FABRIC: An Everywhere Programmable Research Infrastructure for Network Experimentation

Kuang-Ching (KC) Wang, Clemson University
Paul Ruth, RENCI

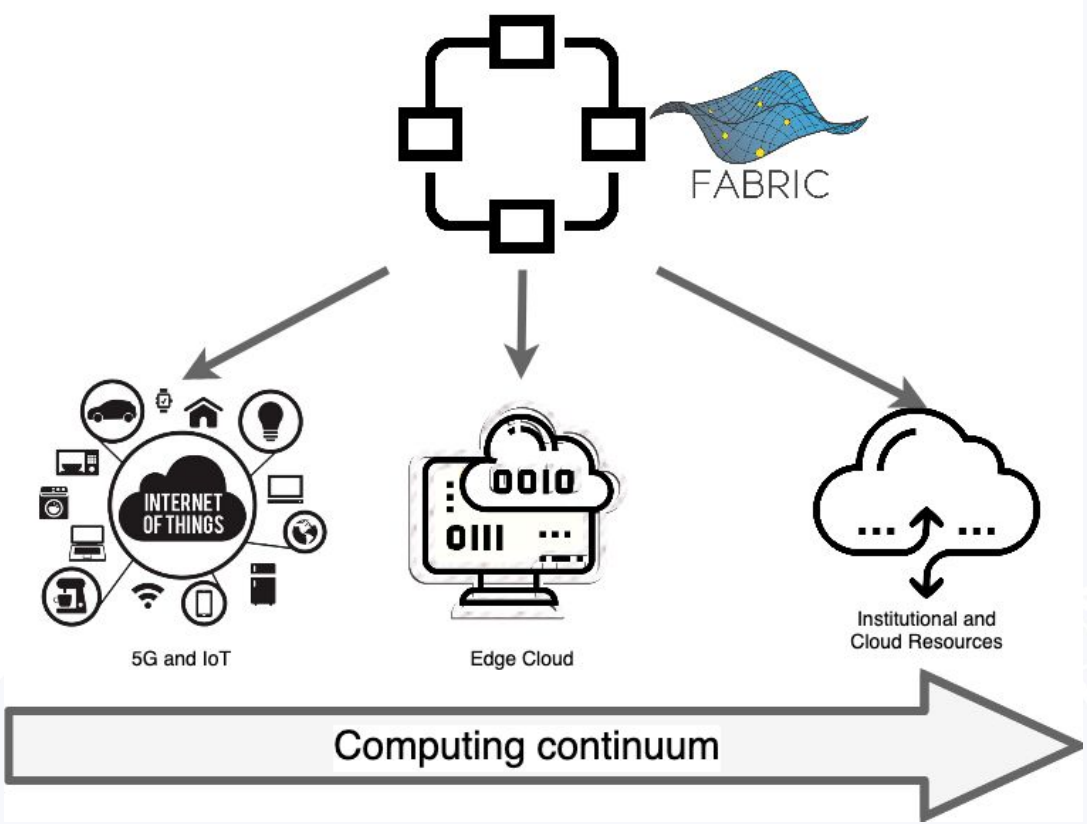
Programmable Switches Workshop, Feb. 16, 2022

Why FABRIC?

- Change in economics of compute and storage allow for the possibility that future Internet is more stateful than we've come to believe
  - "If we had to build a router from scratch today it wouldn't look like the routers we build today"
  - Explosion of capabilities in augmented computing - GPUs, FPGAs
  - Opportunity to reimagine network architecture as more stateful
- ML/AI revolution
  - Network as a 'big-data' instrument: real-time measurements + inferencing control loop
    - Network vendors have caught on to it:
      - "Self-driving network" - Juniper CTO Kireeti Kompella
  - Provisioning, cyber-security, other applications
- IoT + 5G - the new high-speed intelligent network edge
- New science applications
  - New distributed applications - data distribution, computing, storage
- A continuum of computing capabilities
  - Not just fixed points - "edge" or "public cloud"
  - Network as part of the computing substrate - computing, fusing, processing data on the fly

FABRIC

Network as part of computing continuum



FABRIC

FABRIC for everyone

FABRIC Enables New Internet and Science Applications

- Stateful network architectures, distributed applications that directly program the network

FABRIC Advances Cybersecurity

- At-scale realistic research facilitated by peering with production networks

FABRIC Integrates HPC, Wireless, and IoT

- A diverse environment connecting PAWR testbeds, NSF Clouds, HPC centers and instruments

FABRIC Integrates Machine Learning & Artificial Intelligence

- Support for in-network GPU-accelerated data analysis and control

FABRIC helps train the next generation of computer science researchers

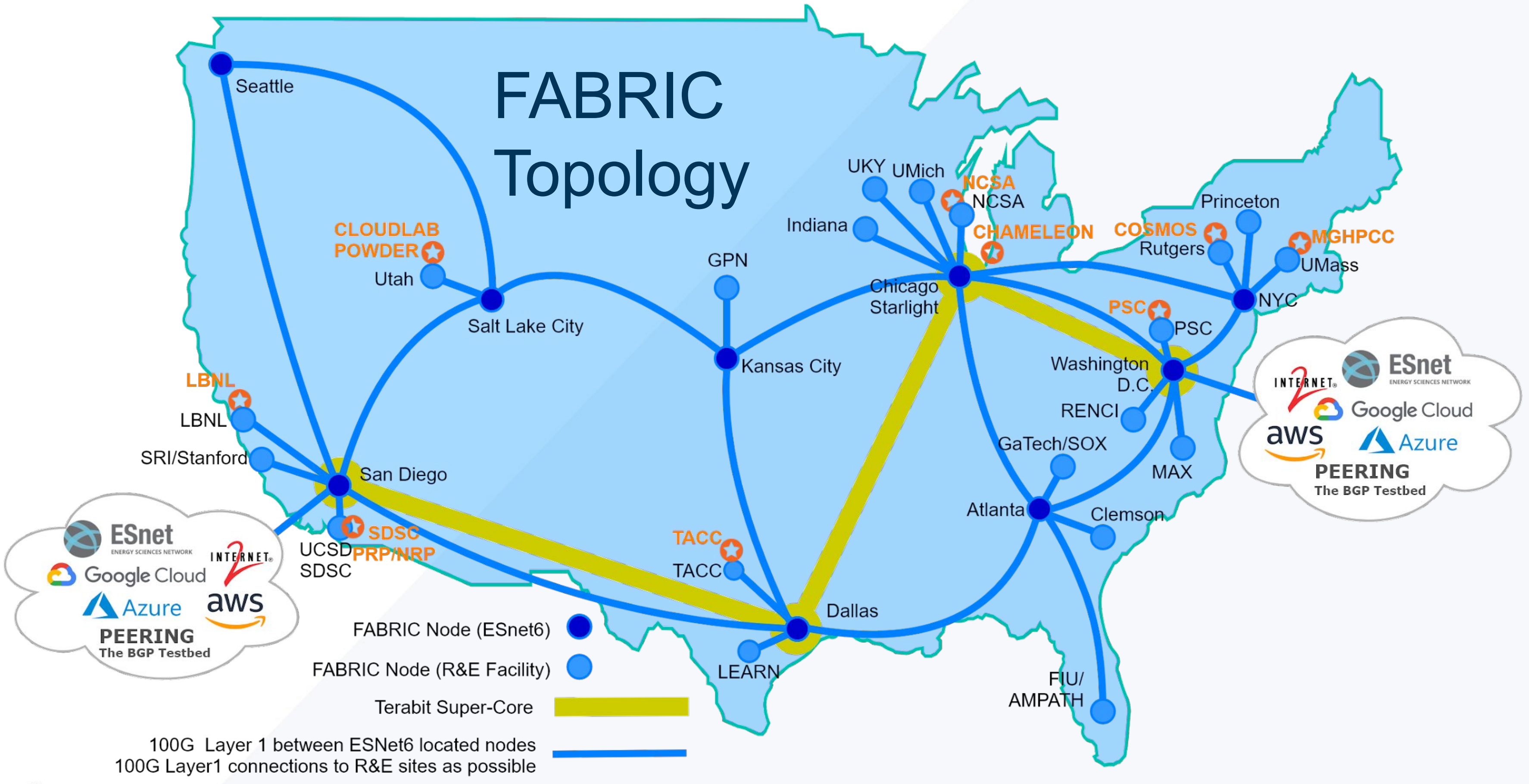
What is FABRIC?

FABRIC enables a completely *new paradigm for distributed applications and Internet protocols and services:*

- A nation-wide programmable network testbed with significant compute and storage at each node, allowing users to run computationally intensive programs and applications and protocols to maintain a lot of information in the network.
- Provides GPUs, FPGAs, and network processors (NICs) inside the network.
- Supports quality of service (QoS) using dedicated optical 100G links or dedicated capacity
- Interconnects national facilities: HPC centers, cloud & wireless testbeds, commercial clouds, the Internet, and edge nodes at universities and labs.
- Allows you to design and test applications, protocols and services that run at any node in the network, not just the edge or cloud.

FABRIC

FABRIC Topology



FABRIC

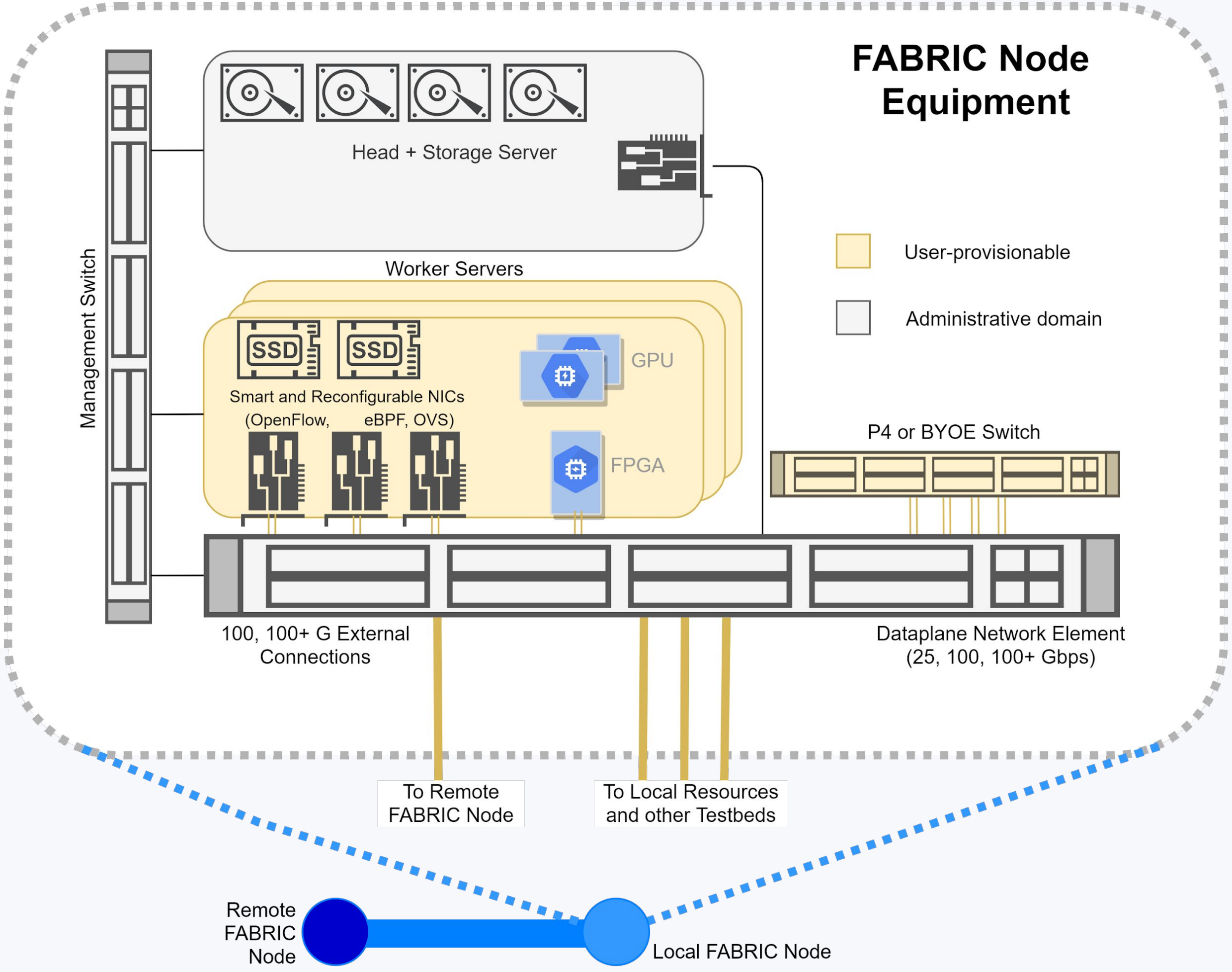
FAB

- FAB (FABRIC Across Borders) is the international expansion of FABRIC to Asia and Europe
  - Funded by NSF IRNC (International Research Network Connections) program
- Led by Anita Nikolich
- Includes sites in Japan (University of Tokyo), UK (University of Bristol), EU (University of Amsterdam and CERN)
  - To be deployed in 2021-2023 timeframe similar to the rest of FABRIC
  - Linked by available capacity on IRNC trans-oceanic links
- Brings new use-cases
  - Astronomy/Cosmology, High-Energy Physics, Urban Sensing/IoT
  - Computer Science: 5G across borders, P4/SDN, Cyber-security/Censorship evasion

