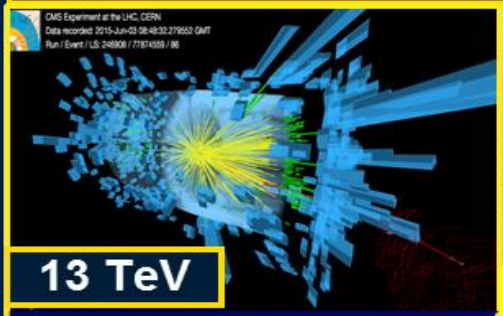
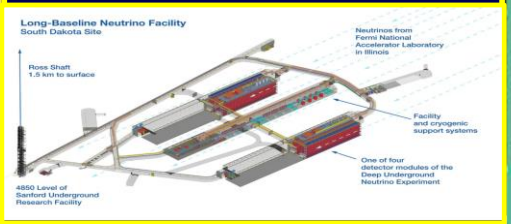
GNA-G DIS WG: Network-Integrated R&D

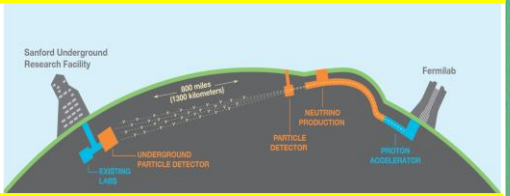
and a New Computing Model for Data Intensive Sciences



LHC

LSST





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

TNC P4 and Intelligent Data Plane BOF

June 18, 2021

https://www.dropbox.com/s/hpdvq7irsjlv53y/NetworkTrendsandRandD_forDataIntensiveScience_TNChbn061821.pptx?dl=0

Data Intensive Sciences Working Group
Key Developments in 2021

- The DIS WG has expanded and is progressing well on several fronts
- The AutoGOLE/SENSE testbed is getting close to the start of persistent operations; a major milestone
  - Adding many sites, interfacing to the GEANT/RARE testbed now, and Bridges and FABRIC across the Atlantic in 2021-22
  - An overlay Layer 2/3 topology of virtual circuits will allow continuous development, and a smooth process of migration into production
- This will include use of upgraded infrastructure, including 400G between SURFnet and CERN; 2 X 100G each Caltech – UCSD and Caltech – SNVL
- Launching the use of P4 and P4 capable switches and Freertr with the help of the GEANT/RARE team: at Caltech, UCSD, Tennessee Tech, Starlight: Edgecore Wedge, APS and Mellanox
- Began work with UCSD/PRP and Reservoir Labs on the use of their Gradient Graph (G2) decision support tools for impactful flow identification, flow and/or path adjustments, congestion detection and resolution, SLAs, etc.
- P4 and G2 integration: Targeting: Stateful user defined labels on flow groups; agile flow steering and load balancing; enforced compatibility with the aggregate of best effort flows on shared links; Machine Learning for prediction and optimization
- SC21: Multiple NRE submissions on topics above: abstracts due this week
- Following requirements reviews, we are working with the "blueprint committees" of CMS and ATLAS, to integrate the DIS WG's network R&D in their workplans

