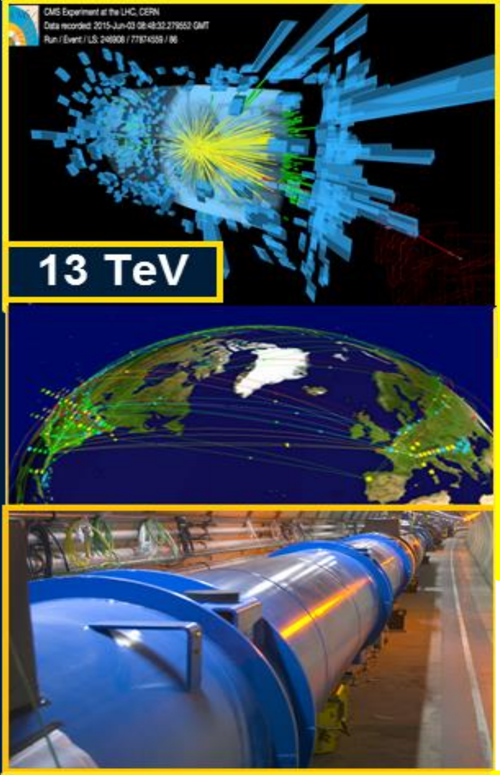
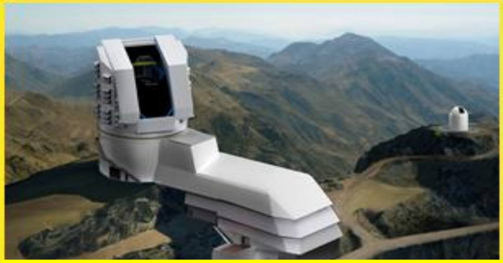
ESnet Requirements Review 2020-21 [US HEP]

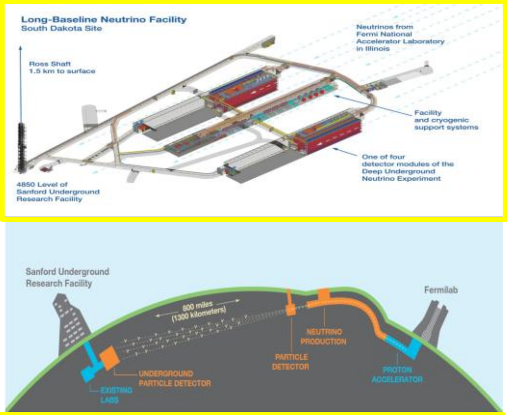
Network Requirements and Associated Issues: Now to HL LHC


CMS Experiment at the LHC, CERN
Data recorded: 2015-Jun-03 08:48:32.279552 GMT
Run / Event / LS: 246908 / 77874559 / 86
13 TeV

LHC



LSST


Long-Baseline Neutrino Facility
South Dakota Site
Neutrinos from Fermi National Accelerator Laboratory in Illinois
Ross Shaft 1.5 km to surface
Facility and cryogenic support systems
One of four detector modules of the Deep Underground Neutrino Experiment
4850 Level of Sanford Underground Research Facility
Sanford Underground Research Facility
Fermilab
800 miles (1300 kilometers)
NEUTRINO PRODUCTION
PARTICLE DETECTOR
UNDERGROUND PARTICLE DETECTOR
EXISTING LABS
PROTON ACCELERATOR

LBNF/DUNE

SKA

LHC Run3 and HL-LHC

+ Fermilab ν, DM, Muons

CMB S4, VRO, SKAO

Gateways to a New Era

California Institute of Technology 1891

Harvey Newman, Caltech

LHCONE/LHCOPN Meeting

September 17, 2020

GNA-G

Global Network Advancement Group

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

newman@hep.Caltech.edu

SENSE DOE-AC02-07CH11359 SANDIE NSF CC* 1659403

HL LHC: Addressing the Challenges
Computing, Storage and Networking

- **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
  [+ Improved Hierarchical Architectures + Caching Strategies]
- **Networking:** Tuned end systems + QoS via virtual circuits,
  + allocated resources with prioritization, policy;
  Interworking with LHCONE and the major R&E networks
- **Workflow:** Computing + storage + networking coordination,
  Full lifecycle services
  overseeing task completion;
  Integrating Rucio/FTS, XRootD, etc.
  into a bigger picture;
  ATLAS & CMS Orchestrators Interacting
  with Network orchestrators;
- **Multi-objective Optimization** with metrics of success
- There are challenges, and opportunities, in all areas