FABRIC

Conceptual FABRIC Node 'Hank' Overview

a.k.a. 'A disaggregated router'



FABRIC

FABRIC Nodes

- Interpose compute and storage into the path of fast packet flows
- Rack of high-performance servers (Dell 7525) with:
  - 2x32-core AMD 7532 with 512G RAM
  - GPUs (RTX 6000 and T4), FPGA network/compute accelerators
  - Storage - experimenter provisionable 1TB NVMe drives in servers and a pool of ~250TB rotating storage at each site.
  - Network ports connect to a 100G+ switch, programmable through control software
- Reconfigurable Network Interface Cards
  - FPGAs (with P4 support)
  - Mellanox ConnectX-5 and ConnectX-6 with hardware off-load
  - Multiple interface speeds (25G, 100G, 200G+(future)
- Kernel Bypass/Hardware Offload
  - VM/Containers sized to support full-rate DPDK for access to Programmable NICs, FPGA, and GPU resources via PCI pass-through

FABRIC

FABRIC Node Design: Measurement Hardware

- GPS-disciplined clock source at most sites using PTP
  - Subject to constraints of the hosting site
- NICs capable of accurate packet sampling/timestamping
  - High touch/ sampling story
- Programmable port mirroring
- Smart PDUs to measure power
- Optical layer measurements (where available)
- CPU, memory, disk, port/interface utilization and other time-series (software)

FABRIC

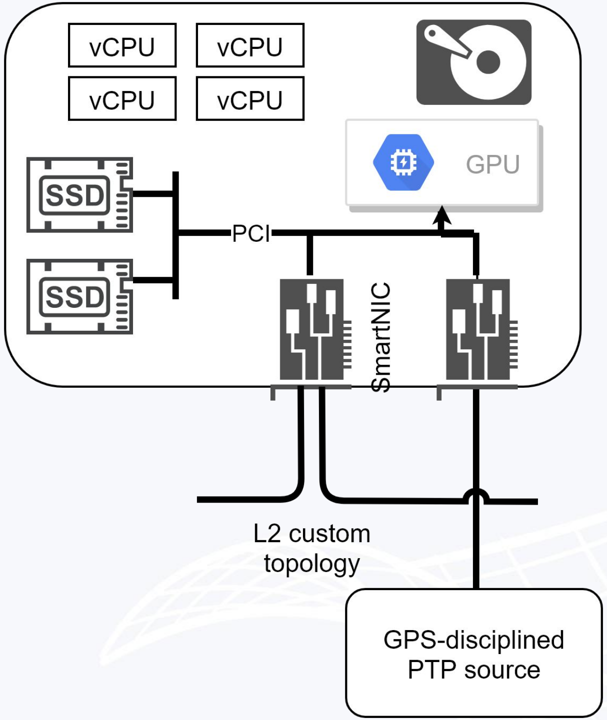
FABRIC Experiment building blocks

- Each experiment is encapsulated in a slice - a topology
- Slices consist of slivers
  - Individually programmable or configurable resources
- Slices can change over time
  - Grow or shrink, adding or shedding resources under programmatic control
- Slice topologies can be
  - Custom L2 using underlying MPLS-SR
  - Rely on persistent routable IPv6 layer in FABRIC
- Basic sliver classes
  - Nodes - can include a selection of PCI-passthrough devices
  - Links - L2 or L3 with QoS and without
  - Measurement points - inside and outside the slice

FABRIC

Bump-in-wire sliver

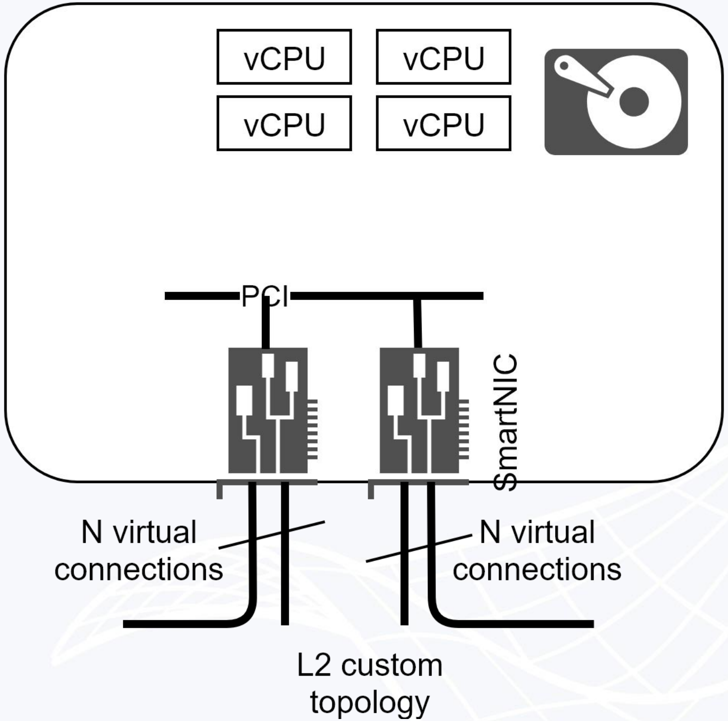
- Useful for collecting and analysing high-volume packet traces
  - Rely on NVMe drive for high-throughput local storage
  - Use GPU to assist in analysis
- Can optionally use a local GPS-disciplined PTP source to achieve millisecond-level accuracy for measurements
  - Multiple ‘bumps-in-wire’ in a slice can help create a snapshot of traffic across the network in a given instant in time



FABRIC

SmartNIC router sliver

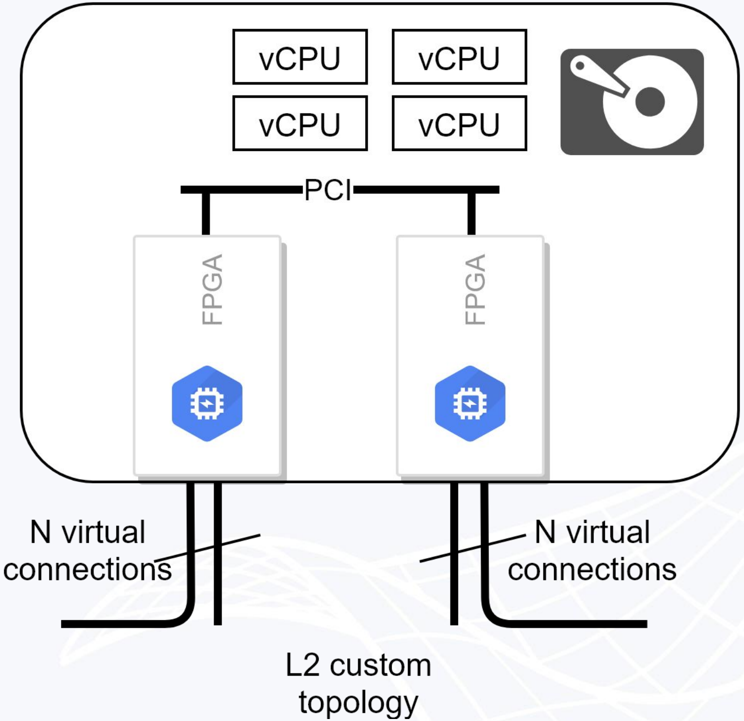
- Can create an small-port-count OpenFlow router with hardware acceleration via Mellanox ConnectX-[5,6] cards
  - Direct access to PCI allows to bypass CPU in many cases.



FABRIC

FPGA or P4 router sliver

- Uses Xilinx FPGAs in a node
- Can build a small port-count FPGA router
- With additional tools support can also serve as a P4 router built on top of the FPGA
- Can route between multiple virtual connections based on e.g. VLAN tags or other header information
-



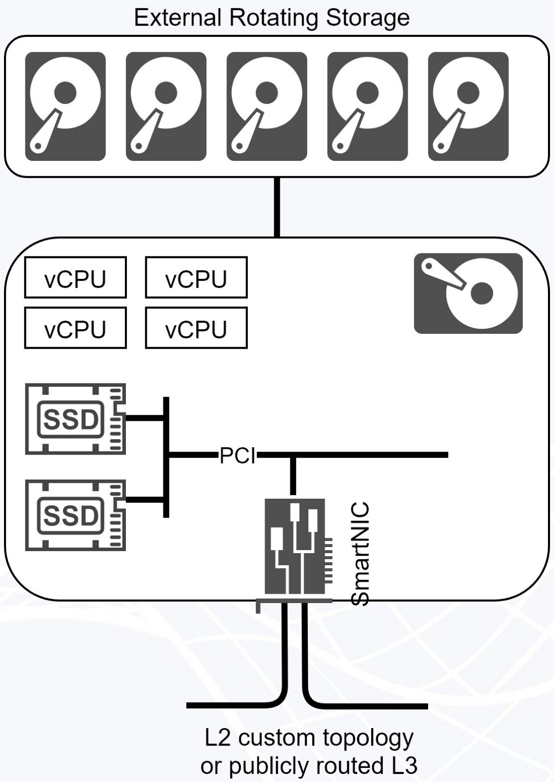
FABRIC

Caching/processing with tiered storage

- Collect in-network measurement data and store using different storage tiers:
  - RAM
  - Attached NVMe drive
  - Local rotating storage
  - External (local to the site) large volume rotating storage