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

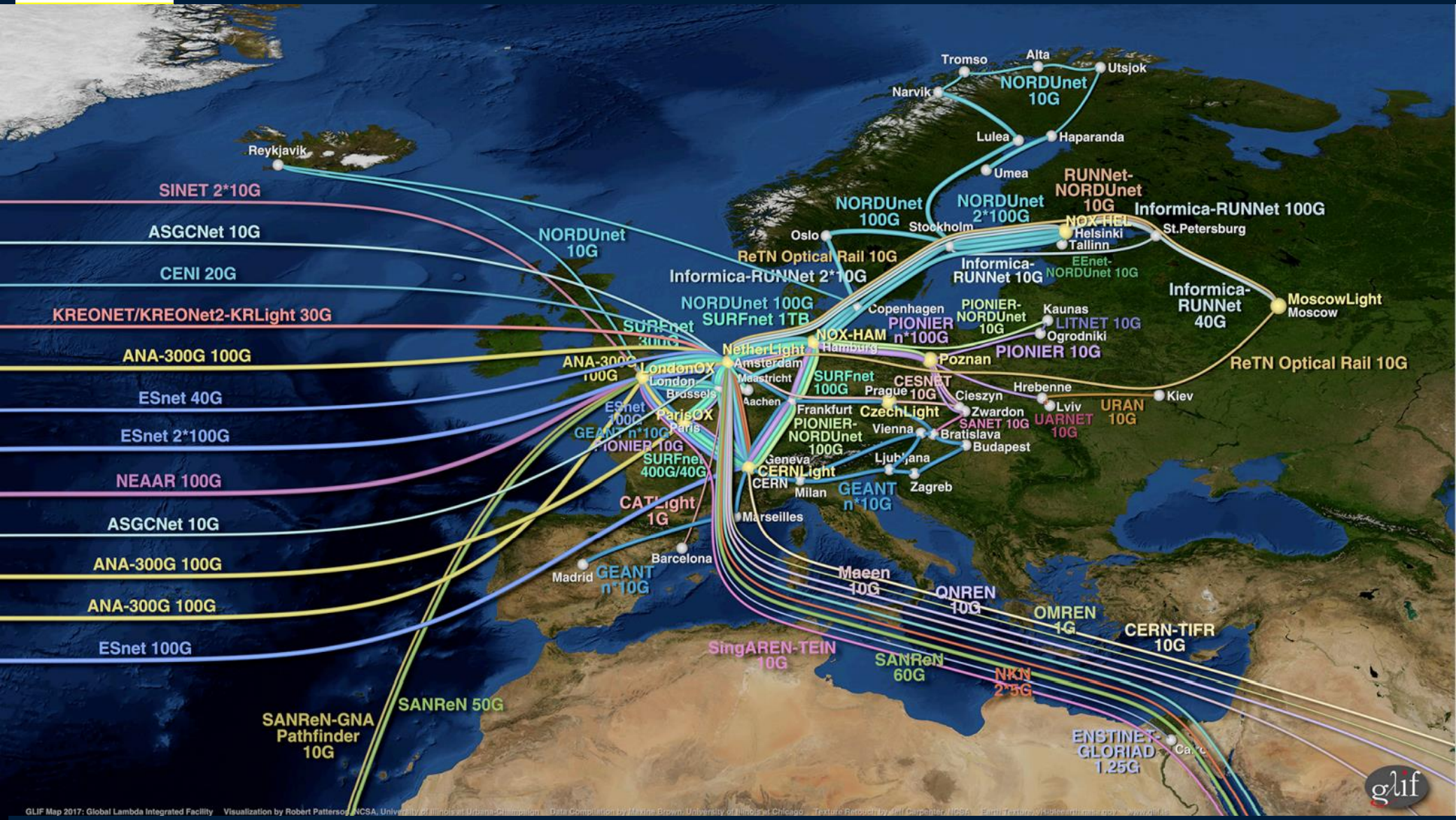
- While sharing and advancing the (new) concepts, tools & systems needed
- Members of the WG will partner in joint deployments and/or developments of generally useful tools and systems that help operate and manage R&E networks with limited resources across national and regional boundaries
- A special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

