte Working Example in Production

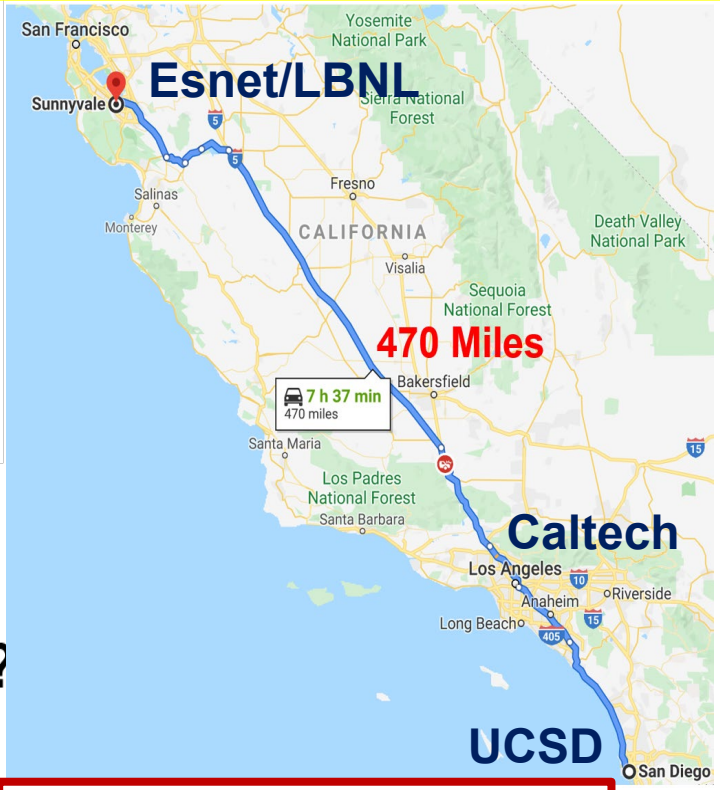


CPU in both places can access storage in both places.

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.

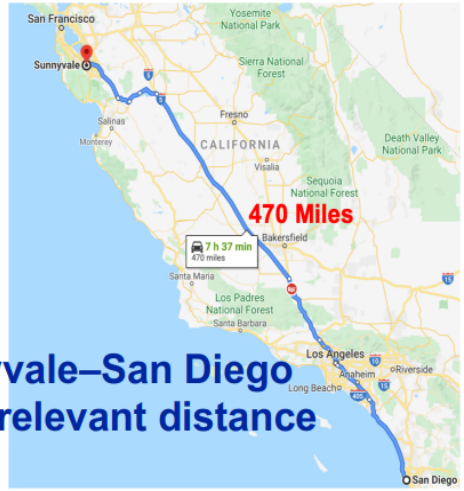
Upgrade by 2Q 2021 with CENIC and Ciena Support:
Caltech (4 X 100G) to UCSD (2 X 100G) and Sunnyvale (2 X 100G)

Application use case with CMS

- R&D Towards HL-LHC
  - High-Luminosity-LHC: the LHC performance to increase the potential for discoveries after 2025
  - All processing done via buffers
  - All analysis done via caches
- High level assumptions of annual volumes and use
  - 384 PB of RAW
  - 240 PB of AOD
  - 30 PB of MINI

  } Mostly kept on Tape => accessed a couple times per year
  - 2.4 PB of NANO

  } Mostly kept on disk => heavily re-used by many researchers
- Petabyte scale cache for CMS in CA
  - Deployed/Operated by UCSD and Caltech
  - To gain experience with MiniAOD reuse
  - Includes the ESnet cache node
  - 500 miles distance for a distributed cache is a socio-politically very relevant distance scale



Sunnyvale–San Diego is the relevant distance scale

Exploring in-network data caching - ESnet-US CMS collaboration study preliminary results

Resources

- Hardware: 40TB storage and 40Gbps networking capability
- Expected network utilization: about 10-20 Gbps

Alex Sim, Katherine Zhang, Ellie Copps, John Wu at LBNL
Chin Guok, Inder Monga at ESnet
Frank Wuerthwein, Edgar Hernandez, Diego Davila at UCSD

Summary

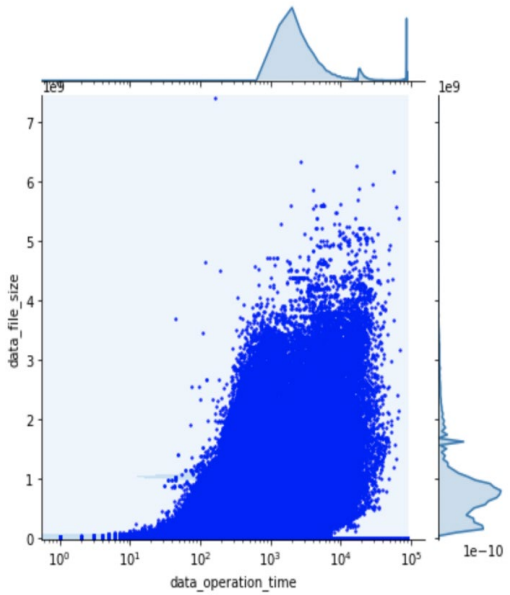
Demonstrated the capability of a network-based temporary data cache