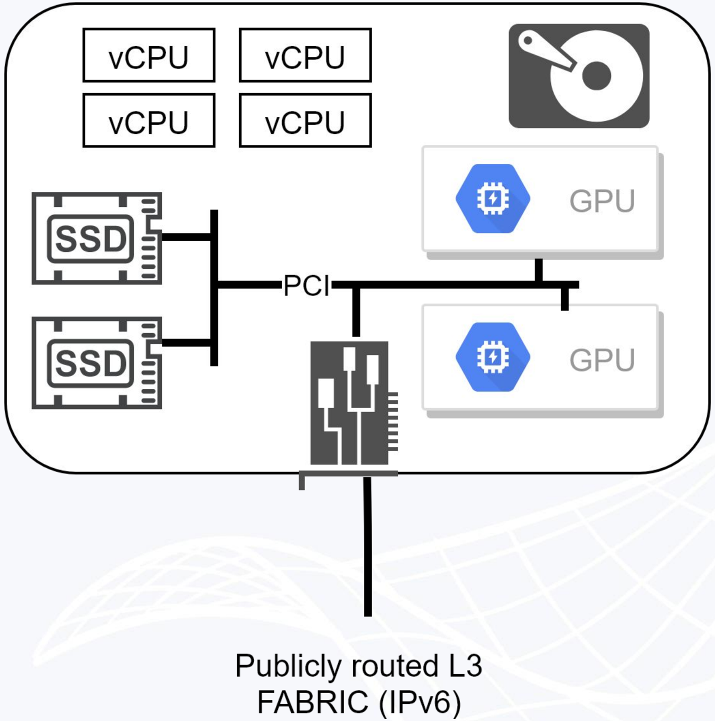
Speed

Available Size



In-network AI/ML

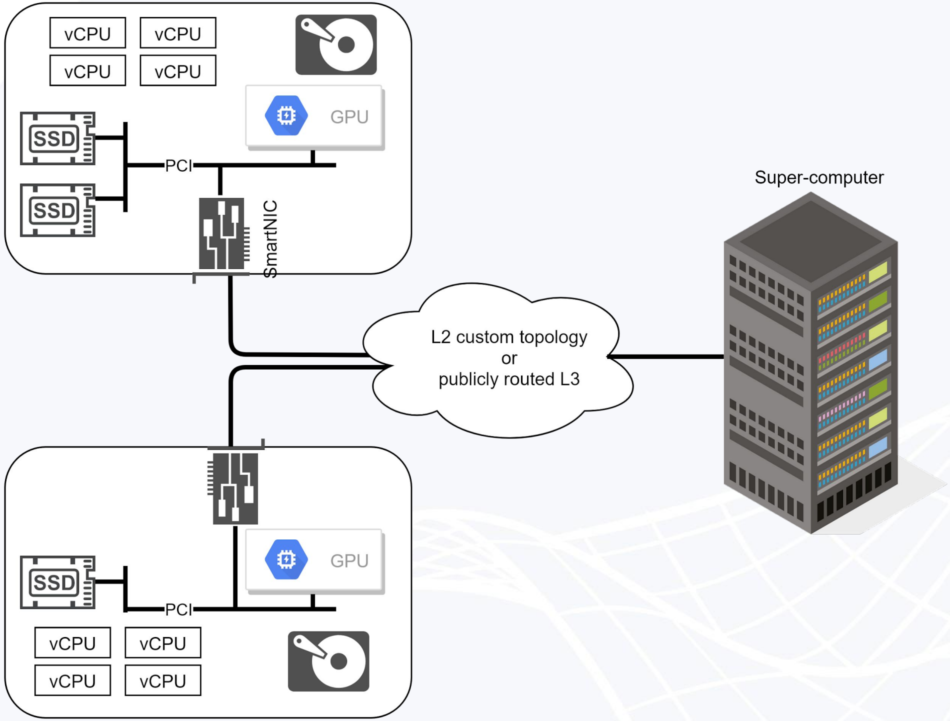
- Investigating autonomous network behavior using in-network GPU support
  - Using RTX6000 for learning and inference using streaming data
- Perform intelligent data fusion/processing in the network
- Implement in-network analytics/security functions



FABRIC

Attaching external facilities

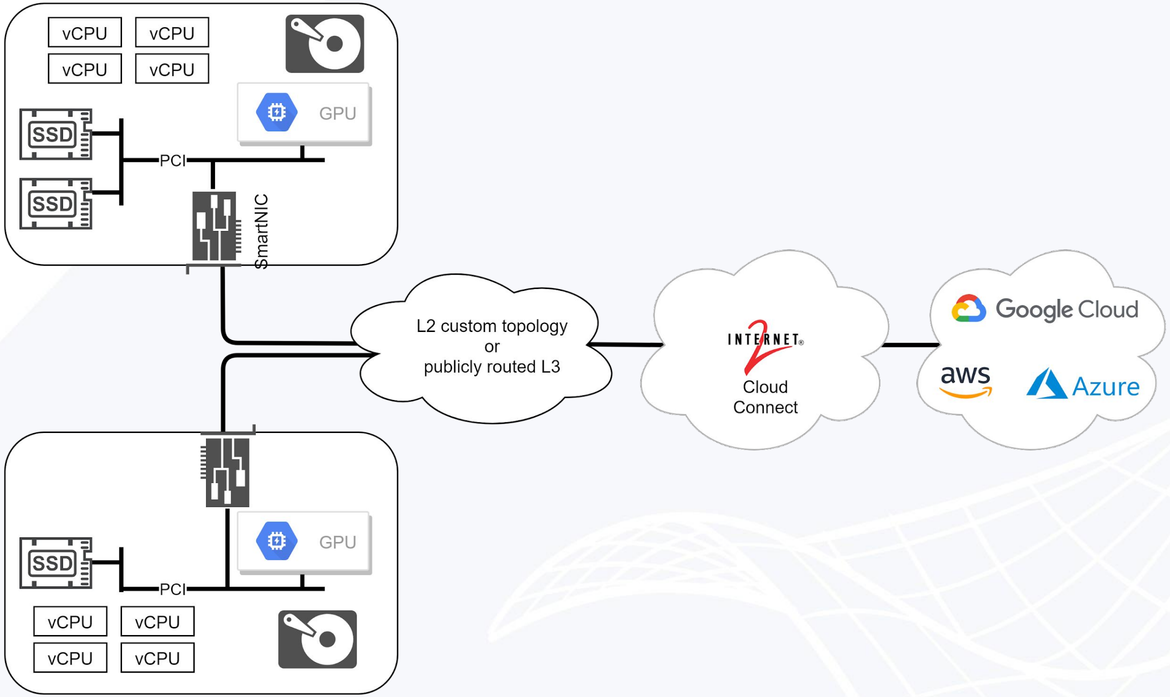
- The US NSF has made significant investments in scientific CI
- Future networks must better support domain science needs
- FABRIC connects to a number of facilities and testbeds to enrich the set of resources that can be used in experiments
  - Supercomputing centers (PSC, NCSA, SDSC, TACC, MGHPCC)
  - Cloud testbeds - CloudLab, Chameleon, Open Cloud Testbed
  - 5G testbeds - COSMOS, Powder
- Through FAB we will also reach
  - University of Bristol, University of Amsterdam, University of Tokyo, CERN



FABRIC

Using public clouds in experiments

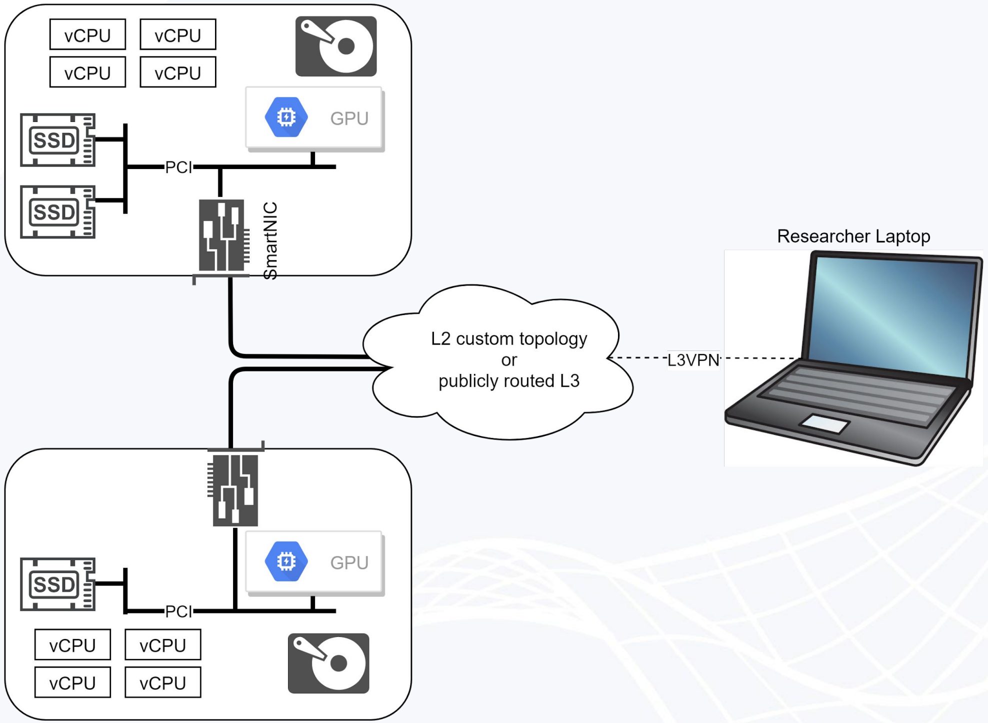
- Future networks will connect clouds and their customers
- 5G+Cloud experiments
- Through partnership with Internet 2 FABRIC will provide connectivity to commercial clouds
    - Utilize I2 CloudConnect system



FABRIC

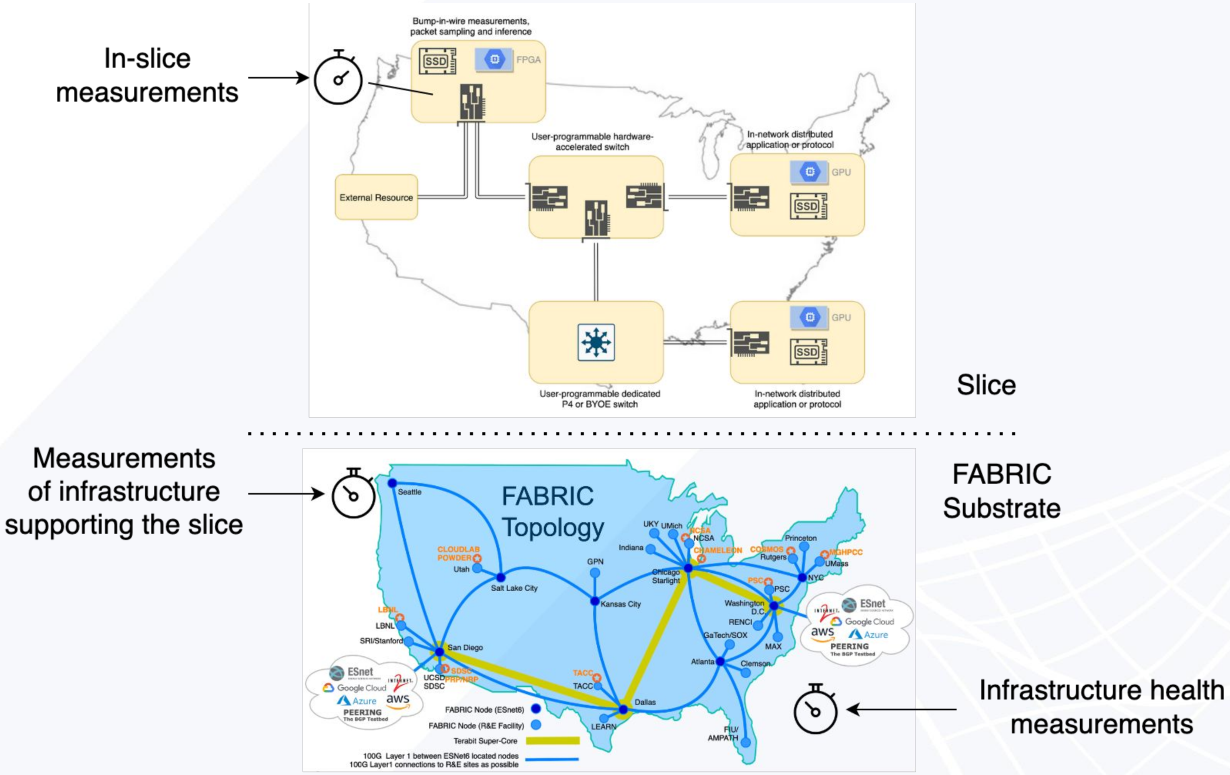
Adding experimenter-owned resources

- Many experimenters may be interested in connecting their own resources to their slice topologies
  - FABRIC may not be able to reach every campus with a dedicated connection
- VPN/VPW options will be available to support these cases.
  - Allow experimenters to offer services to others from their slices