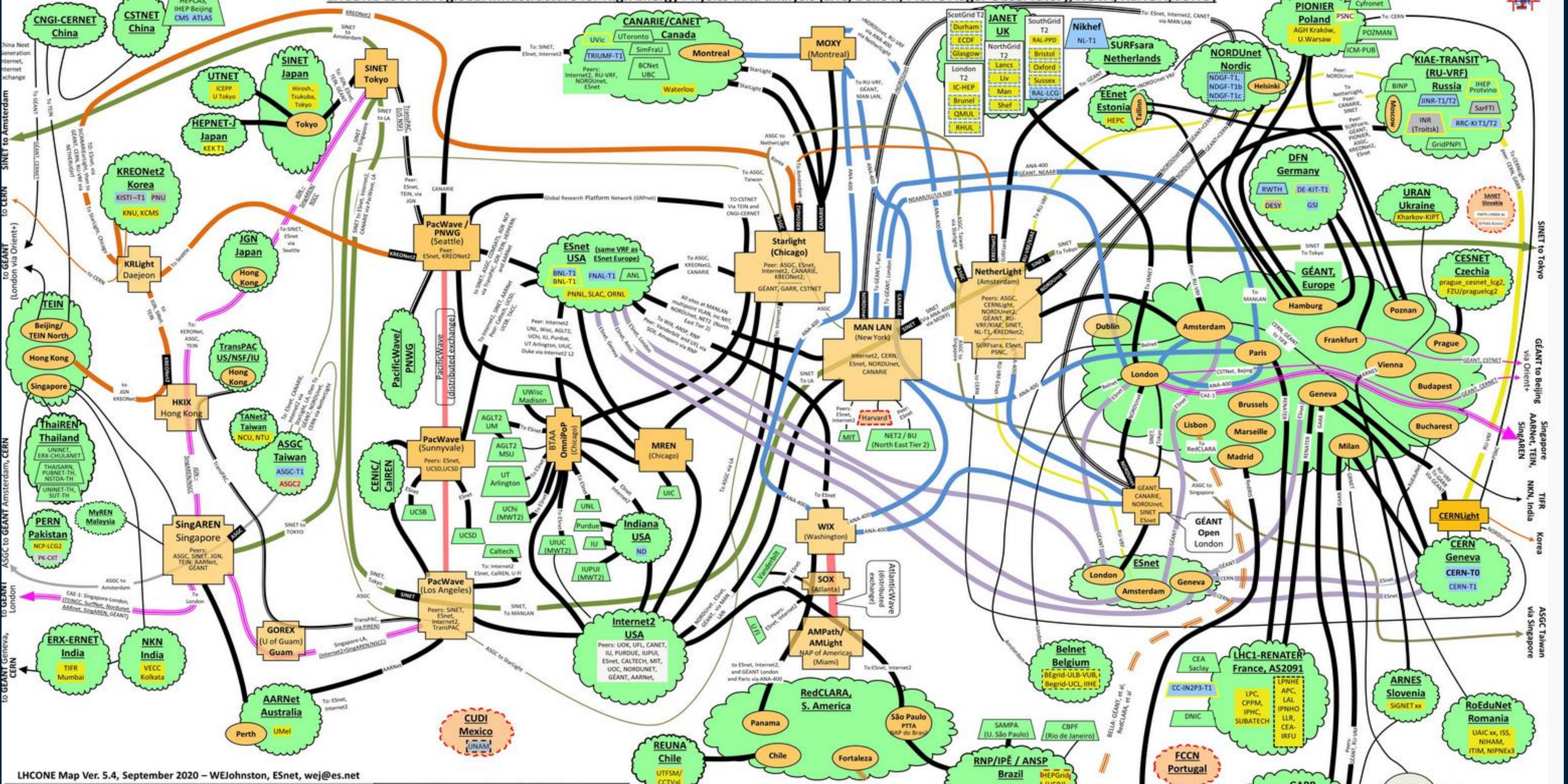
HL LHC Challenges:
Capacity in the Core and at the Edges

- 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 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 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 ESnet 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

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)



LHCONE Map Ver. 5.4, September 2020 – WEJohnston, ESnet, wej@es.net

GARR	LHCONE VRF domain/aggregator - A provider network.
ANSP	Connector network – provides, e.g., an L2 path between VRFs.
London	Provider network PoP router
CUDI	Not currently connected to LHCONE
✚	Exchange point
CANARIE	NREN/site router at exchange point
—	Communication links: 1/10, 20/30/40, and 100Gb/s or N x 100G
JGN – SingAREN/NSCC	Underlined link information indicates link provider, not use
NDGF-T1	Dotted outline indicates distributed site
RAL	Blue dashed outline indicating a WLCG federation site not currently on LHCONE

International infrastructure by provider/collaboration

▬	various	▬	NORDUnet
▬	GÉANT	▬	KIAE, Russia
▬	SINET, Japan	▬	KREONet2, Korea
▬	ASGC, Taiwan	▬	BELLA: GÉANT, et al, RedCLARA, et al
▬	ESnet transatlantic, USA		
▬	JGN, SingAREN (+ TIENCC, SurfNet, NORDUnet, GÉANT on the London link)		
▬	ANA-300/400 - Various links provided by CANARIE, ESnet, GÉANT, Internet2, NORDUnet, SURFnet, SINET,IU/NSF		

