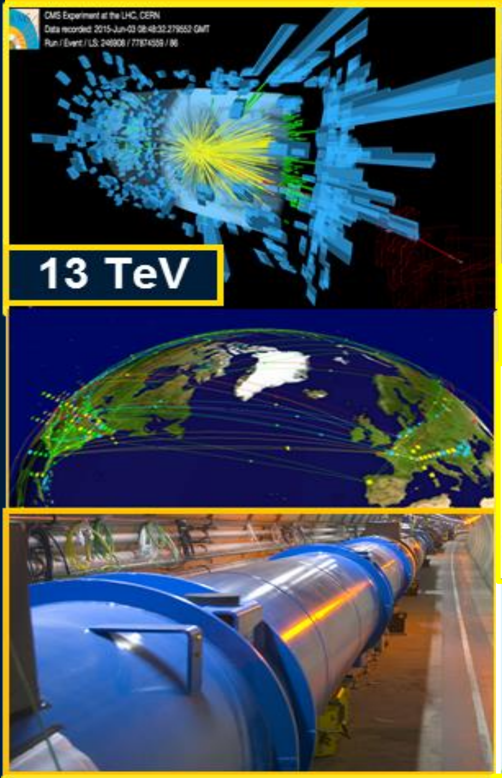
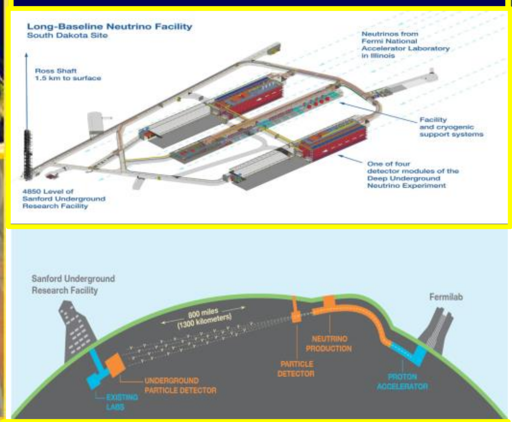
Data Center Requirements, Network-Integrated R&D

Towards a New Computing Model for the HL-LHC Era

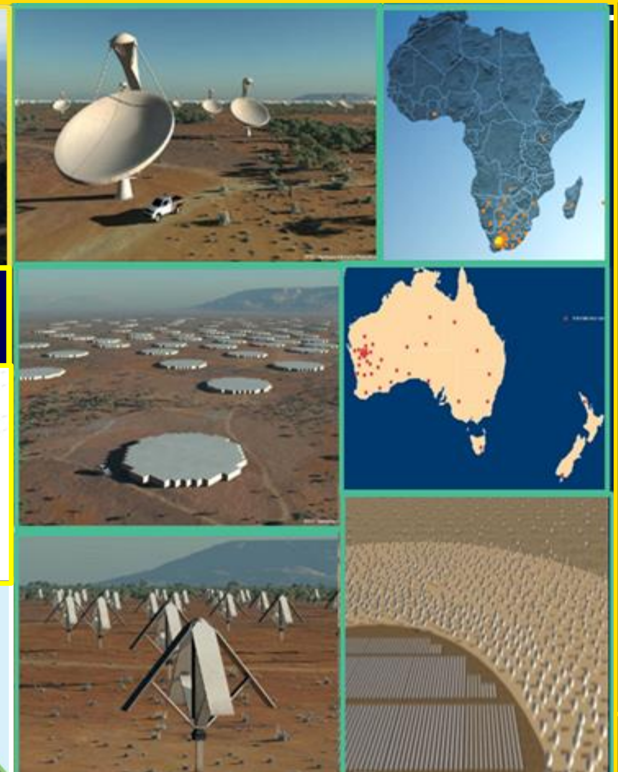


LSST



LHC

LBNF/DUNE



SKA

LHC Run3 and HL-LHC

DUNE

VRO SKA

BioInformatics

Earth Observation

Gateways to a New Era

GNA-G
Global Network Advancement Group

Harvey Newman, Caltech

Pre-GDB Data Center Architecture Workshop

June 8, 2021

California Institute of Technology 1891

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
- 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 requirements, and shortfall follows

Developing the Next Generation Computing Model

- **Top Line Message**

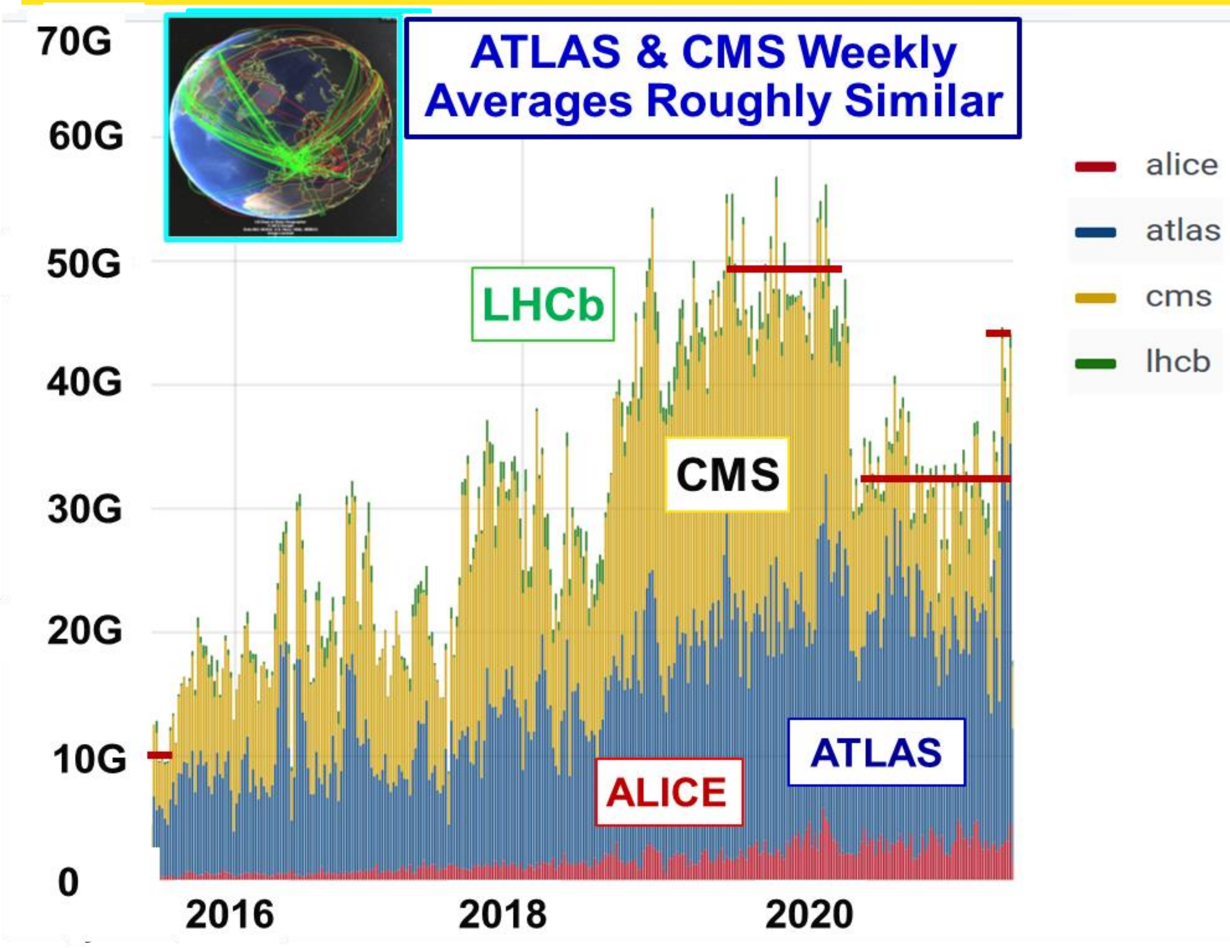
  A comprehensive R&D program to develop the architecture, design, prototyping, scaling and optimization of the HL-LHC Computing Model is required

  - A new system coordinating worldwide networks as a first class resource along with computing and storage
  - 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]
  - Moving towards fully programmable networks (e.g. P4, PINS), system level tools (e.g. Reservoir Labs G2) and ML-based optimization. Site – network real-time interactions are a key part
  - Leveraging regional network developments to form a worldwide fabric supporting OSG/HEP workflow

- The LHC experiments, the GNA-G and the R&E Network community should jointly consider how such an effort should be organized and implemented, to accomplish the paradigm shift by ~2027

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

WLCG Transfers Dashboard: Throughput May 2015 – May 2021



10-58 GBytes/s Week Avg
To 70+ GBytes/s Daily Avg

Complex Workflow

- **~1M jobs (threads) simultaneously**
- **Multi-TByte to Petabyte Transfers;**
- **To ~10 M File Transfers/Day**
- **100ks of remote connections**

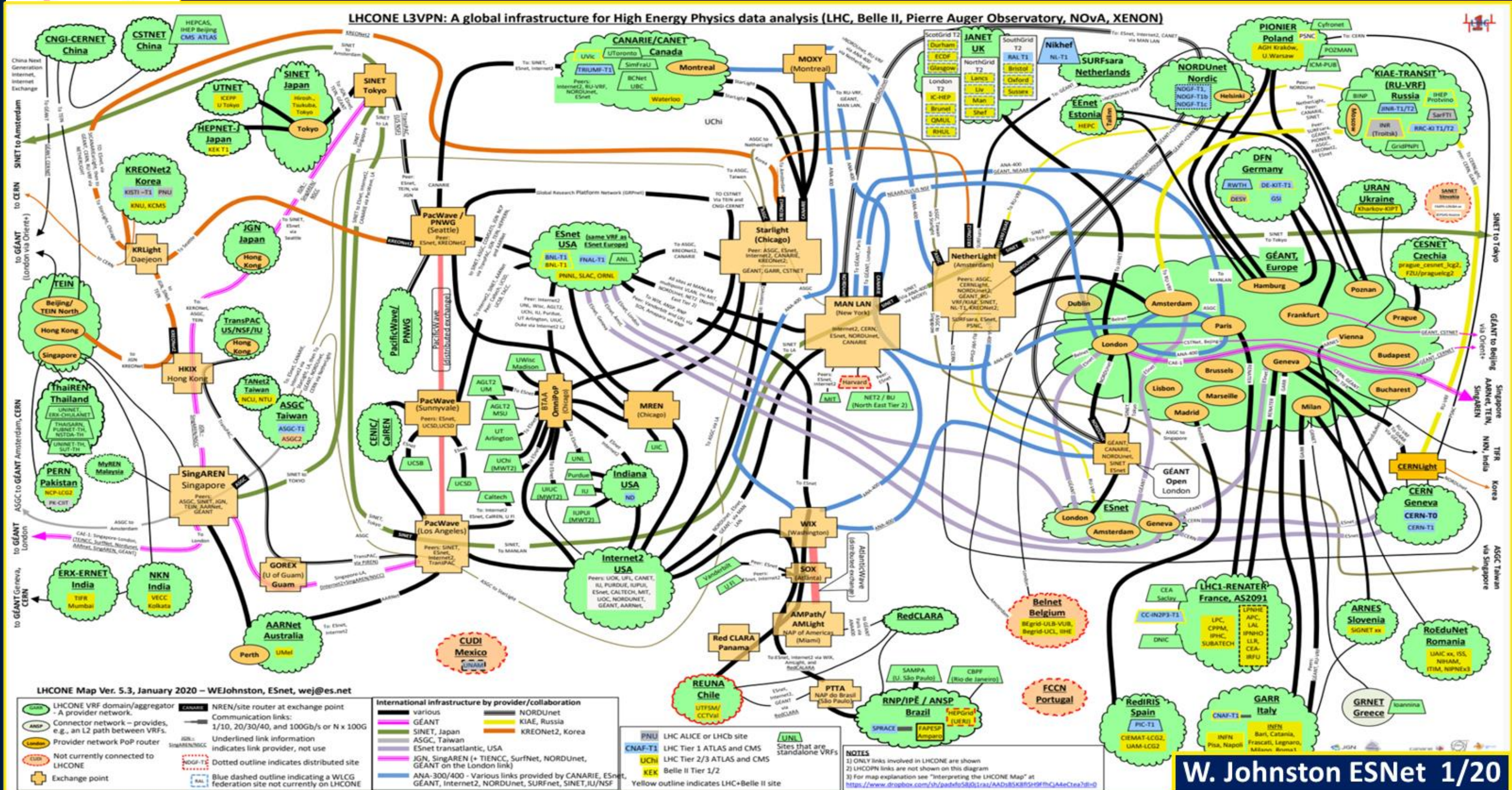
- **The effects of Covid are evident**
- **The recovery is emerging: warrants careful watching**

5X Growth in Throughput in 2016-2020: +50%/Yr; ~60X per Decade

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

LHCONE VRF: The Challenge of Complexity and Global Reach

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



W. Johnston ESNet 1/20

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

Challenge: A complex system with limited scaling properties.

Response: New Mode of Sharing ? Multi-One ?

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

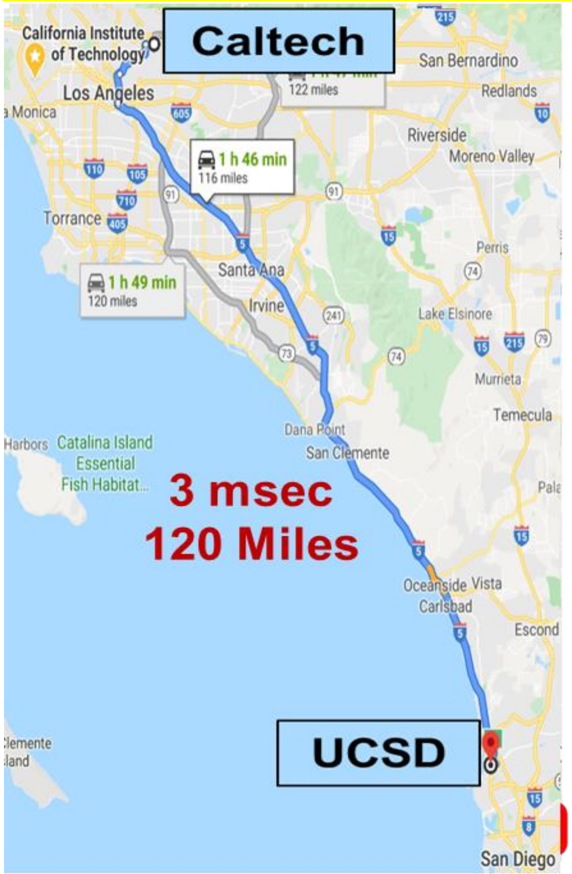
- Export of Raw Data from CERN to the Tier1s (350 Pbytes/Year):
  - 400 Gbps *Flat* each for ATLAS and CMS; +100G each for other data formats; +100 G each for ALICE, LHCb
- "Minimal" Scenario [*]: Network Infrastructure from CERN to Tier1s Required
  - 4.8 Tbps Aggregate: Includes 1.2 Tbps Flat (24 X 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

(Southern) California ((So)Cal) Cache

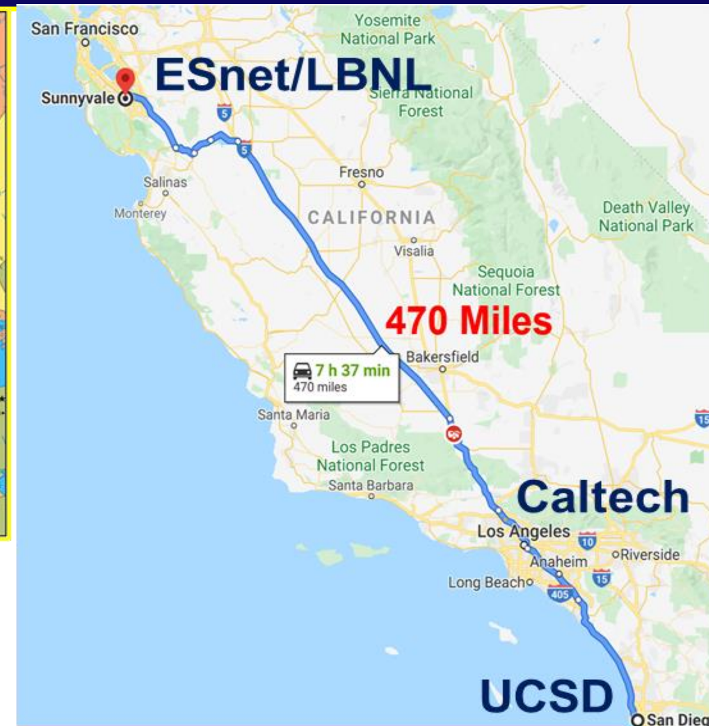
Roughly 20,000 cores across Caltech & UCSD … half typically used for analysis
A 1.5 Pbyte Working Example in Production



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.

