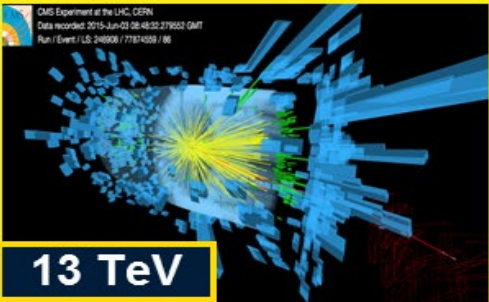


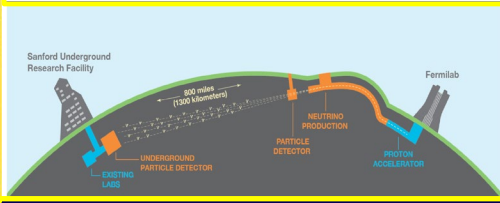
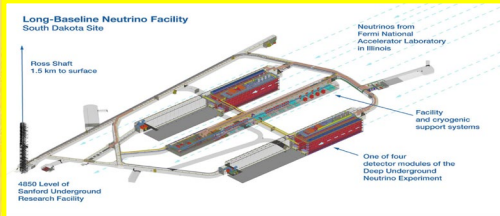
Global Testbeds & System Requirements for Data-Intensive Sciences: A Network-Integrated Paradigm for the HL-LHC Era



LSST



**LHC Run3
and HL-LHC**



LBNF/DUNE



SKA

DUNE

VRO SKA

BioInformatics

**Earth
Observation**

**Gateways
to a New Era**



Harvey Newman, Caltech
Net-Centric Workshop
October 1, 2021



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

- In a manner consistent and compatible with support for the needs of individuals and smaller groups in the at large A&R communities

(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 testbed** as a platform, to foster **ongoing developments** among the science and network communities

- While sharing and advancing the **(new)** concepts, tools & systems needed
- Members of the WG 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**

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**
- **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 use by the at-large A&R communities**
- **Members:**

Alberto Santoro, Azher Mughal, Bijan Jabbari, Brian Yang, Buseung Cho, Caio Costa, Carlos Antonio Ruggiero, Carlyn Ann-Lee, Chin Guok, Ciprian Popoviciu, Dale Carder, David Lange, David Wilde, Dima Mishin, Edoardo Martelli, Eduardo Revoredo, Eli Dart, Eoin Kenney, Frank Wuerthwein, Frederic Loui, Harvey Newman, Heidi Morgan, Iara Machado, Inder Monga, Jeferson Souza, Jensen Zhang, Jeonghoon Moon, Jeronimo Bezerra, Jerry Sobieski, Joao Eduardo Ferreira, Joe Mambretti, John Graham, John Hess, John Macauley, Julio Ibarra, Justas Balcas, Kai Gao, Karl Newell, Kaushik De, Kevin Sale, Lars Fischer, Liang Zhang, Mahdi Solemani, Maria Del Carmen Misa Moreira, Marcos Schwarz, Mariam Kiran, 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*
- **Working Closely with the AutoGOLE/SENSE WG**
 - ★ **Meets Weekly or Bi-weekly; All are welcome to join.**

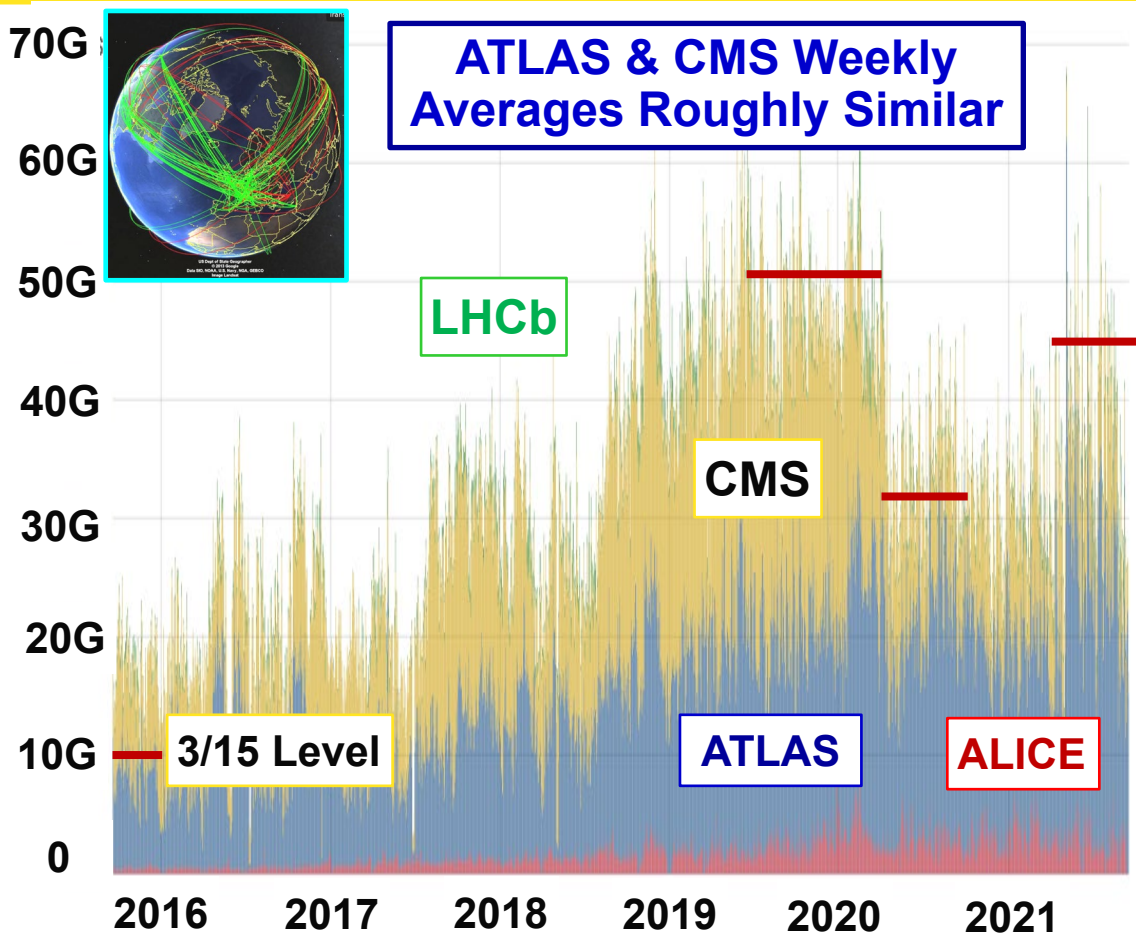
Towards a Computing Model for the HL LHC Era

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
- 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 projected to outstrip the affordable capacity**
- 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
- Needs are further specified in “blueprint” Requirements documents by US CMS and US ATLAS, submitted to the ESnet Requirements Review in August 2020, and captured in a comprehensive DOE Review report for HEP
- Three areas of particular 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 24 X 7 X 365 average throughput 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

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

WLCG Transfers Dashboard: Throughput Sept 2015 – Sept 2021



15 to 58 GBytes/s Week Avg
To 70+ GBytes/s Daily Avg

Complex Workflow

- ~ 900k jobs (threads) simultaneously
- Multi-TByte to Petabyte Transfers;
- To ~10 M File Transfers/Day
- 100ks of remote connections
- The effects of Covid from Spring 2020 are evident
- The recovery is emerging: warrants careful watching

5X Growth in Throughput in 2015-2019: +50%/Yr; ~60X per Decade

<https://monit-grafana.cern.ch/d/AfdonlvGk/wlcg-transfers?orgId=20&from=now-6y&to=now>

Annual CMS Data Volume

	# of collisions	# of events simulated	RAW event size [MB]	AOD event size [MB]	Total per year [PB]
Today	9 Billion	22 Billion	0.9	0.35	~20
HL-LHC	56 Billion	64 Billion	6.5	2	~600

The beams get “brighter” by x6
 Data taking rate goes up by x6
 Simulations go up by x3

**Primary Data volume
 per year goes up by x30**

This talk is about R&D strategies to keep the cost the ~same despite a x30 increase in data volume per year.

Will motivate the R&D via a detour on how science is done.

Conclusion: CMS Data ~Exabyte/Year by ~2028 at HL-LHC



HL-LHC Network Needs and Data Challenges

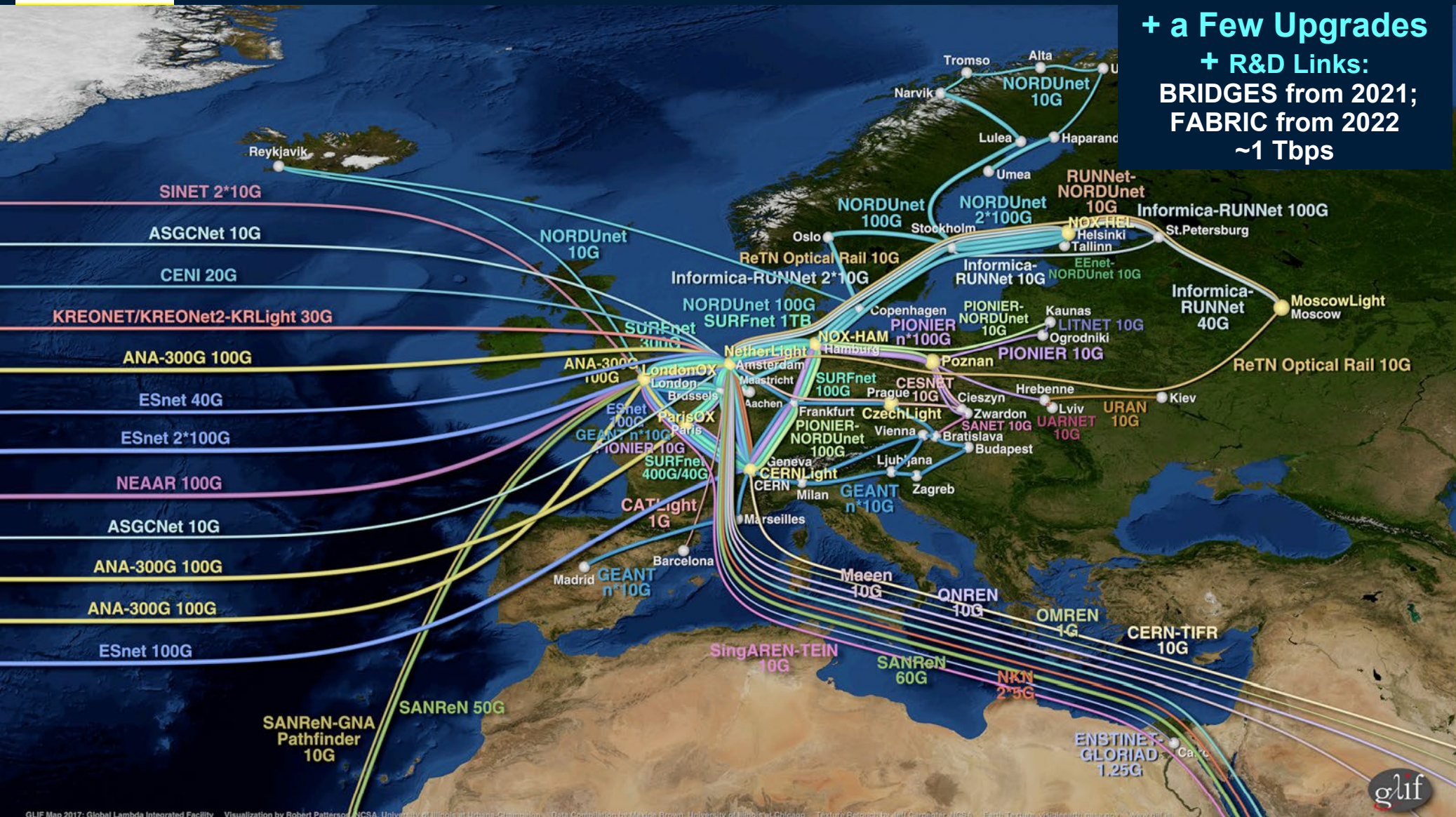
Current Understanding: Q2 2021

- **Export of Raw Data from CERN to the Tier1s (350 Pbytes/Year):**
 - **400 Gbps Flat each** for ATLAS and CMS Tier1s;
+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