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

Shared Network Infrastructure: GLIF 2017 Map

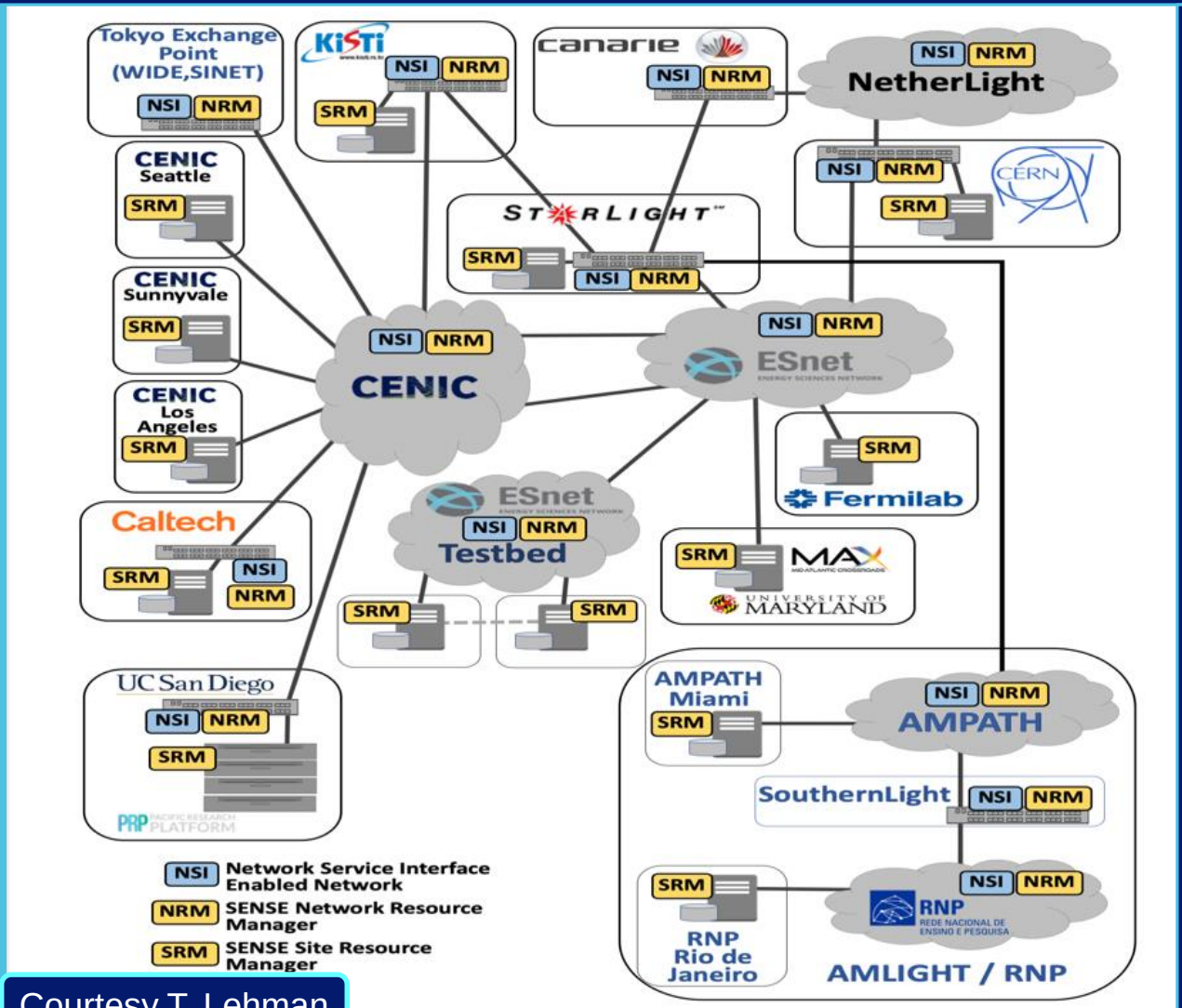


Slow Growth in Capacity at Fixed Cost: ~2 Tbps TA by 2028 ?