ESnet plan to install additional in-network caches near US Tier2s in 2021

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

Developing the Next Computing Model

Trends and Key Elements Affecting Sites and Networks

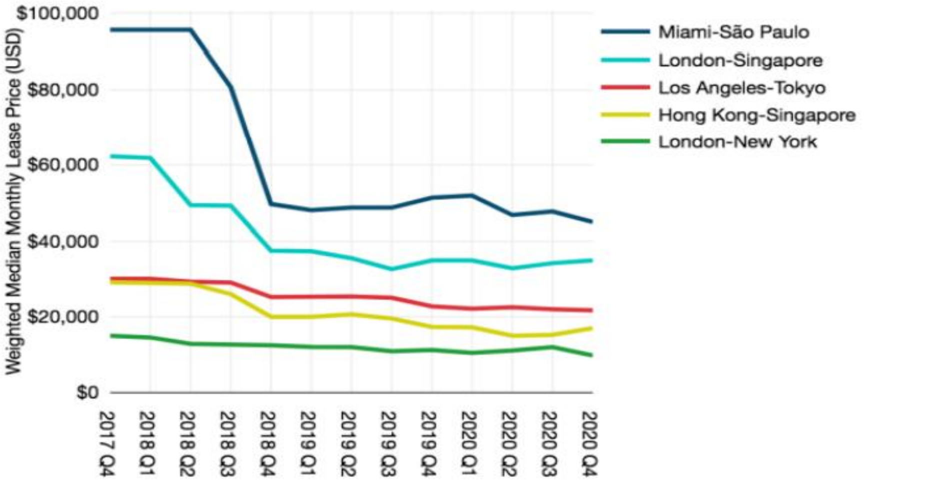
- Tier2 Storage: ~5 - 12 Pbytes (usable) now to ~20-40 Pbytes (?) (with erasure coding) by 2028
- Data Lake Model: from ~1 Pbyte now to 5-10 PByte Working Sets in Caches
- Typical Routine Network Transaction: Petabyte transferred in a shift; requires a 400G link to a Tier2, with heavy use for hours at a time.
  - We need to alert and work with campuses starting now, and plan for evolution/ upgrades starting in 2021-22, to be ready by ~2027
- We need to deploy & develop front end SSD caches (1-3 DWPD), with capacities from ~50 Tbytes now to ~1 Pbyte when affordable
- Wide networks: Are now just starting to move to 400G backbones (it has been nearly 10 years since 100G was first widely deployed)
- The required Tbps links to Tier1s and 400G to Tier2s will be a challenge, also at the start of HL LHC
- Transoceanic networks are a particular challenge due to pricing:
- Reduction only -10% CAGR on mature routes: NYC – London, LA – Tokyo; Equivalent to only a 2X price decrease by 2028.
- According to recent requirements reviews (e.g. ESnet study)

Shortfall may be 2-4X.

International Bandwidth Pricing Trends

Executive Summary (telegeography.com)

Weighted Median 100 Gbps Wavelength Price Trends on Major International Routes

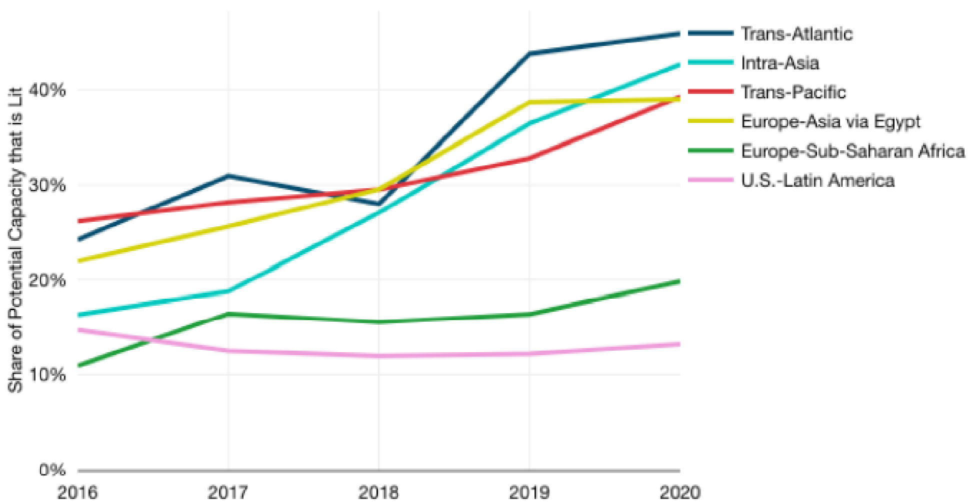


Notes: Each line represents the weighted median monthly lease price for an unprotected 100 Gbps Wavelength on the listed route. Prices are in USD and exclude local access and installation fees.

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



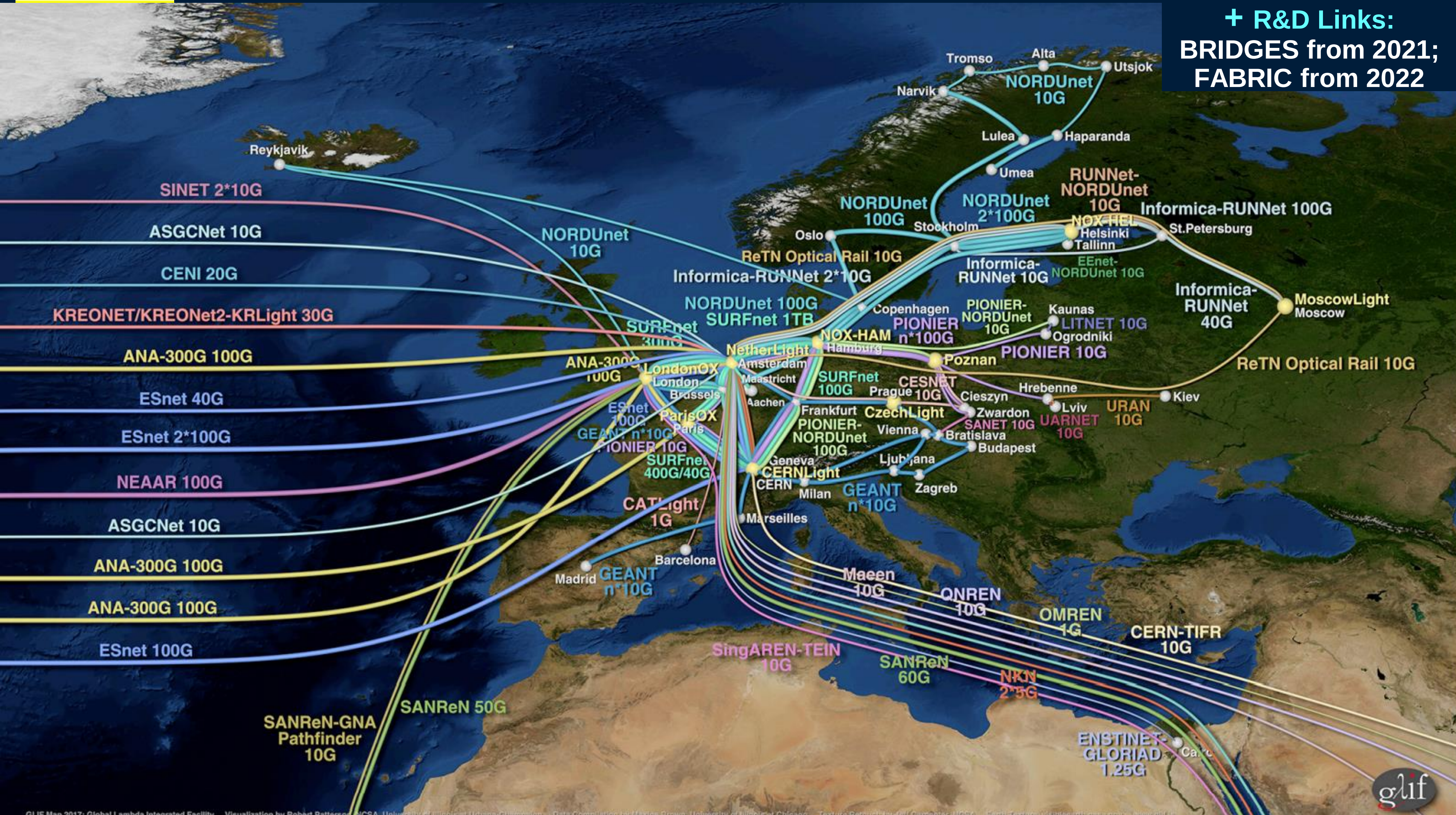
- **Price Evolution 2017-20**
  - **-16% Price CAGR Average**
  - **Only -10 to -13 % CAGR LA-Tokyo and NYC – London**
  - **To -6% 2019-20 due to COVID**
  - **100G/10G Price Multiple: 4.3X, Down from 6.4X in 2015**
    - **Below 4X NYC-London**

Developing the Next Computing Model
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

Shared Network Infrastructure: GLIF Map (2017)

+ R&D Links:
BRIDGES from 2021;
FABRIC from 2022



Slow Growth in Capacity at Fixed Cost: ~2 Tbps TA by 2028

Sharing with the larger academic & research community on several continents

Next Computing Model Outlook

Technology Push: Rising Network Capabilities of Servers + Storage

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

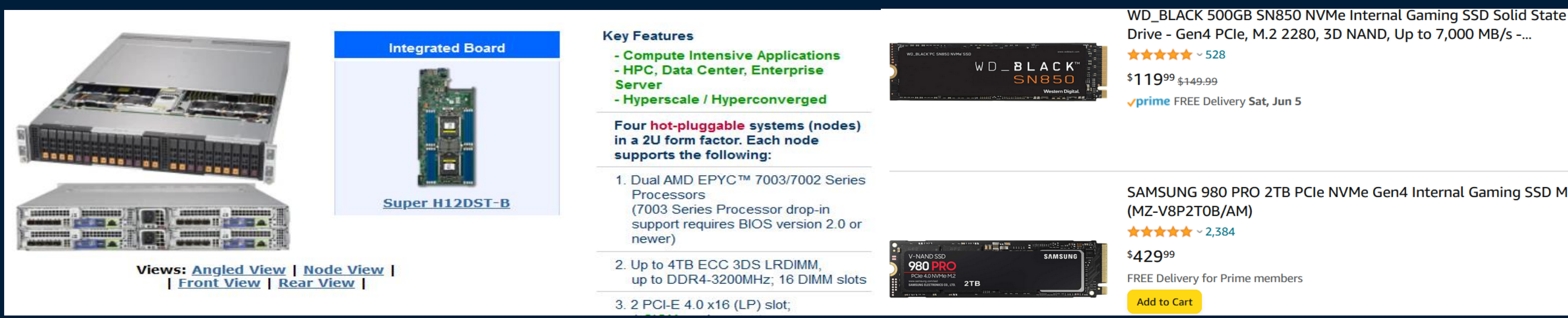
- Commoditization of 200G NICs and 200-400G Switches is well underway

Tofino

Fully P4 Programmable

Tofino2 (25.6 Tbps)

- 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

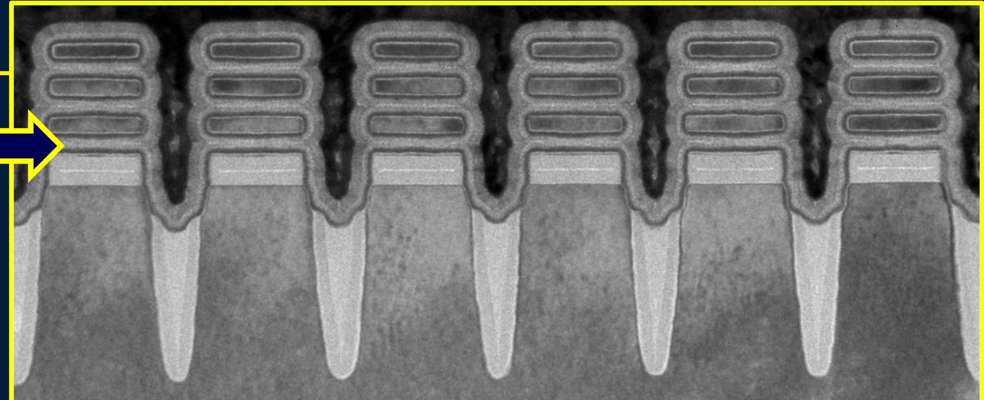


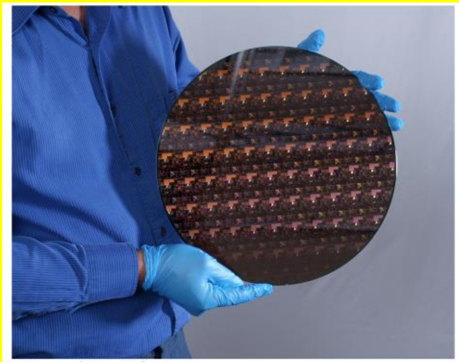
- NOTE: PCIe Standards Clock Now 2 Years: Products: PCIe 5.0 by ~2023; PCIe 6.0 by ~2025; ~2X performance per generation; Multi-Tbps servers possible by HL LHC
- Paralleled/driven by motherboard, chip architecture and interconnect improvement

IBM Research (Albany): First 2 Nanometer Chip Technology

2nm smaller than a DNA strand

- Nanosheet based design
- 50 billion transistors on a chip
- 2nm Relative to 7nm:
  - + 45% in performance, or
  - 75% lower power use
- Application target examples:
- 4X cell phone battery life
- Reducing data center carbon footprint
- Drastically speeding up laptop functions
- Faster object detection + reaction time in autonomous vehicles





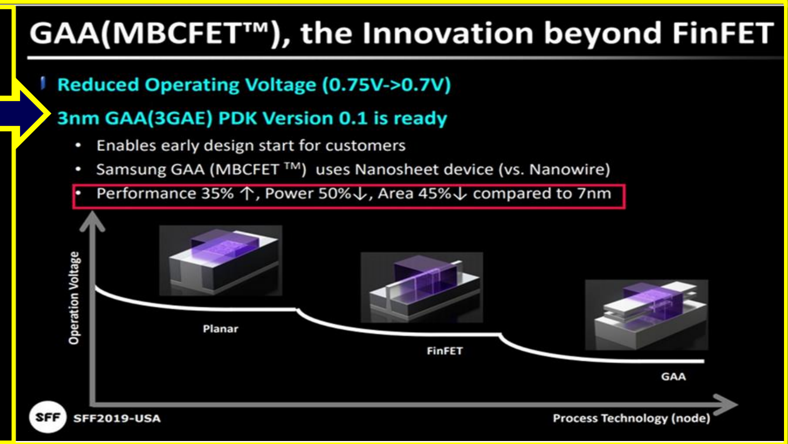
IBM Research 2 nm Wafer

A 2 nm wafer fabricated at IBM Research's Albany facility. The wafer contains hundreds of individual chips. Courtesy of IBM

IBM Research 2 nm

A close-up of a 2 nm wafer fabricated at IBM Research's Albany facility, with individual chips visible to the naked eye. Courtesy of IBM.