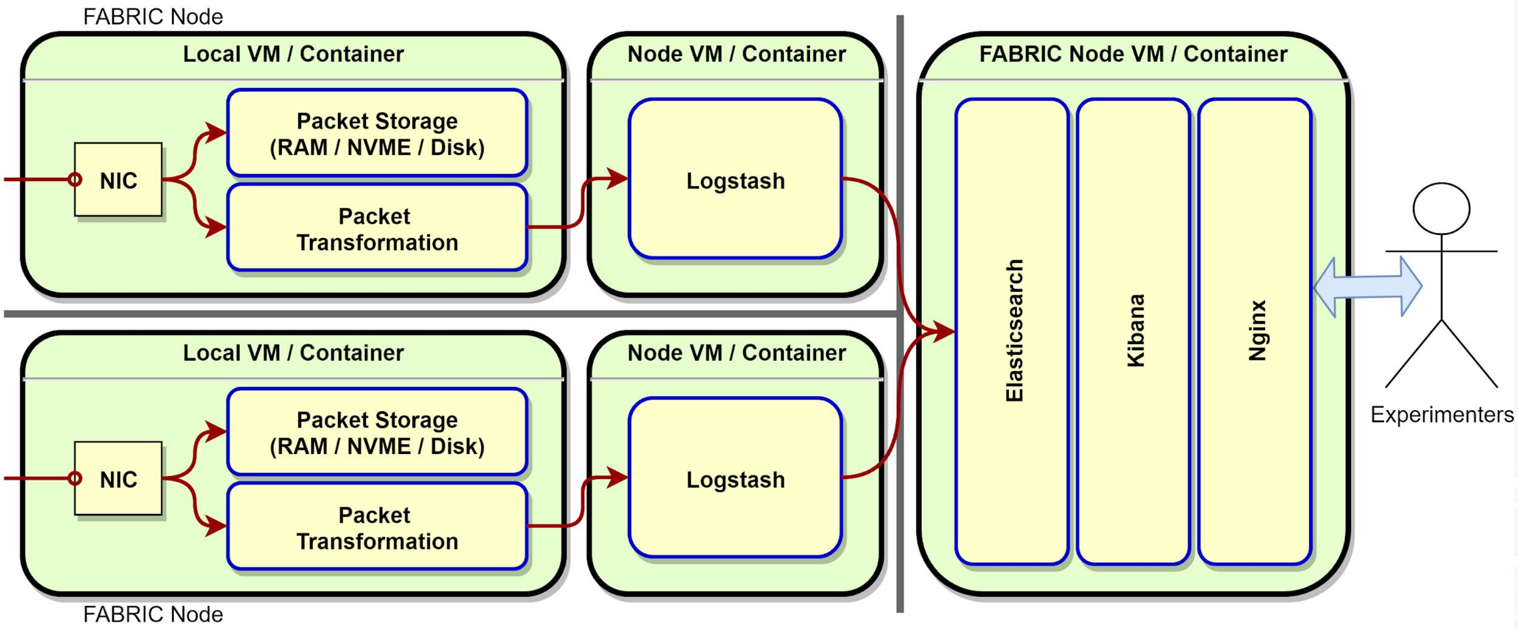
FABRIC Measurement Capabilities

- Key to FABRIC being a scientific instrument
- Provides measurements
  - Inside the slice
  - Outside the slice



FABRIC

FABRIC Measurement Data Processing/Analysis

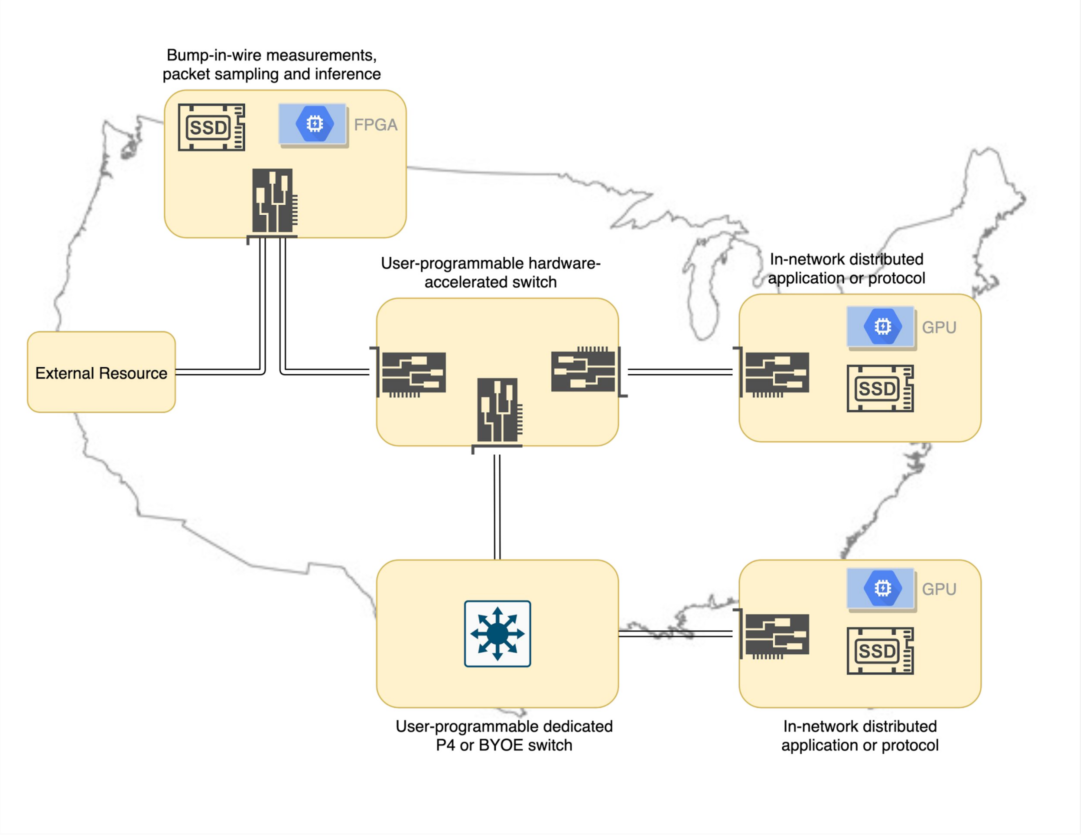


FABRIC

Enabling P4 on FABRIC

Examples of potential uses:

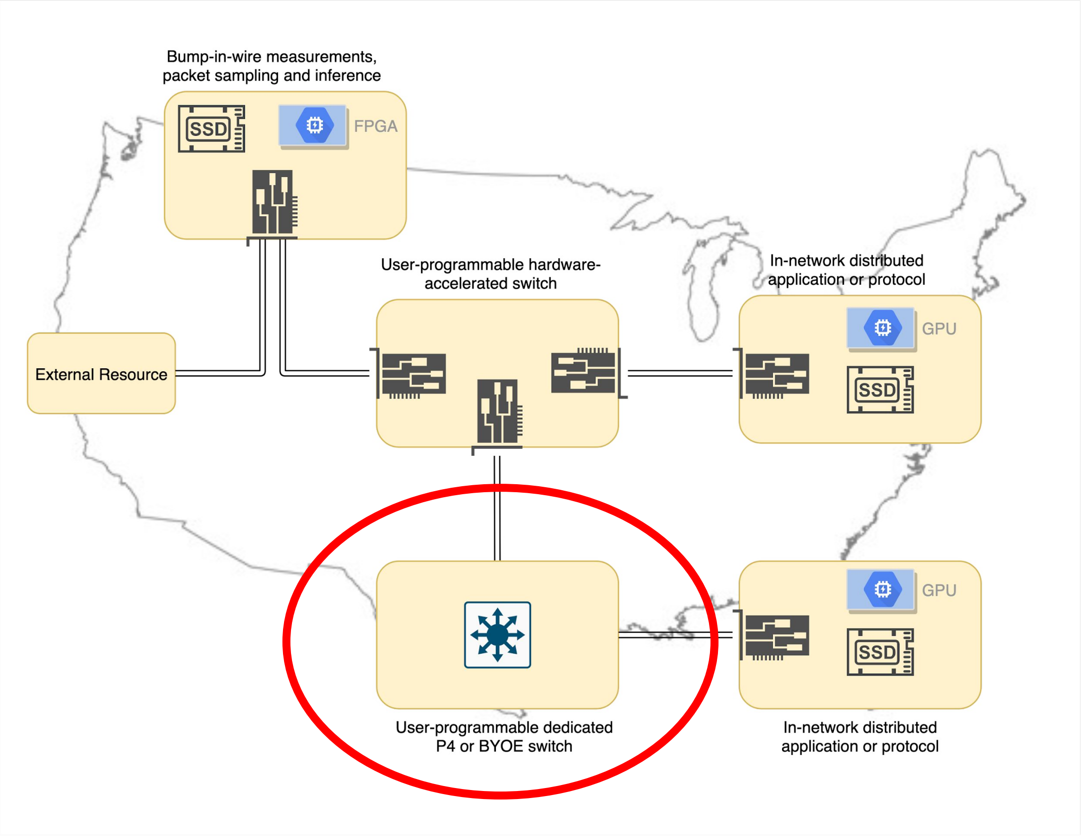
- Bump-in-wire measurements and packet sampling at high bit rates (25, 40, 100, 100+ Gbps)
- Hardware-accelerated switching using Smart NICs, FPGA NICs or P4 switches in individual nodes
- Hosting in-network applications and stateful architectures using a combination of storage and compute resources in individual nodes
- In-network inference, other types of accelerated computing via FPGAs and GPUs
- Connect experiments to external facilities like IoT, 5G, cloud testbeds, public clouds and HPC resources.
- Deploy non-IP protocols on top of wide-area L2 topologies, that may include in-network processing and storage