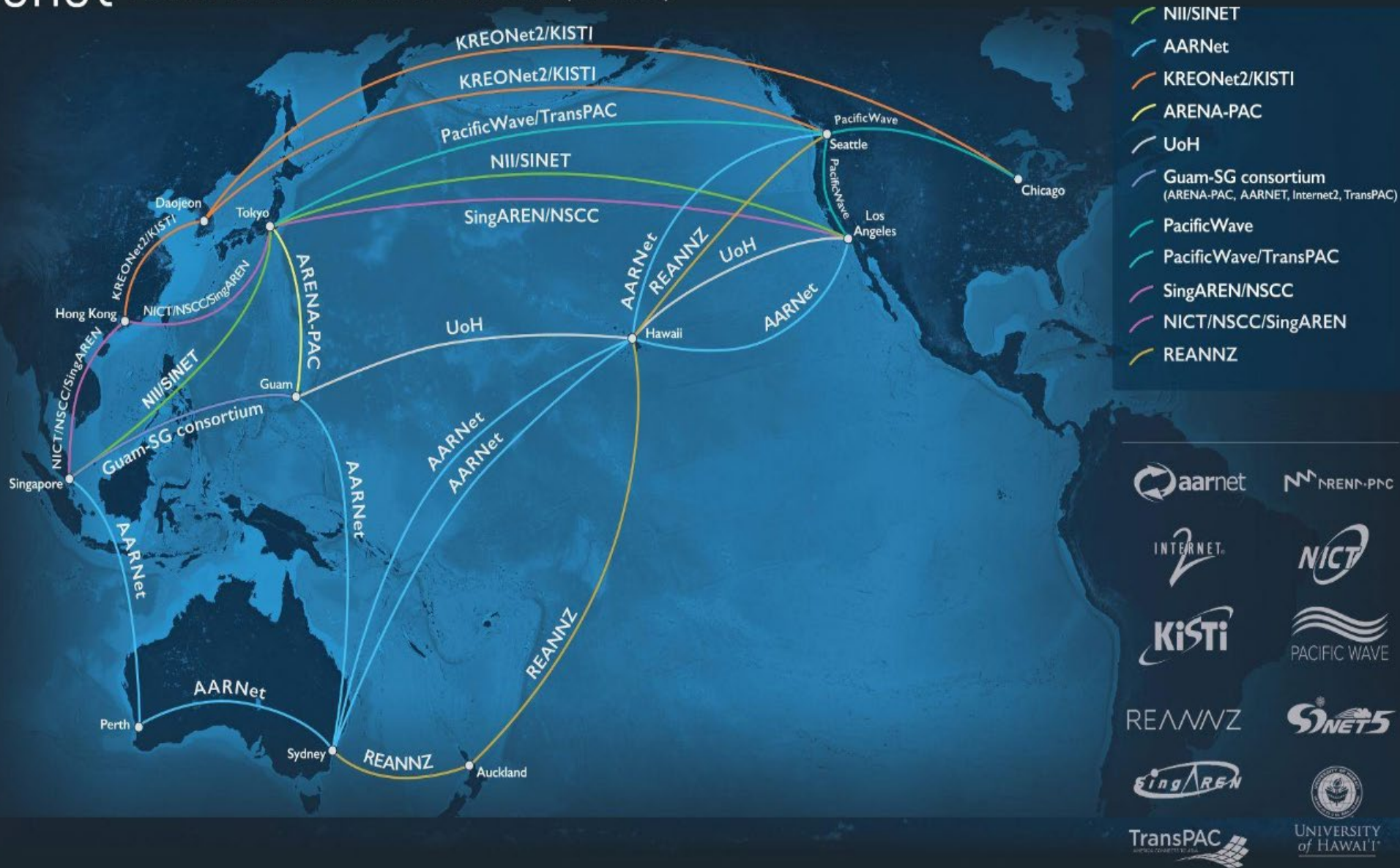
Shared Network Infrastructure: GLIF Map (2017)



**+ a Few Upgrades
+ R&D Links:
BRIDGES from 2021;
FABRIC from 2022
~1 Tbps**

GLIF Map 2017: Global Lambda Integrated Facility Visualization by Robert Patterson, NCSA, University of Illinois at Urbana-Champaign. Data Compilation by Maxine Brown, University of Illinois at Chicago. Texture Retouched by Jeff Carpenter, NCSA. Earth Texture by Kiblenet/earth.google.com. www.glif.org

Slow Growth in Capacity at Fixed Cost: ~2 Tbps TA by 2028
Sharing with the larger academic & research community on several continents

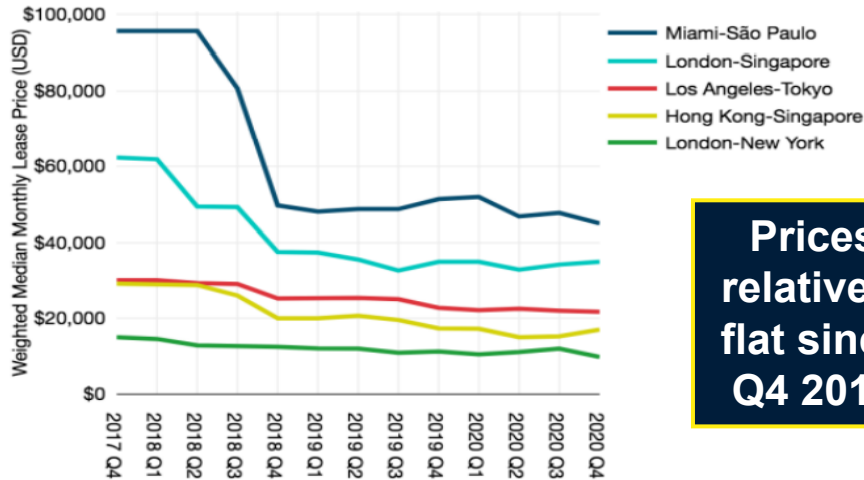


The Rising Transpacific and Asia Pacific Network Community In a Global Context

International Bandwidth Pricing Trends

Executive Summary (telegeography.com)

Weighted Median 100 Gbps Wavelength Price Trends on Major International Routes



Prices relatively flat since Q4 2018

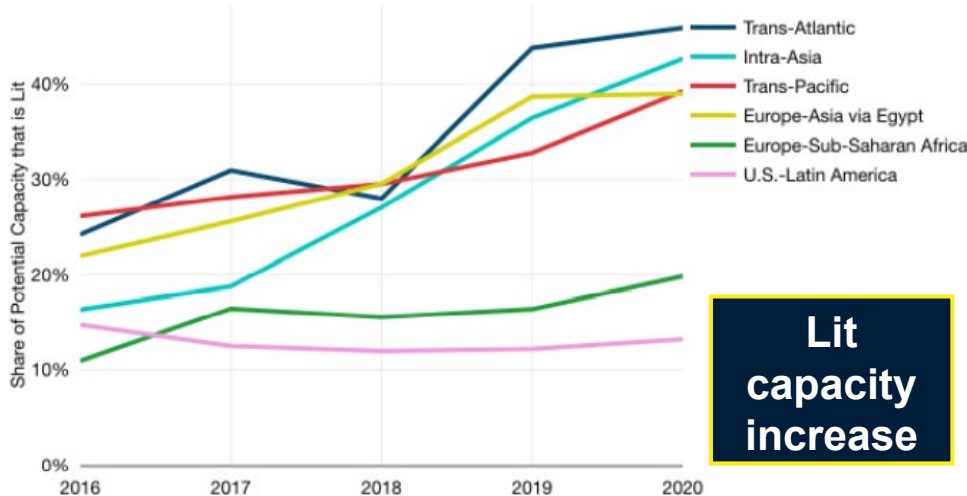
Notes: Each line represents the weighted median monthly lease price for an unprotected 100 Gbps Wavelength on the listed route. Prices are in USD and exclude local access and installation fees.

10 Gbps and 100 Gbps Wavelength Weighted Median Prices and Multiples on Select International Routes



Notes: Each bar represents the weighted median price for an unprotected wavelength for the listed capacity and route. Prices are in USD and exclude local access and installation fees. MRC = Monthly recurring charge. Multiples are derived by dividing the price of the larger circuit by the price of the smaller circuit.

Percentage of Potential Capacity that is Lit on Major Submarine Cable Routes



Lit capacity increase

- **Price Evolution 2017-20**
 - ★ **-16% Price CAGR Average**
 - ★ **Only -10 to -13 % CAGR NYC – London and LA-Tokyo**
 - ★ **To -6% 2019-20 due to COVID**
 - ★ **100G/10G Price Multiple: 4.3X, Down from 6.4X in 2015**
 - ★ **Below 4X NYC-London**

Technology Push: Rising Network Capabilities of Servers + Storage





- The commoditization of 32 X 100G Switches, NICs, transceivers is now mature

<p>** Image may not exactly match product **</p>  <p>Z9100-ON US \$2,260.00</p> <p>Description DELL NETWORKING Z9100-on 32 X 100gbe + Refurbished. In Stock.</p>	 <p>NVIDIA Mellanox MCX515A-CCAT ConnectX®-5 EN Network Interface Card, 100GbE Single-Port QSFP28, PCIe3.0 x16, Tall Bracket #119648</p> <p>Accelerated Switching / Packet Processing / DPDK</p> <p>US\$ 749.00 Import Fees included ⓘ FS P/N: MCX515A-CCAT 260 Sold · 0 Review · 4 Questions</p>	 <p>Dell 407-BCDH Compatible 100GBASE-LR4 QSFP28 1310nm 10km DOM LC SMF Optical Transceiver Module for Data Center #8922</p> <p>US\$ 499.00 FS P/N: Q5FP28-LR4-100G 1,48 Sold · 164 Reviews · 12 Questions</p>	 <p>Edgecore Networks WEDGE100BF-32X</p> <p>Edgecore 32 x 100GbE QSFP28 ToR spine switch, Tofino ASIC, Layer 2/3 switching and routing via additional OS, redundant PSU, 3Y Warranty</p>
---	--	---	--

- Commoditization of 200G NICs and 200-400G Switches is well underway

 <p>NVIDIA Mellanox MCX623106AN-CDAT ConnectX®-6 Dx EN Network Interface Card, 100GbE Dual-Port QSFP56, PCIe4.0 x16, Tall Bracket #119646</p> <p>Virtualization / PTP / Connection Tracking Offload</p> <p>US\$ 1,119.00 Import Fees included ⓘ FS P/N: MCX623106AN-CDAT 181 Sold · 0 Review · 3 Questions</p>	 <p>N9500-32D, 32-Port L3 Data Center White Box Switch, 32 x 400Gb QSFP-DD, Broadcom Chip, Bare-Metal Hardware #96982</p> <p>Spine switch for data centers and large enterprise networks</p> <p>US\$ 9,999.00 Import Fees included ⓘ FS P/N: N9500-32D</p>	 <p>Introducing 400GbE Tofino 2 Switch</p> <p>23,500 USD</p> <p>Edgecore AS9516-32D</p>	<p>Tofino Fully P4 Programmable Tofino2 (25.6 Tbps)</p>
---	---	---	--