Before: Samsung Foundry Forum 2019 Outlook



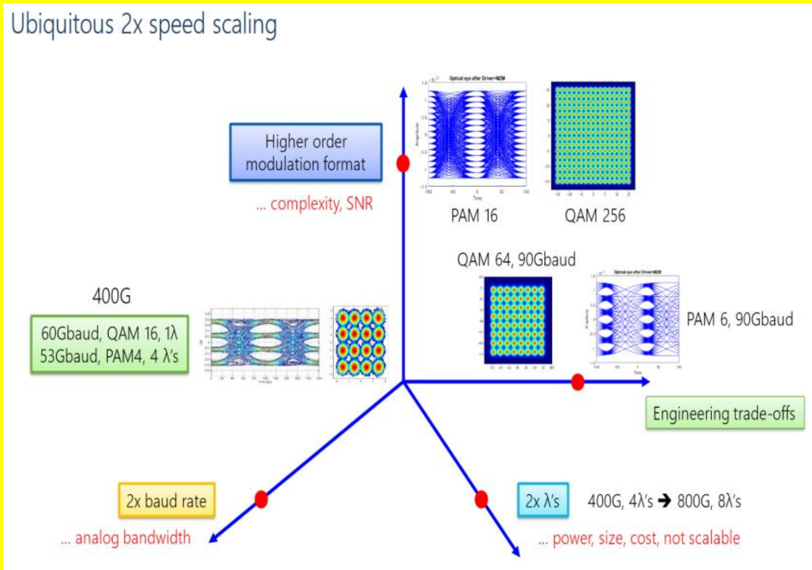
https://newsroom.ibm.com/2021-05-06-IBM-Unveils-Worlds-First-2-Nanometer-Chip-Technology,-Opening-a-New-Frontier-for-Semiconductors

Technology Push: Data Center, Metro, Long Haul

Interconnects: 400G Long Haul + "The Race to 800G"

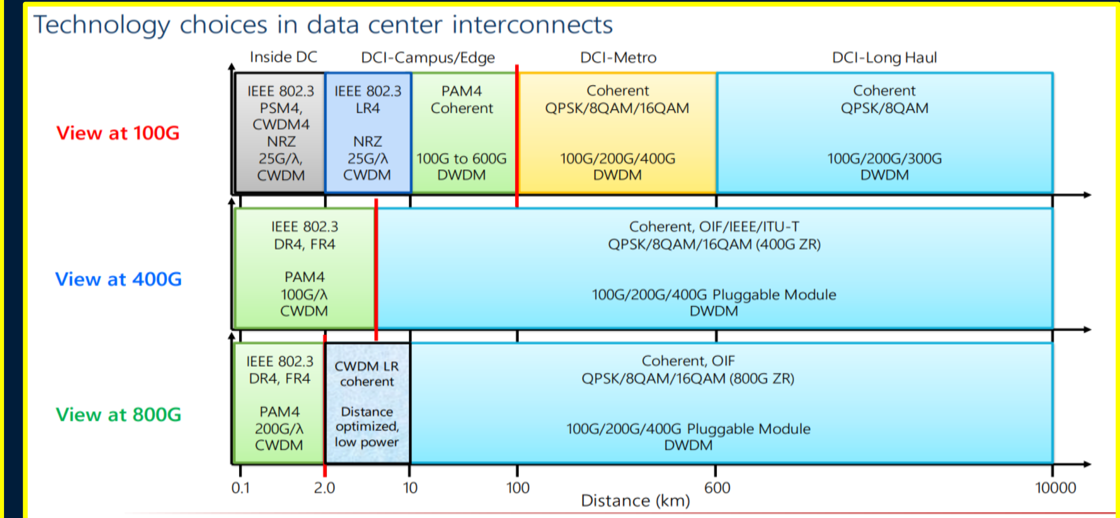
https://www.inphi.com/wp-content/uploads/2021/01/20210113_COBO_RNagarajan_Inphi_v3_distri.pdf

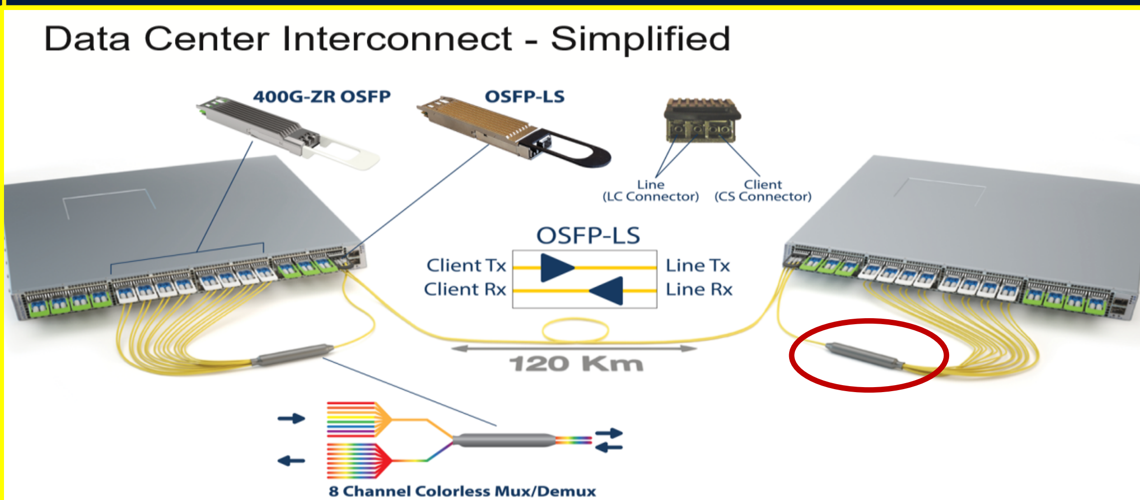
- **New Modulation schemes**



Technology Choices over Distances:

Modulations, Coherent, WDM with 100, 200G channels





Emerging Already in 2021-22:

PluggableTransceiver/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

https://www.dropbox.com/s/hn8c3wjm6mkhwqv/Arista_Optics_Intro_and_Roadmap032021f.pptx?dl=0

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

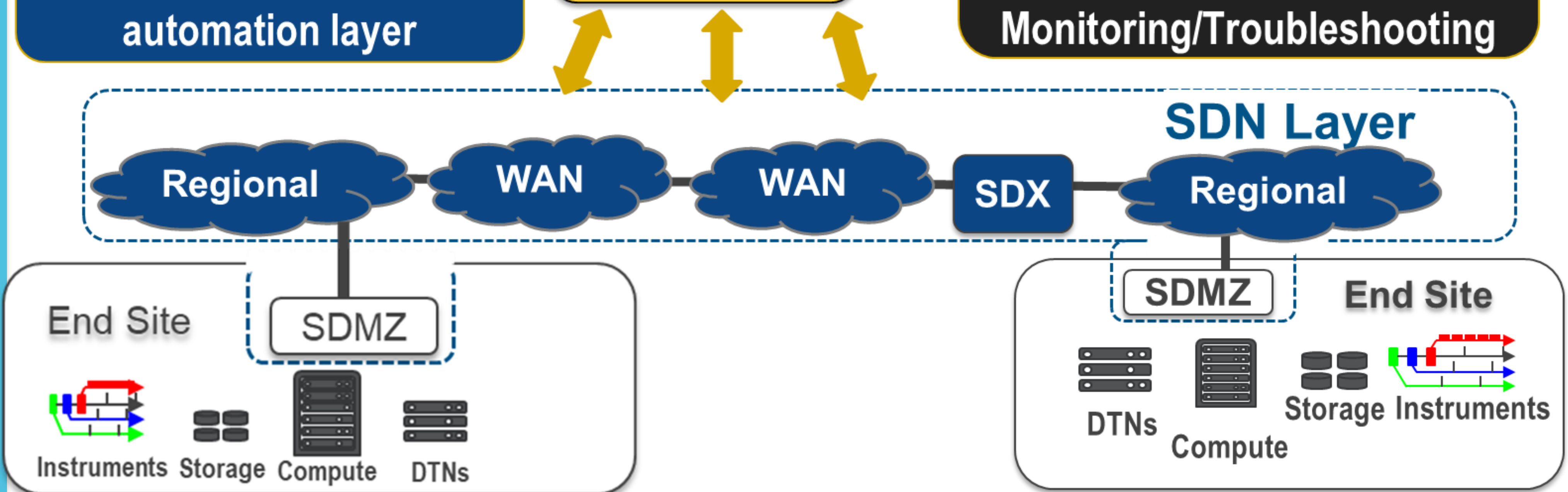
SENSE Native RMs are Available if no current automation layer

Application Workflow Agents

SENSE

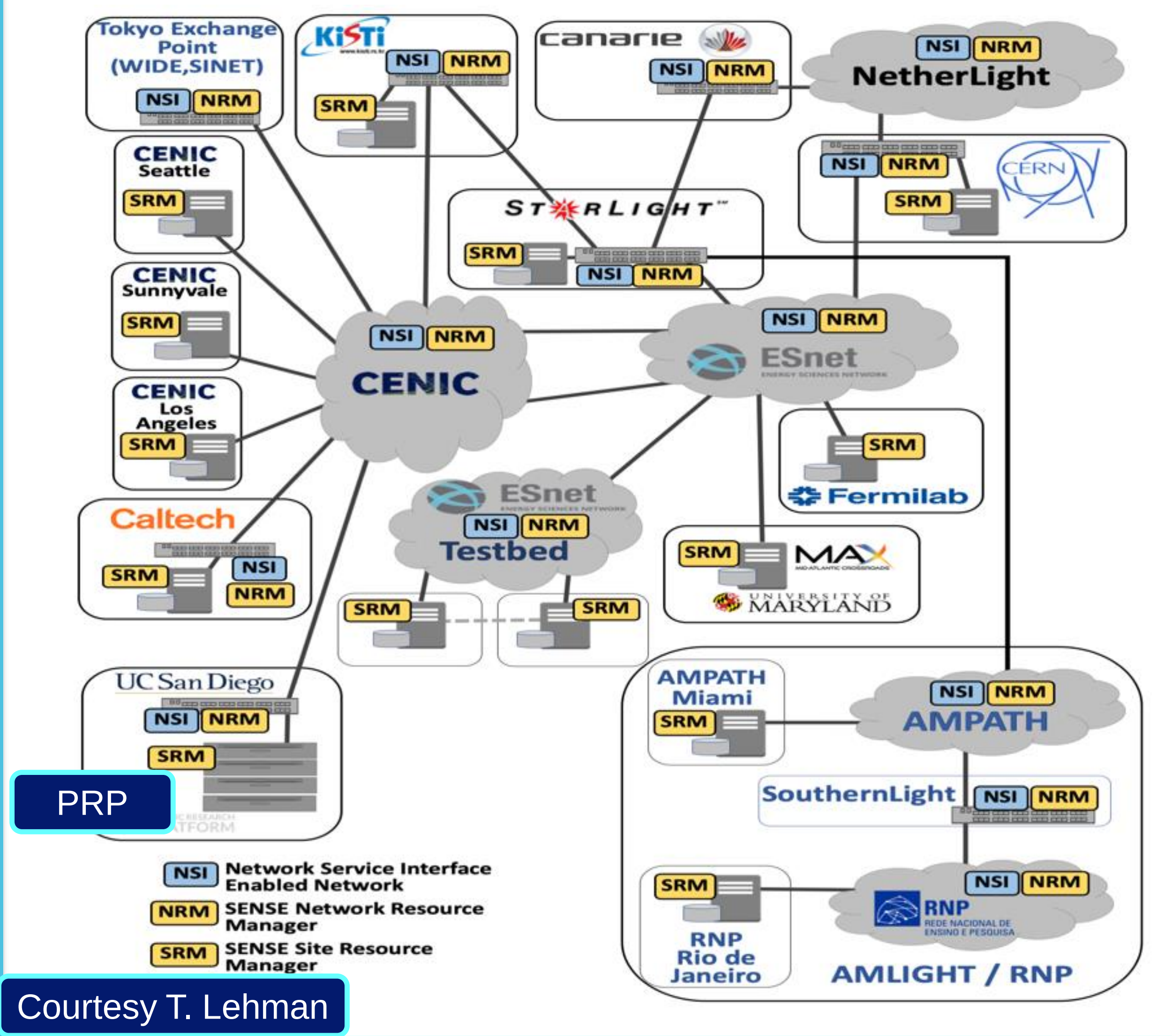
SENSE operates between the SDN Layer controlling the individual networks/end-sites, and science workflow agents/middleware

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, FIU, CERN, Fermilab, UMd, DE-KIT



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

Persistent Operations: ***Beginning this Quarter***

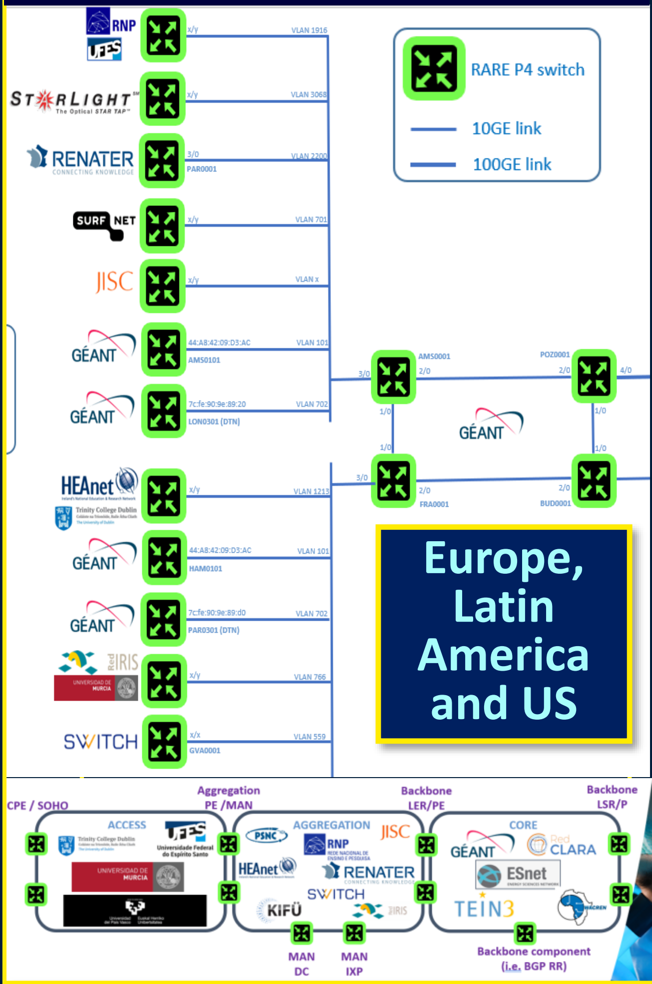
R&D on Network Capabilities: Key Technologies Towards an Intelligent Data Plane Using P4

- **Overlay Networks based on Virtual Circuits across multiple domains:** SENSE and its Orchestrator, Network & Site RMs
  - **Allows emerging paradigms (SENSE, P4 programmable networks, NDN)** to co-exist with traditional networks, migrate into production
- **Programmable (P4-based) production switches: Tofino, Tofino2, Mellanox Spectrum2 and -3**
- **Network telemetry:** precision timestamps, classification of sets of flows, services to handle flows by class
  - **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
- **E.g.** RARE Freertr in **GEANT: Both production-ready open images in inexpensive switches;** and **fully programmable images** for the academic and research community. Also SmartNICs (e.g. Bluefield2), Xilinx accelerators

RARE — GEANT Project

Router for Academia Research & Education

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



Europe, Latin America and US

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

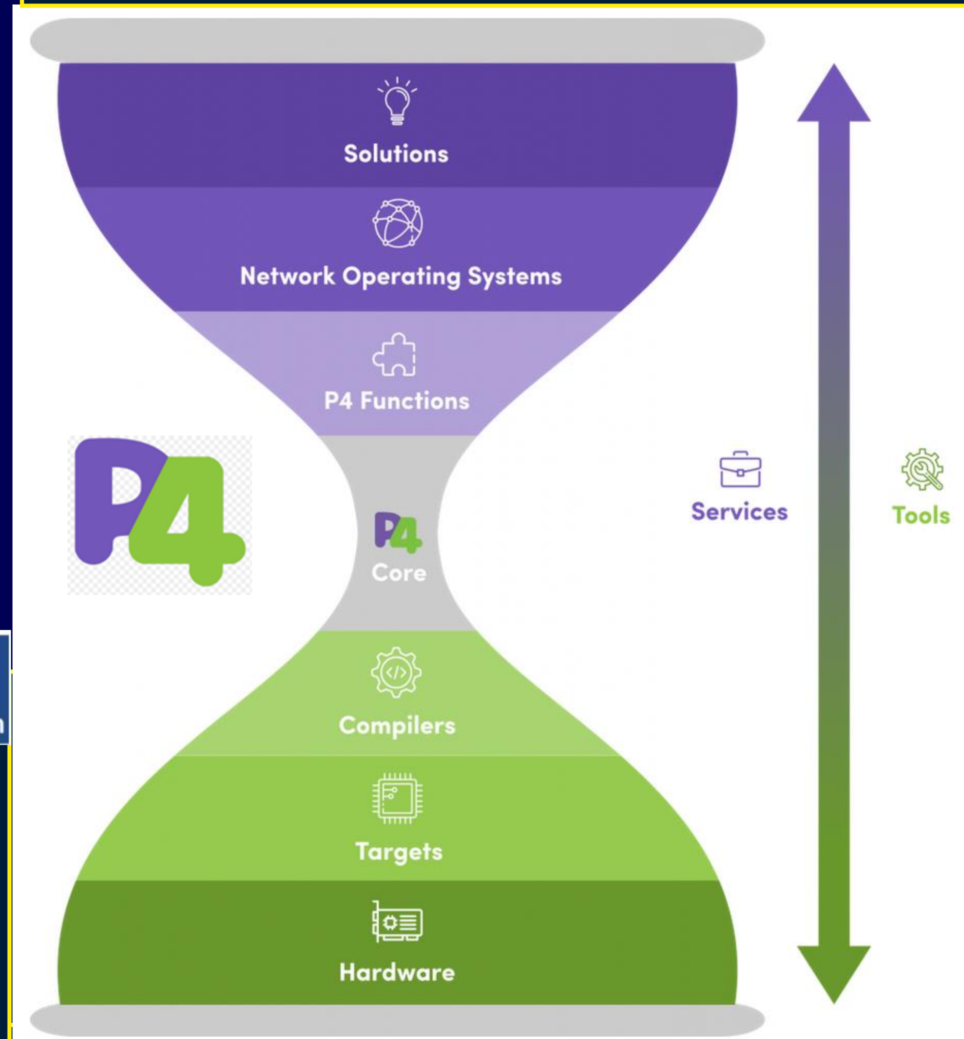
P4.org Open Source Network Programming Ecosystem

- "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."