Shared data caching mechanism

- Reduced the redundant data transfers, saved network traffic volume
- Summary of the 1,286,748 accesses from May 2020 to Oct 2020
  - Total 490.831 TB of client data access (first time reads and repeated reads)
  - Transferred/cached 168.08 TB (from remote sites to cache)
  - Saved 322.748 TB of network traffic volume (repeated reads only)
    - Network demand reduced by a factor of ~3

Further studies

- Cache miss rates
  - How caches affect each other when one or more of the federated caches are down
  - How many time a file needs to be retrieved from remote sites?
  - How are the cache misses affecting the application performance?
  - Regional cache impacting application performance (local vs remote data access)
- Cache utilization
  - How many Xcache installations are good enough?
  - What size of each disk cache would be appropriate?
  - If the number of physicists using the system doubles, how many more cache deployments are needed?



Transfer Size (bytes) vs.Duration (log(sec))

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 (-2020)

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
- 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

See https://www.dropbox.com/s/qsh2vn00f6n247a/GNA-G%20Meeting%20slides%20-%20TechEX19%20v0.8.pptx?dl=0

Working Groups: AutoGOLE/SENSE and Data Intensive Sciences WGs,
+ Monitoring, Security, Routing (TBC), etc.

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.*

Next Generation Networking System for HL LHC, HEP and Data Intensive Sciences

- We require a *comprehensive, forward looking global R&D program*
  - To meet the challenges faced by the LHC and other major science programs, including
    - Petabyte transactions and caching; using 400G to Tbps throughput
  - To coordinate provisioning the feasible capacity across a global footprint, *and* enable best use of the available infrastructure
  - While remaining compatible with uses by the at-large R&E community
- Beyond capacity alone, we need Real-time coordination among the VO (LHC) & Network Orchestrators to enable the workflows within constraints
  - To provide dynamic, adaptive, goal-oriented, policy driven operations among the sites and networks, based on
    - Comprehensive end-to-end monitoring
    - Stable, resilient high throughput flows
    - Controls at the network edges, and in the core
- The OSG, the Experiments, GNA-G and its DIS WG, have key roles in
  - Deciding how the effort to define and implement the new HL LHC Computing Model should be organized, designed and implemented
  - To successfully complete the needed paradigm shift by ~2027

Extra Slides Follow

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

- **Exascale Data for the LHC Experiments**
  - ~1 Exabyte by end of Run2
  - To ~50 EB during HL LHC Era
- **Network Total Flow of >1 EB this Year**
  - 1.6 Exabyte flowed over WLCG in 2019
- **Emergence Now of 400G in Hyper-Data Centers, 100 to 200G in Wide Area**
  - 400G in Wide Area by 2021-22
- **Network Dilemma: Per technology generation (~10 years)**
  - Capacity at same unit cost: 4X
  - Bandwidth growth: 35-70X in Internet2, GEANT, ESnet
- ***During LHC Run3***
  We will 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**

Vision: Next Gen Integrated Systems for Exascale Science: a Major Opportunity

Opportunity: Exploit the Synergy among

1. **Global operations data and workflow management systems developed by HEP programs, *to respond to both steady state and peak demands***
   - **Evolving to work with *increasingly diverse (HPC) and elastic (Cloud) resources***