- Production 2U compute servers (e.g. Supermicro 2124BT-HNTR): PCIe 4.0, 16 200G NICs and 16 Gen4 NVMe SSDs possible in 2U capable of 8 X 200G, ~100 GB/sec IO

 <p>Views: Angled View Node View Front View Rear View </p>	 <p>Integrated Board Super H12DST-B</p>	<p>Key Features</p> <ul style="list-style-type: none"> - Compute Intensive Applications - HPC, Data Center, Enterprise Server - Hyperscale / Hyperconverged <p>Four hot-pluggable systems (nodes) in a 2U form factor. Each node supports the following:</p> <ol style="list-style-type: none"> 1. Dual AMD EPYC™ 7003/7002 Series Processors (7003 Series Processor drop-in support requires BIOS version 2.0 or newer) 2. Up to 4TB ECC 3DS LRDIMM, up to DDR4-3200MHz; 16 DIMM slots 3. 2 PCIe 4.0 x16 (LP) slot; 	 <p>WD_BLACK 500GB SN850 NVMe Internal Gaming SSD Solid State Drive - Gen4 PCIe, M.2 2280, 3D NAND, Up to 7,000 MB/s -...</p> <p>★★★★★ ~ 528</p> <p>\$119⁹⁹ \$149.99</p> <p>✓prime FREE Delivery Sat, Jun 5</p>	 <p>SAMSUNG 980 PRO 2TB PCIe NVMe Gen4 Internal Gaming SSD M.2 (MZ-V8P2T0B/AM)</p> <p>★★★★★ ~ 2,384</p> <p>\$429⁹⁹</p> <p>FREE Delivery for Prime members</p> <p>Add to Cart</p>
--	--	--	--	---

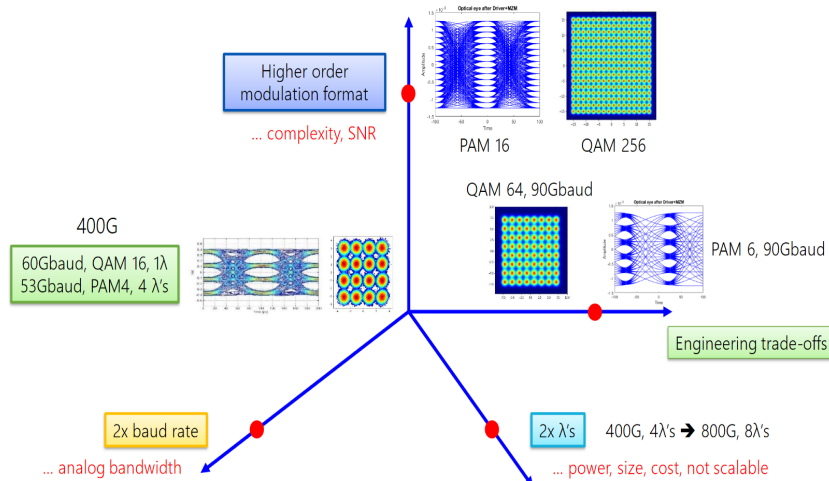
- NOTE: PCIe Standards Clock Now 2 Years: Products: PCIe 5.0 by ~2022-3; PCIe 6.0 by ~2025; ~2X performance per generation; Multi-Tbps servers by HL LHC**
- Paralleled/driven by motherboard, chip architecture and interconnect improvements

Technology Push: Data Center, Metro, Long Haul Interconnects: 400G Long Haul + "The Race to 800G"

https://www.inphi.com/wp-content/uploads/2021/01/20210113_COBO_RNagarajan_Inphi_v3_distri.pdf

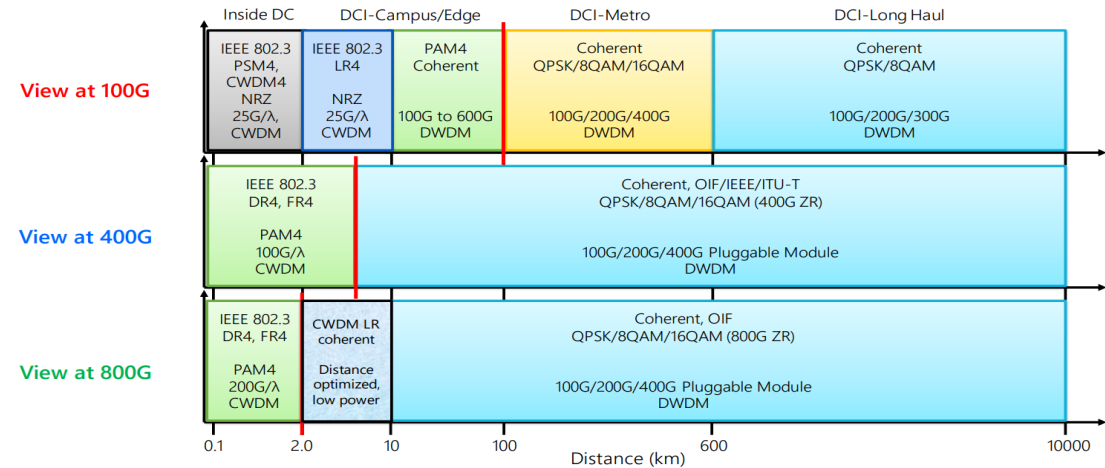
New Modulation schemes

Ubiquitous 2x speed scaling

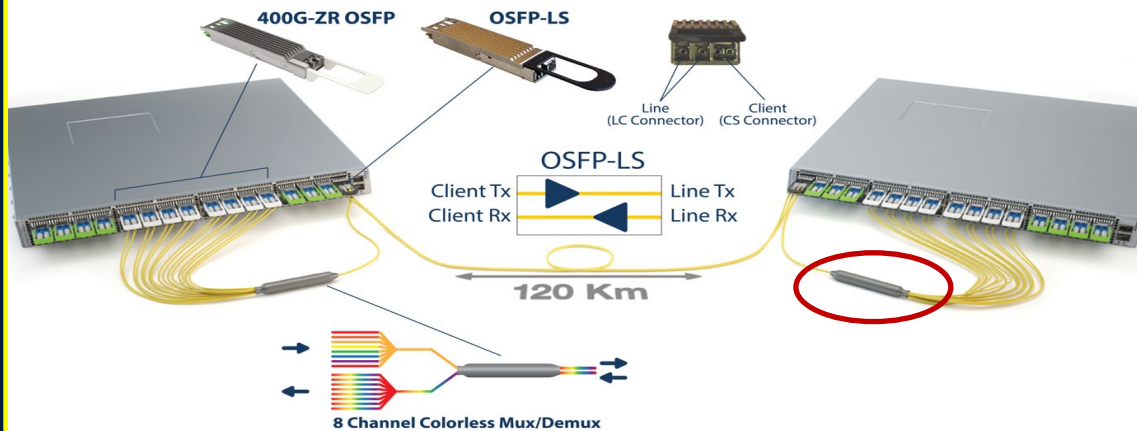


Technology Choices over Distances: Modulations, Coherent, WDM with 100, 200G channels

Technology choices in data center interconnects



Data Center Interconnect - Simplified



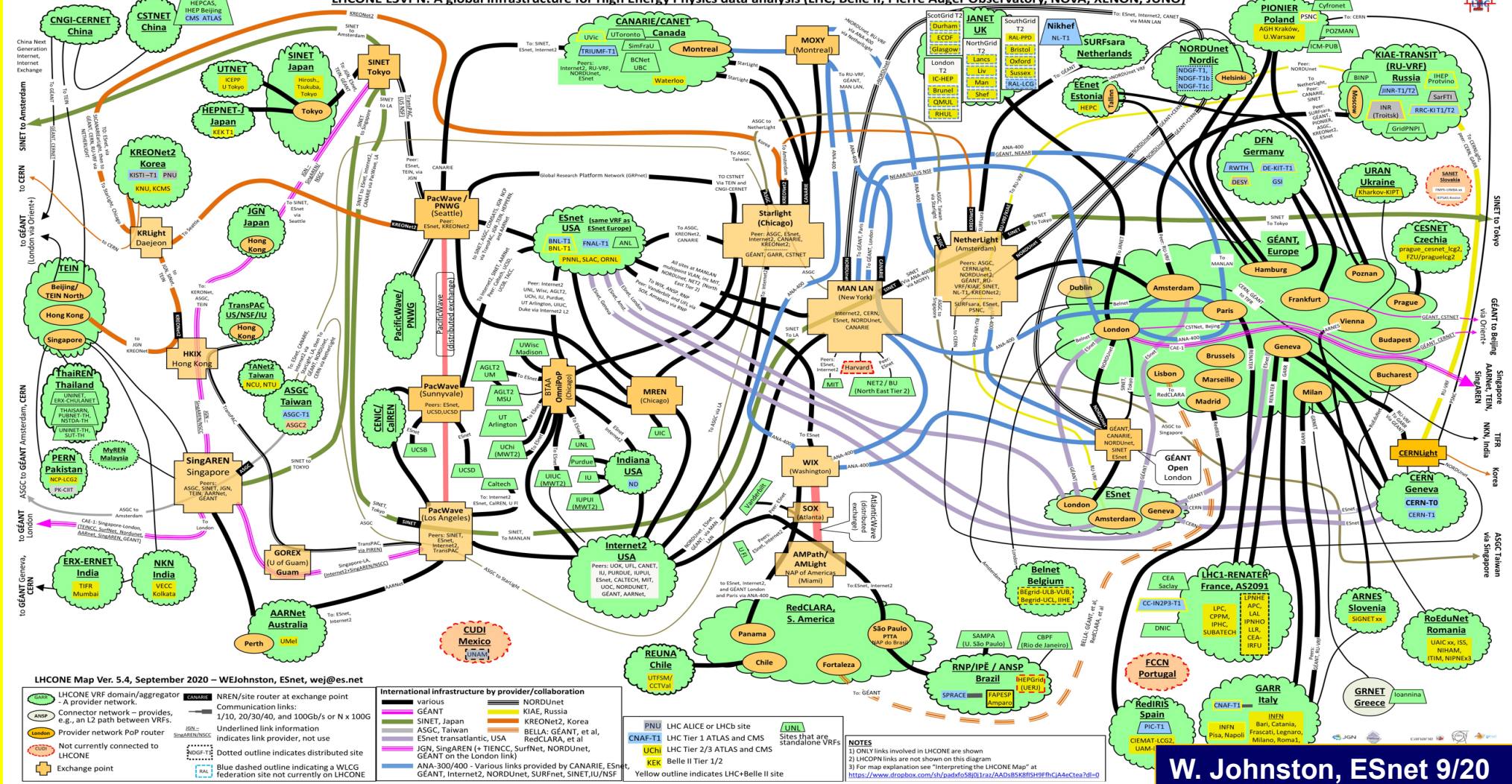
Emerging Already in 2021-22:
Pluggable Transceiver/Transponders
+ SMALL Colorless Mux/Demux
Wave Mixers: 400G ZR for ~100km,
400G ZR+ for 250-500 km+
Eliminating the Optical Line System
in up to 8 or 16 X 400G Use Cases
★ XR Optics: Optical Stat-Muxing



LHCONE VRF: The Challenge of Complexity and Global Reach

Global infrastructure for HEP (LHC, Belle II, NOvA, Auger, Xenon, Juno ...)