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

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



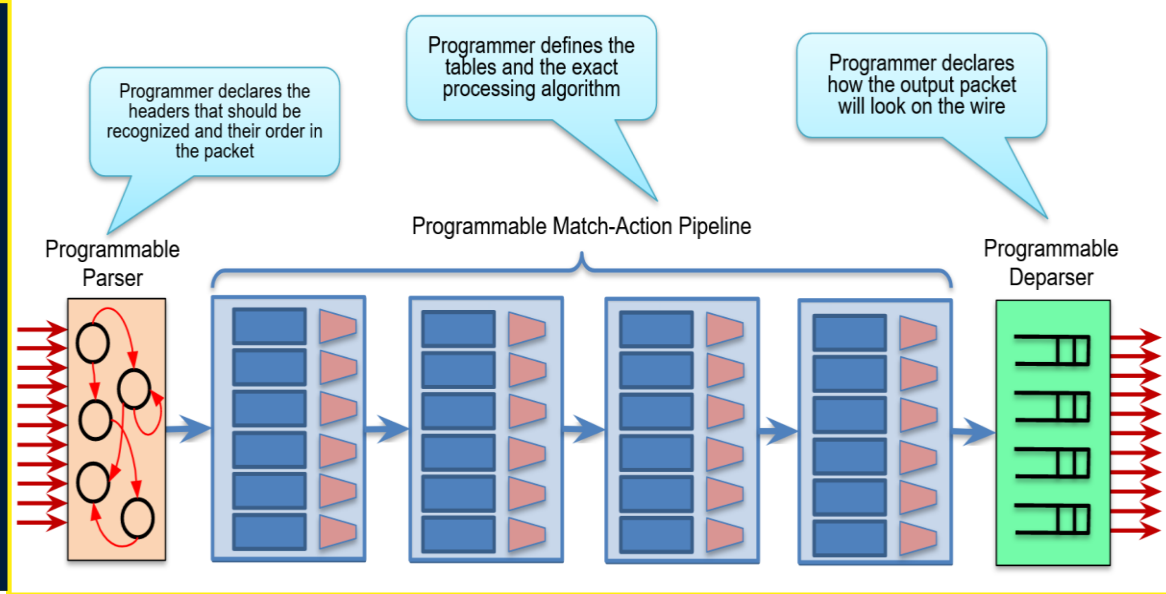
- For Example: RARE — Router for Academia Research & Education
- **GEANT RARE/freeRtr** is a software routing platform with a modular design that uses a message-based API between the control plane and data plane. **RARE** is powered by the **freeRtr control plane and interfaces to multiple data planes** such as P4 BMv2, Intel Tofino, DPDK.

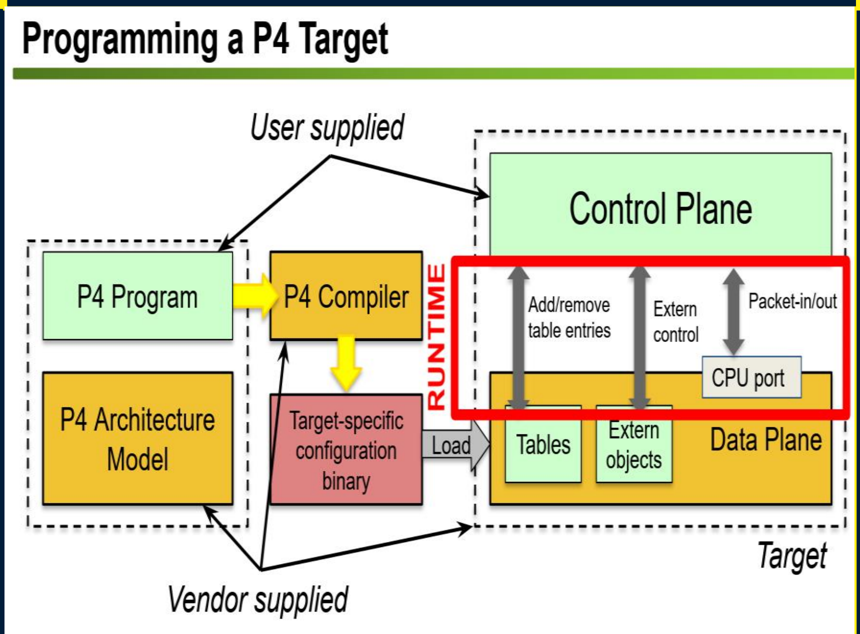
Elements, Tutorials

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 **departed (serialized)**





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

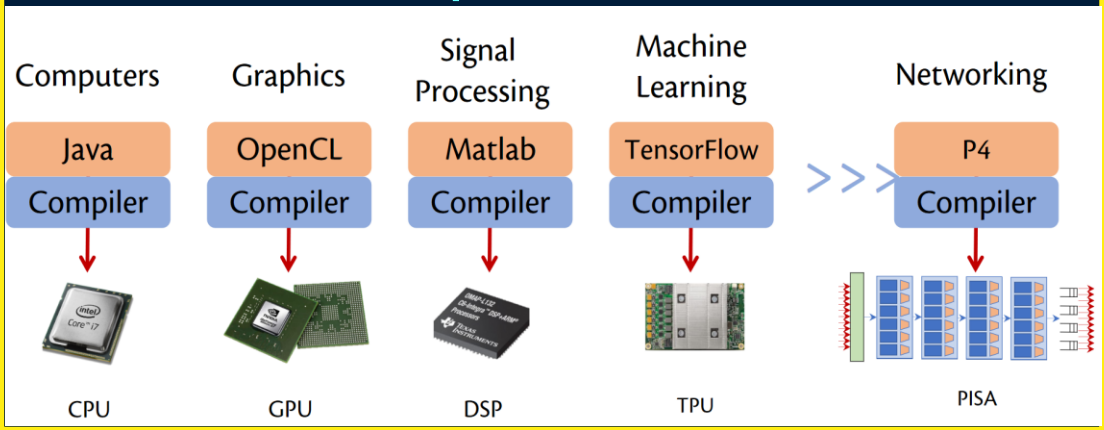
- **Basic forwarding** and **tunneling**
- **P4 Runtime** and **the control plane**
- **Monitoring** and **Debugging (ECN; Route Inspect)**
- Advanced: **INT, Source routing, Load balancing; QoS; Sub-RTT Coordination; In-Network Caching; NDP**
- Stateful Packet Processing: **Link Monitoring, Firewall**
- **Slides available here:** https://docs.google.com/presentation/d/1zliBgsS8lOD4nQUboRRmF_19poeLLDLadD5zLzrTkVc/edit#slide=id.g37fca2850e_6_831
- Annual Tutorials at **P4 Workshop (April or May)**; some at **SIGCOMM**

P4 2021 Workshop: May 2021

https://opennetworking.org/2021-p4-workshop-content/

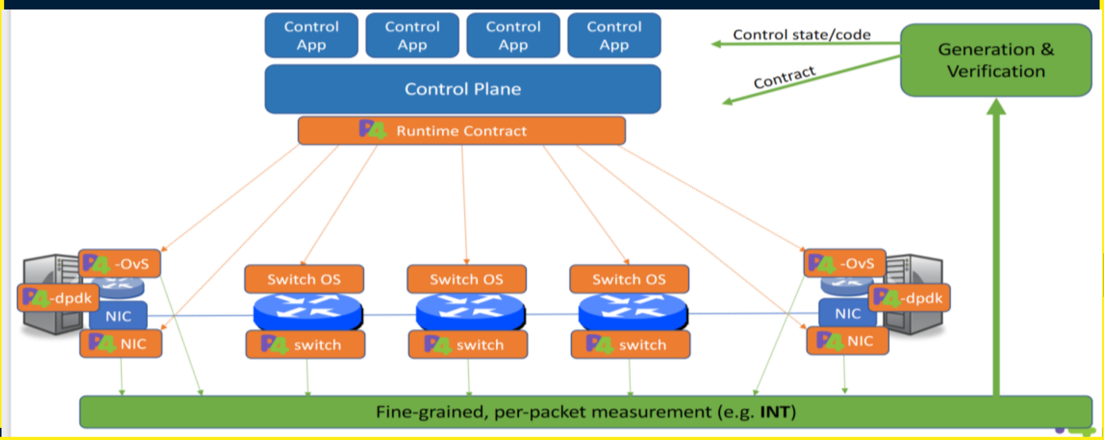
- **Videos** and **Slides:** Keynotes, Invited, Technical, Demo Talks, Tutorials
- P4: Language, Targets, Use Cases

Domain Specific Processors



Deep Programmability: Across the Ecosystem

from Switch to Smart NIC to FPGA to Host (OVS, dpdk)



P4 State in 2021

New Features

- Continued evolution of $P4_{16}$ Language, P4Runtime, and P4 architectures (PSA, PNA, etc.)
- Open-source developers contributing to a growing set of software targets and tools

New Targets

- User-space (e.g., p4-dpdk)
- Kernel networking (e.g., P4-OvS)
- FPGAs and SmartNICs (multiple vendors)

New Applications

- Hardware offloads
- Congestion control
- Security

P4 at Intel: Also **NFV, Middlebox, CEPH Storage Interface** etc.

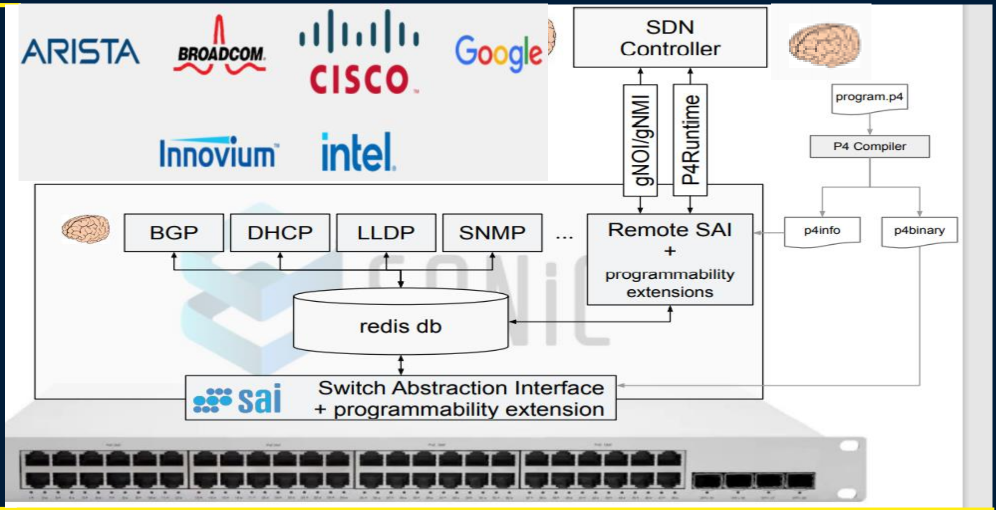
P4 Workshop Keynote: Nate Foster (Cornell)

P4 Integrated Network Stack (PINS)

https://opennetworking.org/pins/
https://opennetworking.org/wp-content/uploads/2021/05/P4-WS-RamanWeitz.pdf

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

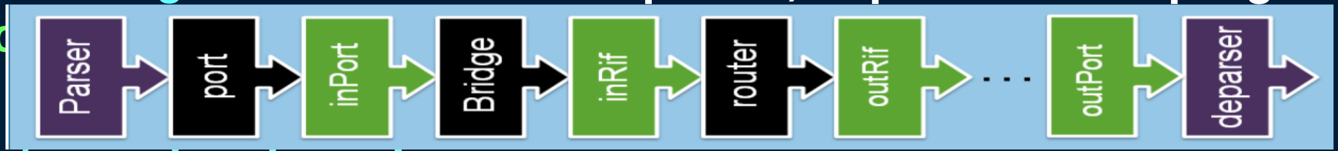
(2) Added a remote Switch Abstraction Interface (SAI), with programmability extensions

(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 programable pipelines [Green bo



in

between fixed pipelines

Beyond Programmability Alone: A Systems Approach
Reservoir Labs Gradient Graph (G2) Analytics

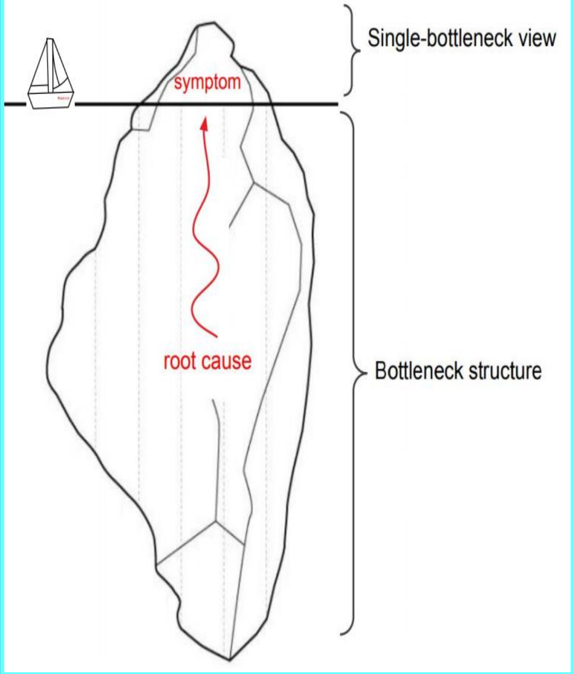
- **Objective:** Flow performance optimization in high speed networks, with fairness
- **Approach:** Built on a new mathematical Theory of Bottleneck Structures and an analytical framework
  - Enables operators understand and precisely control flow and bottleneck performance
- **Value:** Improved capacity 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].
- "A Quantitative Theory of Bottleneck Structures for Data Networks," (in review) submitted to IEEE Transactions on Networking.
- "Designing Data Center Networks Using Bottleneck Structures," accepted for publication at ACM SIGCOMM 2021 (to be announced).