PNU	LHC ALICE or LHCb site
CNAF-T1	LHC Tier 1 ATLAS and CMS
UChi	LHC Tier 2/3 ATLAS and CMS
KEK	Belle II Tier 1/2
UNL	Sites that are standalone VRFs

Yellow outline indicates LHC+Belle II site

NOTES
1) ONLY links involved in LHCONE are shown
2) LHCOPN links are not shown on this diagram
3) For map explanation see "Interpreting the LHCONE Map" at https://www.dropbox.com/sh/padxfo58j0j1raz/AADs85K8fiSH9FfhCjA4eCtea?dl=0

Beyond Capacity Alone: Complexity, Scalability, Global Reach

W. Johnston 9/14/20

US DOE ESnet Requirements Review

Timeline: from June 2020 to February 2021; then Onward

- July – August 2020: Case Study Documents [Partly Joint to Separate] Due 8/14
  - Case Study #10 ATLAS: Science, Experiment Specific Computing/Storage/Software Profiles, Tier1 Ops, Tier2 Ops
  - Case Study #11 CMS: Science, Experiment Specific Computing/Storage/Software Profiles, Tier1 Ops, Tier2 Ops
  - Case Study #12 LHC Computing & Networking Operations ATLAS and CMS Focusing on overlap in Computing, Software, & Storage - examples include Rucio, LHCOPN/LHCONE, traffic marking, Fabric, OSG
  - Case Study #13 HL-LHC ATLAS and CMS - networking, computing, and software R&D for the High Luminosity LHC
- Discussion Meetings held in August: One per Document and per Experiment, with document authors, S&C management, program management
- 4 Focus Groups in September: Bringing LHC & other programs together "Roundtable discussion on common problems & unique solutions. Leveraging our different experiences & approaches, we hope to cross-pollinate, share insight."
  - Group 1: Cosmology Computation [1], LZ DM Experiment [6], Belle II [8], CMS [11]
  - Group 2: Astro DES [2], Intensity Frontier Fermi (ν: DUNE and SBN) [9], ATLAS [10]
  - Group 3: Vera Rubin Observatory [4], CMB-S4 [5], LHC Ops [12]
  - Group 4: DESI [3], Intensity Frontier at Fermilab Muons (Mu2e, g-2) [7], HL-LHC [13]

US DOE ESnet Requirements Review

Timeline: from Now to February 2021; and Onward

Next Steps Toward February 2021 Meeting (in person ?)

- Last Call for edits to the case studies after the focus groups
  - Editing to be completed in October
- ESnet team will combine the case studies, and any notes from 1:1 meetings and focus groups into a final report
  - Target is to complete draft of final report in early December
  - Edits to final report in December, with a target of publishing final report in early 2021
- While the in-person meeting depends on the COVID-19 situation and travel approvals, the team wishes to ensure that the final report is published early enough
  - To be useful for Snowmass and other HEP activities and meetings in 2021

ATLAS: Four Main Requirements for HL LHC [10]

- **Capacity:** Run-3 is moving to multiple 100G links for large sites, while Run-4 (HL-LHC) is targeting Tbps links
  - Capacity is fundamental for us to do our science at HL-LHC scales
  - As noted, in a capacity constrained environment, it will be important to manage the capacity we do have to do the most science possible.
- **Capability:** We need to understand the impact of new features in networking (SDN/NFV) by *testing, prototyping* and *evaluating* impact
  - We will need to evolve our applications, facilities and computing models to meet the HL-LHC challenges; it will take time.
  - Traffic shaping activities underway in the Research Network Technical Working Group are a good capability example
  - Network orchestration between sites is another good example [e.g. with SENSE] & something that could help us more effectively exploit our resources
- **Visibility:** As the ESnet Blueprinting meetings have shown, our ability to understand our WAN network flows is too limited
  - We need new methods to mark and monitor our network use
    - Packet-marking is viewed as a high priority so that everyone (experiments, sites and R&E networks) can understand the origin & intent of network flows
    - We need better mechanisms to coordinate the available monitoring resources from the sites, experiments and networks, to allow us to better understand how our complex infrastructures are using the network

ATLAS: Main Requirements for HL LHC (II) [10]

- **Testing: We need to be able to develop, prototype and test network features at suitable scale**
  - **Networks of the future will likely have new features, capabilities and services that could be leveraged to do more with the resources we have available**
  - **Our challenge is identifying which features might be beneficial to try to integrate into our operations, noting that such integration can require significant effort to realize**
  - **Having at-scale network testbeds will be very important to understand the potential impact changes might have on our operations**
- **These requirements should also motivate and guide our network infrastructure upgrades and replacements**
- **It will be critical to understand the bigger picture of R&E networking and its evolution, To ensure we are able to take advantage of the capacity and services available**

CMS Data-Related Resource Constraints [11]