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

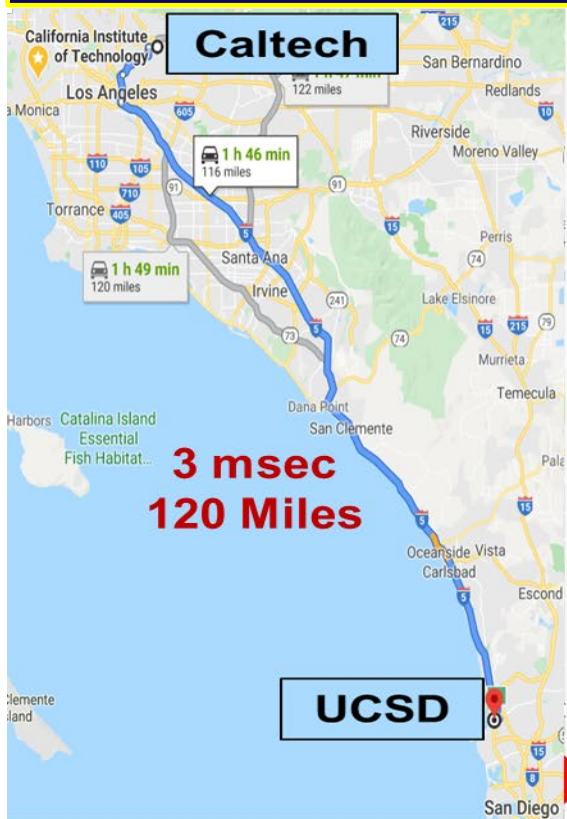


W. Johnston, ESnet 9/20

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 ?

(Southern) California ((So)Cal) Cache

Roughly 20,000 cores across Caltech & UCSD ... half typically used for analysis
A 1.5 to 2 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



Plan to include Riverside and other SoCal Tier3s; ESnet plan to install additional in-network caches near US Tier2s; Move to Ceph

Scaling to HL LHC: ~ 20-30 Pbytes Per Tier2, ~5-10 Pbyte Caches, ~1 Petabyte Refresh in a Shift Requires 400G Link. Still relies on use of compact event forms, efficient managed data transport

ESnet HEP Network Requirements Review 2020*

CMS Ideas for Future R&D with ESnet

- 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 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.

* DOE Offices of HEP, ASCR, 2021 HEP Requirements Report: <https://escholarship.org/uc/item/78j3c9v4>
Report LBNL-2001398. Also see <https://www.es.net/science-engagement/science-requirements-reviews/requirements-review-reports/>

Developing the Next Computing Model

System Prerequisites and Proposed Paradigm

- **The new Computing Model must do more than make best use of limited network resources:**
- **It must also ensure that our use does not overly impede other traffic**
 - ★ **We must remain a friendly partner of the R&E networks**
- **Corollaries: (1) Experiments must account for and manage *all* operations requiring wide area network resources**
 - (2) **We cannot assume that many smaller transfers can be left unmanaged: in aggregate they can also damage shared networks**
- **Any defined level of service requires VO-network communication**
 - **Examples:** BW allocation with QoS, deadline scheduling, flow-group classification + prioritization, taking back of unused net resources, etc.
 - **Sufficient information exchange is needed to deal with:** service adjustments in flight, compromises, what-ifs, hard choices
- **Model: A distributed data center analog, with adaptive real-time responses**
 - **Keys:** intelligent, software driven control & data planes; ML optimization
- **We need to embark on the recommended R&D program now**
 - **To learn and adapt to the actual requirements and constraints**
 - **Evaluate the complexity versus capacity (funding) tradeoffs if needed**



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

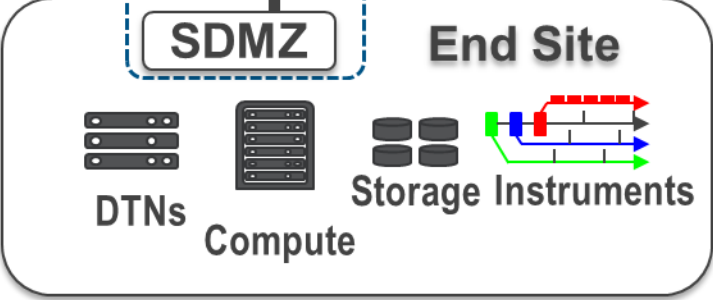
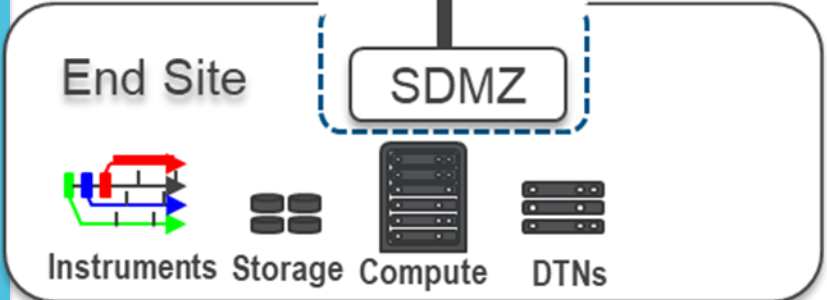
SENSE Native RMs are Available if no current automation layer

