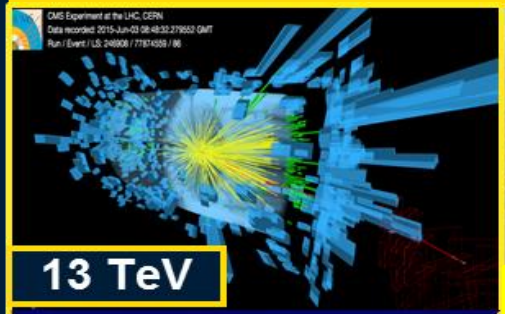
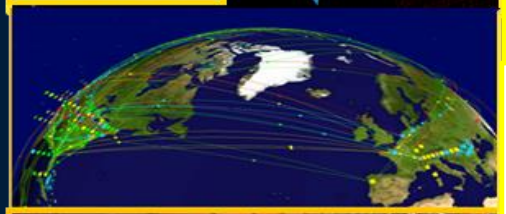
Networking Requirements Outlook
A New Computing Model for the HL-LHC and DUNE Era

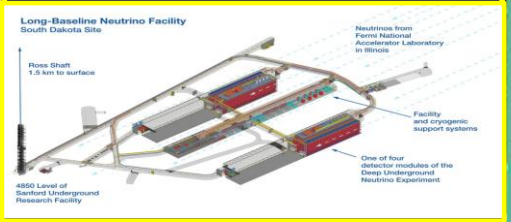


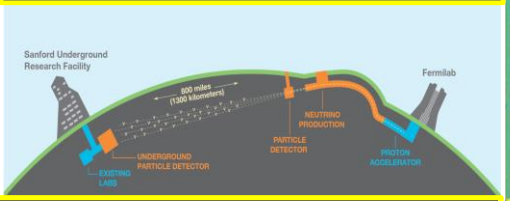
LHC



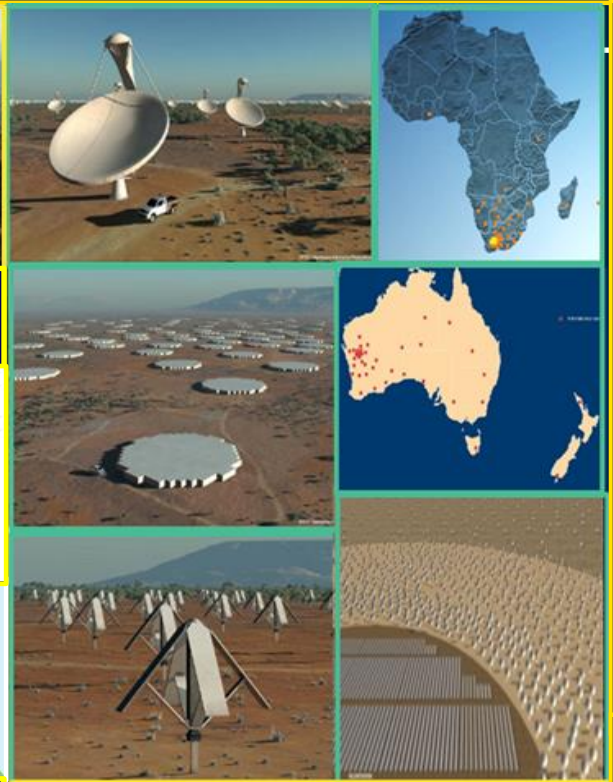
LHC

LSST





LBNF/DUNE



SKA

LHC Run3 and HL-LHC

DUNE

LSST SKA

BioInformatics

Earth Observation

Gateways to a New Era

Harvey Newman, Caltech

Huge Data Workshop
April 14, 2020

newman@hep.Caltech.edu

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

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

- **Exascale Data for the LHC Experiments**
  - ~1 Exabyte by 2019; to ~50 EB during HL LHC Era
- **Network Flow: 45-60 Gbytes/sec**
  - 1.6 Exabyte flowed over WLCG in 2018
- **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