- **Transatlantic links:** CMS is a major user
  - For archival of raw data, initial copies of analysis data samples, and user-derived data sets using CERN or other non-U.S. facilities
  - RAW data transferred to FNAL alone are expected to average more than 10 GB/second during HL-LHC operations.
- **CMS tools need to change to prioritize site proximity (in the networking sense)** when scheduling data transfers
- **Streaming data across the Atlantic for now is allowed** (even if discouraged); it might no longer be allowed by HL LHC
- **Streaming to sites beyond the CMS infrastructure:** Reliable high capacity streaming of RAW or pileup simulation, would considerably reduce the disk requirements of CMS at HPC and other non-dedicated computing facilities
- **If the current growth rate in transatlantic use by CMS continues,** the capacity these links will become a major limitation already during in Run 3 (2022-24)
- **Disk vs. network tradeoffs:** As with HPCs, reliable networking can be used to reduce disk replica requirements either by use of tape recall or caching
- **Understanding caching use cases and needs** are part of on-going R&D

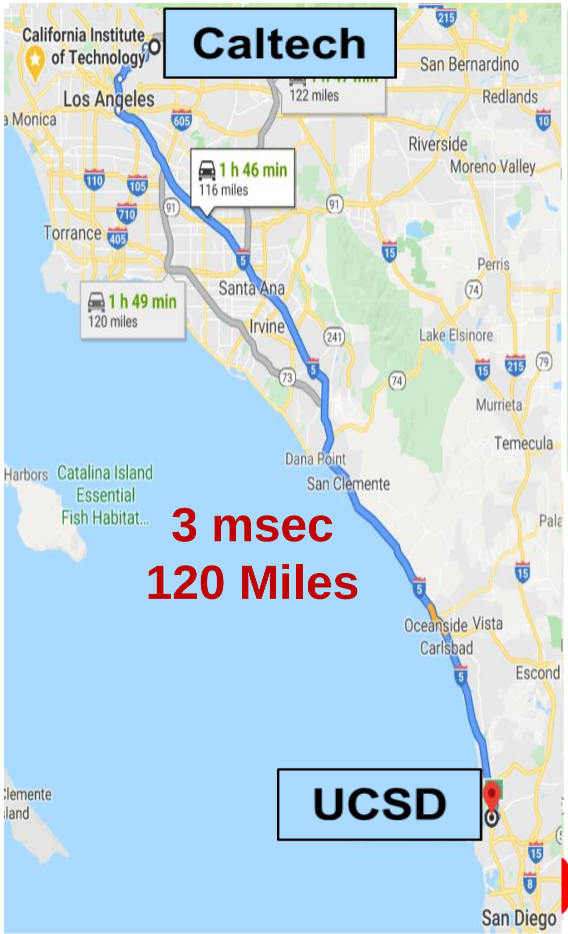
CMS Ideas for Future R&D with ESnet [11]

In the HL-LHC “Exabyte per year” era, CMS:

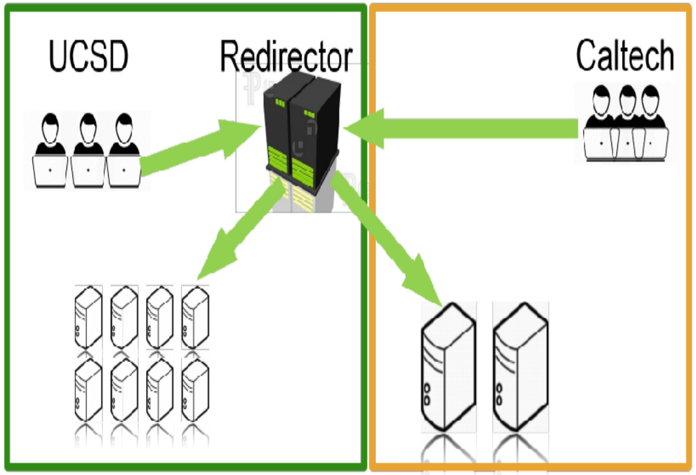
- Envisions to both make more aggressive use of networks, and Account & manage its use of networks much more carefully
- Expects to trade investments in disk space at Tier-2s against network utilization; CMS welcomes collaboration with Esnet as it learns about best uses of caching
- The main cache R&D effort now is the production cache deployment across Caltech and UCSD, which includes a cache in Sunnyvale on ESnet hardware
  - The hardware owned by ESnet and operated by the UCSD Tier-2 team is an integral part of the production cache
- Have started internal efforts to validate & improve the transfer accounting in software layers, following regular meetings on transfer accounting with ESnet
- Long term goal is to match network layer accounting with our higher level accounting, to gain confidence that its network usage is understood
- The extent to packet and/or flow tagging is needed is unclear at this point. CMS is looking to ESnet leadership in this area
- CMS intends to explore with ESnet + R&D projects including SENSE/ AutoGOLE: How to transition to managed network usage in production operations