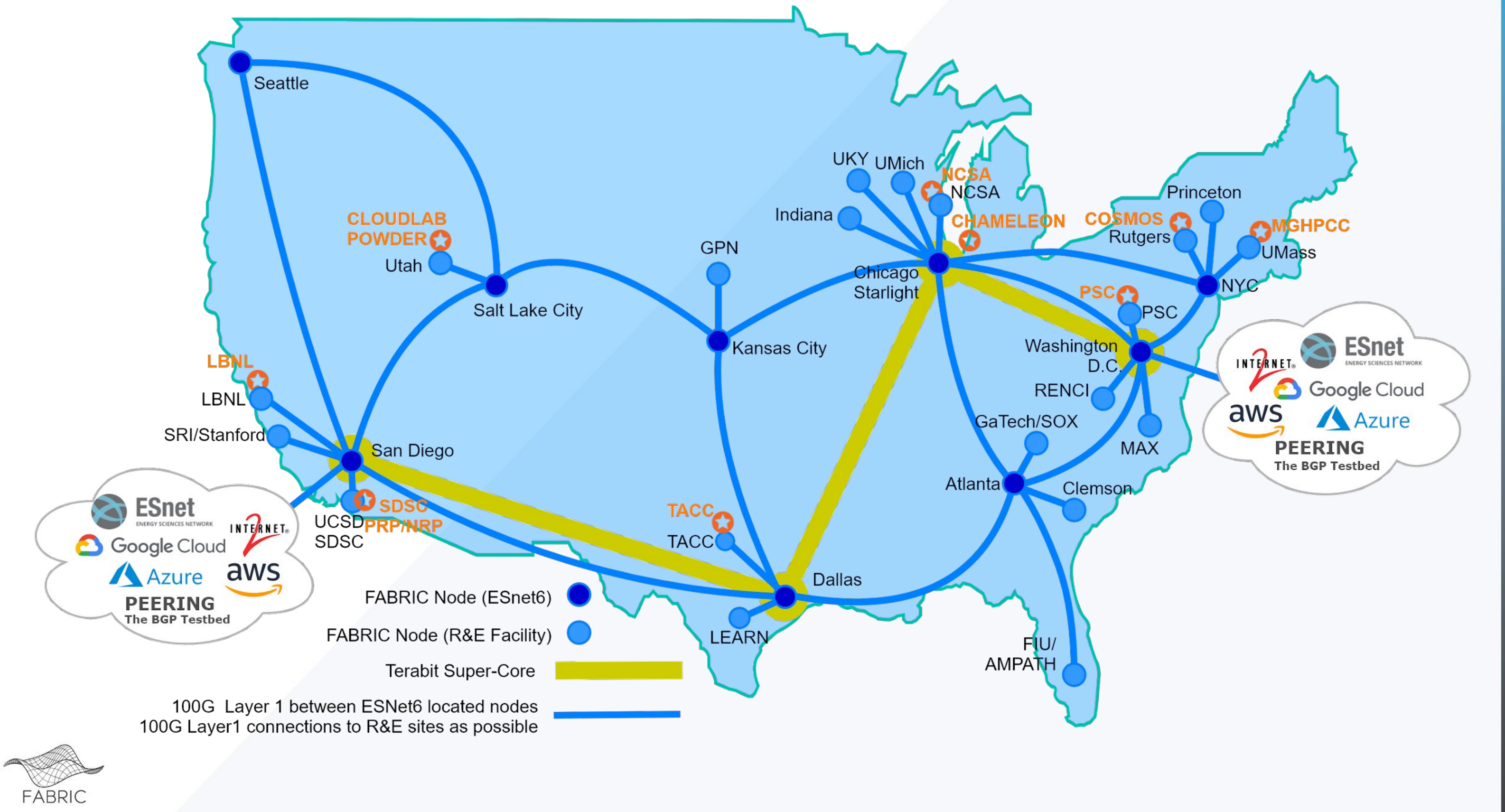


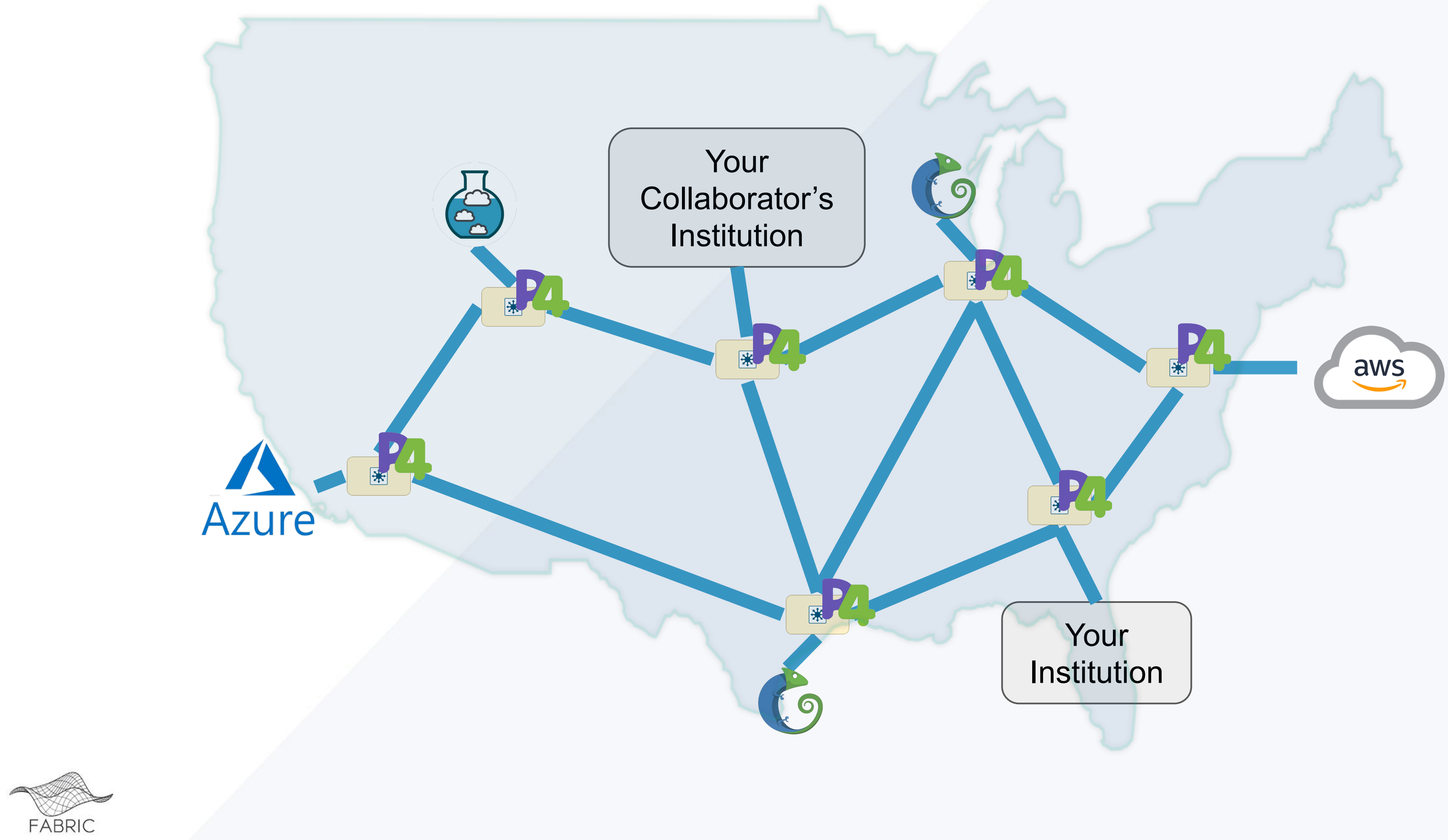
Enabling P4 on FABRIC

Examples of potential uses:

- Bump-in-wire measurements and packet sampling at high bit rates (25, 40, 100, 100+ Gbps)
- Hardware-accelerated switching using Smart NICs, FPGA NICs or P4 switches in individual nodes
- Hosting in-network applications and stateful architectures using a combination of storage and compute resources in individual nodes
- In-network inference, other types of accelerated computing via FPGAs and GPUs
- Connect experiments to external facilities like IoT, 5G, cloud testbeds, public clouds and HPC resources.
- Deploy non-IP protocols on top of wide-area L2 topologies, that may include in-network processing and storage

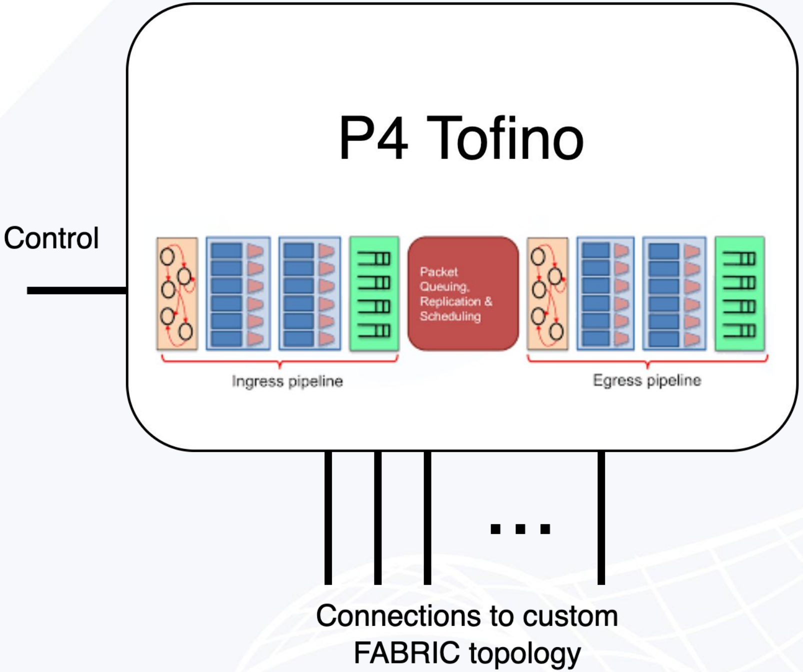



Seattle
CLOUDLAB
POWDER
Utah
Salt Lake City
GPN
Kansas City
UKY
UMich
NCSA
NCSA
Indiana
CHAMELEON
Chicago
Starlight
COSMOS
Rutgers
Princeton
MGHPCC
UMass
NYC
PSC
PSC
Washington
D.C.
RENCI
GaTech/SOX
MAX
Atlanta
Clemson
FIU/
AMPATH
LBNL
LBNL
SRI/Stanford
San Diego
UCSD
SDSC
SDSC
PRP/NRP
TACC
TACC
Dallas
LEARN
ESnet
ENERGY SCIENCES NETWORK
INTERNET2
Google Cloud
Azure
aws
PEERING
The BGP Testbed
FABRIC Node (ESnet6)
FABRIC Node (R&E Facility)
Terabit Super-Core
100G Layer 1 between ESNet6 located nodes
100G Layer1 connections to R&E sites as possible
FABRIC


Your
Collaborator's
Institution
Azure
aws
Your
Institution
FABRIC

P4 Resources: Tofino (Not yet available)

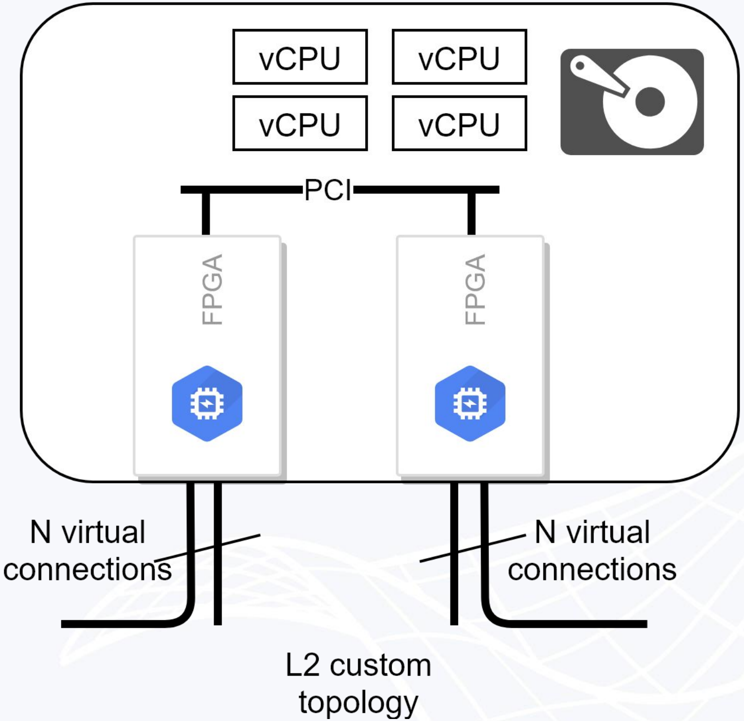
- Dedicated Tofino switches controlled by the user
- Initial hardware is being acquired
- Licensing and NDAs complicate sharing with arbitrary users.
  - User workflow is being designed



FABRIC

P4 Resources: Xilinx FPGA (Not yet available)

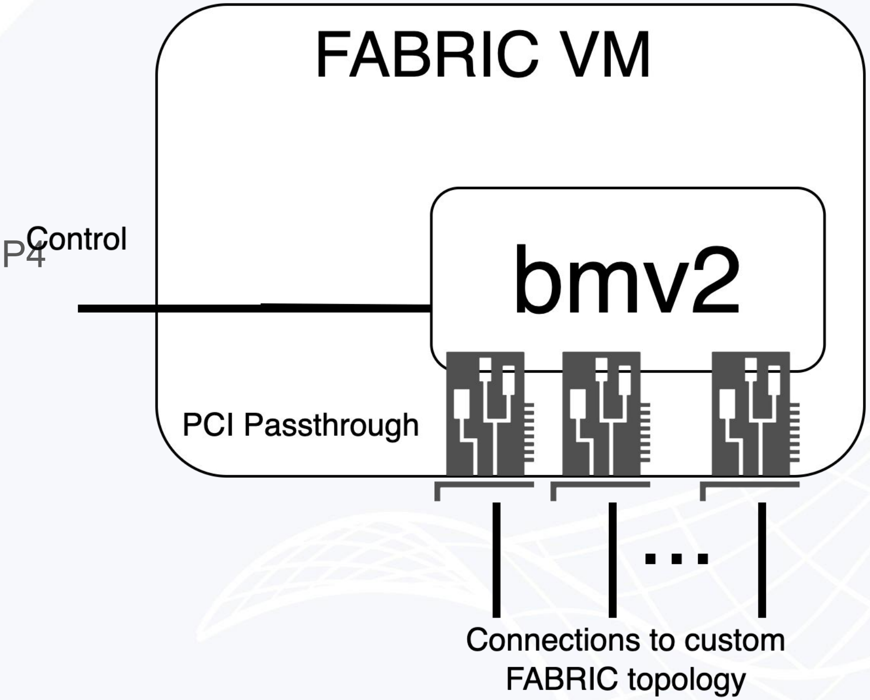
- Uses Xilinx FPGAs in a node
- Can build a small port-count FPGA router
- With additional tools support can also serve as a P4 router built on top of the FPGA
- Xilinx P4-SDNet extensions
- FABRIC P4 bit code currently under development at Northeastern (Miriam Leeser)



FABRIC

P4 Resources: Software P4 (available)

- P4 Behavioral Model (BMV2)
  - Opensource
  - Jupyter notebook available
- Tofino Native Architecture (possible)
  - User must provide Tofino Model and Intel P4 Studio
  - Intel NDA Required
- Primarily for education and development