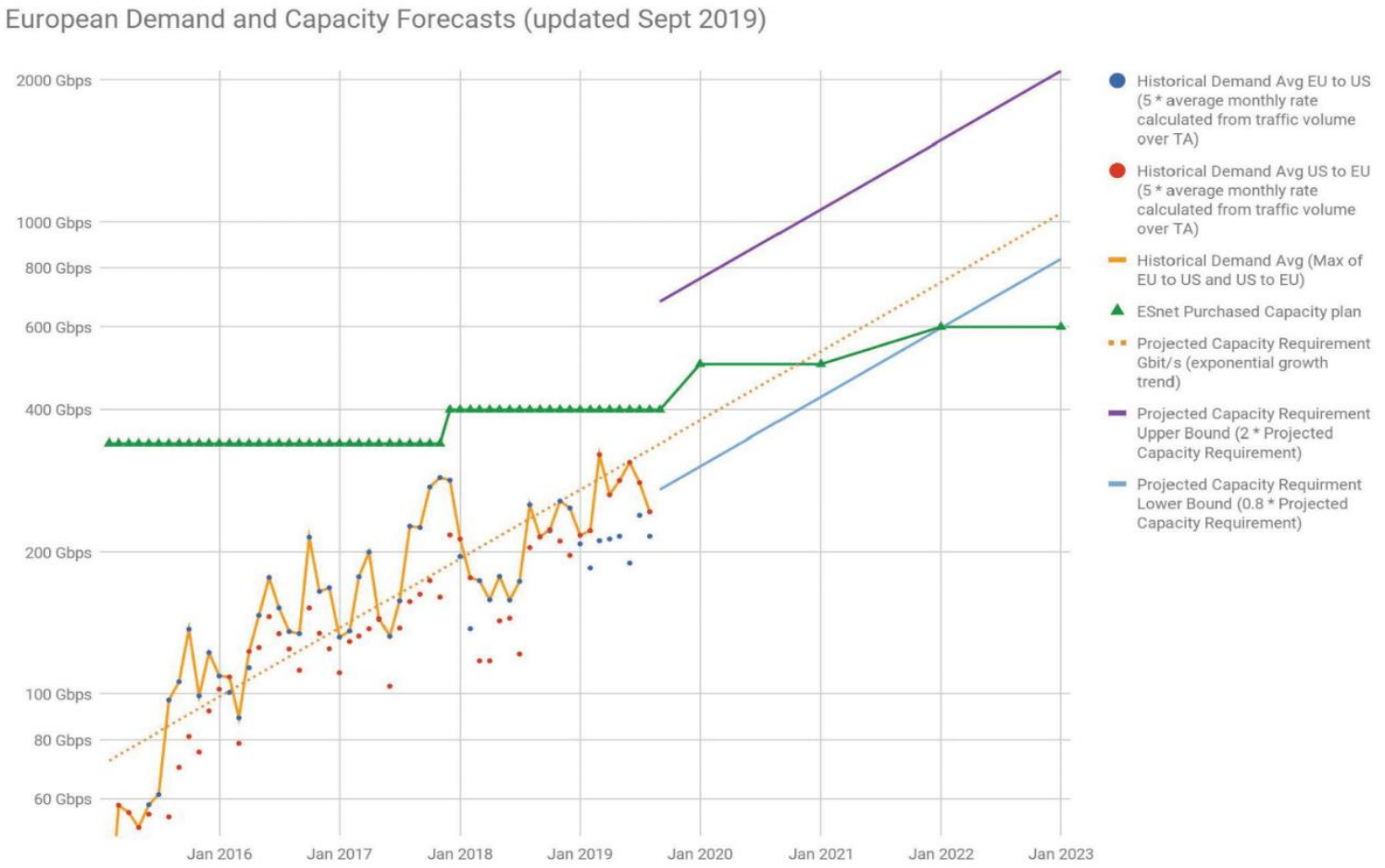
- **Global data distribution, processing, access and analysis**
- **Coordinated use of massive but still limited *diverse* compute, storage and network resources**
- **Coordinated operation and collaboration *within and among* scientific enterprises**

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

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.