www.reservoir.com/gradientgraph/

System Wide information to identify, deal with root causes



Key Components

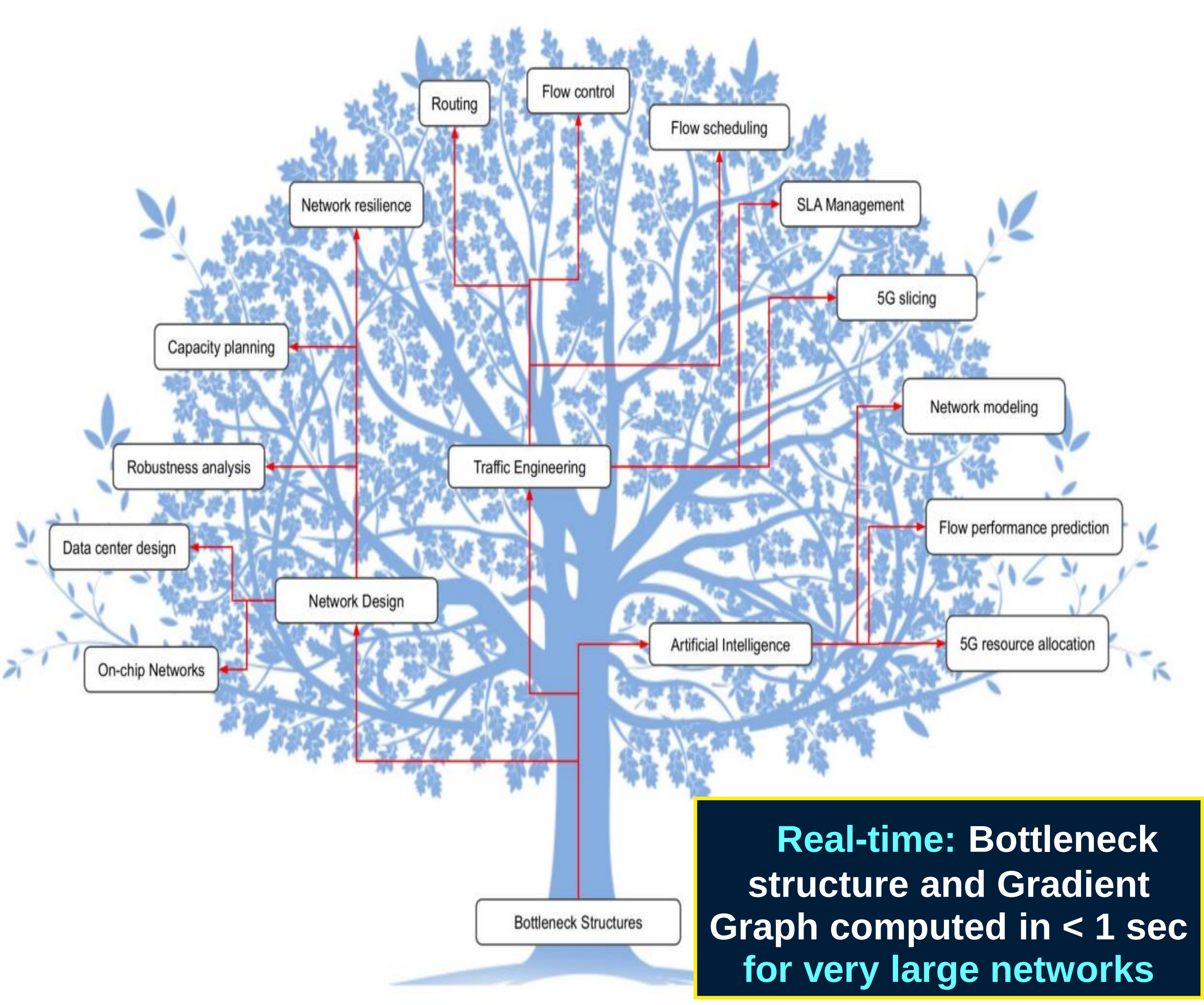
- Bottleneck precedence + flow gradient graphs
- Impactful flow and flow group ID
- Alternate path recommendations

Reservoir Labs Gradient Graph (G2): Systems Approach
Bottleneck Structures to Application Areas

Network Design
- Network Resilience
- Capacity Planning
- Robustness Analysis
- Data Center Design
- On Chip Networks

Traffic Engineering
- ✸ Routing
- ✸ Flow Control
- ✸ Flow Scheduling
- ✸ SLA Management
- 5G Slicing

Artificial Intelligence
- Network Modeling
- Flow Performance Prediction
- Resource Allocation



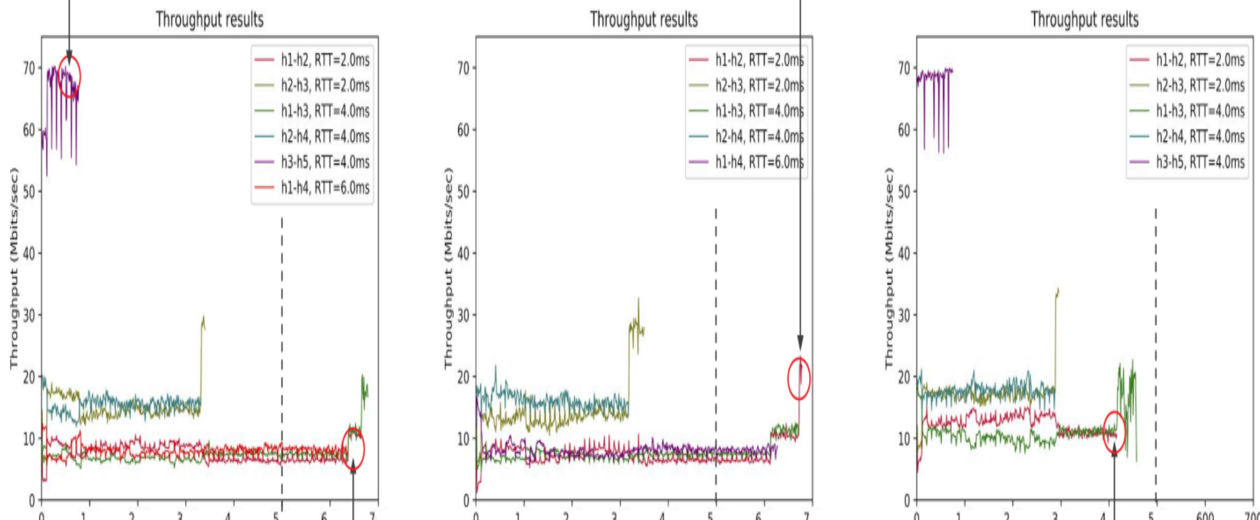
Real-time: Bottleneck structure and Gradient Graph computed in < 1 sec for very large networks

www.reservoir.com/gradientgraph/

Operational Use Case: Scheduling of Deadline-Bound Data Transfers

(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

Flow Gradient Graph:

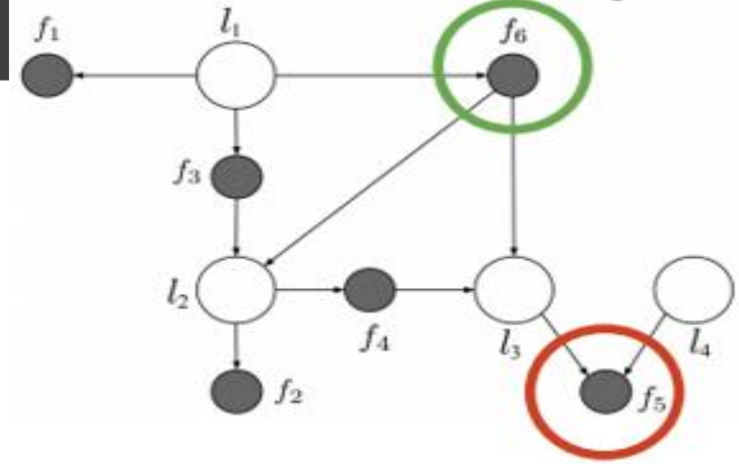
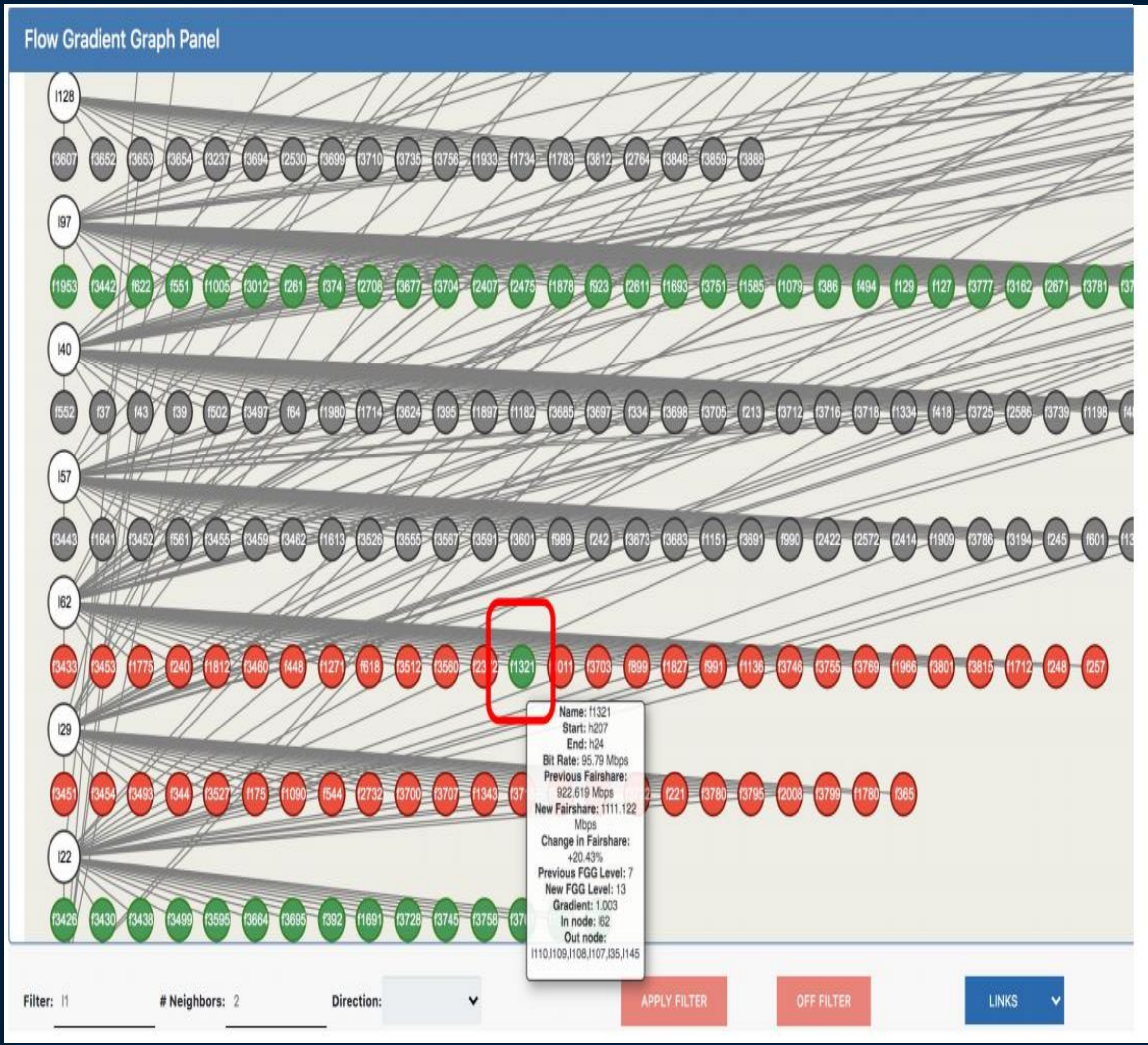


Table 3: As predicted by the theory of bottleneck ordering, flow f_6 is a significantly higher impact flow than flow f_5.

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

Pacific Research Platform: Running Gradient Graph on Federated Kubernetes Clusters



Designing Data Center Networks Using Bottleneck Structures RL, Yale, Columbia

ABSTRACT

This paper provides a mathematical model of data center performance based on the recently introduced Quantitative Theory of Bottleneck Structures (QTBS). Using QTBS, we prove that if the traffic pattern is *interference-free*, there exists an optimal design that both minimizes maximum flow completion time and yields maximal system-wide throughput for that traffic pattern. We use these theoretical insights to study three widely used interconnects—fat-trees, folded-Clos and dragonfly topologies. We show that common production traffic patterns are *interference-free* for these three topologies, and we derive equations that describe the optimal design for each as a function of the traffic pattern. Our model predicts, for example, that a 3-level folded-Clos interconnect with radix 24 that routes 10% of the traffic through the spine links can reduce the number of switches and cabling at the core layer by 25% without any performance penalty. We present experiments using production TCP/IP code to empirically validate the results and provide tables for network designers to identify optimal designs as a function of the size of the interconnect and traffic pattern.

Paper Accepted for SIGCOMM 2021

R&D on Network Capabilites

Approach: Develop a stateful network management system to address the issues

⇨ **Comprehensive R&D**

Key System Features include:

- **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 multi-objective optimization strategy, with constraints
- **Identification, diversion and assignment to alternate, additional, or privileged paths for large flows** when available, OR
- **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**
- **Constraining the allocations and the aggregate, so as not to impede others' existing best effort traffic on the major shared routes**
- **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%*

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***
- ***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, Brian Yang, Buseung Cho, Caio Costa, Carlos Antonio Ruggiero, Carlyn Ann-Lee, Chin Guok, Ciprian Popoviciu, Dale Carder, David Lange, David Wilde, 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, Mahdi Solemani, 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***

✷ ***Meets Weekly or Bi-weekly; All are welcome to join.***

Next Generation Networking System for Data Intensive Sciences

- A comprehensive, forward looking global R&D program is needed:
  - To meet the challenges faced by the major science programs, including Petabyte transactions, caching, 400G to Tbps flows
  - To coordinate provisioning the feasible capacity globally, in a way compatible with the overall use by the at-large R&E community
- Beyond capacity alone, we need a Real-time System Coordinating the VO (LHC) & Network Orchestrators, Site and Network Resource Managers
  - Providing dynamic, adaptive, goal-oriented, policy and priority driven operations among the sites and networks
  - Beginning to understand how to operate, manage and optimize this new class of systems via prototypes of increasing scale and scope
- The LHC Experiments, WLCG Sites, GNA-G and its DIS WG, and R&E network community have key roles in:
  - Considering how the effort to design and build the new Computing Model should be organized and carried out
  - To successfully complete the paradigm shift required by ~2027
- The GNA-G and R&E network community should pursue feasible capacity increases (e.g. via spectrum) to frame the capacity/complexity tradeoff

Extra Slides Follow

P4 + Reservoir Labs + SENSE 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

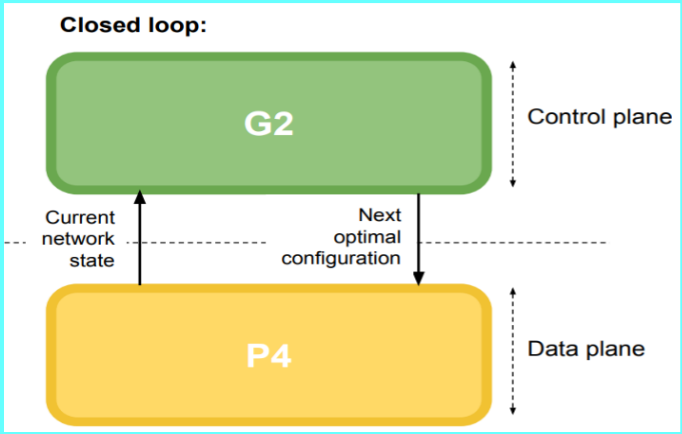
- **Congestion detection**
- **Impactful flow-group identification**
- **Agile flow steering or moderation**
- **Verification of congestion mitigation**
- **Load balancing**

(2) Subsequently

- **Tune the sequence of steps and decision parameters**
- **Begin to develop + evaluate success metrics**
- **Predict and optimize using machine learning**

P4 + Reservoir Labs G2 + SENSE System Design
Factors and Model

P4 – G2 Closed Loop: Leverage INT capabilities – standardized specs on header format/content/placement to match multiple protocols, and INT reporting standards



Stateful User Defined Headers/Labels: Sufficient information to:

- **Set short- and longer-term priorities, deadlines and other characteristics to adjudicate among competing SLAs**
- **Know attributes, performance and reliability records of segments and of sites when choosing among path options, task assignment, data location, etc.**

Data Center Analog Model with 4 to 6 Transaction Classes
Assign resources; Send incoming requests to each class; Monitor class progress; Adjust among and within each class

P4 + Reservoir Labs G2 + SENSE System Design

Decision Classes and Engines

(1) Tactical: Proceed at will based on G2 “Next optimal configuration”; + validation: response-adjustments if effect of a change is not as expected, within bounds

(2) Policy-driven based on short term SLAs: respecting deadlines for the delivery of a limited set of privileged flows.