(Southern) California ((So)Cal) Cache

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



3 msec
120 Miles

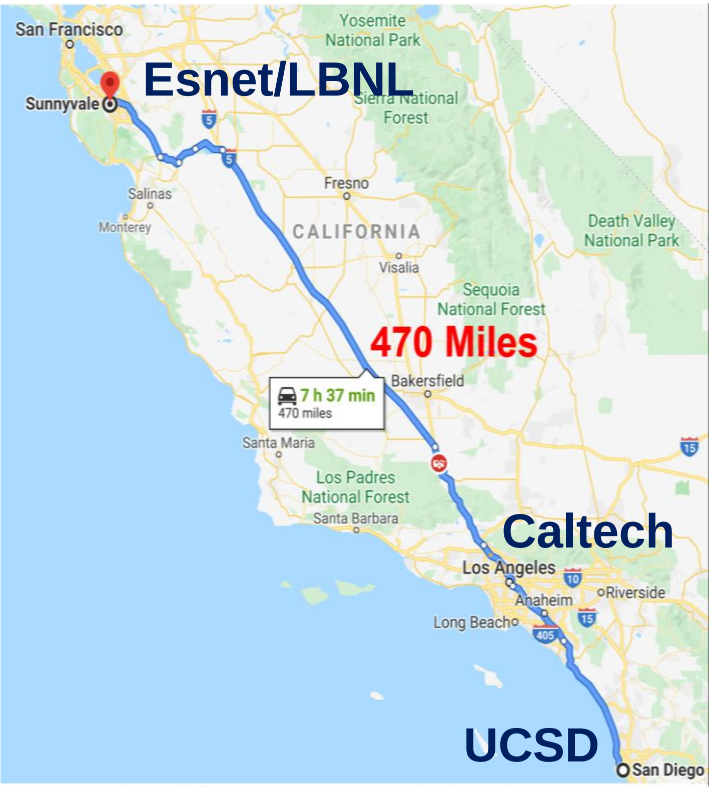


CPU in both places can access storage in both places.

How much disk space is enough?

Cache MINI and measure working set accessed:

0.45 Petabytes in October 2019



470 Miles

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

Examples in Production:
UCSD + Caltech; INFN

SoCal Cache Expanding to Include UC Riverside Tier3;
Other SoCal US CMS Tier3s possible in near future

CMS Ideas for Future R&D with ESnet [11] (II)

- Traditionally, we have treated the global network of CMS sites as a mesh with identical links when it comes to bulk transfers
- The XRootd data federation was designed from the beginning to be cognizant of the transatlantic link being limited, but treated links within the U.S. as identical
- The Data Lake model currently discussed in WLCG makes clean regional distinctions. We expect that at least the existence of the Atlantic will become an architectural feature of our data distribution architecture.
- We would like to develop a program of transfer tests both to benchmark our methods at increased capacity, and integrate new functionality; We would like to do such tests in collaboration with ESnet and FABRIC
- We believe that national and international collaboration bringing together researchers, data management experts and networking experts is important for making better use of network resources as usage levels of research networks increase.
  - In HEP, these collaborations include the WLCG Networking Throughput working group, or more broadly groups including the Global Network Advancement Group.

CMS Infrastructure Software Products &Tools; Scaling of Network Use Vs. What is affordable [12]