Application Workflow Agents

SENSE 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

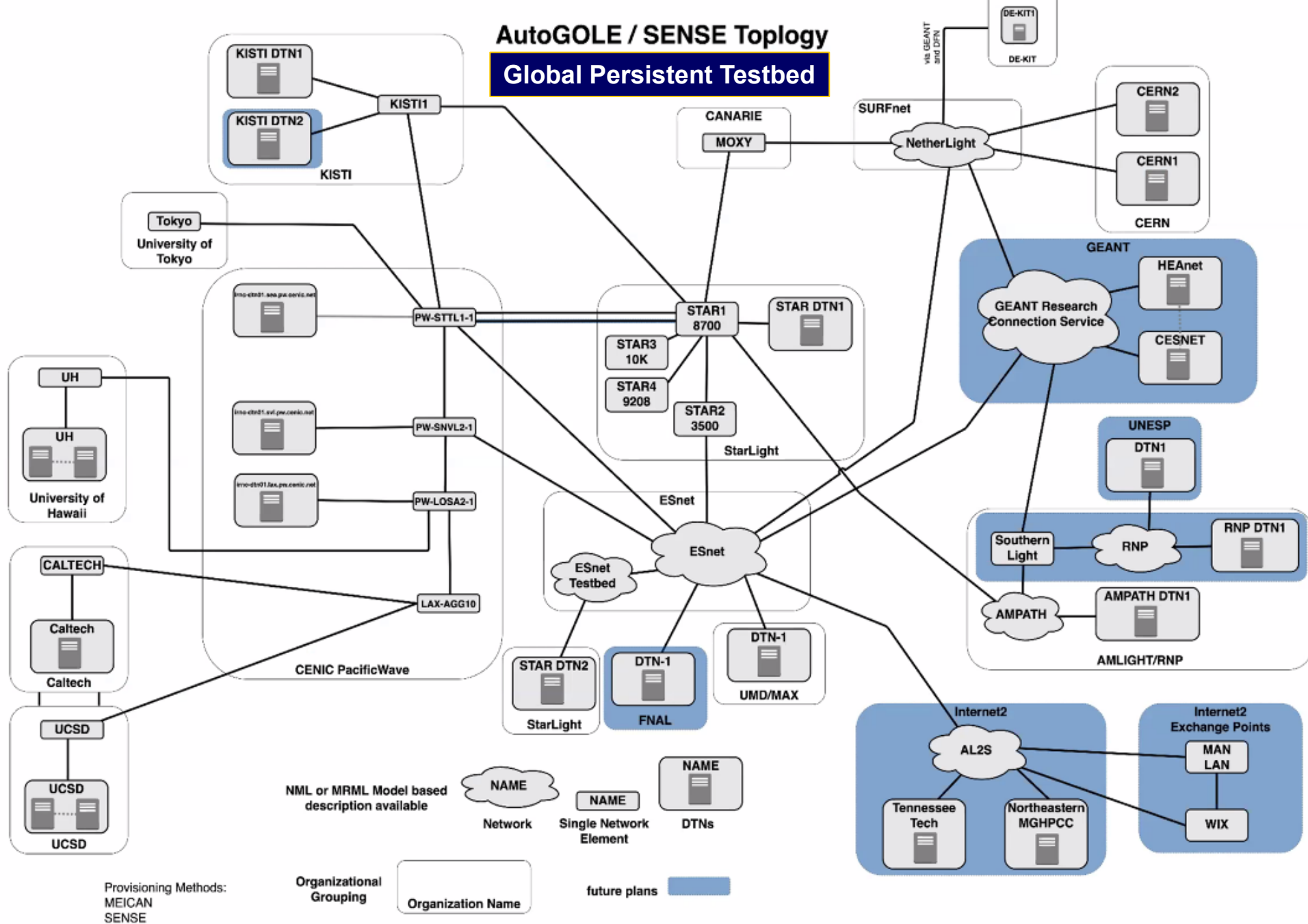
SENSE



Tom Lehman Talk

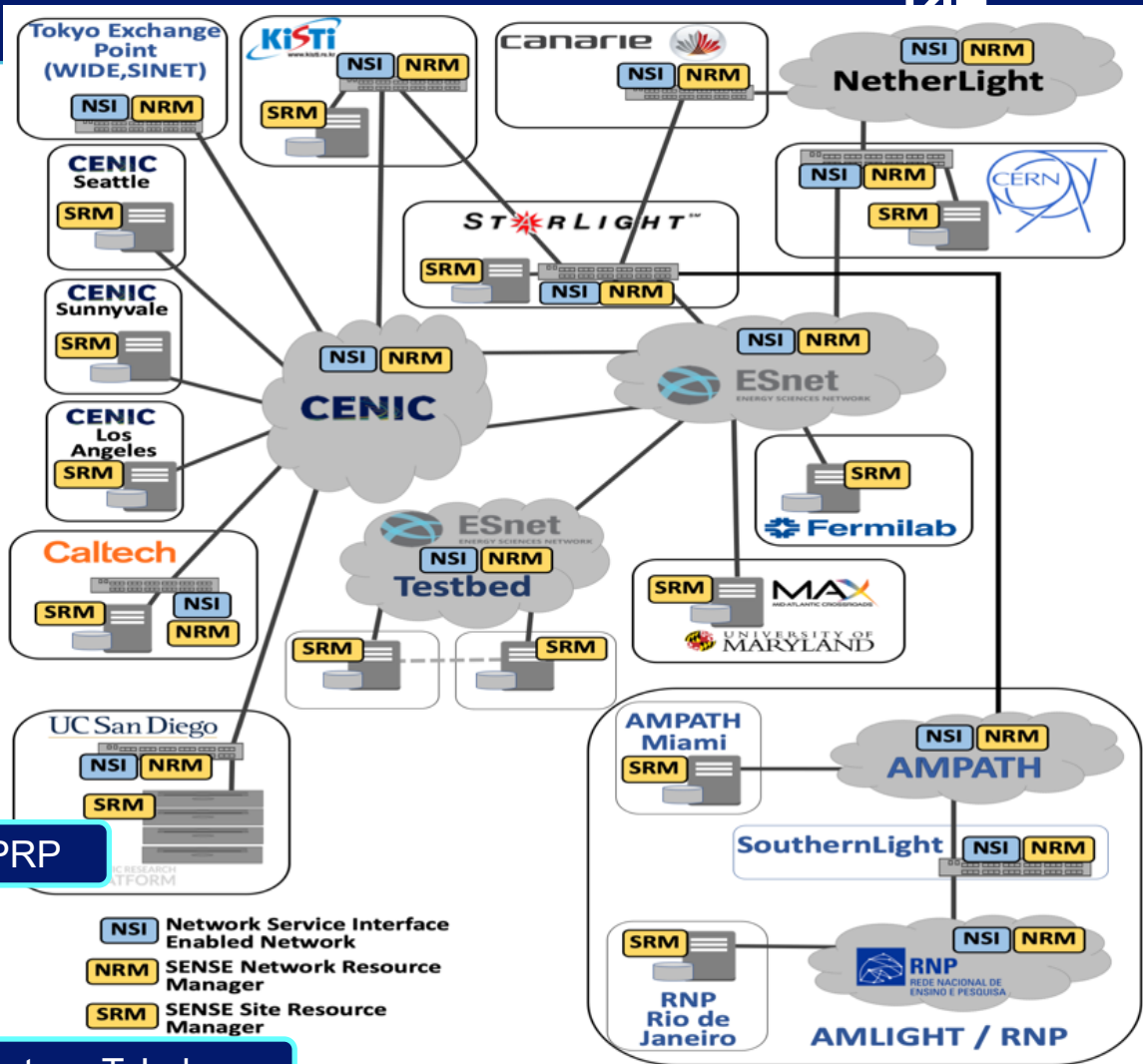
AutoGOLE / SENSE Topology

Global Persistent Testbed



SC20 to 21: AutoGOLE/SENSE Persistent Testbed:

ESnet, SURFnet, Internet2, StarLight, CENIC, Pacific Wave, AmLight, RNP, GridUNESP, KISTI, Tokyo, Caltech, UCSD, PRP, FIU, CERN, Fermilab, UMD, DE-



PRP

- NSI** Network Service Interface Enabled Network
- NRM** SENSE Network Resource Manager
- SRM** SENSE Site Resource Manager

Courtesy T. Lehman

2021 Outlook
ESnet6/
High Touch
FABRIC
BRIDGES

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

Federation with
the StarLight
GEANT/RARE
& AmLight
P4 Testbeds

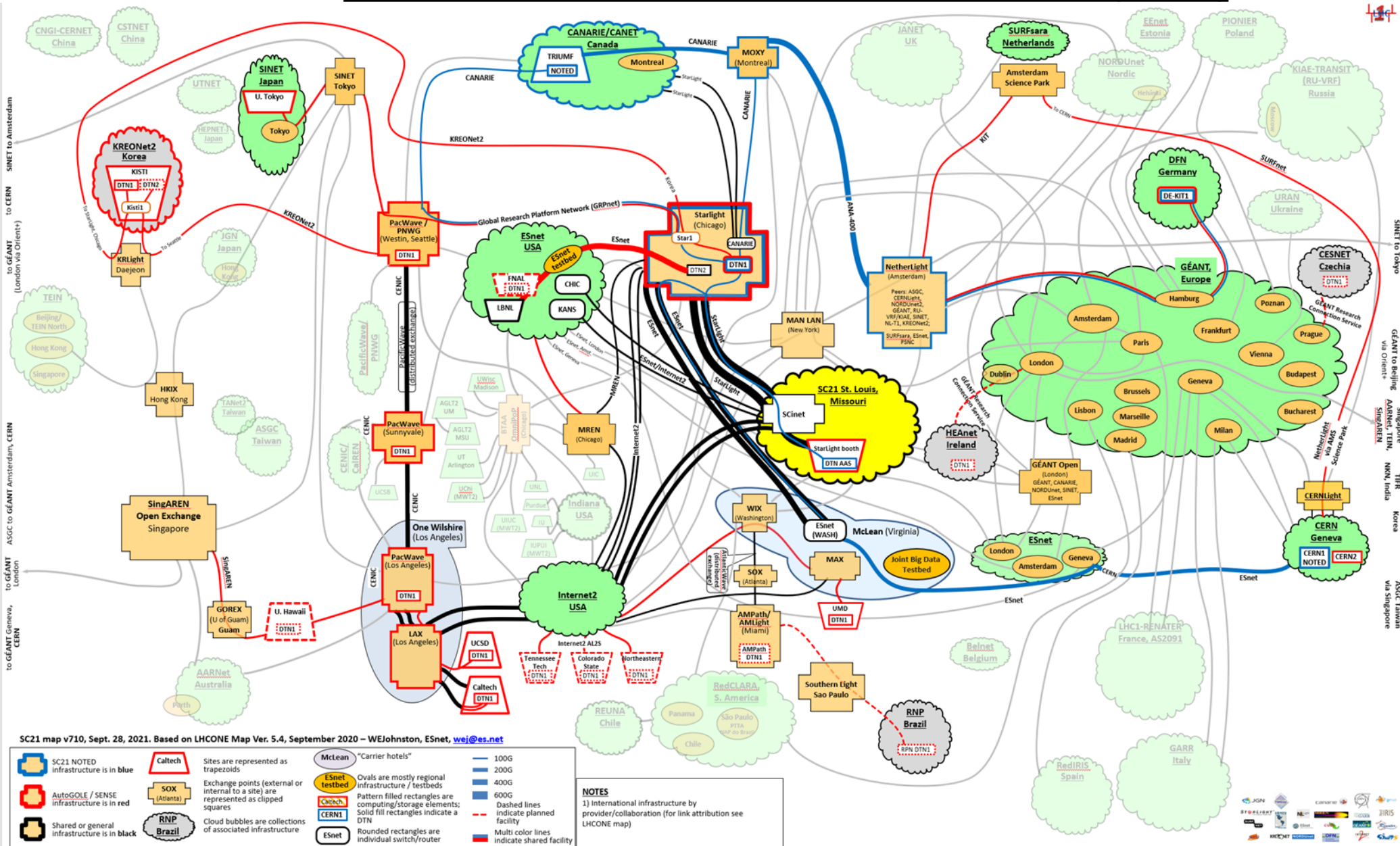
400G
Link(s)
NetherLight-
CERN

Caltech/
UCSD/
Sunnyvale
400G/
2 X 200G
with CENIC
Enhanced
for SC21

Automation
Following
Atlantic
Wave SDX

Persistent Operations: Beginning this Year

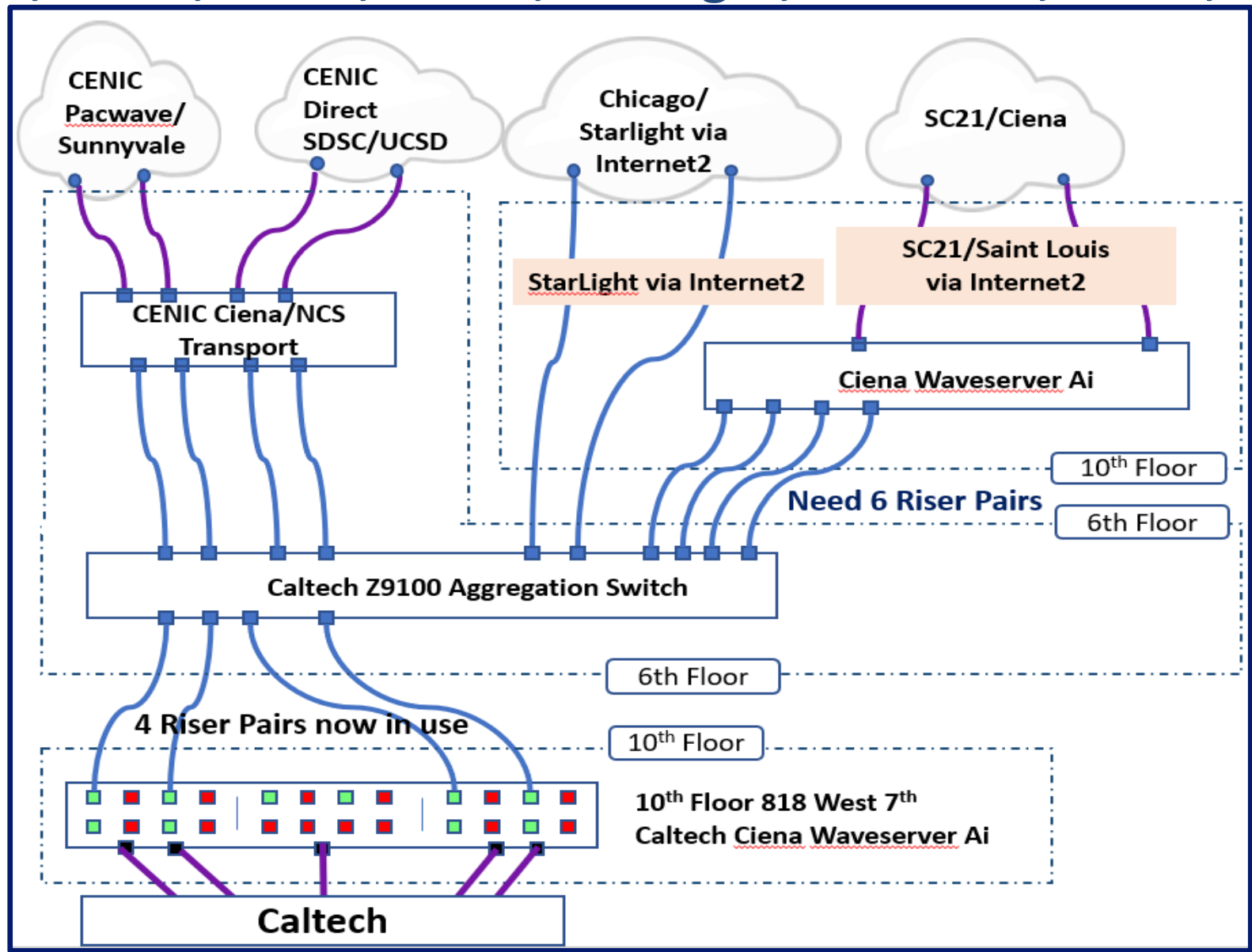
SC21 NREs: AutoGOLE / SENSE and NOTED (Draft)



Global footprint. Multiple 400G and some 600G Optical Links

Bill Johnston 9/21

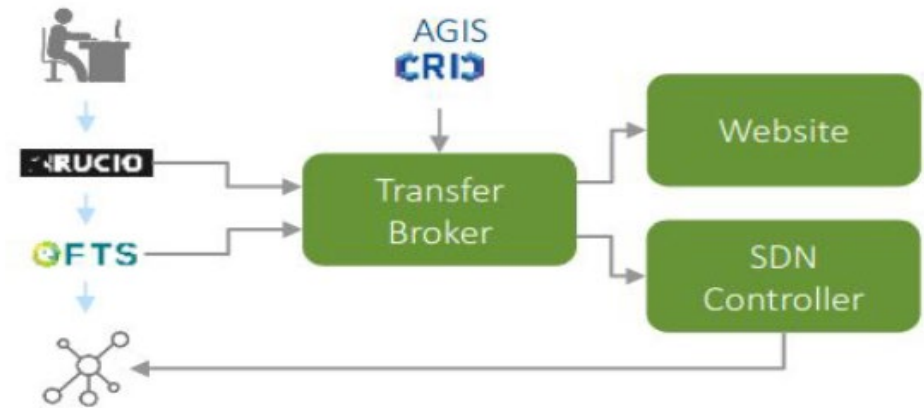
SC21: Next Step Advanced Advanced Network Infrastructures: Caltech, UCSD, Ciena, CENIC, StarLight, Internet2, ESnet, SCInet