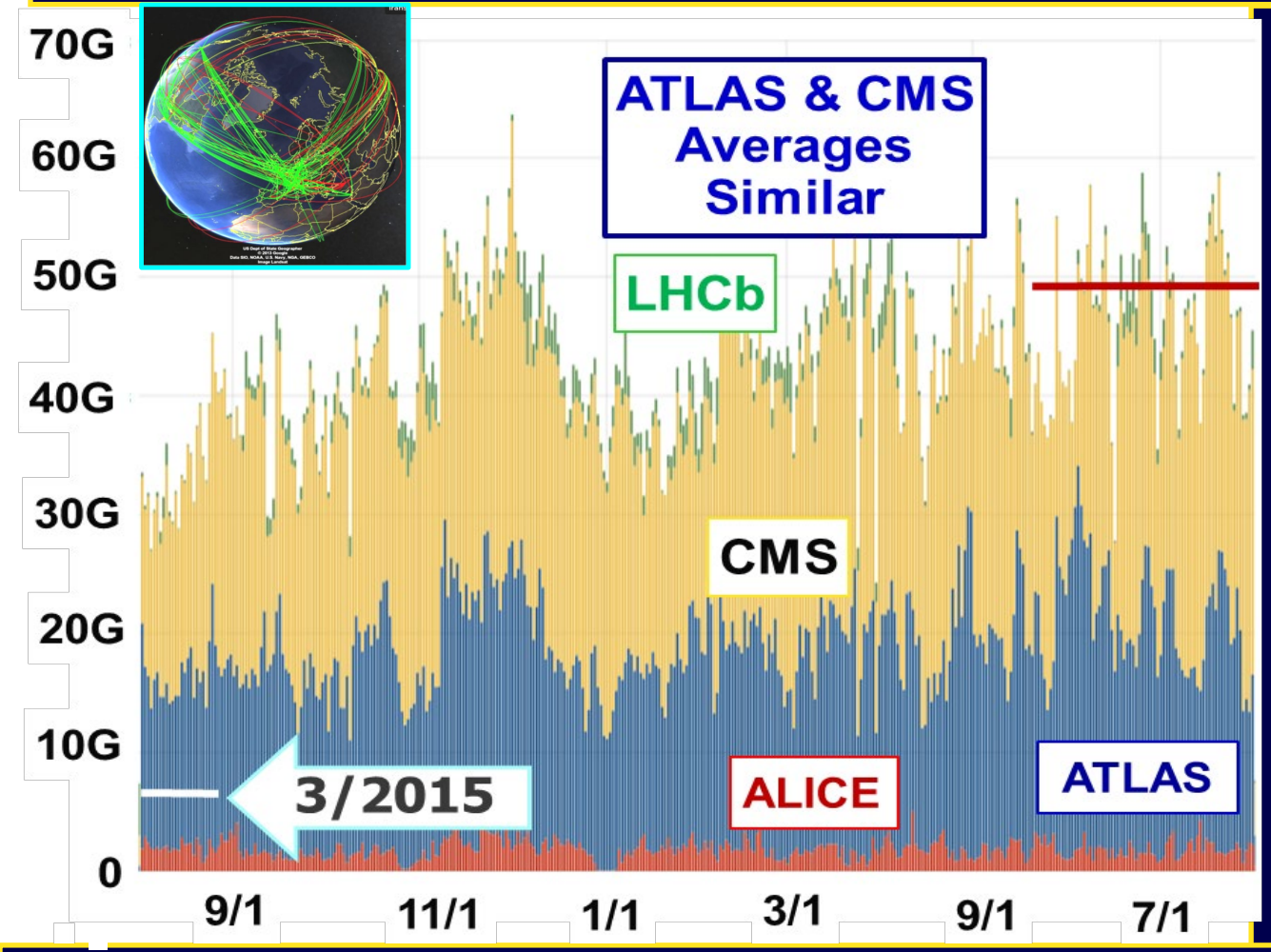
2. **Deeply programmable, agile software-defined networks (SDN), emerging as multidomain network operating systems (e.g. SENSE & SDN NGenIA; AutoGOLE; AmLight SDX; Yale Mercator, Carbide, KISTI VDN)**

3. **Machine Learning, modeling; Ai: Extract key variables; move to real-time self-optimizing workflows with Reinforcement Learning. Stable optimal solutions according to complex metrics: Game theory**

Internet2's NGI, ESnet6, and NSF's Fabric and Nat'l Ai Initiative are Pivotal Elements in this Transition

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



49 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

Network Requirements Update for the HL-LHC Era

LHC Experiments Awaken

- In January, at the 43rd LHCOPN/LHCONE meeting at CERN https://indico.cern.ch/event/828520/, the LHC experiments expressed the need for **Terabit/sec links** by the start of HL-LHC operations in 2027-28, preceded by the usual Computing and Storage (and Network) challenges starting during LHC Run3 (2021-4)
- This was reinforced by the requirements presented by the DOMA project which *"foresees requiring 1 Tbps links by HL-LHC (ballpark) to support WLCG needs. This is for the network backbones and larger sites…"*
  - References: (1) E. Martelli, S. McKee LHCOPN-LHCONE Report to the Grid Deployment Board, (2) DOMA project presentation at the LHCONE meeting https://indico.cern.ch/event/828520/contributions/3570904/attachments/1968554/3274036/LHCONE-DOMA-01-2020.pdf
- NB: The quoted network capacity requirements are **an order of magnitude greater than what is available now** through the present national and transoceanic networks based on 100GE links.
  - As discussed at the LHCONE meeting, in the GNA-G Leadership group meeting that followed, and in the HEPIX Techwatch technology tracking group, **these requirements cannot be accommodated solely through the exploitation of technology evolution within a constant budget.**

The GNA-G Data Intensive Sciences WG

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

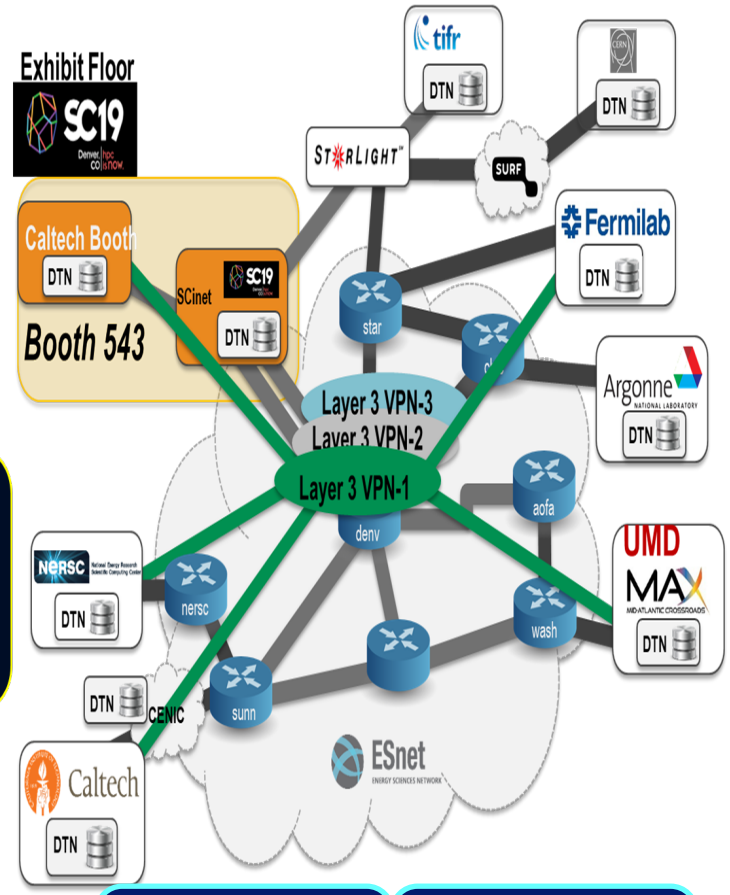
- While sharing and advancing the (new) concepts, tools & systems needed
- Members of the WG will partner in joint deployments and/or developments of generally useful tools and systems that help operate and manage R&E networks with limited resources across national and regional boundaries
- A special 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