Esnet Transatlantic Network Traffic Projections



- 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

- Current Requirements: 0.35 – 0.85 Tbps
  [0.8 to 2X 2016-19 traffic projection]
- Growth Rate 1.4X per year, hence 16X capacity requirement in 2028
- Capacity Requirement = 5.6 to 13.6 Tbps;
  Since this is an Esnet 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
- Implied Shortfall: 3.7 to 5.2X
- Naïve Implementation Outlook by 2028: 68 200G links across the Atlantic
  (for example 17 links on each of 4 disjoint 4 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: This will 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

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

  - **Including the worldwide network as a first class resource** coordinated with distributed computing and storage
  - **Including innovative approaches** in several areas
  - **Leveraging, coordinating and pushing forward** several key developments: from ML methods to computing to regional caches to SDN networks
  - **Integrating or mediating among regional developments** to form a **worldwide fabric** supporting HEP workflow

- **ICFA should consider how the program** to review, design and develop the HL-LHC Computing Model **should be organized**

SDN Enabled Networks for Science at the Exascale

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

Designed for adaption to available "SDN" systems

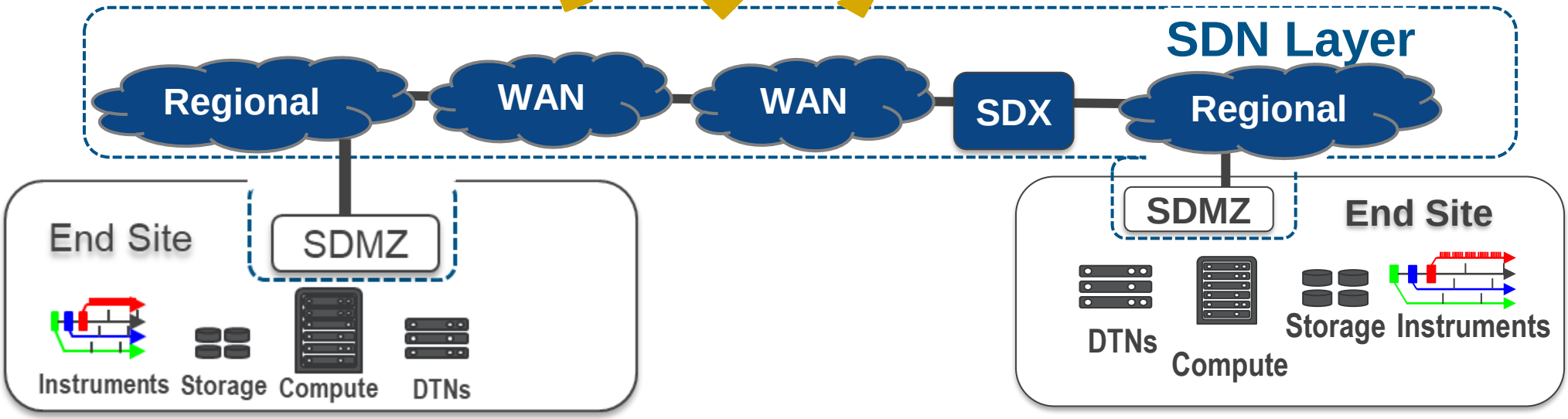
Application Workflow Agents

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

SENSE native "Resource Manager" is available if no current automation layer

SENSE

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

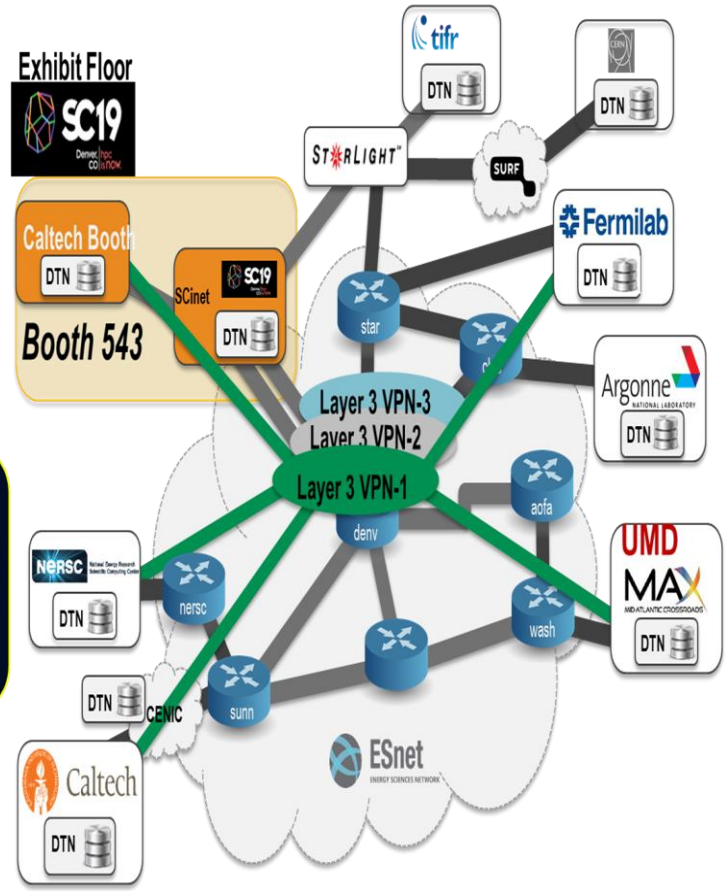


SENSE SC19 Demonstration Topology

SENSE Testbed and L3 VPN Service

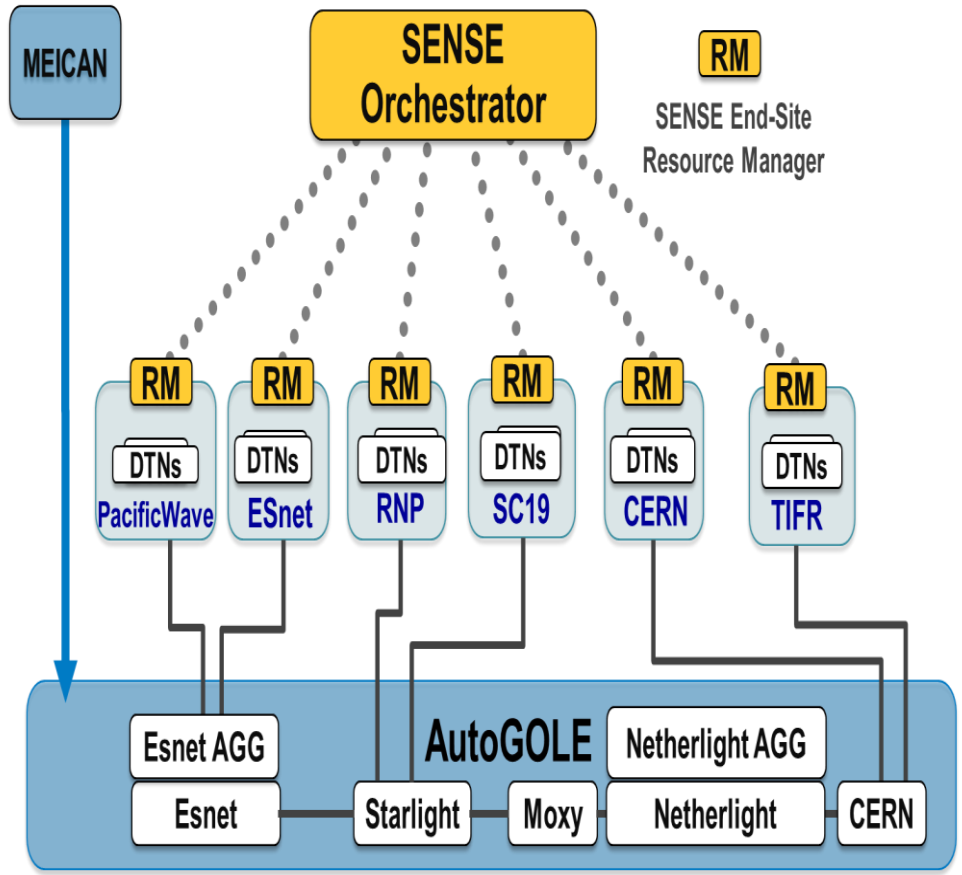
SENSE enabled resources at DOE Labs, Universities, Research Facilities, + SC19