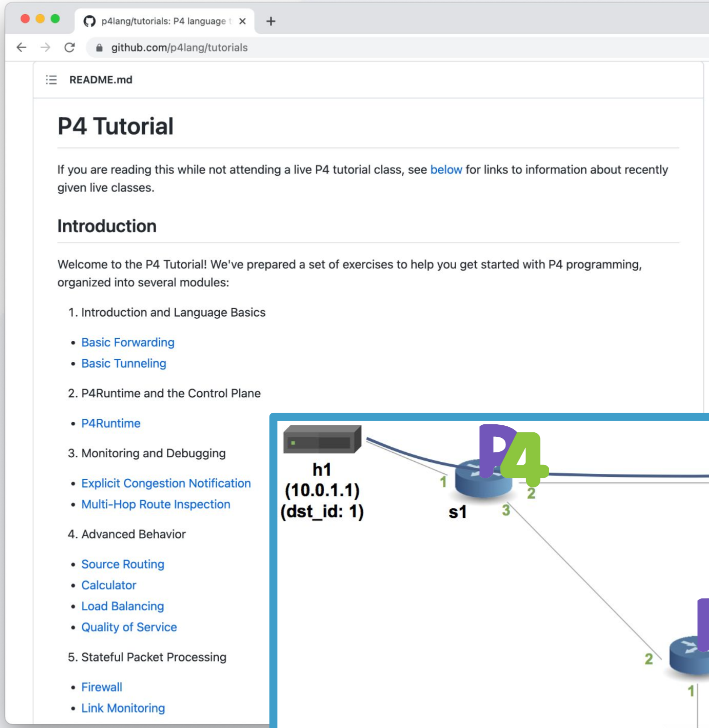


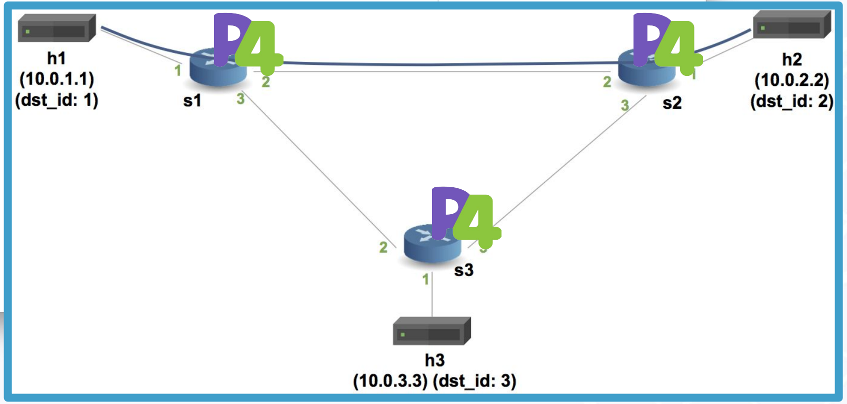
FABRIC

https://github.com/p4lang/tutorials

Example: P4Lang Tutorials

- P4Lang Tutorials
  - Great place to start with P4
  - Hands-on exercises
  - Github repository
- Resources and Topology
  - One virtual machine
  - Mininet



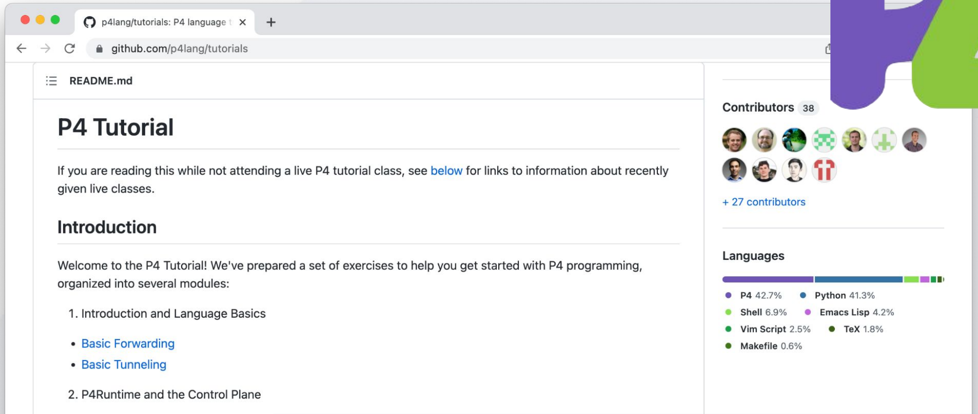


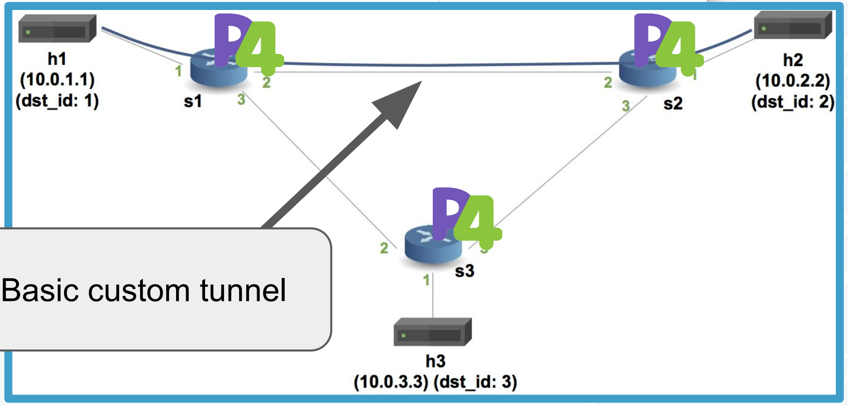
FABRIC

Example: P4Lang Tutorials

https://github.com/p4lang/tutorials

- P4Lang Tutorials
  - Great place to start with P4
  - Hands-on exercises
  - Github repository
- Resources and Topology
  - One virtual machine
  - Mininet



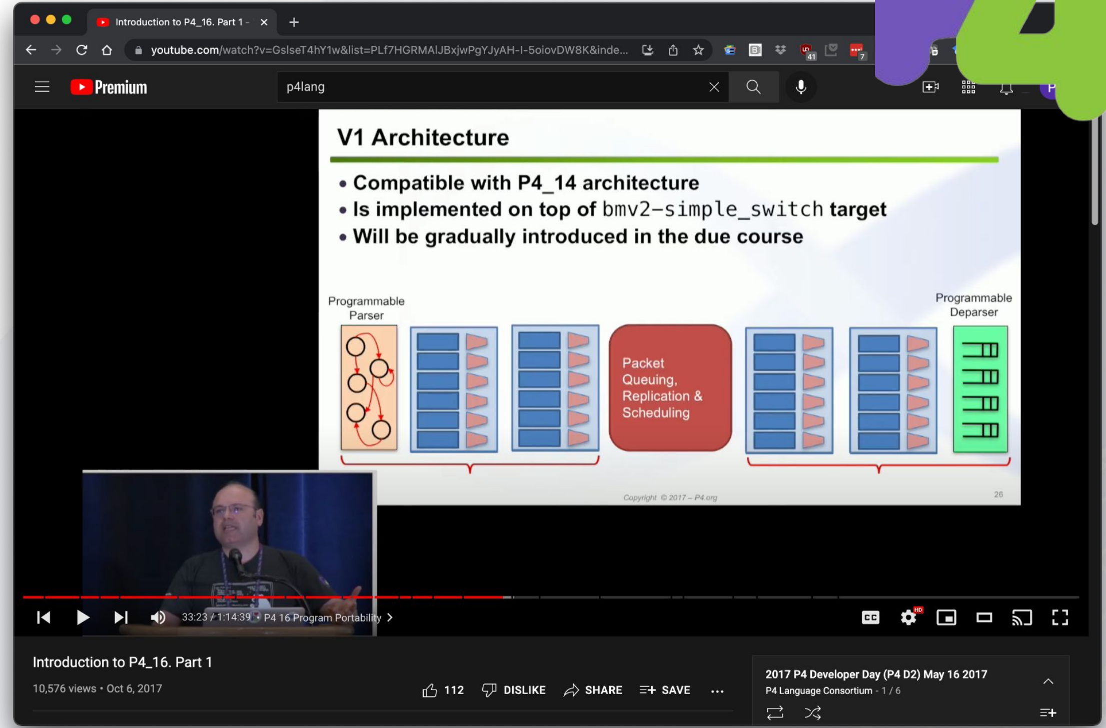


Exercise: Basic custom tunnel

https://github.com/p4lang/tutorials

Example: P4Lang Tutorials

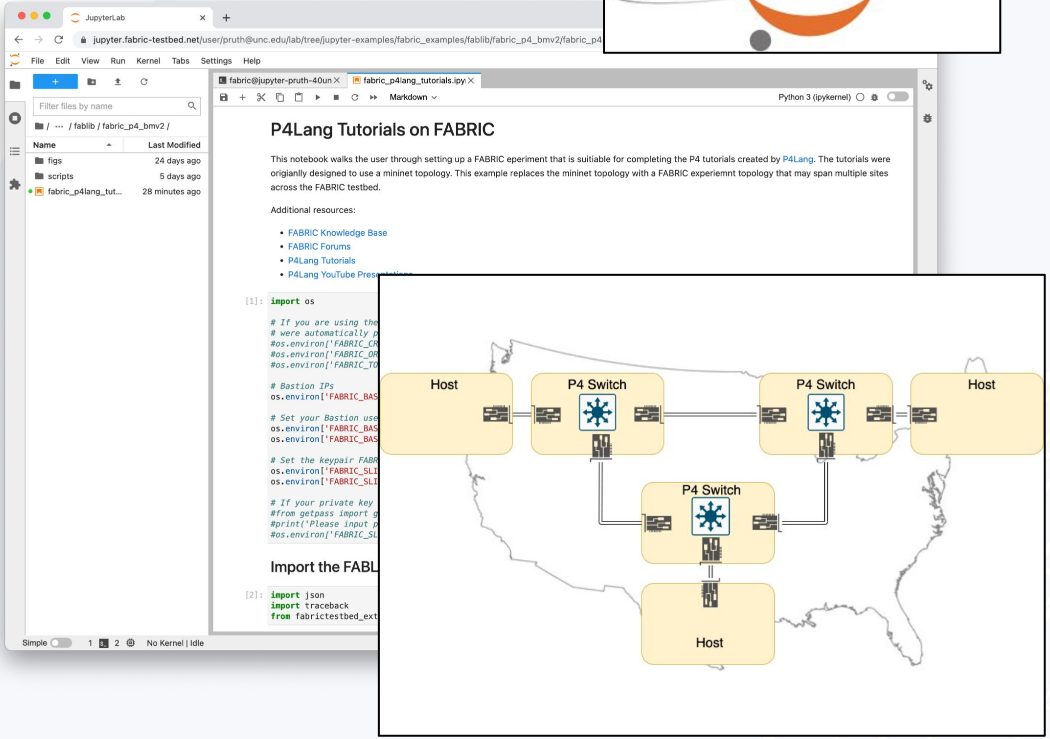
Great videos on YouTube featuring your next speaker! (Vladimir Gurevich)