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

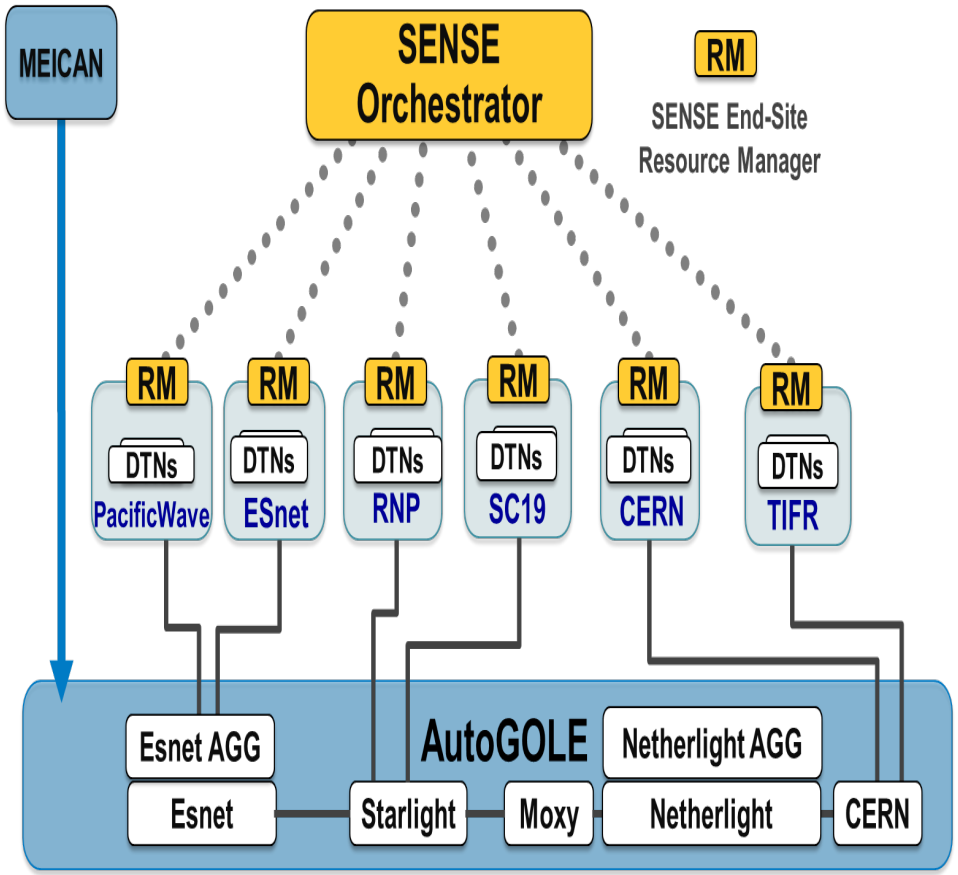


Provisioning SENSE

AutoGOLE Topology

SC19-NRE-020 Intercontinental Demonstration

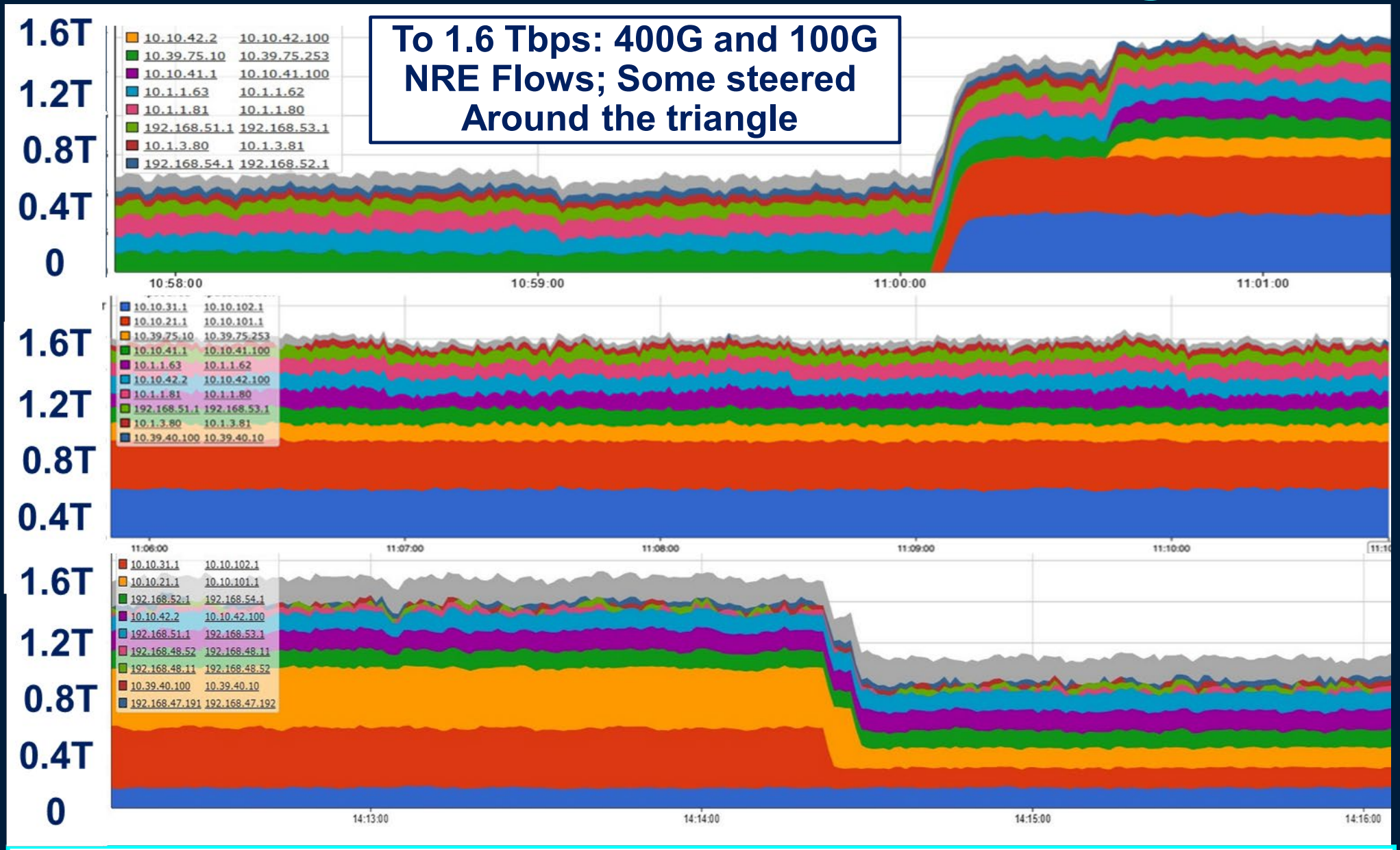
Multi-Resource Orchestration via AutoGole and SENSE



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

SC19 Results on the 400G Triangle



Microcosm: Creating the Future of SCinet and of Networks for Science

Interfacing to Multiple VOs With FTS/Rucio/XRootD

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

CMS

ATLAS EXPERIMENT

LHCb

ALICE

IceCube South Pole Neutrino Observatory

Xe XENON Dark Matter Project

ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

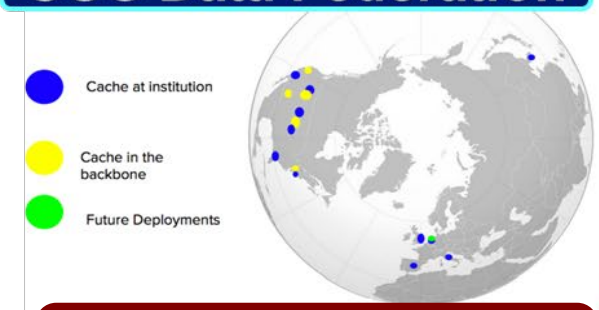
FAIR Particle Physics Facility for Antiproton and Ion Research

cta cherenkov telescope array — the observatory for ground-based gamma-ray astronomy

EST european solar telescope

ESO

OSG Data Federation



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

SKA SQUARE KILOMETRE ARRAY