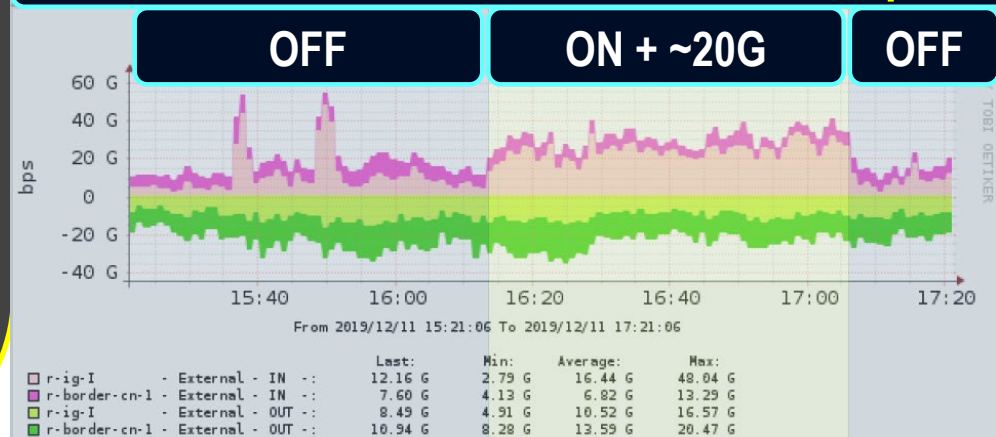
NOTED: Network Optimized Transfer of Experimental Data CERN/IT Project (J. Waczynska, E. Martelli et al.)

- 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.
- 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



SC21 NRE Involving the LHC Tier1s and the LHCOPN

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

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%**

SENSE Services for LHC/Open Science Grid (OSG)

Workflows: with T. Lehman (ESnet), F. Wuerthwein (UCSD)

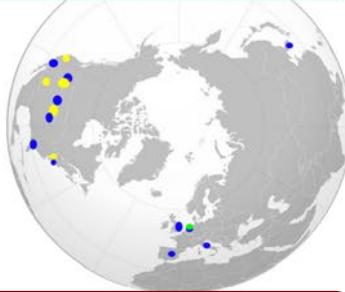
- ★ **SENSE provides the mechanisms to enable multi-domain orchestration for a wide variety of network and other cyberinfrastructure resources**
 - Via Layer 2 and 3 point-to-point VPNs and multipoint network topologies
 - Customizable for individual domain science workflow systems
 - + a variety of interactive services allowing application workflows to ask open-ended questions about capabilities, negotiate with the networked infrastructure
 - Or request network services in a highly abstract and workflow-centric manner
- ★ **The Open Science Grid (OSG) provides a distributed petascale national facility where Consortium members provide guaranteed and opportunistic access to shared computing and storage resources across 100s of autonomous sites**
- ★ **Goal: SENSE providing key network-related capabilities to applications in OSG:**
 - ★ **Developing mechanisms for an application workflow to obtain information** regarding the network services, capabilities, and options; to a degree similar to what is possible for compute resources, is the primary motivation for this work.
 - ★ **Giving applications in OSG the ability to interact with the network:** to exchange information, negotiate performance parameters, discover expected performance metrics, and receive status/troubleshooting information in real time.
- ★ **Key pathways: (1) Interfacing and interaction with RUCIO, FTS and XRootD data federation and storage systems; engagement with those development teams**
(2) Enabled by the ESnet6 plan with – up to 75% of link capacity available for policy-driven dynamic provisioning and management by 2027
- ★ **4 Phase 1 Year Schedule:** from system evaluation and design, to prototype deployments and tests, to a plan for the transition to operations and the additional R&D needed, by mid-2022

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

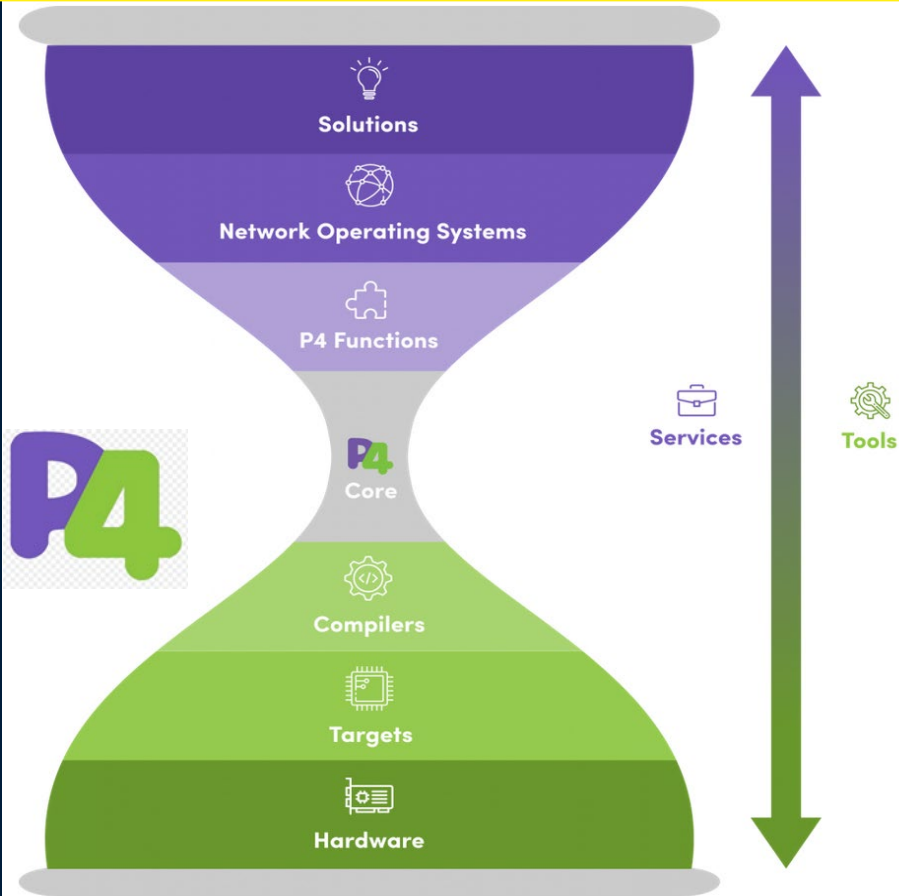


P4.org Open Source Network Programming Ecosystem



- “Application developers and network engineers can now use P4 to implement specific behavior in the network. Changes can be made in minutes instead of years.”

A large and growing P4 Ecosystem of P4-related products, projects, services



See Marcos Schwarz talk on Operationalizing Programmable Networks

P4 Workflow

- Programs and compilers are target-specific; Target can be hardware-based (FPGA, Programmable ASICs) or software (on x86 CPU, DPU ...)
- Program (prog.p4) classifies packets by header and the actions to take on incoming packets (e.g., forward, drop, insert, *other*)
- A P4 compiler generates the runtime mapping metadata to allow the control and data planes to communicate using P4Runtime (prog.p4info).
- A P4 compiler also generates an executable for the target data plane (target_prog.bin), specifying the header formats and corresponding actions for the target device

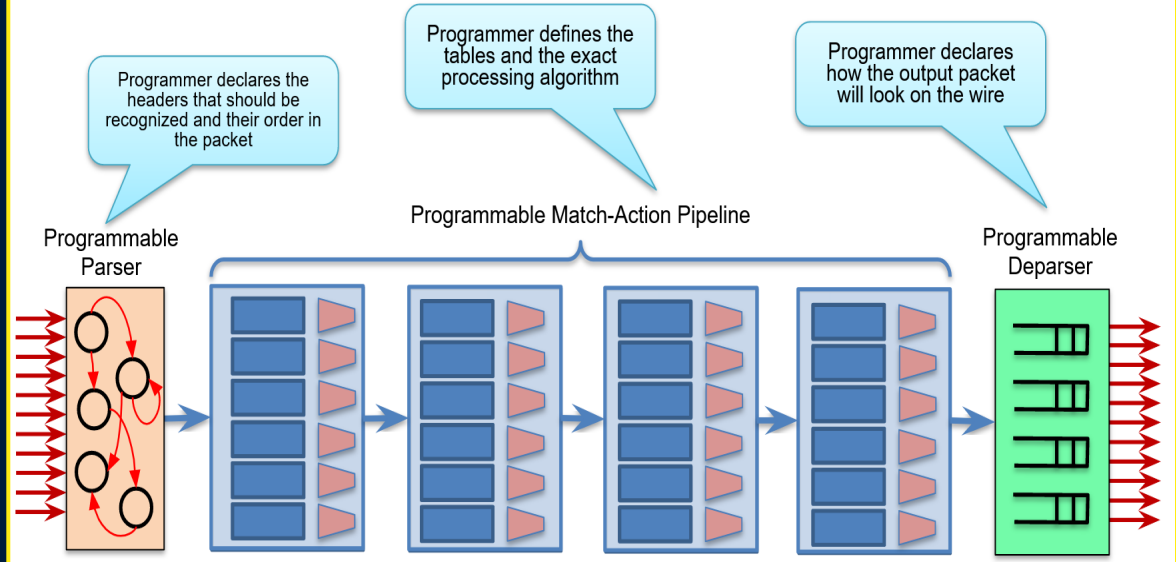
RARE

- For Example: **R**OUTER FOR **A**CADEMIA **R**ESearch & **E**duCation
- GEANT RARE/freeRtr is a software routing platform with a modular design that uses a message-based API between the control plane and data plane. RARE is powered by the freeRtr control plane and interfaces to multiple data planes such as P4 BMv2, Intel Tofino, DPDK.