(3) Reactive decisions based on: [lack of] progress in classes of flows; network events (link or site failure/impairment/...); injection of large higher priority flows; adjusting priorities for transactions pending or incomplete for too long; congestion avoidance to impact on the aggregate of "best effort flows”

(4) Strategy-based adjustments, such as: resource sharing among client VOs; efficient use of site computing resources; dataset placement/caching; regionality (limit flows to a given country or continent or a defined link set.

(5) Long term (days to weeks) decisions based on: optimizing an overall synthetic metric that considers: throughput, efficiency of network use, efficiency of site resource use, SLA and priority profile matching.

(6) Longer term optimization and evaluation (months to years): Use ML and performance records to formulate and tune recommendations:

- Part of the task is to develop the metrics themselves
- Balance among the various requirements and constraints
- Dev cycles: Consider, discuss, adjust what the metric delivers once “optimized”

GradientGraph Analytics: Technology Status

- Completed base theory of bottleneck structures. Work published at SIGMETRICS 2020, SC INDIS 2019/2020, recent submission accepted for publication at SIGCOMM 2021.
- Completed base implementation of G2 Analytics: computation of bottleneck structure.
- Support for NetFlow, sFlow and Mininet integration plugins.
- Support for REST API for scripting G2 Analytics.
- Support for base Graphical User Interface.
- Implemented first in-depth analytical apps (1) for optimal flow routing (Google maps for networking) and (2) capacity planning.
- Deployed in production at the Pacific Research Platform (a global REN network) and in DOE ESnet (US nation wide network) for initial testing and evaluation.
- Resolved and implemented the computation of flow/link gradient for very large networks in a fraction of a second.
- Developing new analytical frameworks to tackle other key networking problems: capacity planning, data center designs, optimal flow scheduling, etc.

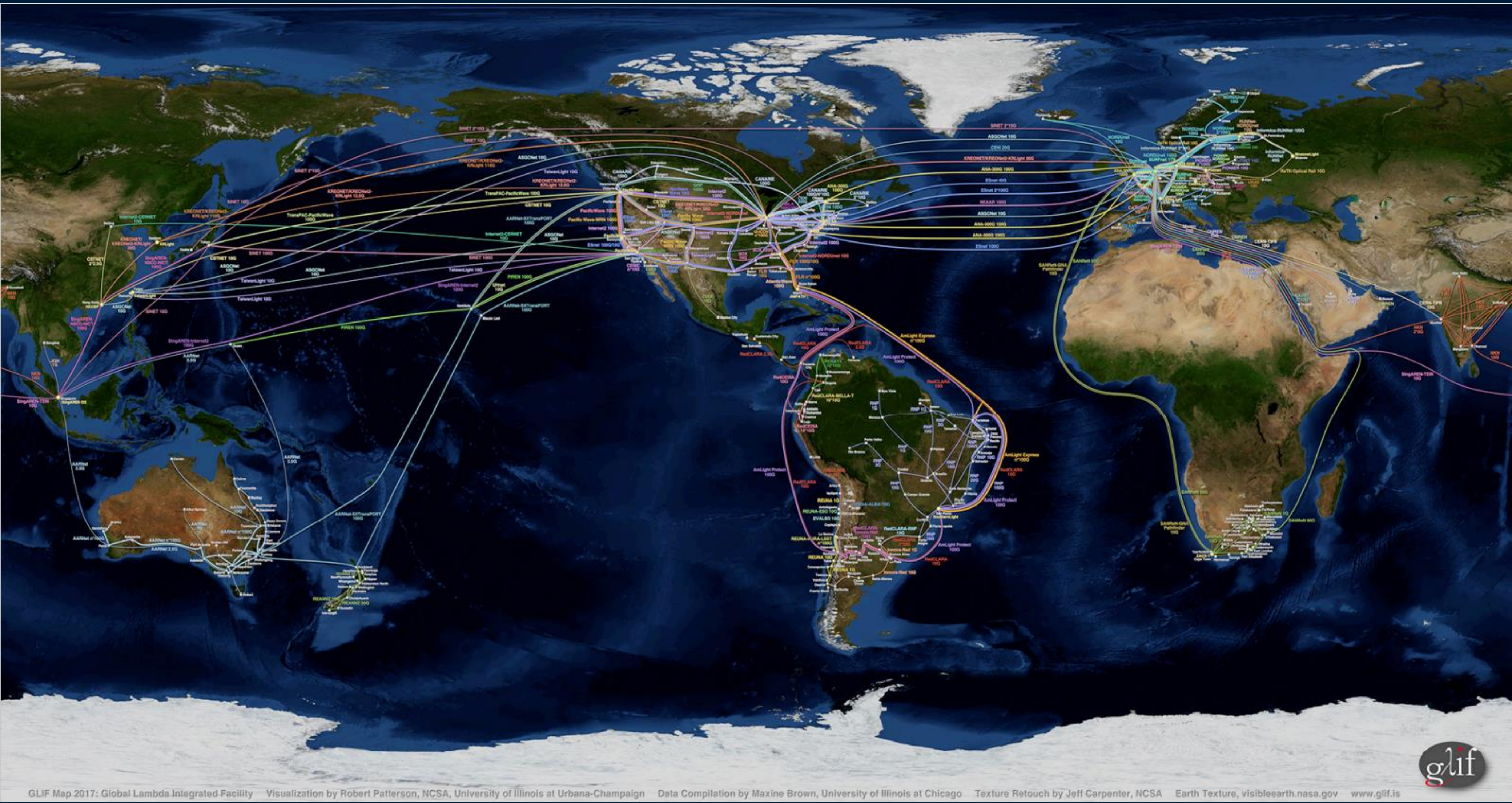
Reservoir Labs Gradient Graph (G2) Analytics

www.reservoir.com/gradientgraph/

Use Cases

- Scheduling of deadline-bound flows
- Network performance baselining
- End-to-end Multi-resource flow optimization. Modeling bottlenecks for links, storage and compute
- Capacity planning
- Risk / network failure analysis
- Flow admission control
- Optimal load balancing
- Optimal design of fat tree and Clos networks
- Bandwidth tapering and bandwidth steering
- Identification of optimal flow/circuit placement
- Troubleshooting of routing misconfigurations
- Bottleneck identification in heterogeneous networks
- SLA management

Shared Network Infrastructure: GLIF 2017 Map



Slow Growth in Capacity at Fixed Cost: ~2 Tbps TA by 2028 ?

Sharing with the academic and research community on several continents

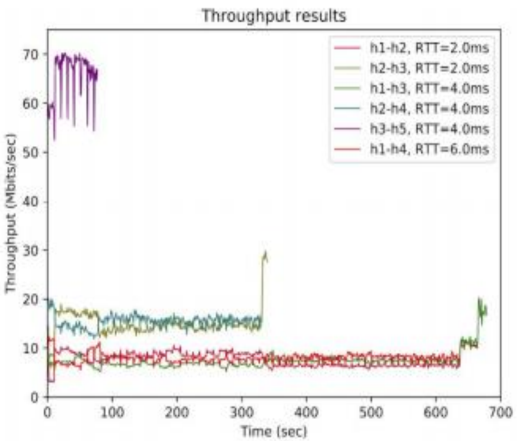
P4 Useful Solution Piece Elements

https://github.com/p4lang/tutorials

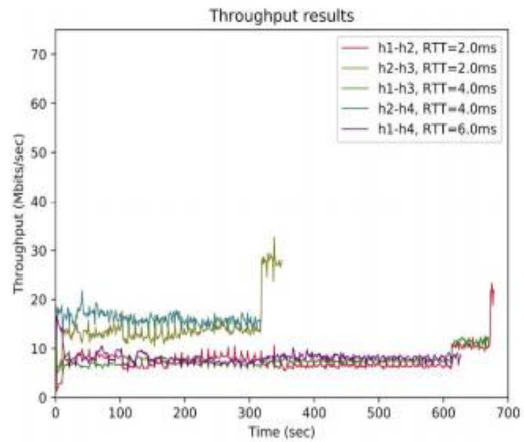
- **Layer 4 Load Balancer – SilkRoad[1]**
- **Low Latency Congestion Control – NDP[2]**
- **In-band Network Telemetry – INT[3]**
- **In-Network caching and coordination – NetCache[4] / NetChain[5]**
- **Aggregation for MapReduce Applications [7]**

[1] Miao, Rui, et al. "SilkRoad: Making Stateful Layer-4 Load Balancing Fast and Cheap Using Switching ASICs." SIGCOMM, 2017.
[2] Handley, Mark, et al. "Re-architecting datacenter networks and stacks for low latency and high performance." SIGCOMM, 2017.
[3] Kim, Changhoon, et al. "In-band network telemetry via programmable dataplanes." SIGCOMM. 2015.
[4] Xin Jin et al. "NetCache: Balancing Key-Value Stores with Fast In-Network Caching." To appear at SOSP 2017
[5] Jin, Xin, et al. "NetChain: Scale-Free Sub-RTT Coordination." *NSDI*, 2018.
[6] Dang, Huynh Tu, et al. "NetPaxos: Consensus at network speed." SIGCOMM, 2015.
[7] Sapio, Amedeo, et al. "In-Network Computation is a Dumb Idea Whose Time Has Come." *Hot Topics in Networks*. ACM, 2017.

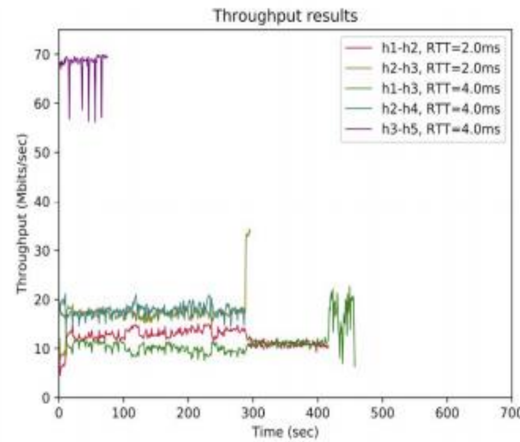
Are all Elephant Flows Heavy Hitters?



(a) Without removing any flow.



(b) Removing the heavy-hitter flow f_5.

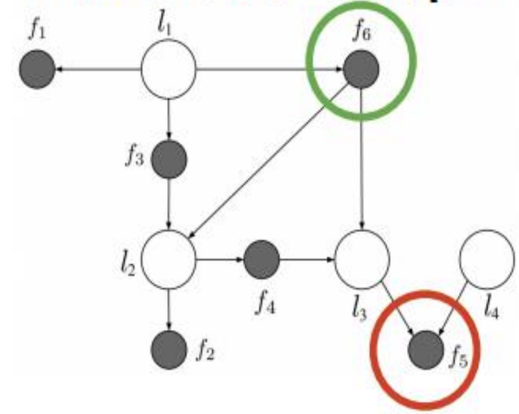


(c) Removing a low-hitter flow f_6.

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

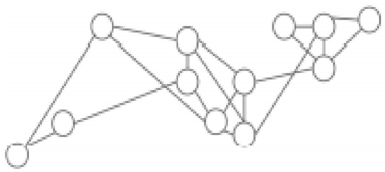
Comp. time (secs)	f_1	f_2	f_3	f_4	f_5	f_6	Slowest
With all flows	664	340	679	331	77	636	679
Without flow f_5	678	350	671	317	—	611	678
Without flow f_6	416	295	457	288	75	—	457
Avg rate (Mbps)	f_1	f_2	f_3	f_4	f_5	f_6	Total
With all flows	7.7	15.1	7.5	15.4	65.8	8.1	119.6
Without flow f_5	7.5	14.5	7.6	16.1	—	8.3	54
Without flow f_6	12.2	17.2	11.1	17.7	68.1	—	126.3

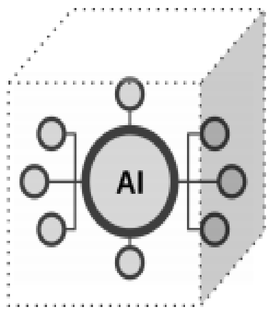
Flow Gradient Graph:

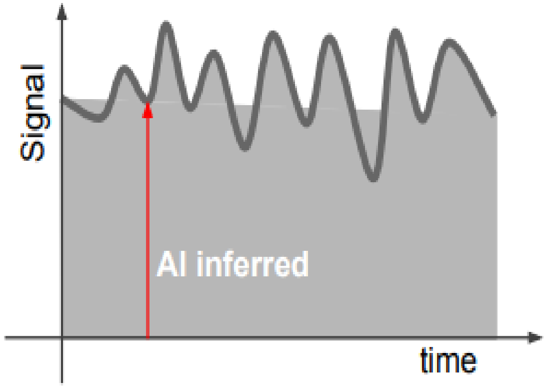


GradientGraph and AI for 5G Network Modeling

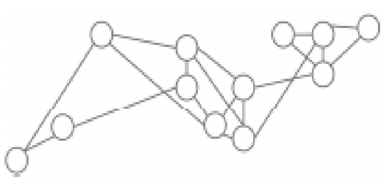
Network: | **AI (unexplained) model:** | **Prediction:**

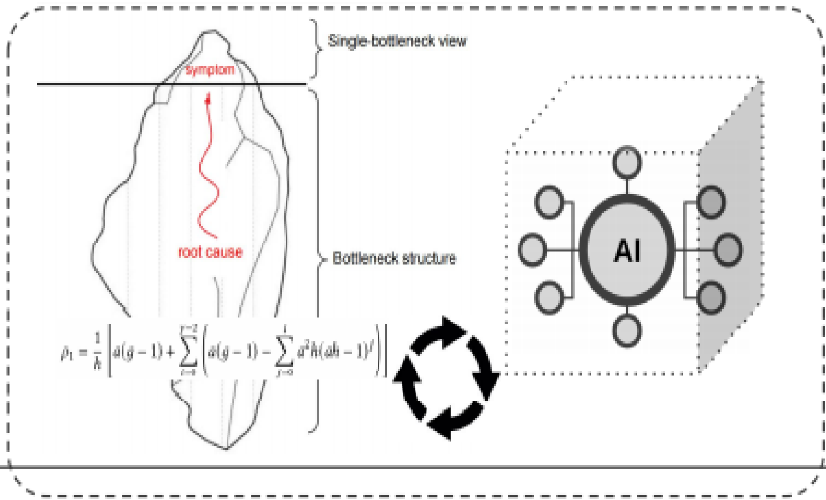




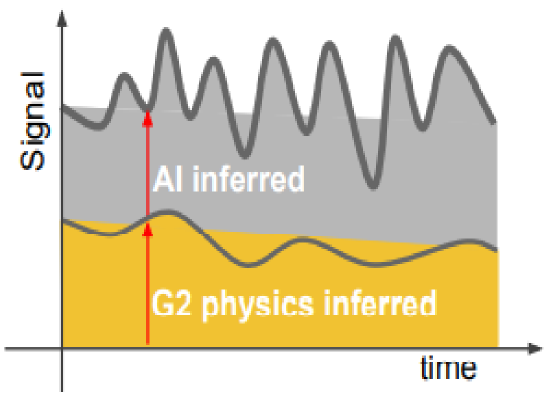


Network: | **G2 Physics (explained) model + AI (unexplained) model:** | **Prediction:**