eneas Advanced European Network of E-infrastructures for Astronomy with the SKA

European Science Data Center

LSC LIGO Scientific Collaboration

VIRGO

KAGRA

Vera Rubin Observatory

LSST CORPORATION INVESTING IN DISCOVERY

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

Courtesy
R. Hughes-Jones

SKA Phase1 Data Flows

SKA1-LOW Australia

~ 2 Pbit/s

LOW-FREQUENCY APERTURE ARRAY

~ 7 Tbit/s

CENTRAL SIGNAL PROCESSOR

9.3Tbit/s

~130 Pflops
300 PB/y

SCIENCE DATA PROCESSOR

100 Gbit/s

SKA
Regional Centres

SKA1-MID South Africa

~ 20 Tbit/s

CENTRAL SIGNAL PROCESSOR

7.9Tbit/s

~130 Pflops
300 PB/y

SCIENCE DATA PROCESSOR

100 Gbit/s

Uploads to Facebook 180PB

SKA Phase1 Science Archive 600PB

Observatory Data Products pushed from SDP → SRC

www.geant.org GÉANT

10

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:	Protocol:
Visibility, Transients 80* 100 Gigabit Bursts	UDP/IP
VLBI 100 Gigabit continuous	UDP/IP
Pulsar Search 740 * 1 Gig = 8 * 100 Gigabit Bursts	TCP/IP
Pulsar Timing 1 * 100 Gigabit Bursts	TCP/IP