Sharing with the larger academic & research community on several continents

[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



Courtesy T. Lehman

2021 Outlook
ESnet6/
High Touch
FABRIC
BRIDGES

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

Federation with the StarLight GEANT/RARE & AmLight P4 Testbeds

400G Link(s) NetherLight-CERN

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

Automation Following Atlantic Wave SDX

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

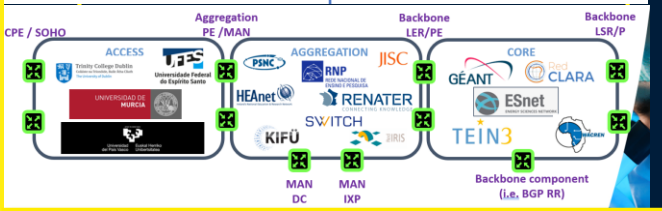
R&D on Network Capabilities: Key Technologies Towards "an Intelligent Data Plane"

- **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**
  - **Traffic Engineering Functions: Impactful flow ID, agile flow steering, load balancing, congestion avoidance.**
- **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 **A**cademia **R**esearch & **E**ducation

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



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

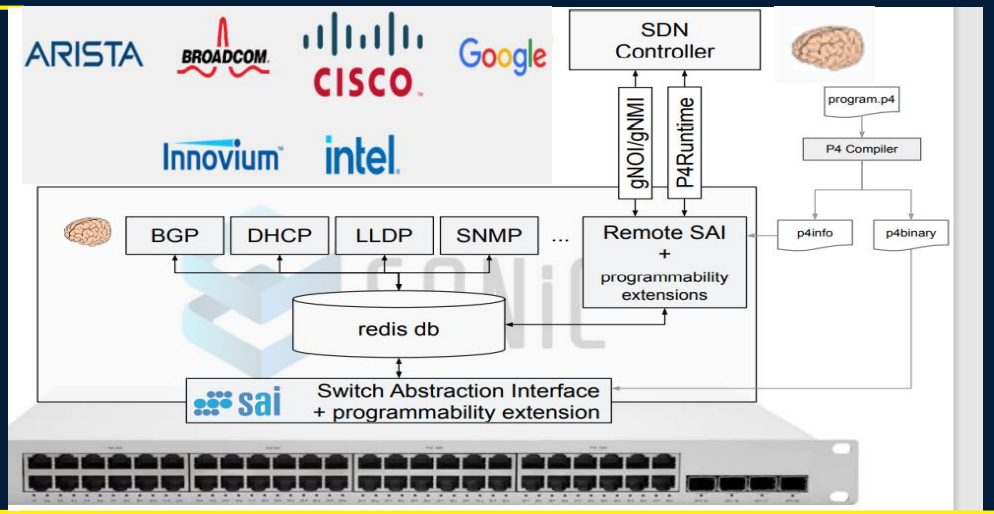
P4 Integrated Network Stack (PINS)