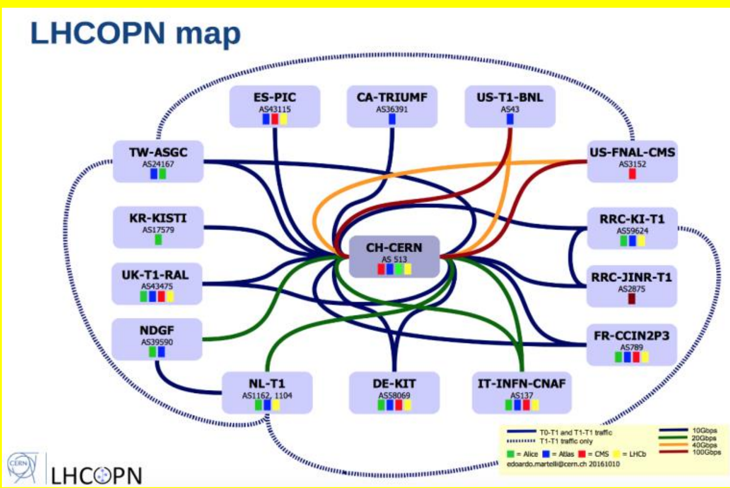


Explained AI model



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

LHCOPN: Simple & Reliable Tier0+1 Ops

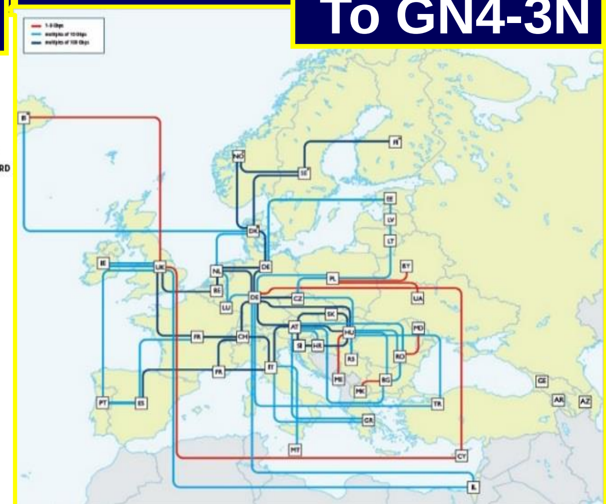


Internet2

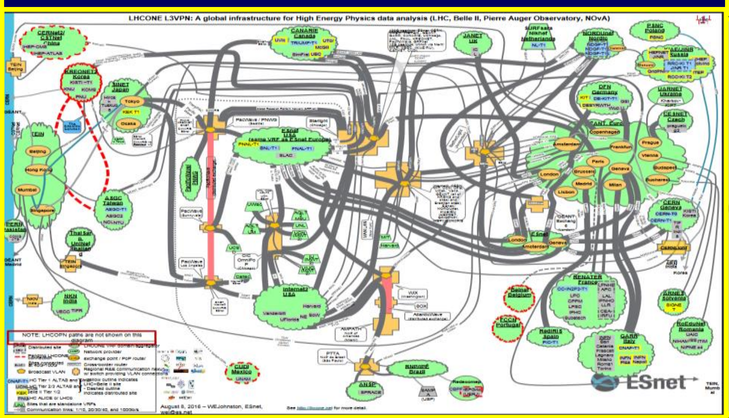
To NGI

GEANT

To GN4-3N



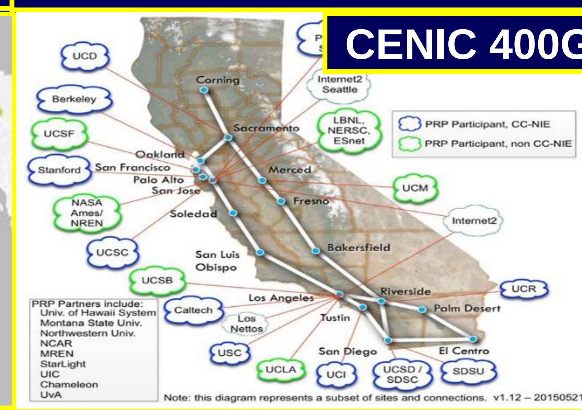
LHCONE VRF: 170 Tier2s



ESnet (with EEX)

CENIC and PRP

CENIC 400G



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

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

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

Where were we?

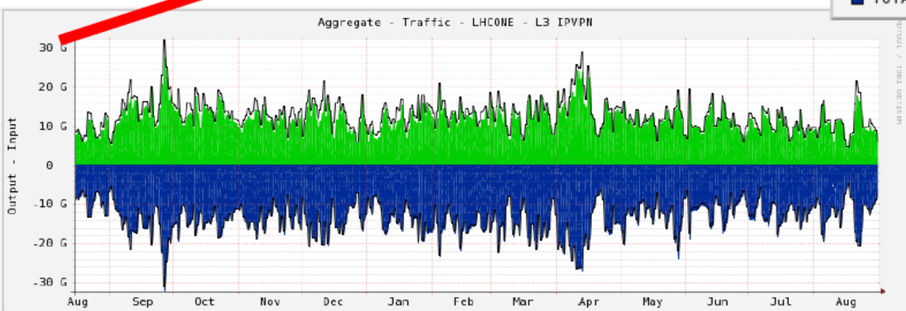
LHCONE

LHCONE in Europe
GEANT

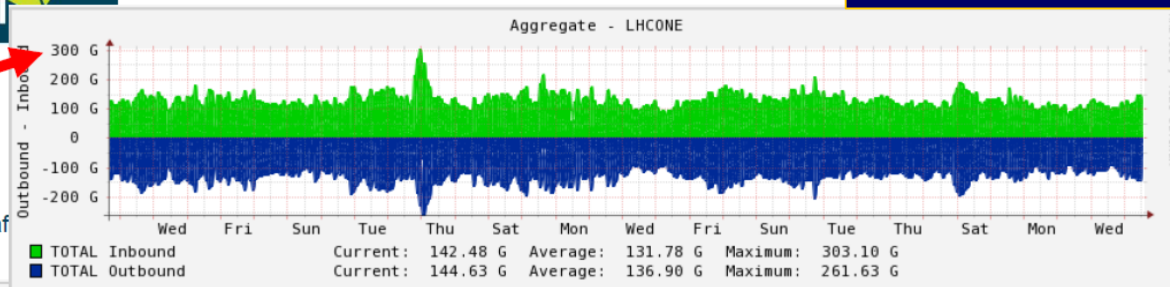
GÉANT

- 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 traffic is much higher

10x

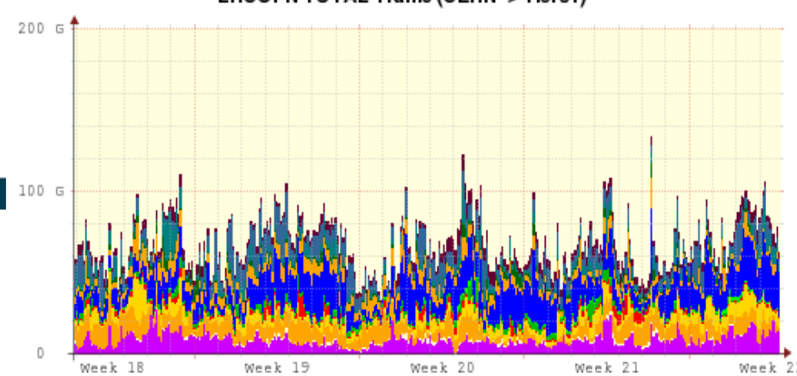


LHCONE



+ LHCOPN

LHCOPN TOTAL Traffic (CERN -> Tiers1)



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 possible challenge already in LHC Run3 (2022-4)

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

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

- Useful distance metric: 10% IO penalty among merged caches
- EU example: ~500 miles
- Advanced protocol, caching methods: could extend distance

F. Wuerthwein (UCSD) et al

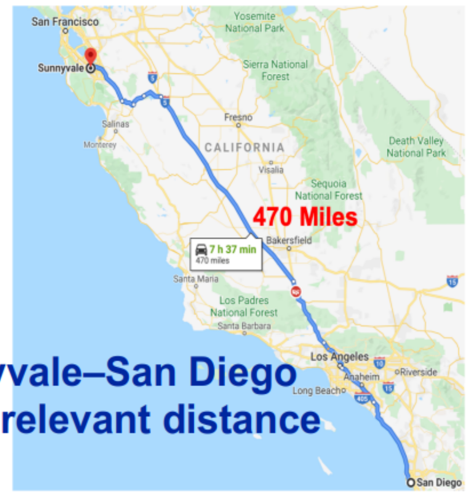
Application use case with CMS

- **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
  - 2.4 PB of NANO

(AOD, MINI) Mostly kept on Tape => accessed a couple times per year

(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

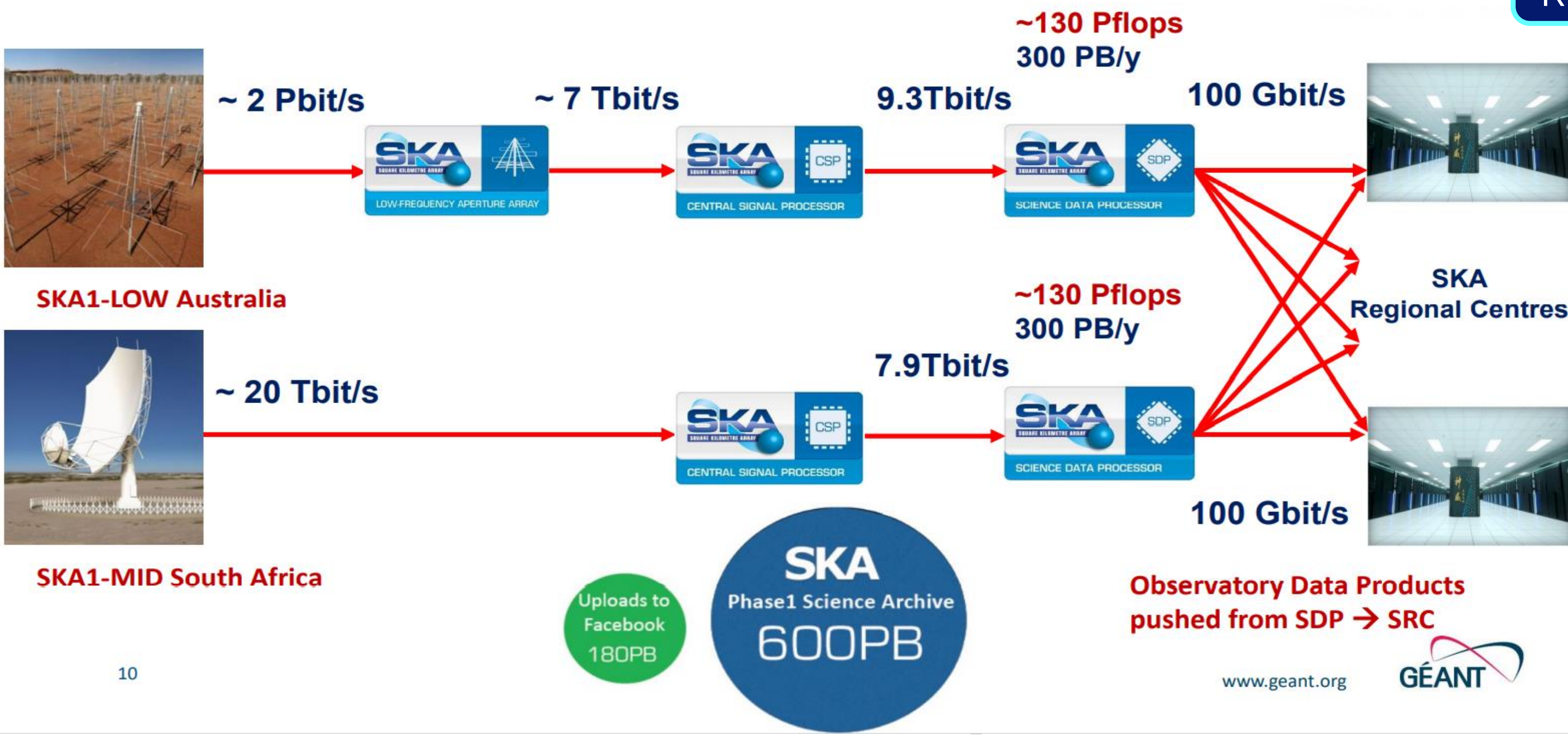
- Hardware: 40TB storage and 40Gbps networking capability
- Expected network utilization: about 10-20 Gbps

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

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

SKA Phase1 Data Flows

Courtesy
R. Hughes-Jones



CSP – SDP Network

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

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

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

Interfacing to Multiple VOs With FTS/Rucio/XRootD

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

CMS

ATLAS EXPERIMENT

LHCb

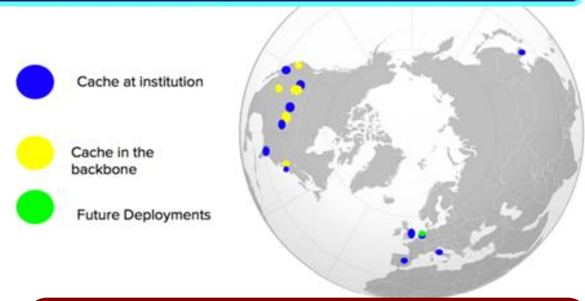
ALICE

ICECUBE South Pole Neutrino Observatory

Xe XENON Dark Matter Project

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

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

FAIR Facility for Antiproton and Ion Research

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

EST european solar telescope

ESO

SKA SQUARE KILOMETRE ARRAY

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

European Science Data Center

LSC LIGO Scientific Collaboration

VIRGO

KAGRA

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

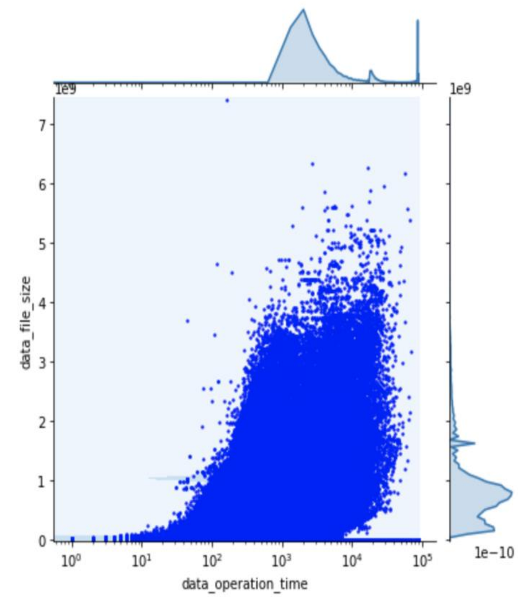
- While sharing and advancing the (new) concepts, tools & systems needed
- Members of the WG will partner in joint deployments and/or developments of generally useful tools and systems that help operate and manage R&E networks with limited resources across national and regional boundaries
- A special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

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**



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

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?

ESnet plan to install additional in-network caches near US Tier2s in 2021

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
To April 2021

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.