Design for peak rates

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 miles**
- **Advanced protocol, caching methods: could extend distance**

Examples in Production:
"SoCal" (UCSD + Caltech); INFN

F. Wuerthwein (UCSD) et al

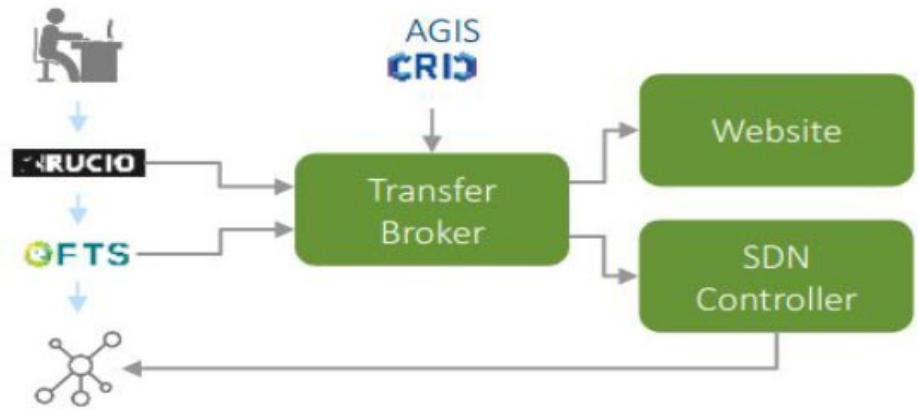
ICFA SCIC Perspective and Outlook

- **Missions**
  - Inform and enable the global community to use networks effectively in support of the communities' science goals
  - Track advanced computing, storage, network and associated software technologies; highlight opportunities and coming issues
  - With a focus on major programs: LHC to HL-LHC, LSST, SKA, DUNE et al
  - Track and help understand and set requirements via both community meetings (e.g. LHCONE/LHCOPN) and agency reviews (e.g. ESnet in July)
  - Bring Issues to the attention of ICFA
- **Activities**
  - Work with R&E network partners to help develop the continental, transoceanic and regional network infrastructures
  - Beyond the basic infrastructures: Formation of a global fabric supporting data intensive research Learning from and going beyond the LHCONE experience
  - Developing ntegrated systems including networks as a first class resource, across a global footprint
- **Engagement**
  - With all of the experiments' computing managements, the major R&E network organizations, major network projects supporting major science programs Also leading edge development projects: SENSE, AutoGOLE, SANDIE etc.
  - Engage in proof of concept, prototype, pre-production exercises and demonstrations to test and prove requirements

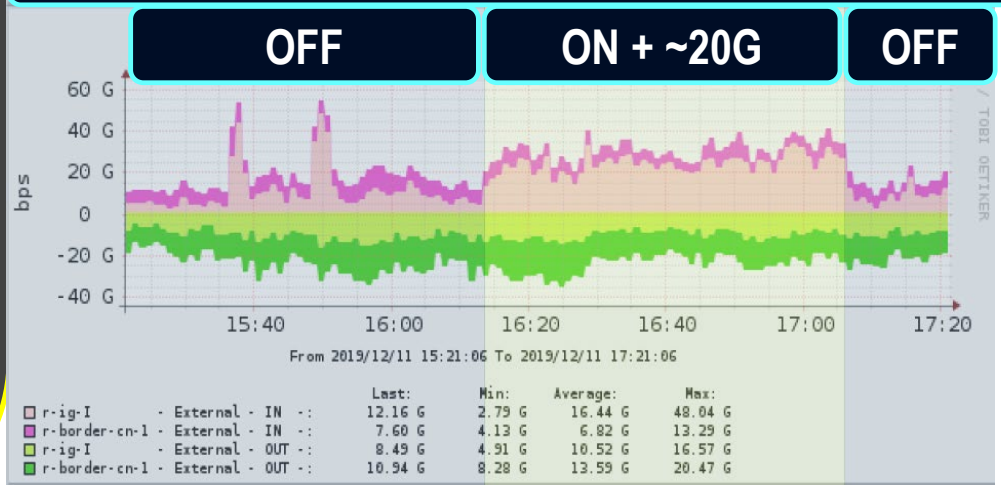
NOTED: Network Optimized Transfer of Experimental Data CERN/IT Project (C. Busse-Grawitz)

- 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).