Dynamic attachment of End Site resources to L3VPNs advertised by ESnet



SC19-NRE-020 Intercontinental Demonstration

Multi-Resource Orchestration via AutoGole and SENSE



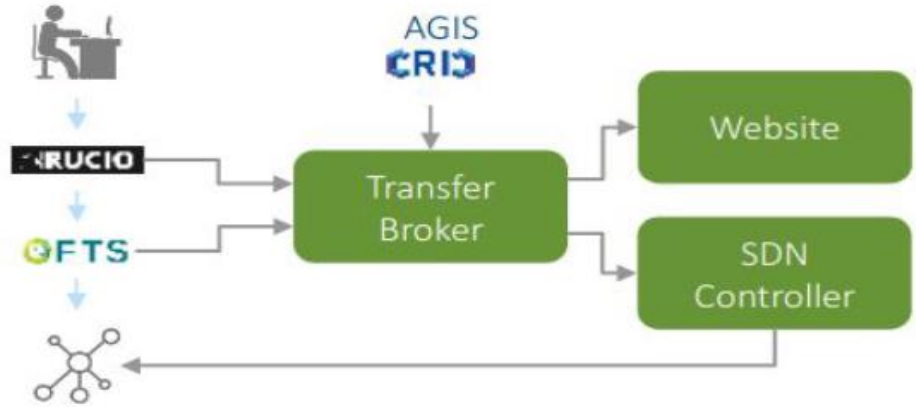
SENSE – AutoGOLE Joint Interworking Demo → Candidate **Inter-regional Mediation Layer for Global Workflows** (as discussed in GNA-G)

For a global fabric, including Australia and Africa we would need to include genomics, LSST, SKA, and others in the overall concept along with HEP

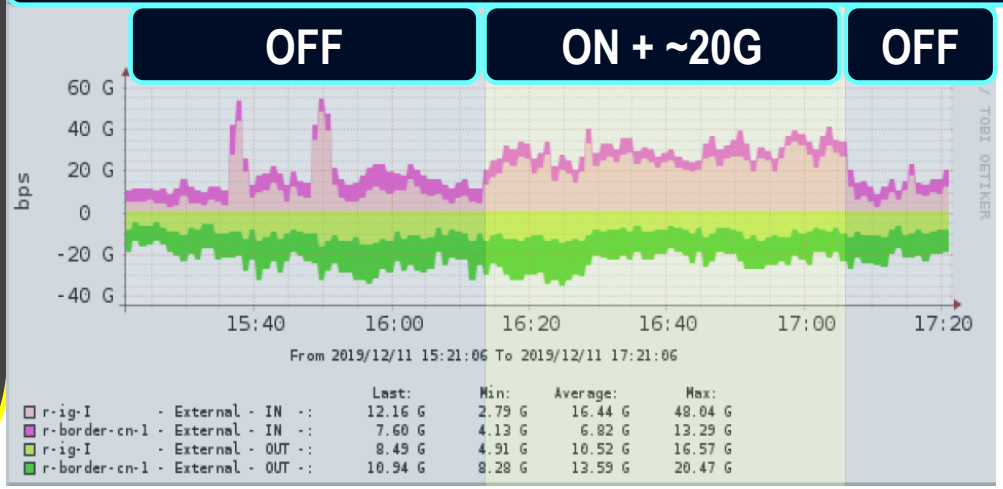
NOTED: Network Optimized Transfer of Experimental Data CERN/IT Project (C. Busse-Grawitz)