CMS Case Study 12 Provides

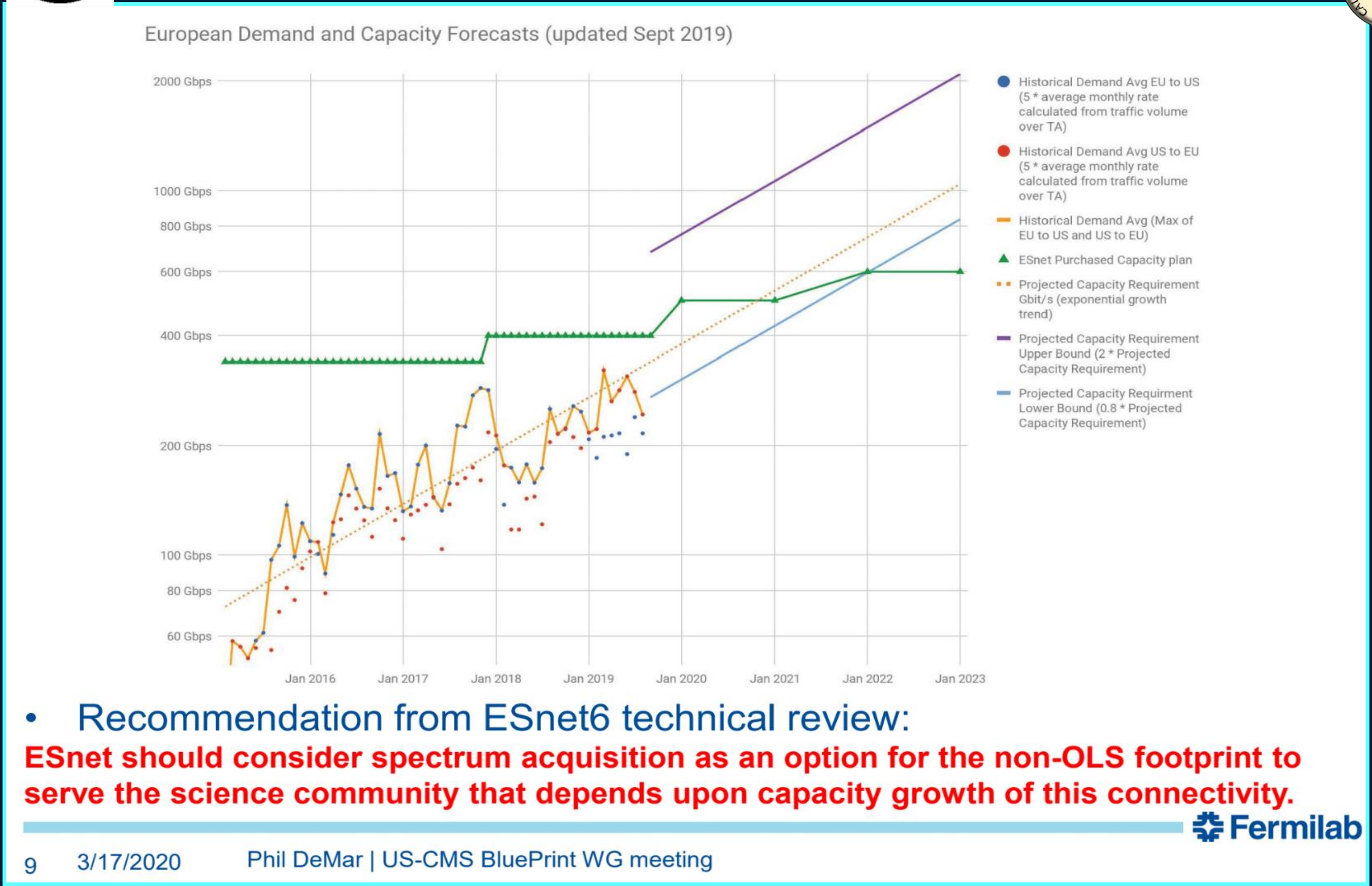
(1) A complete list of infrastructure software products and tools relevant to data movement and/or access, and how the various tools relate to the process of science in CMS as described in [11] and [13]

(2) How CMS network use is scaling, contrasted with what is thought to be "affordable" within a ~constant budget

Findings

- There is a significant gap between needs projections based on past experience, and projections of natural growth based on past investments (and trends) in networking infrastructure expansion
  - This gap is sizeable even ignoring the "step function" increase in data volume per year that the HL-LHC era is expected to bring
  - In light of this step-function increase, an historical projection approach may significantly underestimate the actual needs
- This leads to the conclusions that:
  (1) we need to fundamentally rethink our use of networking resources and (2) we would like to engage with ESnet in R&D towards substantial improvements in effectiveness of our network bandwidth usage

Esnet Transatlantic Network Traffic Projections



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.