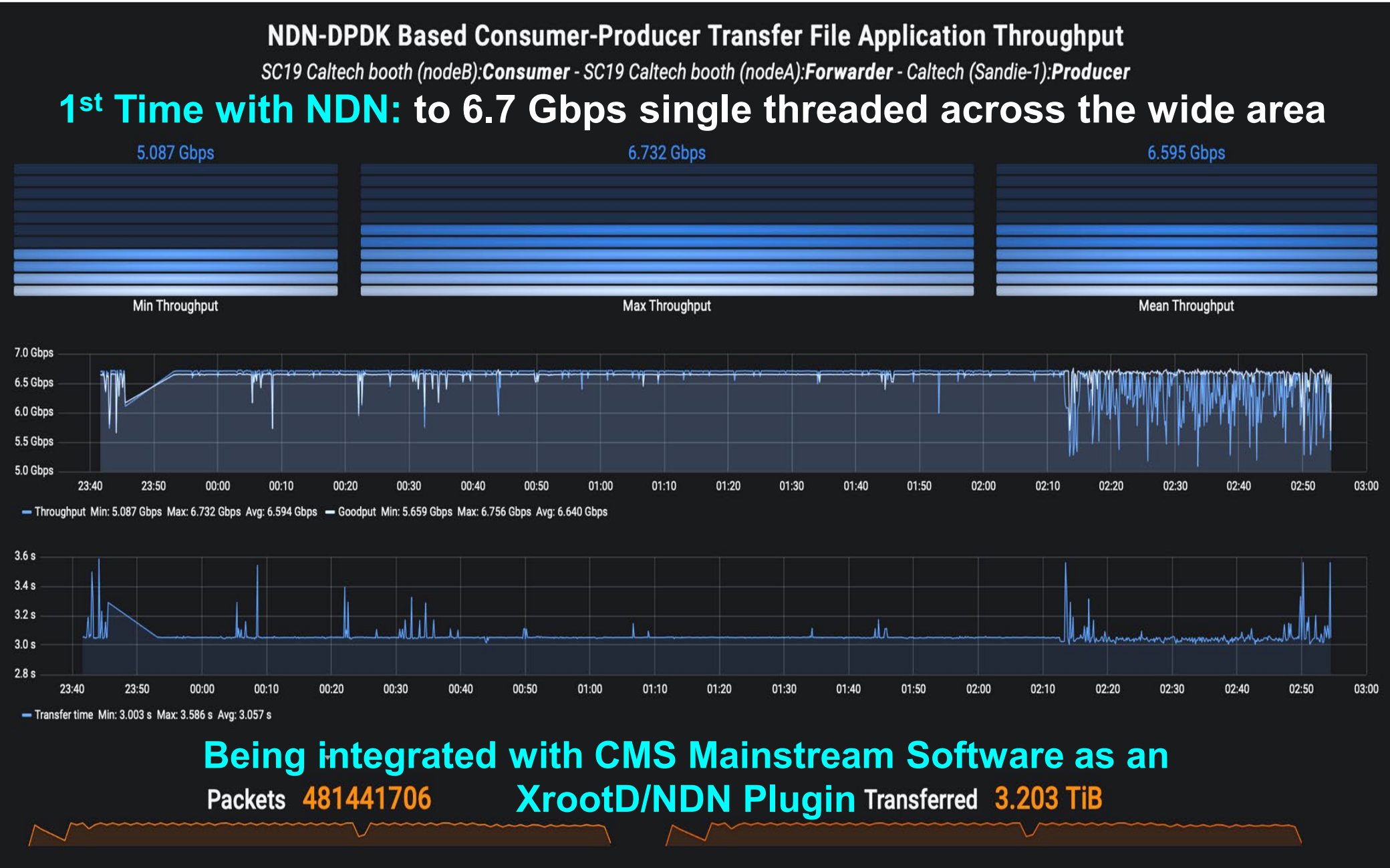
Transfer Broker Interfaces to Job Queues, SDN Controller, WLCG Database