- 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



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

Next Generation Computing and Networking
System for LHC and Data Intensive Sciences

- *To meet the challenges of globally distributed Exascale data and computation faced by the LHC and other major science programs*
- *New approaches required: Ai /ML Algorithms: Trigger, Pattern Rec, Analysis; Experiment operations, Online and Offline*
- *A new "Consistent Operations" paradigm:* **SDN goal-oriented policy-governed end-to-end operations, founded on**
  - *Stable, resilient high throughput flows (e.g. FDT); Controls at the network edges, and in the core*
  - *Real-time dynamic, adaptive operations among the sites and networks*
  - *Increasing negotiation, adaptation, with built-in intelligence*
  - *Coordination among VO and Network Orchestrators*

✸ **Bringing Exascale, pre-Exascale HPC and Web-scale 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**

✸ **Engaging with the full range of technologies and many partners**

✸ **We require a *comprehensive, forward looking global R&D program:***

✸ ***ICFA should consider how this should be organized***

Extra Slides

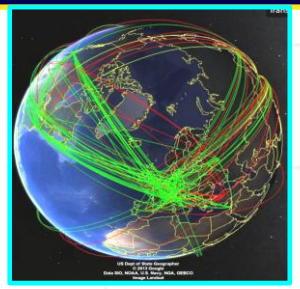
Follow

Towards a Global Network Fabric

- **Following discussions and presentations at the Americas and Global Research Platform workshops in September and the Internet2 Tech Exchange in December, it became clear that the services in the SENSE project could be further developed to serve as a mediator among the intelligent network software systems being developed in the various world regions including Europe (AutoGOLE), Latin America (AmLight), and Asia (Virtual Dedicated Networks). A living example of this was demonstrated at SC19 [*] where interoperation of the SENSE and AutoGOLE network service frameworks, and integral control of the DTN systems and network systems by the SENSE Resource Managers developed by Caltech and ESnet were shown.**
- **This led to plans for a persistent national and global R&D testbed as a venue for ongoing and future network developments in the context of the HL-LHC Computing Model. These developments are also planned to leverage NSF's major investment in FABRIC, "a unique national research infrastructure to enable cutting-edge and exploratory research at-scale in networking, cybersecurity, distributed computing and storage systems, machine learning, and science applications".**
- **[*] SC19 Network Research Exhibition: "LHC Multi-Resource, Multi-Domain Orchestration via AutoGOLE and SENSE", https://sc19.supercomputing.org/app/uploads/2019/11/SC19-NRE-020.pdf**

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

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric

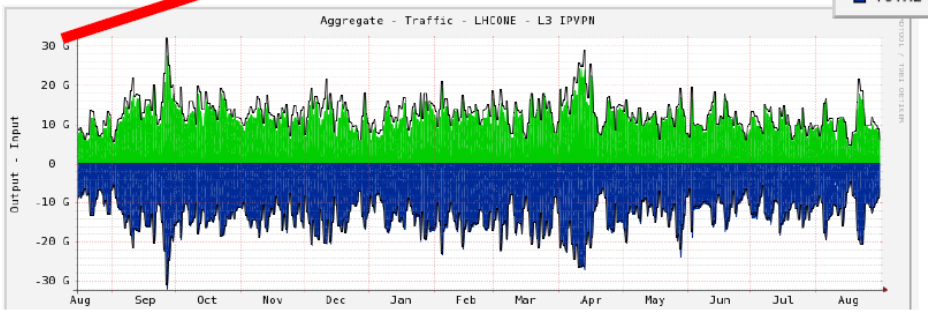
Global infrastructure for *HEP (LHC, Belle II, NOvA, Auger, Xenon)* data flows