FABRIC

Example: P4Lang on FABRIC

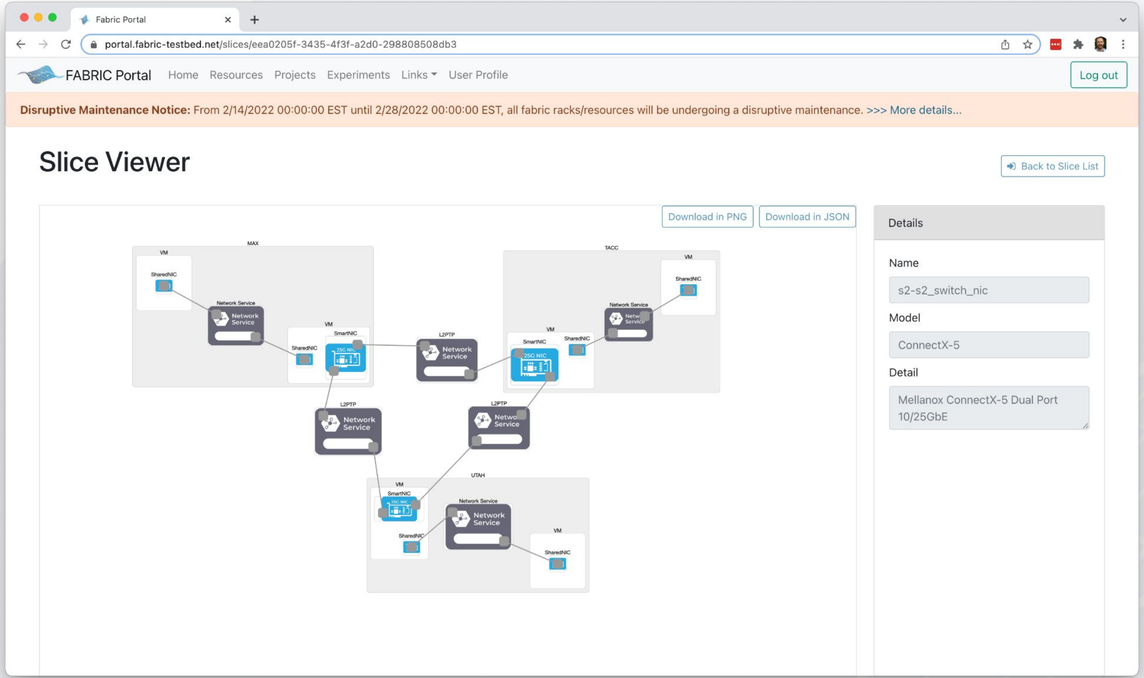
- P4Lang Tutorials
  - Same tutorials
  - JupyterHub notebook using FABlib library
- Resources and Topology
  - Individual BMV2 switches
  - Uses any FABRIC site
  - Dedicated WAN L2 links

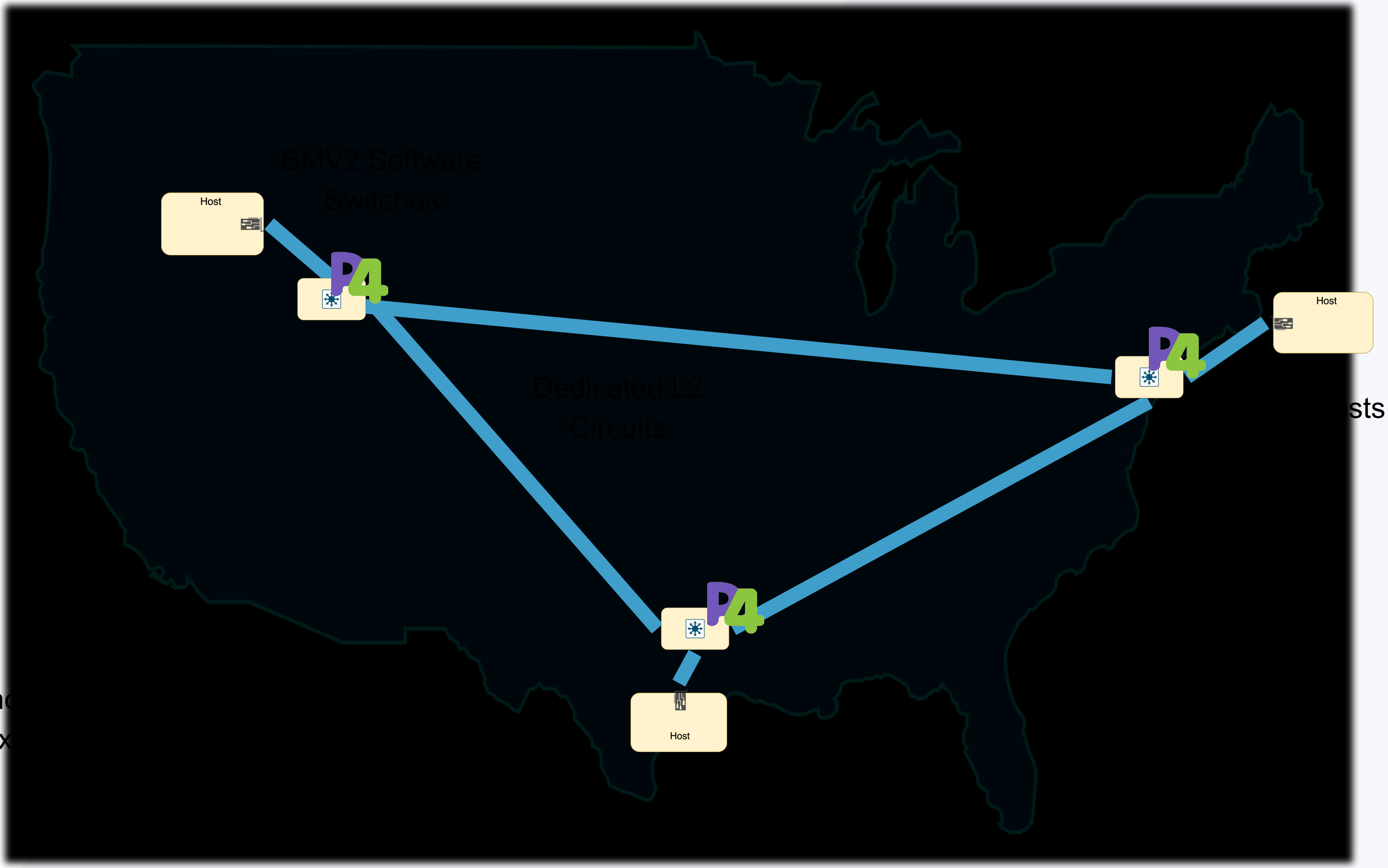


Goal: Replace Mininet with FABRIC resources

FABRIC: P4Lang Tutorials

- Logical View from FABRIC Portal
- Sites
  - TACC: Texas Advanced Computing Center (Austin)
  - MAX: Mid-Atlantic Crossroads (U. Maryland)
  - UTAH: (U. Utah)

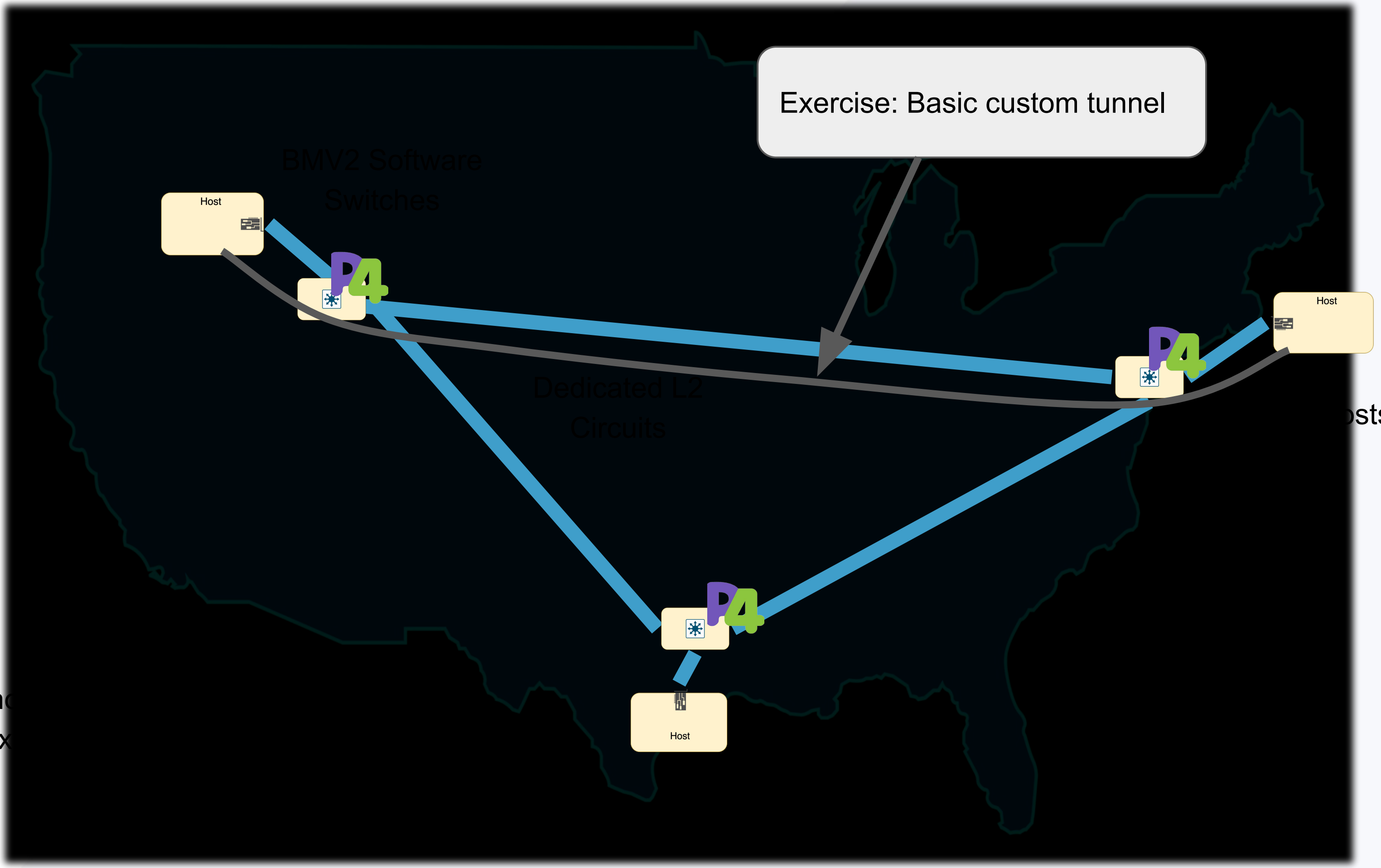




sts

- Tofin
- Xilinx

FABRIC



osts

- Tofin
- Xilinx

Summary: P4 Experiments on FABRIC

- User controlled P4 switches in the network core
- Dedicated L2 circuits
- Hardware P4 (Tofino and Xilinx) coming soon
- Software P4 (BMV2 and Tofino) possible now
- Jupyter notebooks to streamline deployment

FABRIC

Thank You!

Questions?