Fermilab

9 3/17/2020 Phil DeMar | US-CMS BluePrint WG meeting

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

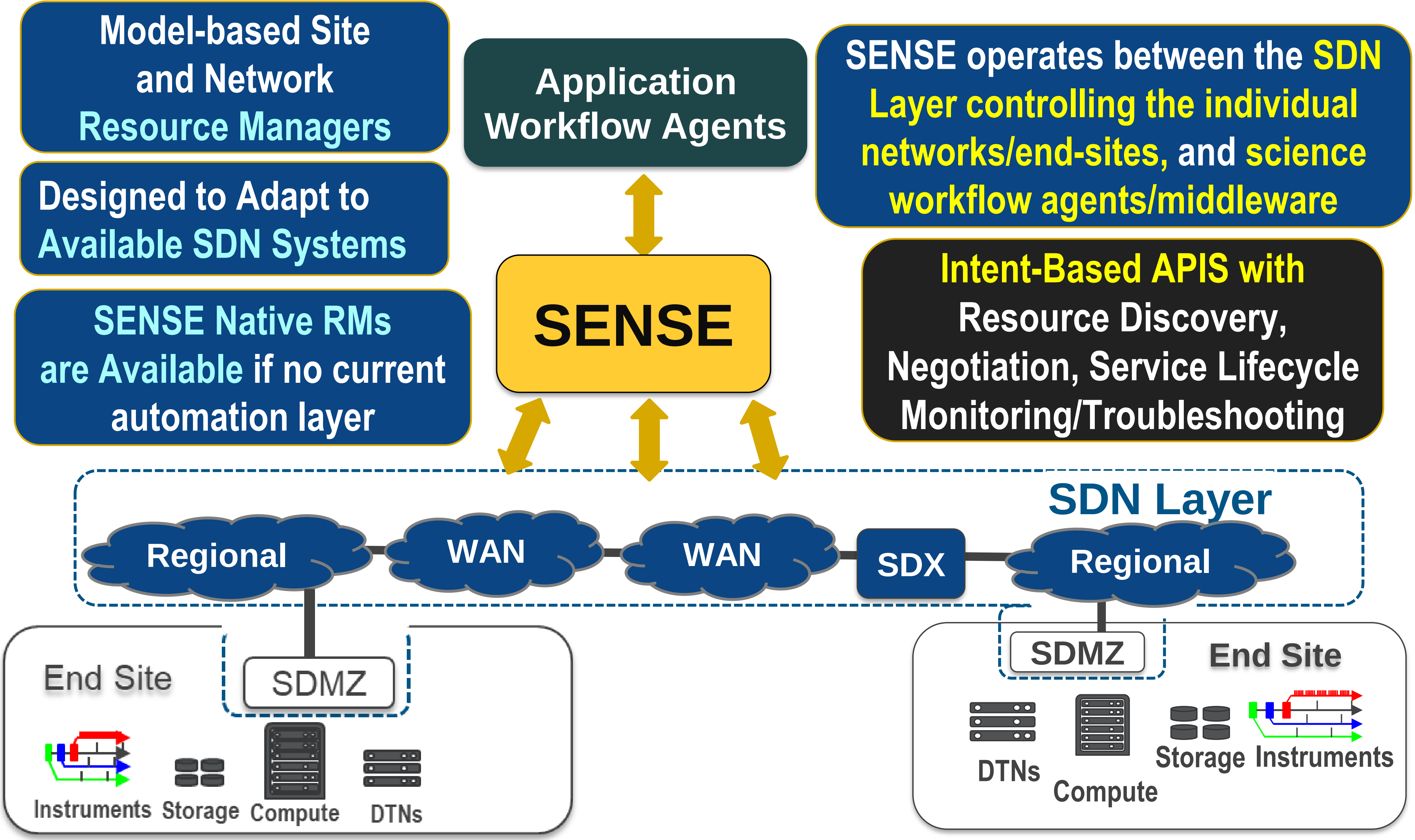
- Requirements based on recent traffic: 0.35 – 0.85 Tbps
  [based on 0.8 to 2X the 2016-19 traffic projection]
- Growth Rate 1.4X per year, or 2X every two years on average
- Hence *16X capacity requirement in 2028* = 5.6 to 13.6 Tbps;
  Since this is an ESnet only, and not a global projection,
  the 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 IRUs (spectrum) 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

Network Management Tools for Future Production Use:
Beyond the present-day monitoring & data movement tools [12]

- Presents ideas for the kinds of functionality needed in production use in the future; As a reference implementation refers to the SENSE software
- Given constrained network resources, the intelligent network services to be provided should allow for the management and best use of available resources
  - Including the coordination of network resource allocations with the corresponding computing and storage resource allocations in the context of a set of workflows
- The basic network services needed should be able to:
  - Allocate guaranteed bandwidth between source and destination
  - Control the characteristics of a given allocation and an associated transfer: such as immediate vs scheduled, transfer a certain amount of data before a deadline, choose a path between A and B depending on policy or network state and/or performance, etc.
- The SENSE architecture, models, and demonstrated prototype define the mechanisms needed to dynamically build end-to-end virtual guaranteed circuits across administrative domains, with no manual intervention. Plus:
  - A highly intuitive "intent"-based interface, allows applications to express their high-level service requirements
  - An intelligent, scalable model-based software orchestrator converts that intent into appropriate network services, across multiple types of devices