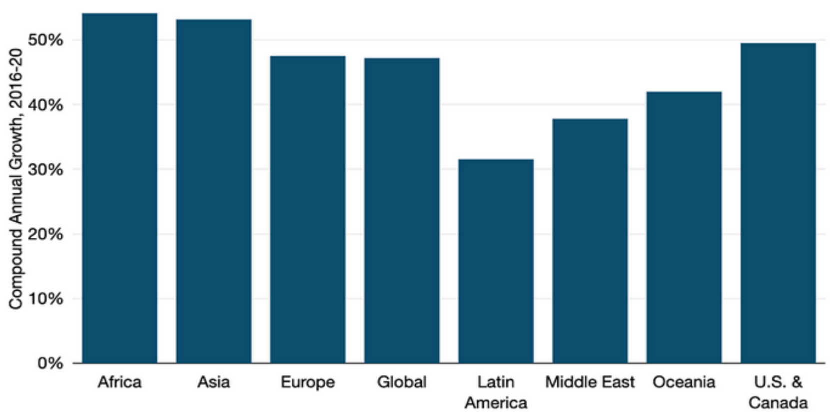
✸ The GNA-G needs to be a data intensive research & science engager that facilitates and accelerates global-scale projects by
(1) enabling high-performance data transfer, and
✸ (2) acting as a partner in developing next generation intelligent network systems that support the workflow of data intensive programs

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

International Bandwidth Trends Telegeography June 2021

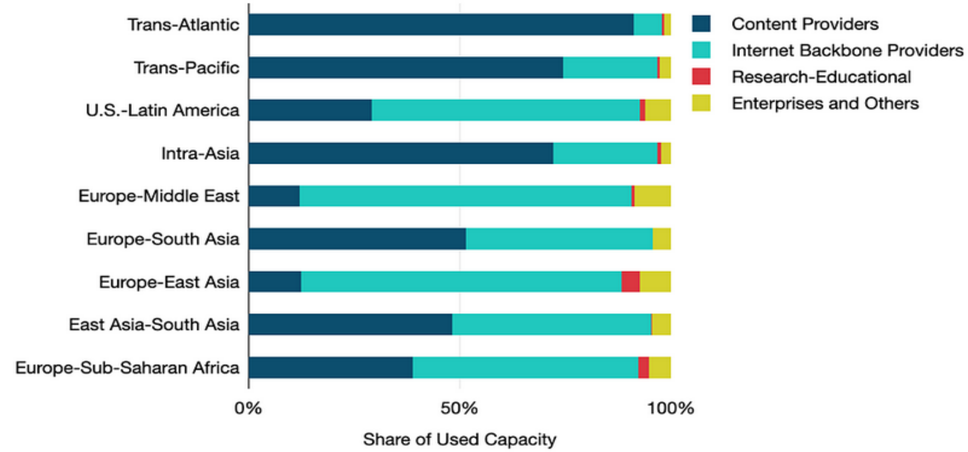
https://blog.telegeography.com/2021-international-bandwidth-trends-demand-global-networks

Used International Bandwidth Growth by Region



Source: TeleGeography

Share of Used Bandwidth by Category for Major Routes

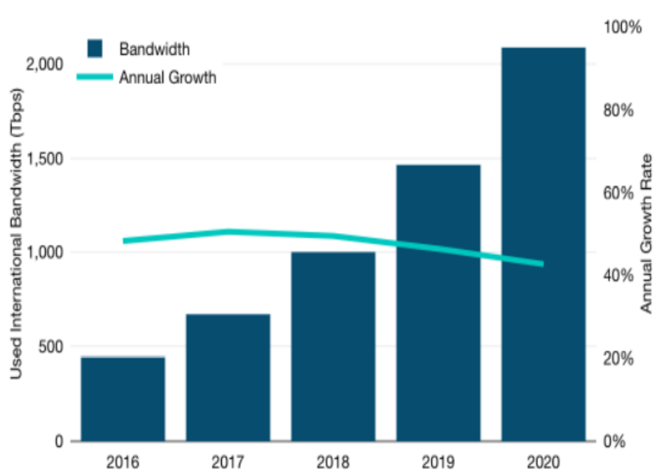


Notes: Data shows used bandwidth as of year-end 2020.

Global Bandwidth Exec Summary 2021: 45% CAGR; Lit Capacity Keeping Pace

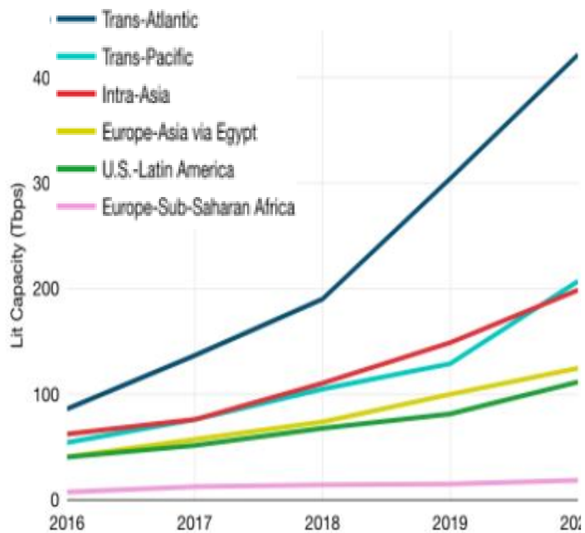
FIGURE 1

Worldwide International Bandwidth Growth



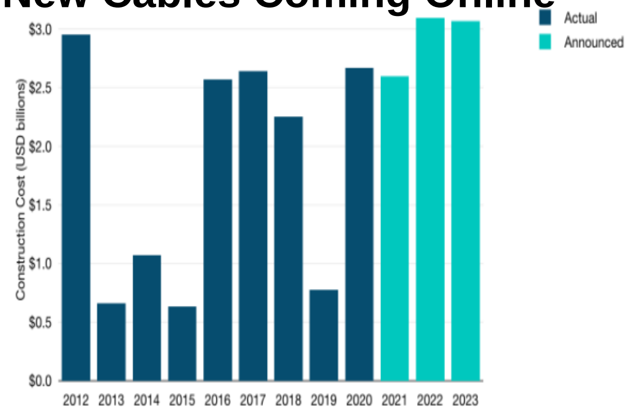
Source: TeleGeography

Lit Submarine Cable Supply by Route



Construction Cost of Submarine Cables

New Cables Coming Online

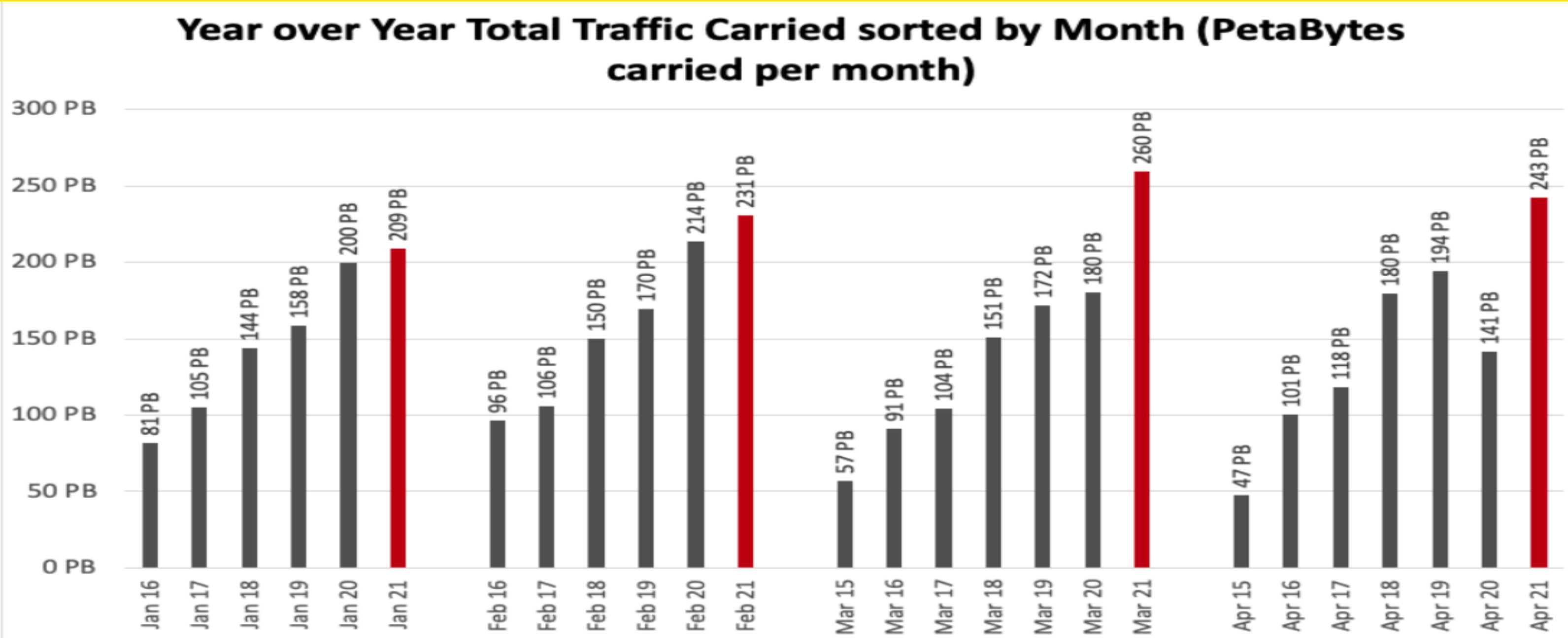


Notes: Total construction costs of all international and domestic submarine cables entering service in designated years. Construction costs exclude the cost of subsequent capacity upgrades and annual operational costs. 2021-2023 construction costs based on announced contract values and TeleGeography estimates. Not all planned cables may be constructed.

Source: TeleGeography

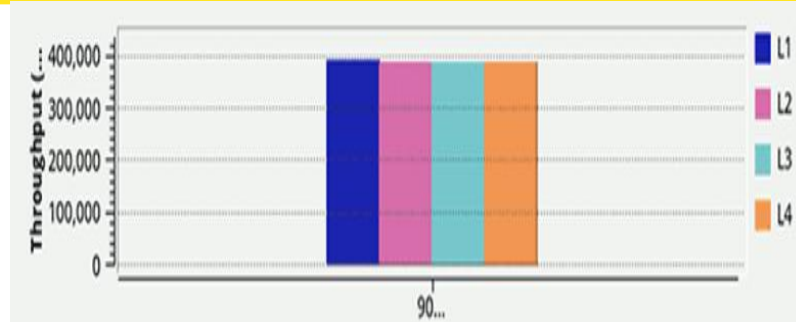
Internet2 Network Milestone: 1 Exabyte moved in 5 Months

https://internet2.edu/internet2-network-milestone-1-exabyte-of-data-moved-between-january-and-may-2021/



Internet2 NGI: 396 Gbps moved Coast to Coast: May 2021

Bidirectional on 400G links



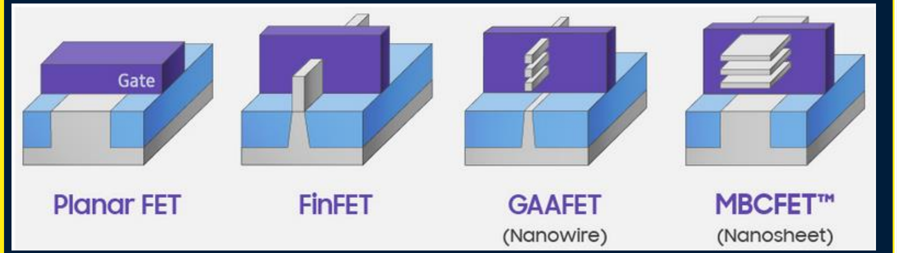
Pass/Fail	Frame Length (Bytes)	Measured L1 Rate (Mbps)	Measured L2 Rate (Mbps)	Measured L3 Rate (Mbps)	Measured L4 Rate (Mbps)	Measured Rate (frms/sec)	Pause Detect	Cfg Rate (L1 Mbps)
Pass	9000	396000.0	395122.0	394331.7	393453.7	5,487,804	No	396,000

Transistor Architecture: How far can one go ?

Samsung Foundry Forum 2019

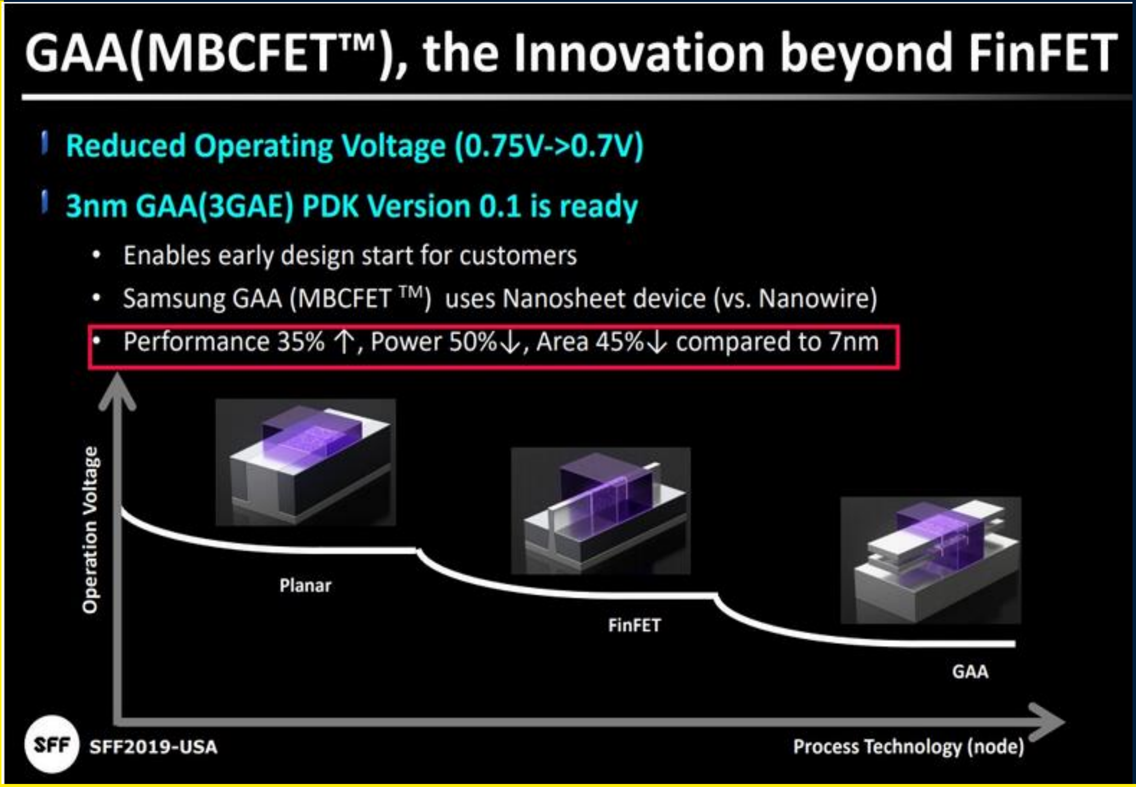
https://www.extremetech.com/computing/291507-samsung-unveils-3nm-gate-all-around-design-tools

- **Planning to launch many process node lines, with development tracks for 7nm, 6nm, 5nm, 4nm, and yes, 3nm. 3nm design kit now in alpha**



- **14nm, 10nm, and 7nm nodes use FinFETs — vertical "fins" above the formerly 2D channel structure, to increase the contact area between transistor channel and the gate.**
- **New Gate All Around (GAA) Architecture with nanowires or nanosheets. From the slide:**

  "3nm increases performance by 35% while reducing power by 50% percent and area by 45% compared to 7nm"



- **Expect 5nm in mass production by 2020 (predicted gains of 10% performance or 20% power consumption over 7nm)**
- **Consumer shipments of products built on 5nm expected in 2021. Samsung's GAA FinFET is planned for volume production in late 2021.**
  Consumer shipments expected in early 2023.

✸ IBM Announces 2nm in May 2021

FS.com 32 X 400G Switch Commoditization of 400G

STORDIS Tofino2 Switch: Fully P4 Programmable 32 X 400G

N9500-32D, 32-Port L3 Data Center White Box Switch, 32 x 400Gb QSFP-DD, Broadcom Chip, Bare-Metal Hardware #96982

Spine switch for data centers and large enterprise networks

US$ 9,999.00

Import Fees included

FS P/N: N9500-32D

11 Sold | 6 Reviews | 8 Questions

The Edgecore AS9516-32D switch is a programmable leaf or spine switch for large scale data centers.
Based on the Intel Tofino 2 (formerly known as Barefoot Tofino) ASIC it allows for advanced P4 Programmability.

In a 1RU form factor, the switch provides, amongst others, line-rate L2 switching and L3 routing, deep packet inspection and traffic load balancing.

KEY FEATURES

- 32 x QSFP-DD switch ports, each supporting 1 x 400 GbE, or via breakout cables 4 x 100 GbE, 8 x 50 GbE or 16 x 25 GbE or 40 GbE or 10 GbE
- Quad-pipe programmable packet processing pipeline for 12.8 Tbps total bandwidth
- Layer 2 or Layer 3 forwarding of 12.8 Tbps (full duplex)
- Pre-Loaded with Open Network Install Environment (ONIE)
- P4 Programmability

USE CASES

- Deploys as leaf or spine switch supporting 25/50/100 GbE to servers, with 100 or 400 GbE uplinks.
- Deploys as spine switch supporting 100 or 400 GbE leaf and spine interconnects.

KEY COMPONENTS

- Intel/Barefoot Tofino 2 BFN-T20-128Q (U series) ASIC
- Intel Pentium D-1517 (4 core@1.6GHz) CPU
- 8GB SO-DIMM RAM
- 128GB M.2 SSD Storage

SUPPORTED SOFTWARE

ONL Open Network Linux | SONiC | onie | P4

Special Pricing for Academia & Research Projects

23,500 USD*

*Only valid for projects registered and approved until March 31, 2021.