Where were we? LHCONE

LHCONE in Europe
GEANT

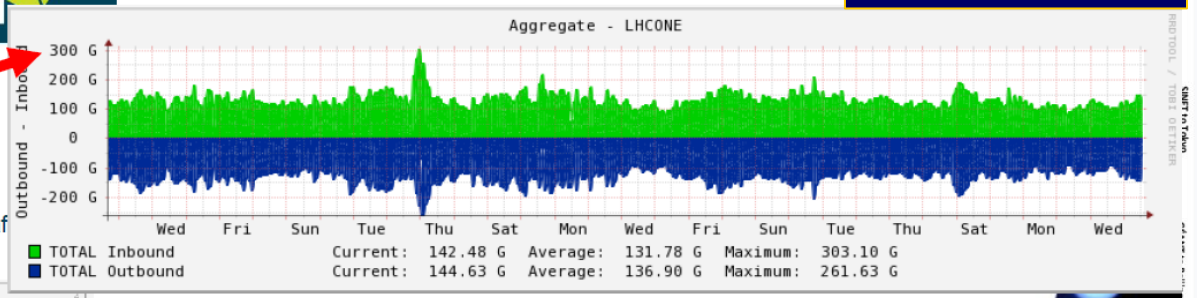
- Aggregate LHCONE traffic from all the NRENs and Peers
  - Average traffic ~25Gbps
  - Sustained Peaks ~35Gbps
  - Trans-Atlantic Traffic ~ 20Gbps (Peak)
- Graphs shows 1 day average traffic over last 12 months the peak traf is much high