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

SDN Layer

Regional

WAN

WAN

SDX

Regional

End Site

SDMZ

Instruments

Storage

Compute

DTNs

SDMZ

End Site

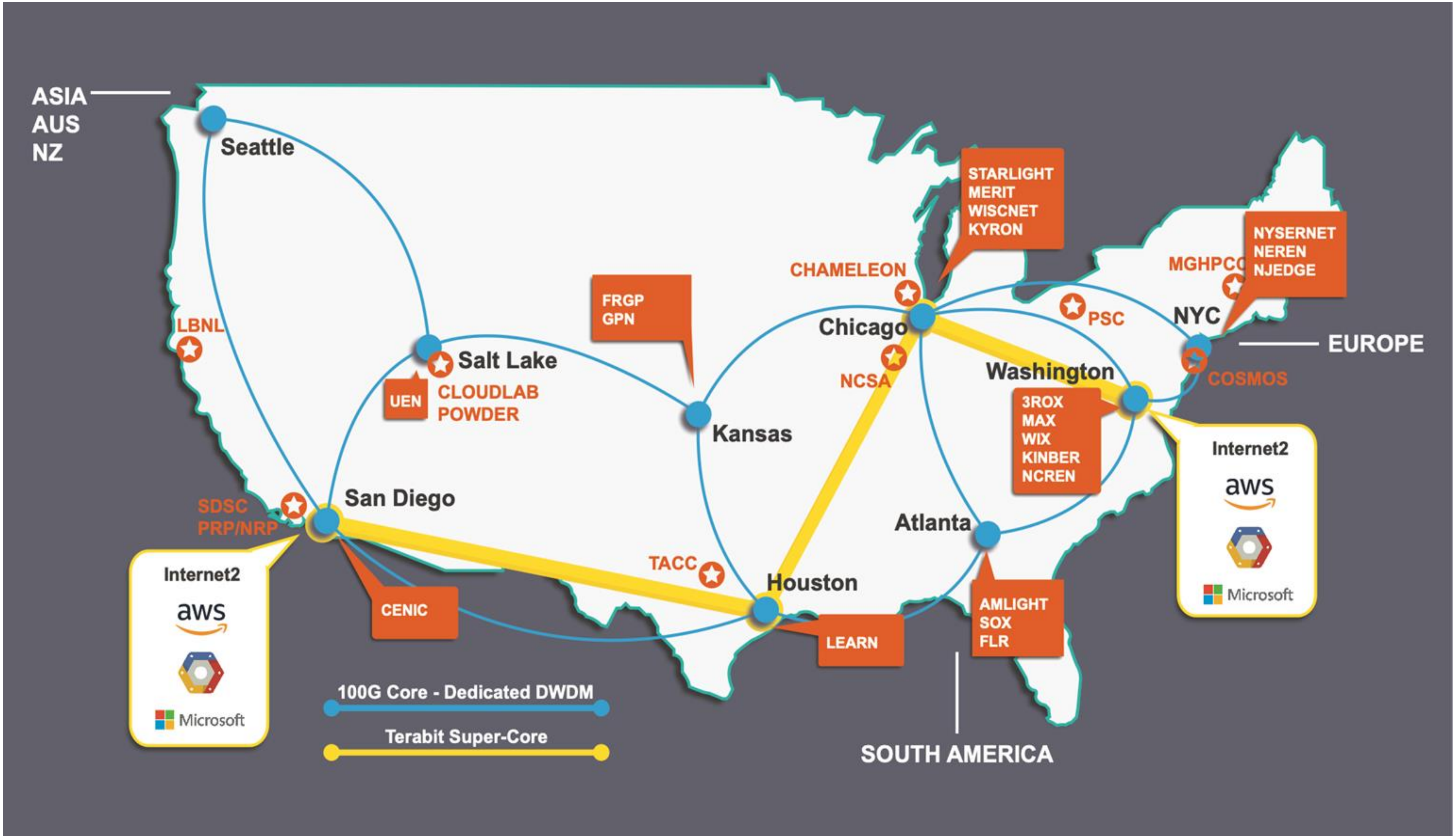
DTNs

Compute

Storage

Instruments

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



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

CMS and ATLAS and ESNet R&D Collaboration [12]

- With the SENSE capabilities + development of sufficient network-aware, interactive software by CMS and ATLAS, the experiments will be able to work with ESnet to develop a coordinated workflow system, including the network as a first-class, managed resource along with computing and storage resources
- We expect that the level of coordination between CPU, storage, and network capacity provisioning and use will increase as we gain operational experience with SENSE and its impact on our operations
- In preparation for Run 3, U.S. CMS would like to engage with ESnet and other partners on transitioning some of the SENSE functionality from R&D to production
- We see “starting the transition to managed production networks” as the most important initial step
- This can enable well-defined and highly-tuned complex workflows that require close coupling of resources spread across a vast geographic footprint, in domains including both high-energy physics and other data intensive sciences
- We would like to identify both appropriate links, and appropriate production-ready functionality, in SENSE, and integrate that into CMS tools for production use
- We expect that US CMS together with ESnet will define appropriate metrics: To measure progress towards the goal of managing the FTS and XRootD transfers across LHCONE, as well as for success of individual managed transfers

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 the development of 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

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

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, Chin Guok, Ciprian Popoviciu, Dale Carder, Dale Finkelson, David Lange, David Wilde, Edoardo Martelli, Eduardo Revoredo, Eli Dart, Frank Wuerthwein, 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, Kaushik De, Kevin Sale, Lars Fischer, Marcos Schwarz, Michael Stanton, Mike Hildreth, Ney Lemke, Phil Demar, Raimondas Sirvinskas, Richard Hughes-Jones, Rogerio Iope, Sergio Novaes, Shawn McKee, Siju Mammen, Susanne Naegele-Jackson, Tom de Fanti, 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, 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.*

Additional Information is available at:

https://www.dropbox.com/s/clzsi8pkixppm3a/ESNetRequirementsReview_NetworkIssuesNowtoHLLHCExtras_hbn091620.pptx?dl=0