Switch some traffic to DE-KIT LHCOPN path



SANDIE Demo with 2.5 Gbyte CMS Files

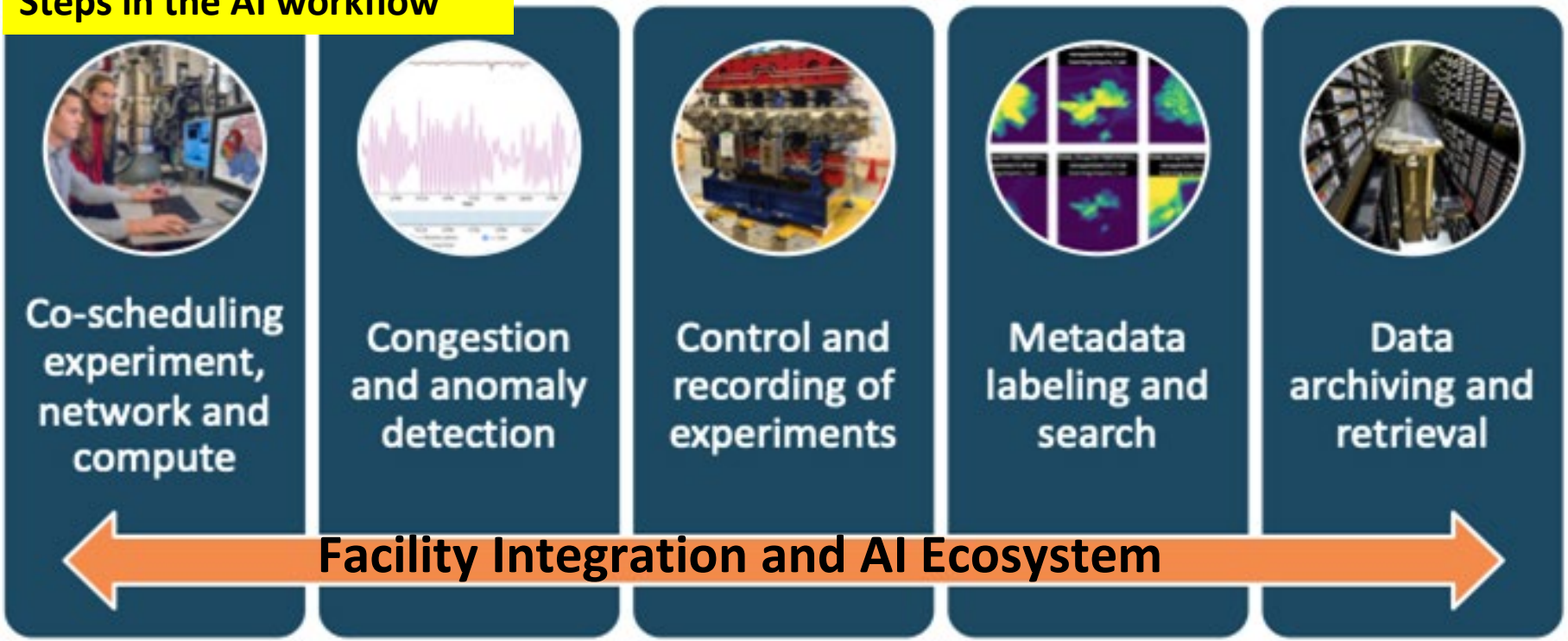


Ai for Science Berkeley Town Hall Breakout: I. Monga et al.

AI is essential for facilities, and facilities are essential for AI

Without the integration of facilities in the AI workflow, AI for science is impossible.

Steps in the AI workflow

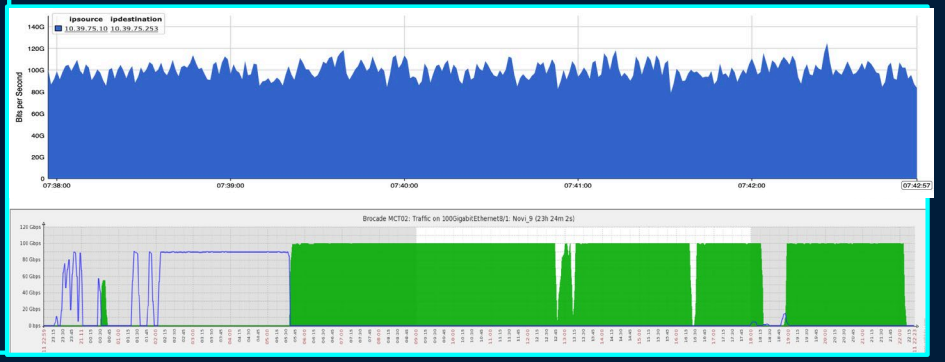


Ai for Networks; Networks for Ai; Ai for Ai

AmLight-ExP at SC19 100G Int'l Data Flows in Support of LHC, LSST, Data Intensive Sciences

- Emulated La Serena – Miami – NCSA (LSST) path through spectrum on the Monet cable system (100G): *Miami, Brazil, Chile, Miami, then to Denver*
  - New Data Center in La Serena, but 100G transponder delayed
- Solid 100G from SC19 Caltech Booth to AmLight (Denver to Miami via Brazil (after adjustments and network stack tuning)
- Also HEPGrid in Rio: 9G (10G Link)
- Key Points: Team was able to run 100G data transfer experiments using the Monet spectrum to Brazil and Chile over the AmLight SDN backbone.

UDP Traffic MIA-FTLZ-SPO-SLC-Denver



Major Changes for 2026 and Beyond:
Technology/Physics Barriers are Being Approached

- *The "End of CMOS" is predicted by 2030, or Earlier*
- *Processor designs: 10 nm feature size in processors is difficult/delayed; 7 nm or perhaps 5 nm feature size appears to be a practical limit*
- *Disk Storage: Below ~10nm: "superparamagnetic limit", large investments will be needed for new techniques.*
- *Network price/performance improvement has similarly slowed.*
  - *Time for 400GE to be at same price point as the present 100GE generation likely to be 10 years (Price/perf. ~ -13%/year)*
  - *Network usage is increasing much faster than this over the medium and long term: ~7X since 2015 (near 60% per year)*
  - *By ~2026 we will likely hit an energy density limit in switches*
- *Conclude: By around 2026-2029 the need for nanoscale processes, 2D materials (and huge investments) will likely be felt throughout industry*
- *Bottom Line: We need to carefully follow these technology developments; Work with some prototype and 1st-generation systems*

Facts of Life Beyond Data Centers, Processing and Analysis: Societal and Cultural Changes

- There are already major changes underway that will have a great impact on our designs and plans
- Between now and 2026, and then beyond
- Emergence of machine-to-machine (MTM) communication
  - To 27B devices by ~2020; pervasive before HL LHC
  - Digital Assistants, AI smartphones are the tip of the iceberg
- IOT: Expanding intelligence, autonomy, and coordination
  - Emergence of smart apartments/homes, buildings, neighborhoods and cities, AI/ML in handhelds
- Real-time *low power* AI devices: responsive on microsecond to millsecond timescale: already on the radar for the Phase II Trigger
  - Can we use this for real-time event cleanup ? Pre-processing ?
  - Can we (for example) reconstruct all < 2 GeV tracks in real-time ?
- Autonomous tracking of workflows: real-time remediation if workflow is delayed; managing user/system interactions ?