Visit https://fabric-testbed.net

Learn more, and Join the Forum at https://learn.fabric-testbed.net

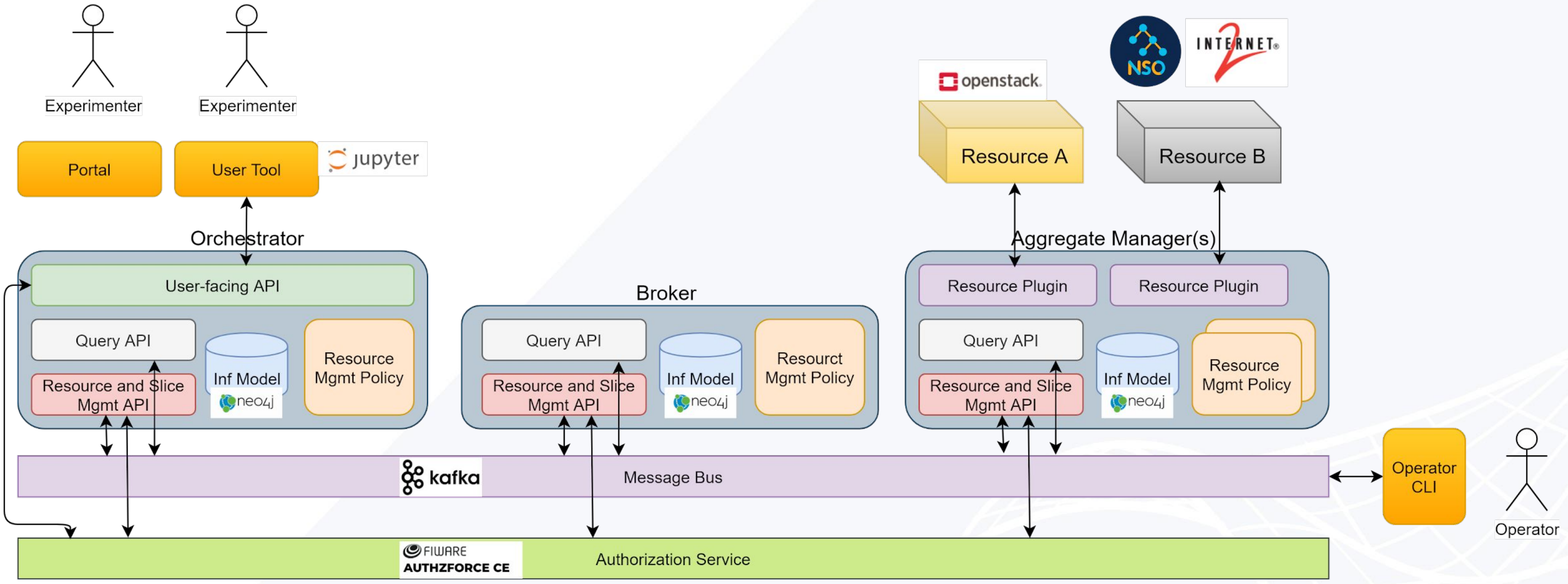
Ask info@fabric-testbed.net

FABRIC Software: https://github.com/fabric-testbed

This work is funded by
NSF grant CNS-1935966

FABRIC

Control Framework (CF) Components

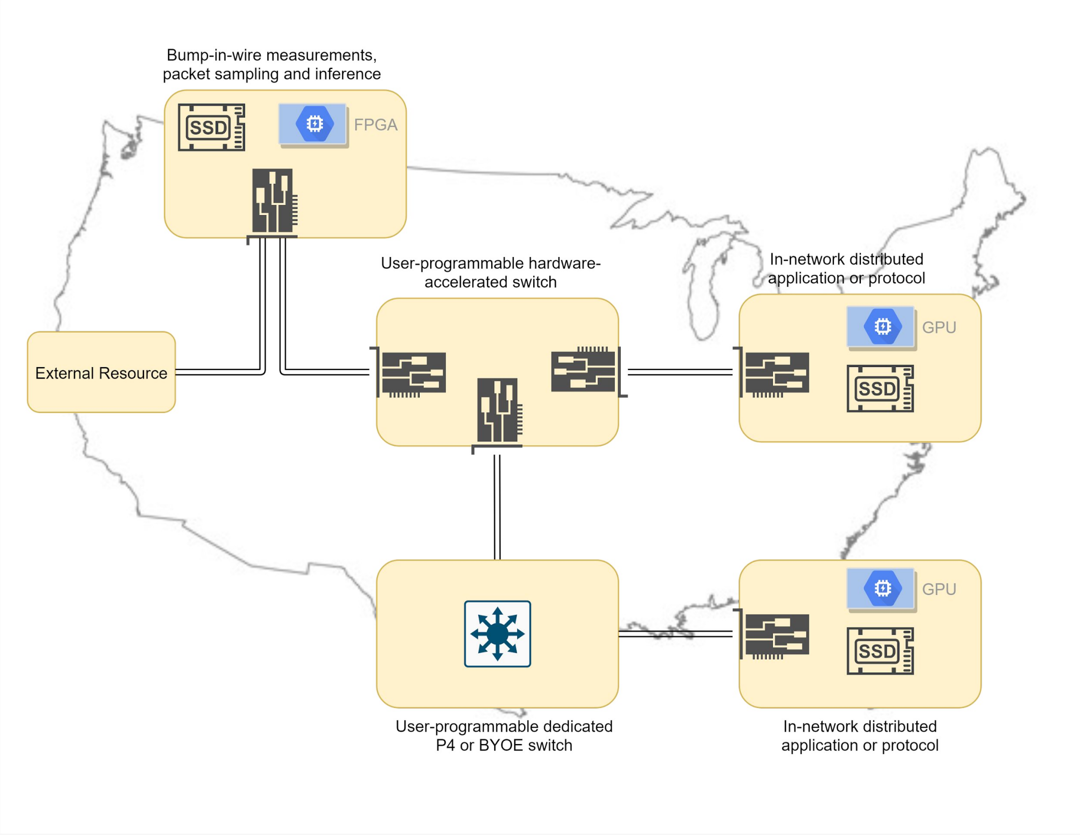


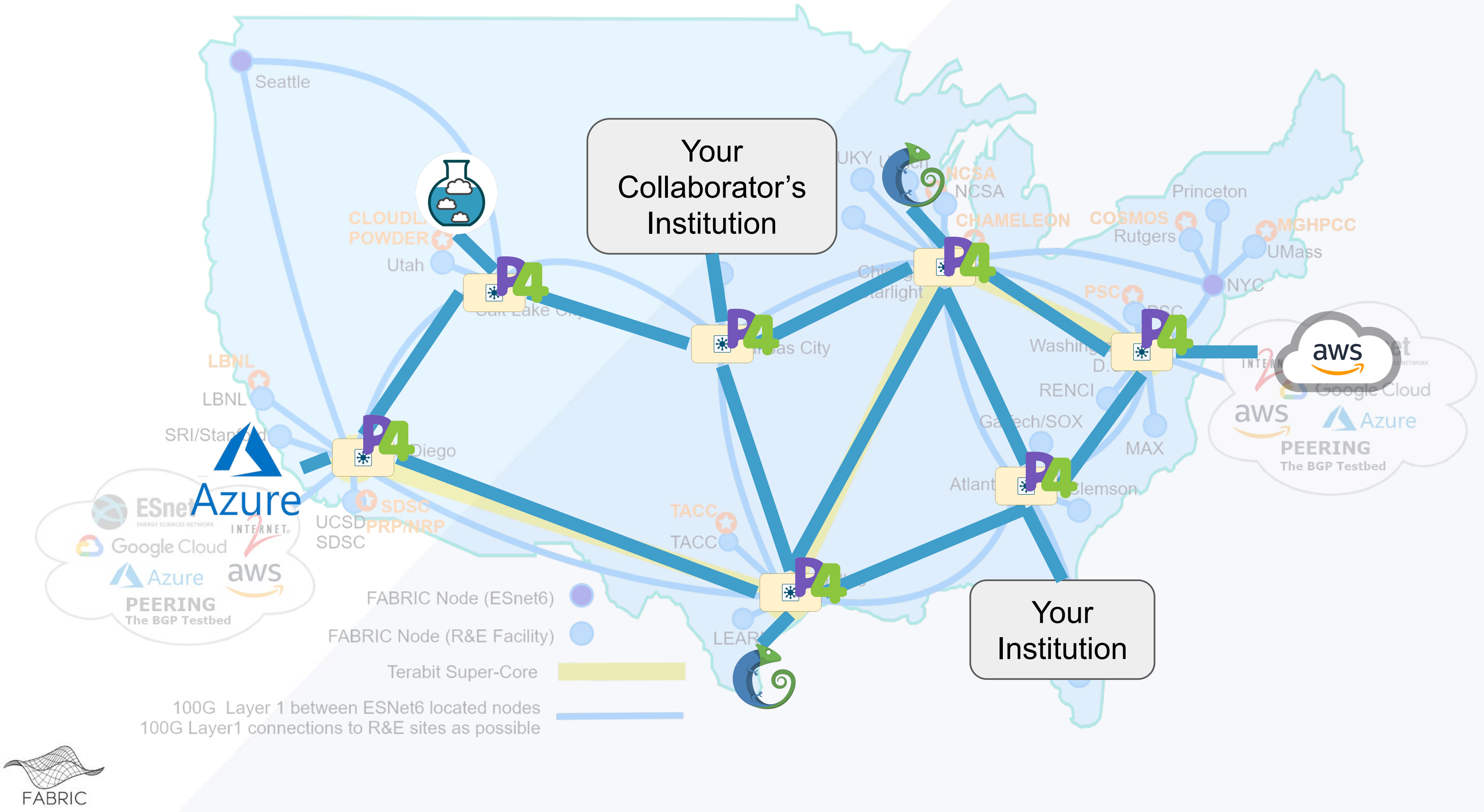
FABRIC

Example FABRIC Use-case Scenarios

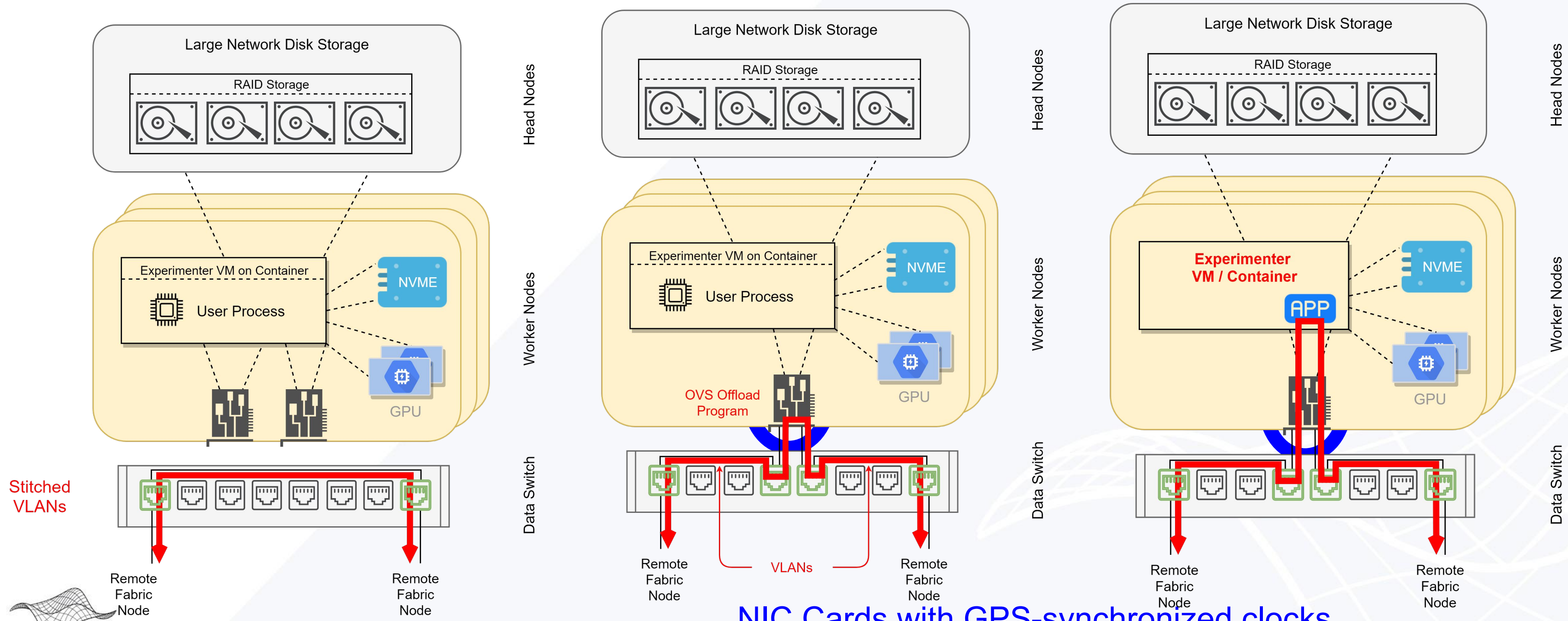
Examples of potential uses:

- Bump-in-wire measurements and packet sampling at high bit rates (25, 40, 100, 100+ Gbps)
- Hardware-accelerated switching using Smart NICs, FPGA NICs or P4 switches in individual nodes
- Hosting in-network applications and stateful architectures using a combination of storage and compute resources in individual nodes
- In-network inference, other types of accelerated computing via FPGAs and GPUs
- Connect experiments to external facilities like IoT, 5G, cloud testbeds, public clouds and HPC resources.
- Deploy non-IP protocols on top of wide-area L2 topologies, that may include in-network processing and storage




Your Collaborator's Institution
Your Institution
Seattle
CLOUDLAB
POWDER
Utah
Salt Lake City
Kansas City
UKY
NCSA
CHAMELEON
COSMOS
Rutgers
Princeton
MGHPCC
UMass
NYC
Starlight
PSC
Washington D.C.
RENCI
GaTech/SOX
MAX
Atlanta
Clemson
LBNL
SRI/Stanford
San Diego
SDSC
UCSD
PRP/NRP
TACC
LEARN
Azure
ESnet
ENERGY SCIENCES NETWORK
INTERNET2
Google Cloud
aws
PEERING
The BGP Testbed
FABRIC Node (ESnet6)
FABRIC Node (R&E Facility)
Terabit Super-Core
100G Layer 1 between ESNet6 located nodes
100G Layer1 connections to R&E sites as possible
FABRIC

FABRIC Use Cases (Revisited)



NIC Cards with GPS-synchronized clocks capture and accurately (ms) timestamp packets

FABRIC