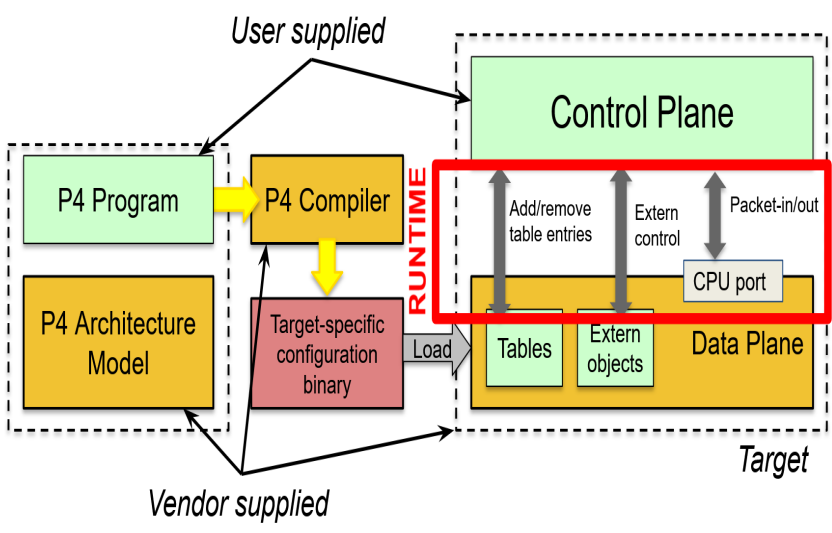
PISA: Protocol Independent Switch Architecture

Flexible, Stateful Packet Handling

- ★ Packet is parsed into individual headers (parsed representation)
- ★ Headers and intermediate results can be used for matching and actions
- ★ Headers can be modified, added or removed
- Packet is deparsed (serialized)



Programming a P4 Target



Tutorials: <https://github.com/p4lang/tutorials>

- Basic forwarding and tunneling
- P4 Runtime and the control plane
- Monitoring and Debugging (ECN; Route Inspect)
- Advanced: INT, Source routing, Load balancing; QoS; Sub-RTT Coordination; In-Network Caching; NDP
- Stateful Packet Processing: Link Monitoring, Firewall
- Slides available here: https://docs.google.com/presentation/d/1zliBqsS8IOD4nQUboRRmF_19poeLLDLadD5zLzrTkVc/edit#slide=id.g37fca2850e_6_831
- Annual Tutorials at P4 Workshop (April or May); some at SIGCOMM



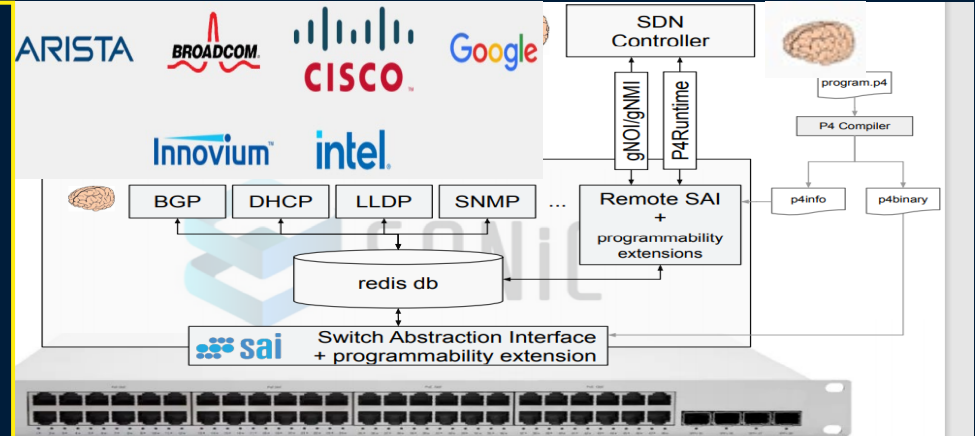
P4 Integrated Network Stack (PINS)



<https://opennetworking.org/pins/>

Network Architecture Evolution:

- Disaggregation of network stack + white box switches led to rise of Open Source NOS's
- Switch OS landscape became fragmented Stratum, SONiC, FBOSS, DANOS, DENT, ...
- While different open source communities have different use cases, they are often solving the same problems



Response: bring SDN capabilities to Open Source NOS

- (1) Remoted the Switch Hardware Abstraction Layer (HAL) under SDN Control
- (2) Added a remote Switch Abstraction Interface (SAI), with programmability extensions
- (3) Modeled the SAI in P4; exposed it in P4 Runtime

Key Design Decisions: Open Source

- Opt In: Existing SONiC use cases see no overhead/impact
- Mix & Match: Mix SDN with local control
- Familiar Interfaces: Reuse SAI, P4, P4Runtime, and gNMI/gNOI
- P4Runtime remotes SAI, not SONiC: Low Level interfaces give full flexibility to the SDN controller

SAI Target Architecture: a P4 parser, deparser and 4 programmable pipelines [Green boxes],



in between fixed pipelines [Black Boxes]

8 High Level Design Documents: https://github.com/pins/SONiC/blob/pins-hld/doc/pins/pins_hld.md

Targeting November 2021 Release

Marcos Schwarz talk

Beyond Programmability Alone: A Systems Approach

Reservoir Labs Gradient Graph (G2) Analytics



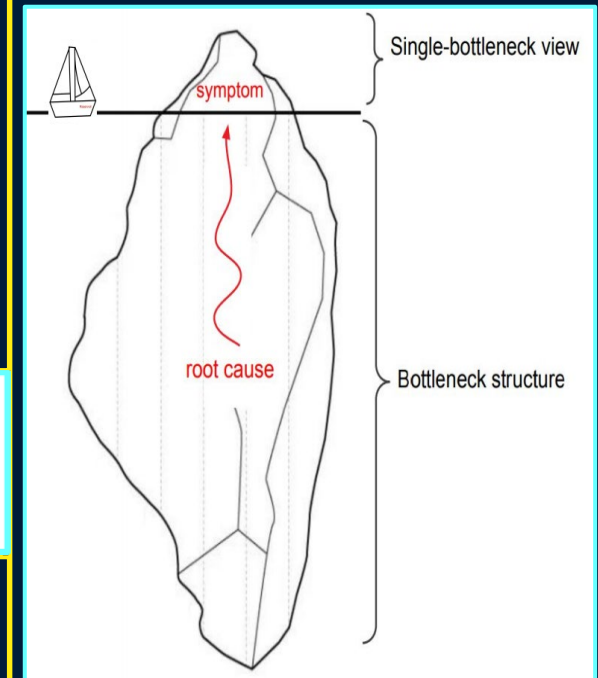
- **Objective:** Flow performance optimization in high speed networks, with fairness
- **Approach:** Built on a new mathematical Theory of **Bottleneck Structures** and an analytical framework
 - Enables operators to understand and precisely control flow and bottleneck performance
- **Value:** Improved **capacity planning, traffic engineering**
 - **Greater, more effective network throughput and stability as a function of capacity and cost**

• **Applications:** 5G Networks, artificial intelligence, large scale data centers (e.g., Google Jupiter), R&E Networks (e.g., DOE ESnet), cloud computing (e.g., AWS), SDN-WAN (e.g., Google B4), Supercomputers (e.g., DOE NERSC Cori), Telco networks, the Internet itself.

- "On the Bottleneck Structure of Congestion-Controlled Networks," ACM SIGMETRICS, Boston, June 2020 [<https://bit.ly/3eGOPrb>]
- "Designing Data Center Networks Using Bottleneck Structures," accepted for publication at ACM SIGCOMM 2021 [<https://bit.ly/2UZCb1M>]
- "Computing Bottleneck Structures at Scale for High-Precision Network Performance Analysis," SC 2020 INDIS, November 2020 [<https://bit.ly/3BriwaB>]
- "A Quantitative Theory of Bottleneck Structures for Data Networks", Technical Report, available upon request [<https://bit.ly/38u8ARs>]

www.reservoir.com/gradientgraph/

System Wide information to identify, deal with root causes



Key Components

- **Bottleneck precedence + flow gradient graphs**
- **Impactful flow and flow group ID**
- **Alternate path recommendations**

Operational Use Case: Scheduling of Deadline-Bound Data Transfers

Flow Gradient Graph:

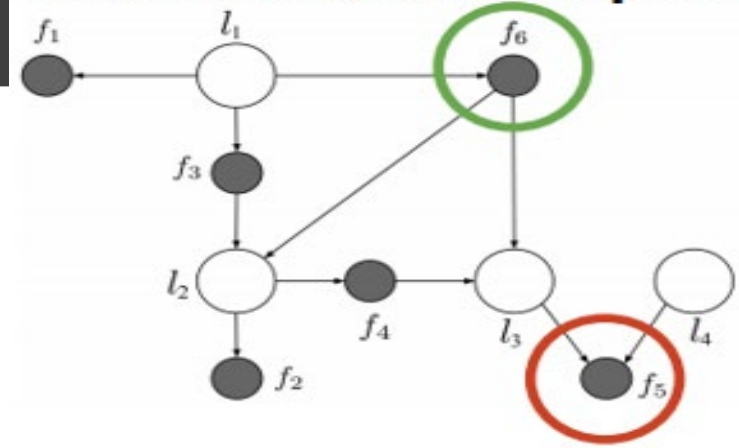


Table 3: As predicted by the theory of bottleneck ordering, flow f_6 is a significantly higher impact flow than flow f_5 .

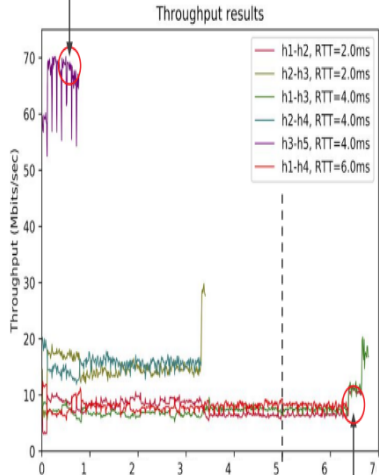
Comp. time (secs)	f_1	f_2	f_3	f_4	f_5	f_6	Slowest
With all flows	664	340	679	331	77	636	679
Without flow f_5	678	350	671	317	-	611	678
Without flow f_6	416	295	457	288	75	-	457

Avg rate (Mbps)	f_1	f_2	f_3	f_4	f_5	f_6	Total
With all flows	7.7	15.1	7.5	15.4	65.8	8.1	119.6
Without flow f_5	7.5	14.5	7.6	16.1	-	8.3	54
Without flow f_6	12.2	17.2	11.1	17.7	68.1	-	126.3

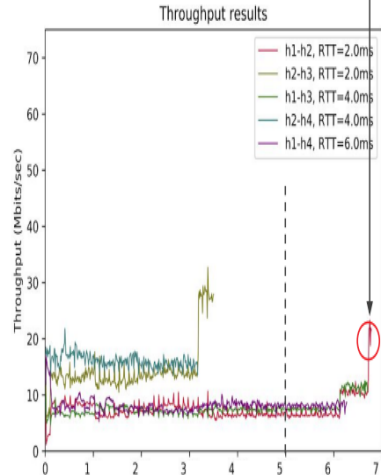


(2) Traditional approach: look at heavy hitters

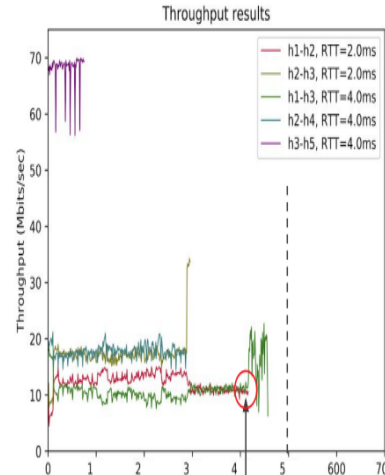
(3) Traditional approach yields no benefit



(a) Without removing any flow.



(b) Removing the heavy-hitter flow f_5 .



(c) Removing a low-hitter flow f_6 .

(1) Goal: deliver red flow (h1-h2) by 5 am, two hours ahead

(4) GradientGraph reveals the solution to meet the deadline-bound constraint

Future: additional decision factors: policy, priority, network and workflow state, cost and marginal benefit of operations, multiple SLAs by VO

P4 + Reservoir Labs + SENSE/AutoGOLE Use Case



“Laboratory use case” to start, using SENSE services, the PRP federated k8s clusters and the running Reservoir Labs G2 instance

- (1) Generate several long-lasting impactful flows;
Also generate background traffic as a set of many smaller flows
- (2) Create congestion on one or more segments
- (3) Identify via the RL G2 and other monitoring tools, the impactful flows, including the ones we created
- (4) Group (in one to three groups) the impactful flows
- (5) Use the Flow Gradient Graph (fgg) and other monitoring to get alternate path recommendations
- (6) Divert a flow group onto an alternate path
- (7) Validate that the impact of changing the path for an impactful flow-group is as predicted (or nearly)
- (8) After handling all the impactful flow groups, verify that the congestion has been relieved.