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

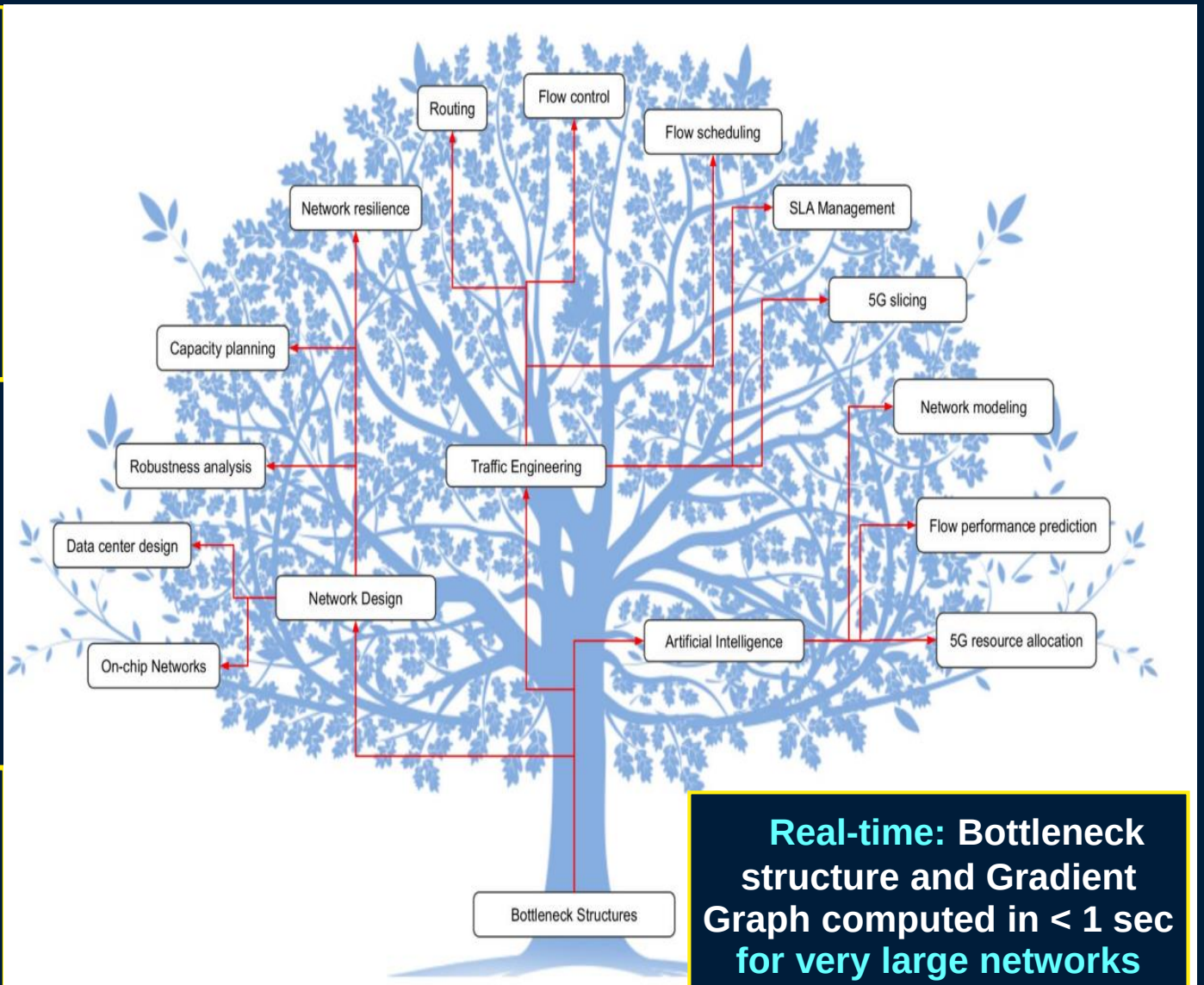
between fixed pipelines

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

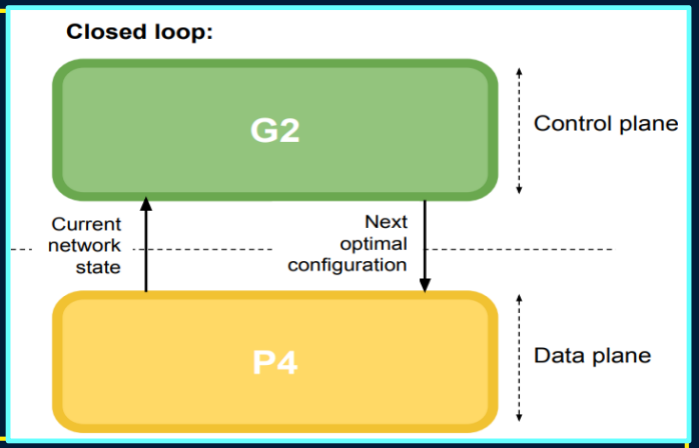


Real-time: Bottleneck structure and Gradient Graph computed in < 1 sec for very large networks

www.reservoir.com/gradientgraph/

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:

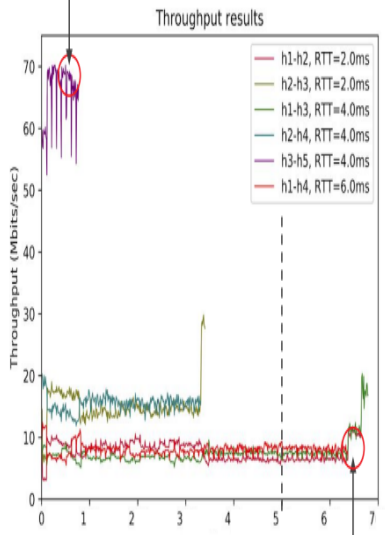
- **Set short- and longer-term priorities, deadlines and other characteristics to adjudicate among competing SLAs**
- **Know attributes, performance and reliability records of segments and of sites when choosing among path options, task assignment, data location, etc.**

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

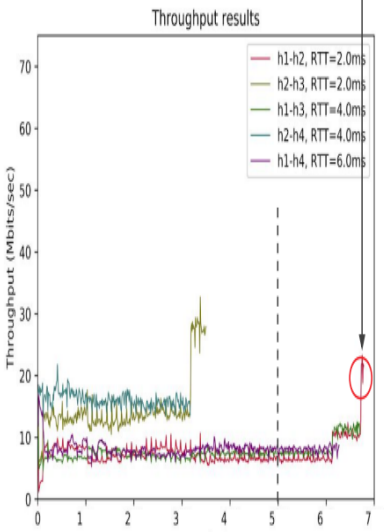
Operational Use Case: Scheduling of Deadline-Bound Data Transfers

(2) Traditional approach: look at heavy hitters

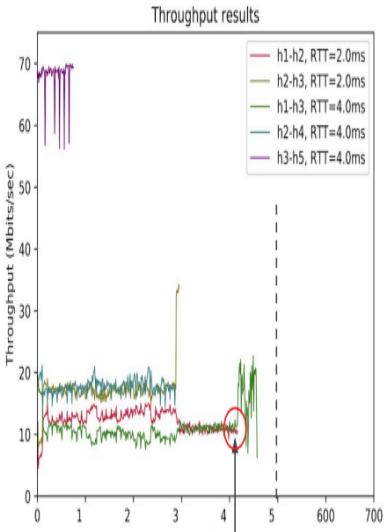
(3) Traditional approach yields no benefit



(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

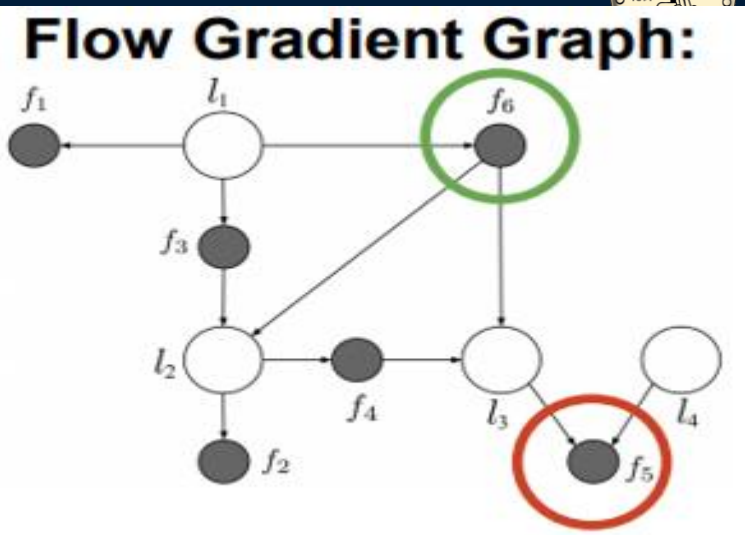


Table 3: As predicted by the theory of bottleneck ordering, flow f_6 is a significantly higher impact flow than flow f_5.