10x

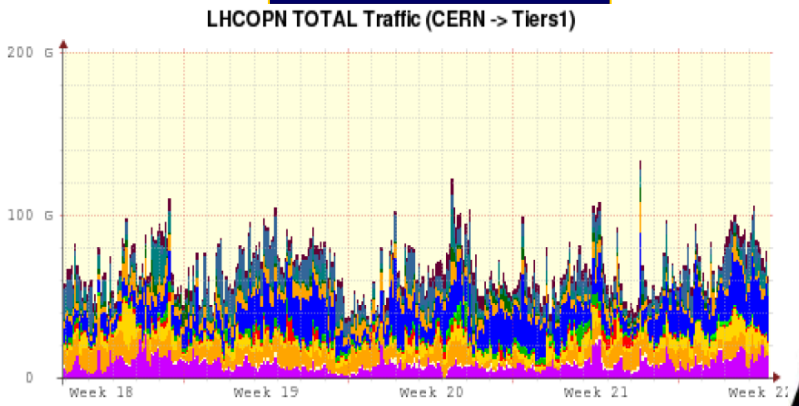


Connect | Communicate | Collaborate 17

LHCONE



+ LHCOPN



3 | www.geant.org

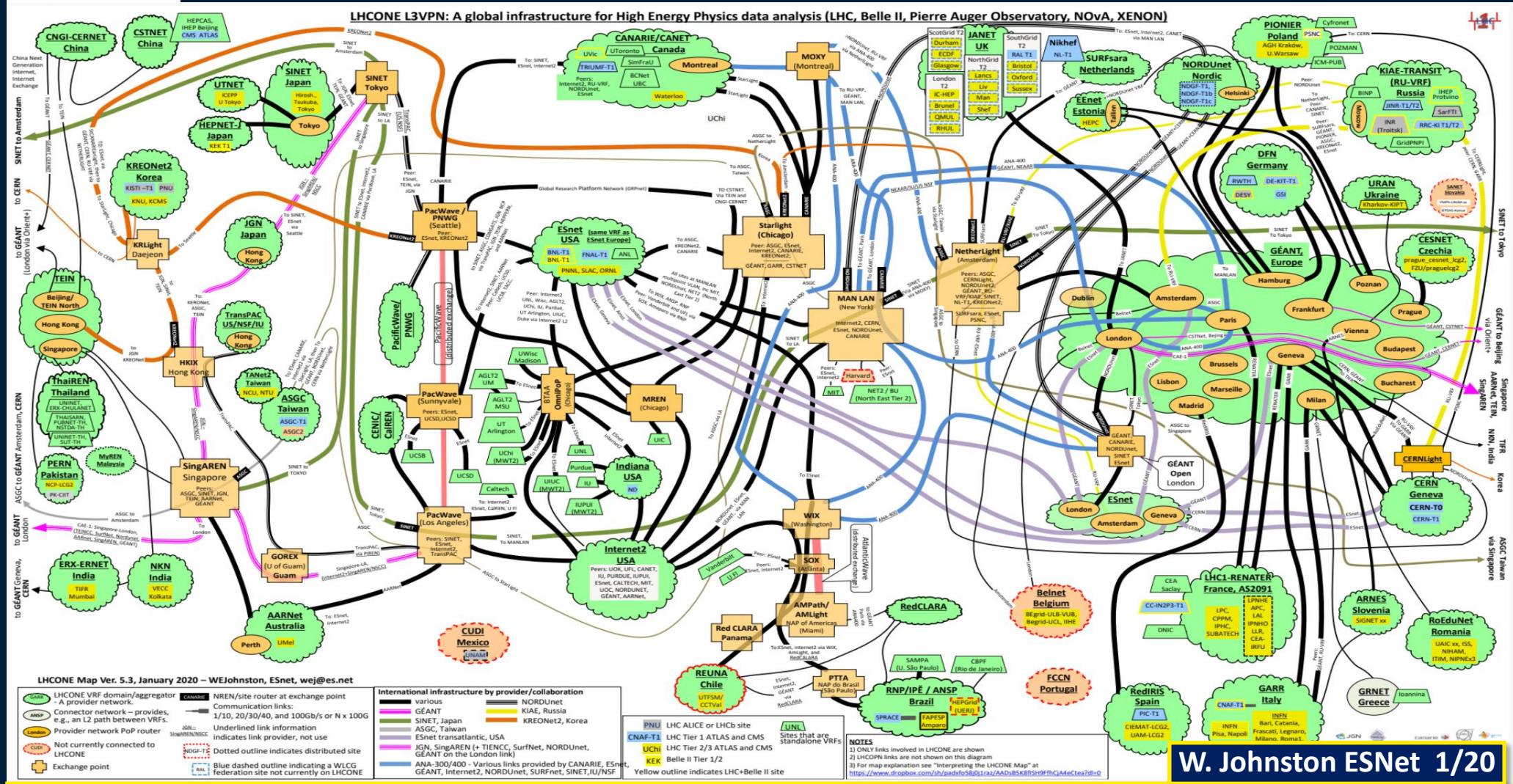
Good News: The Major R&E Networks Have Mobilized on behalf of HEP

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

LHCONE traffic growing by 60-70%/Yr: a challenge already in LHC Run3 (2021-4)

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric

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

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

LHCONE traffic growing by 60-70%/Yr: a challenge already in LHC Run3 (2021-4)

Directions Towards the HL-LHC Computing Model

- It was agreed in subsequent discussions that the HEPIX Technology Watch WG and/or the Global Network Advancement (GNA-G) leadership group that was formed in the fall of 2019 [*], can help define how much of it can be satisfied through technology evolution by 2027, and by 2024 in the preparatory phase
  [Evolution to 400G links; nearing technology limits across oceans]
- The rest will involve a change in paradigm including a system composed of end-to-end services involving coordinated operation among sites and networks, and orchestration:
  - We can leverage developments underway in projects such as SENSE, NOTED and SANDIE.
  - Ongoing discussions should continue to define what the new services and classes required entail.
    - Solutions will vary by region and by network.
- An important part of this is the persistent testbed being deployed by SENSE 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, UCSD, and several other sites in the US, Europe, Latin America and Asia

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