Near Future Following Steps

- (1) *Embed the 8-step sequence in an ongoing set of persistent operations, with*
 - Congestion detection
 - Impactful flow-group identification
 - Agile flow steering or moderation
 - Verification of congestion mitigation
 - Load balancing
- (2) *Subsequently*
 - Tune the sequence of steps and decision parameters
 - Develop + evaluate success metrics, with Multiple SLAs
 - Predict and optimize using machine learning

Self-Driving Network

Adaptive Routing (e.g., real-time data for routing decisions)

Learns to avoid congestion

congestion free -> loss free network

100% utilization

Proactive fault repair

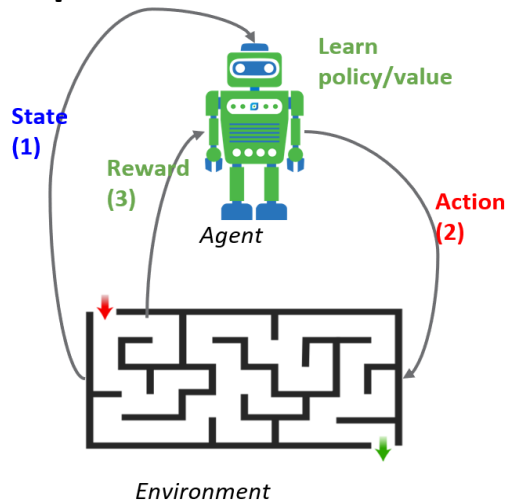
M. Kiran, C, Guok Talk at APAN2021

Kiran et al. Intelligent Networks DOE Project

Self-Driving Network for Science

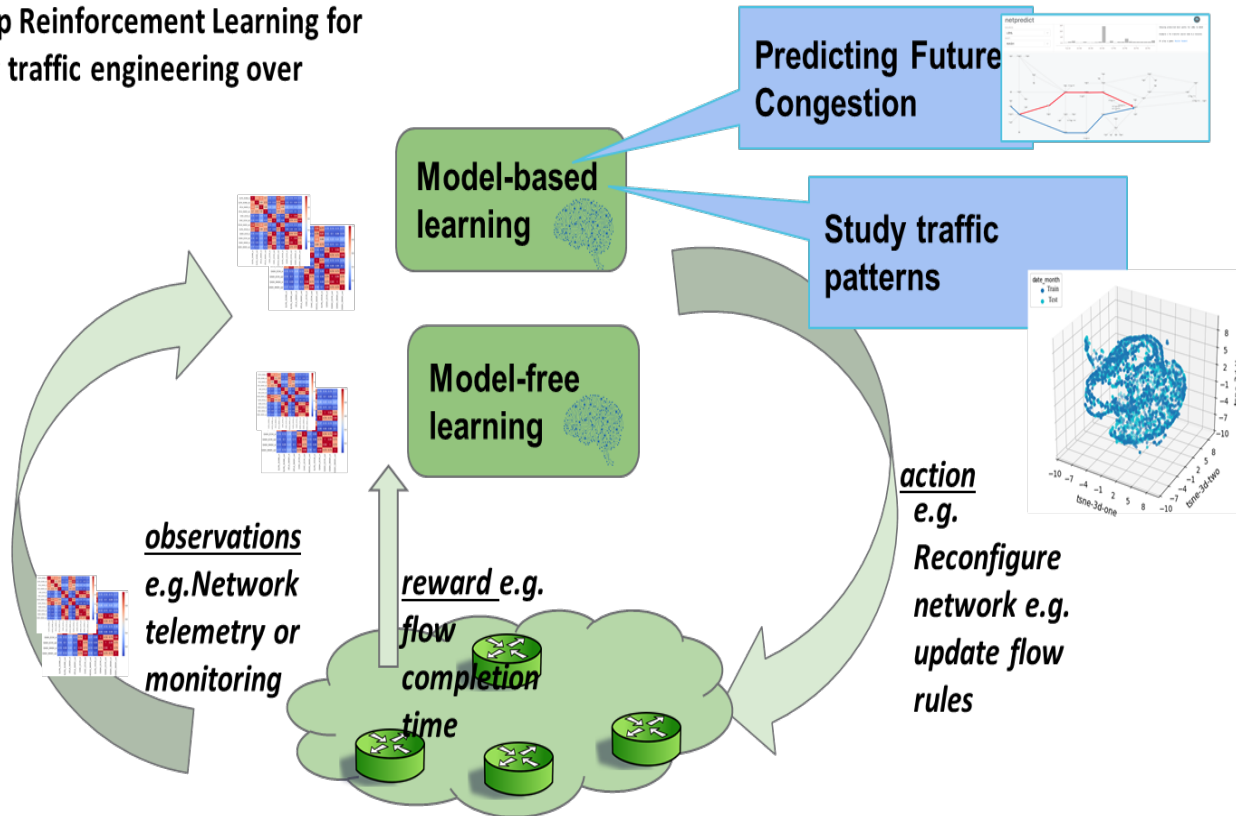
Using Deep Reinforcement Learning for optimizing traffic engineering over networks

Deep Reinforcement Learning



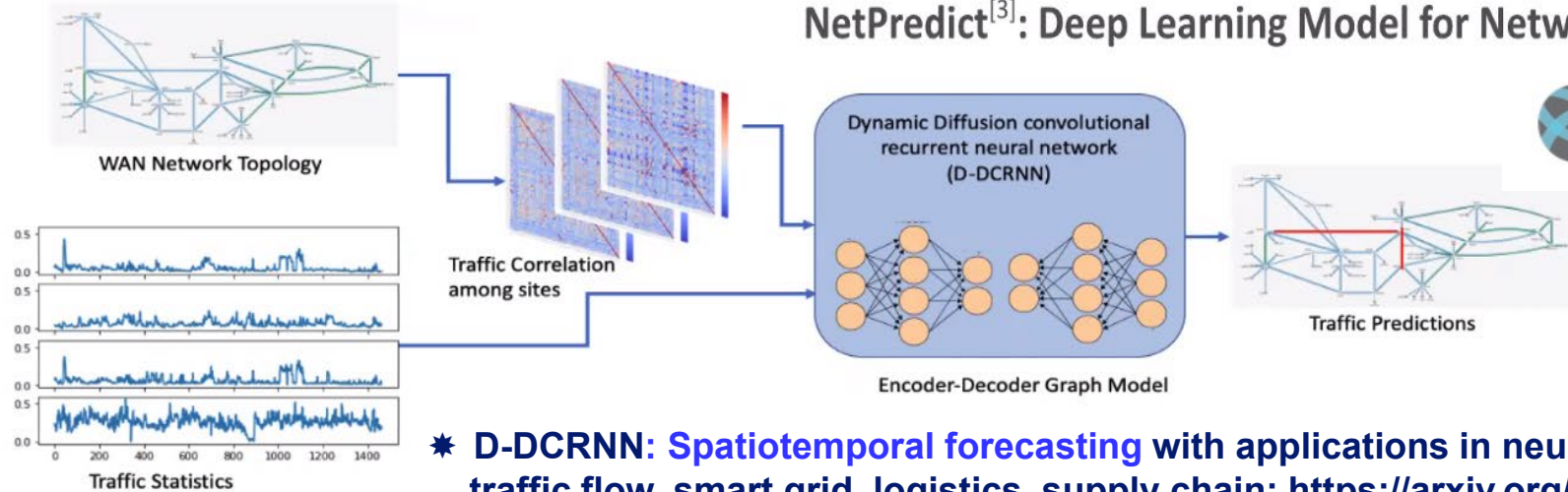
Case Studies:

1. **Model free:** Path selection for large data transfers: better load balancing
2. **Model Free:** Forwarding decisions for complex network topologies: Deep RL to learn optimal packet delivery policies vs. network load level
3. **Model Based:** Predicting network patterns with **Netpredict**



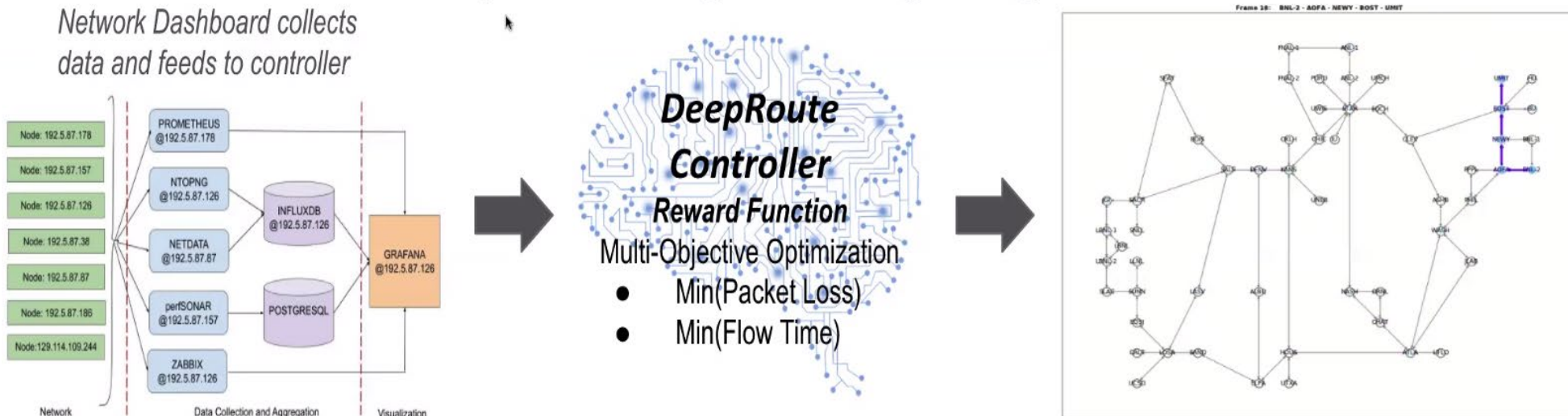
DAPHNE^[2] - Deep and Autonomous High Performance Networks

NetPredict^[3]: Deep Learning Model for Network Congestion



★ **D-DCRNN: Spatiotemporal forecasting with applications in neuroscience, climate, traffic flow, smart grid, logistics, supply chain: <https://arxiv.org/pdf/1707.01926.pdf>**

DeepRoute^[4]: Intelligent Traffic Engineering



★ **Possible Future: Hooks for a Richer, More Stateful Objective Function**

John Macauley Talk

■ Top Line Message

We are poised to make a milestone transition: to a new generation of intelligent, dynamic and adaptive software-driven networks

- ★ **Coordinating worldwide networks as a first class resource along with computing and storage, across multiple domains**
- ★ **Simultaneously supporting major DIS programs and the worldwide academic and research community**
- ★ **A global fabric dynamically & flexibly allocating and conserving resources**
- ★ **Building on and advancing key developments:** from regional caches/data lakes to intelligent control and data planes to ML optimization [E.g SENSE/AutoGOLE, NOTED, ESN Net HT, GEANT/RARE, AmLight, Bridges ...]
- ★ **Moving towards: A fully programmable network ecosystem** (e.g. P4, PINS), with **system level tools** (e.g. Reservoir Labs G2), workflow platforms (OSG, PRP) and **ML-based optimization** (e.g. DeepRoute)
 - ★ **With VO – network real-time interactions at the center**
- ★ **Now is the time for R&E networks to engage,** and join those already leading the transition and defining the next generation
- ★ **The GNA-G + its WGs, together with the GRP, AmRP and APRP are natural venues to enable this happen.**