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

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
- GNA-G and its DIS WG, the LHC, VRO and other major science programs, and the R&E network community should come together to:
  - Consider how the effort to design and build the new Computing Model should be organized and carried out
    - 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

For More Information:
https://www.dropbox.com/s/iz2e5ilio9oop3d/DataCenterArchitectureWorkshop_HLLHCRequirementsTrends_NetworkRandD_hbn06052021s4.pptx?dl=0

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

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 limit 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.

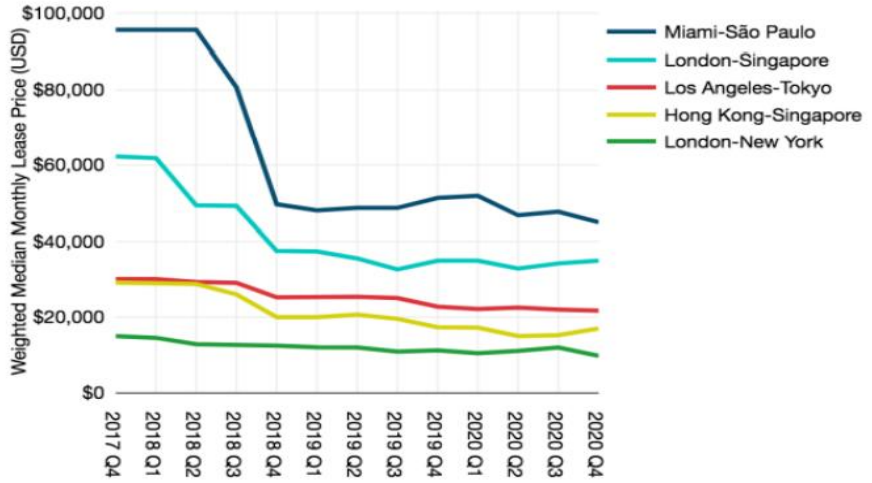
(6) Longer term optimization and evaluation (months to years): Use ML and performance records to formulate and tune recommendations:

- 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"

International Bandwidth Pricing Trends

Executive Summary (telegeography.com)

Weighted Median 100 Gbps Wavelength Price Trends on Major International Routes

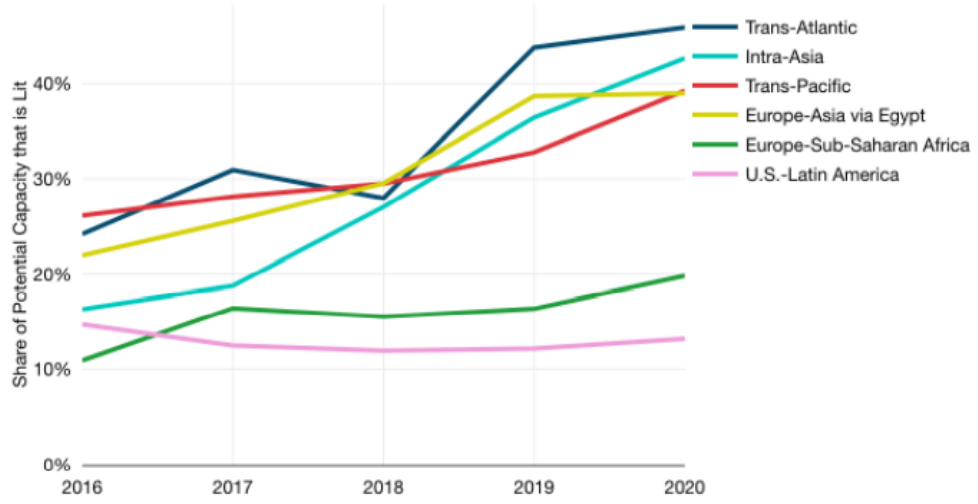


Notes: Each line represents the weighted median monthly lease price for an unprotected 100 Gbps Wavelength on the listed route. Prices are in USD and exclude local access and installation fees.

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**

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, 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, 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.***

https://www.gna-g.net/join-working-group/data-intensive-science/