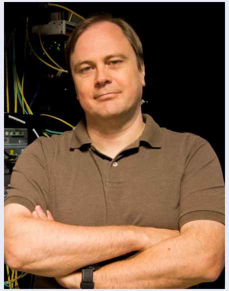
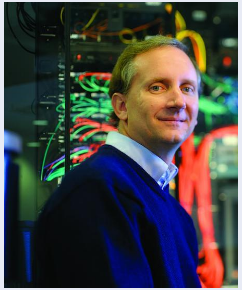
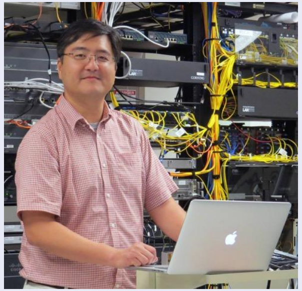
FABRIC

FABRIC: Adaptive Programmable Research Infrastructure for Computer Science and Science Applications

Ilya Baldin, Inder Monga

SC20, XNET, November 13, 2020

FABRIC Leadership Team

Ilya Baldin (RENCI)	Anita Nikolich (UIUC)	Inder Monga (ESnet)	Jim Griffioen (UKY)
			
KC Wang (Clemson)	Tom Lehman (Virnao)	Paul Ruth (RENCI)	Zongming Fei (UKY)
			

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

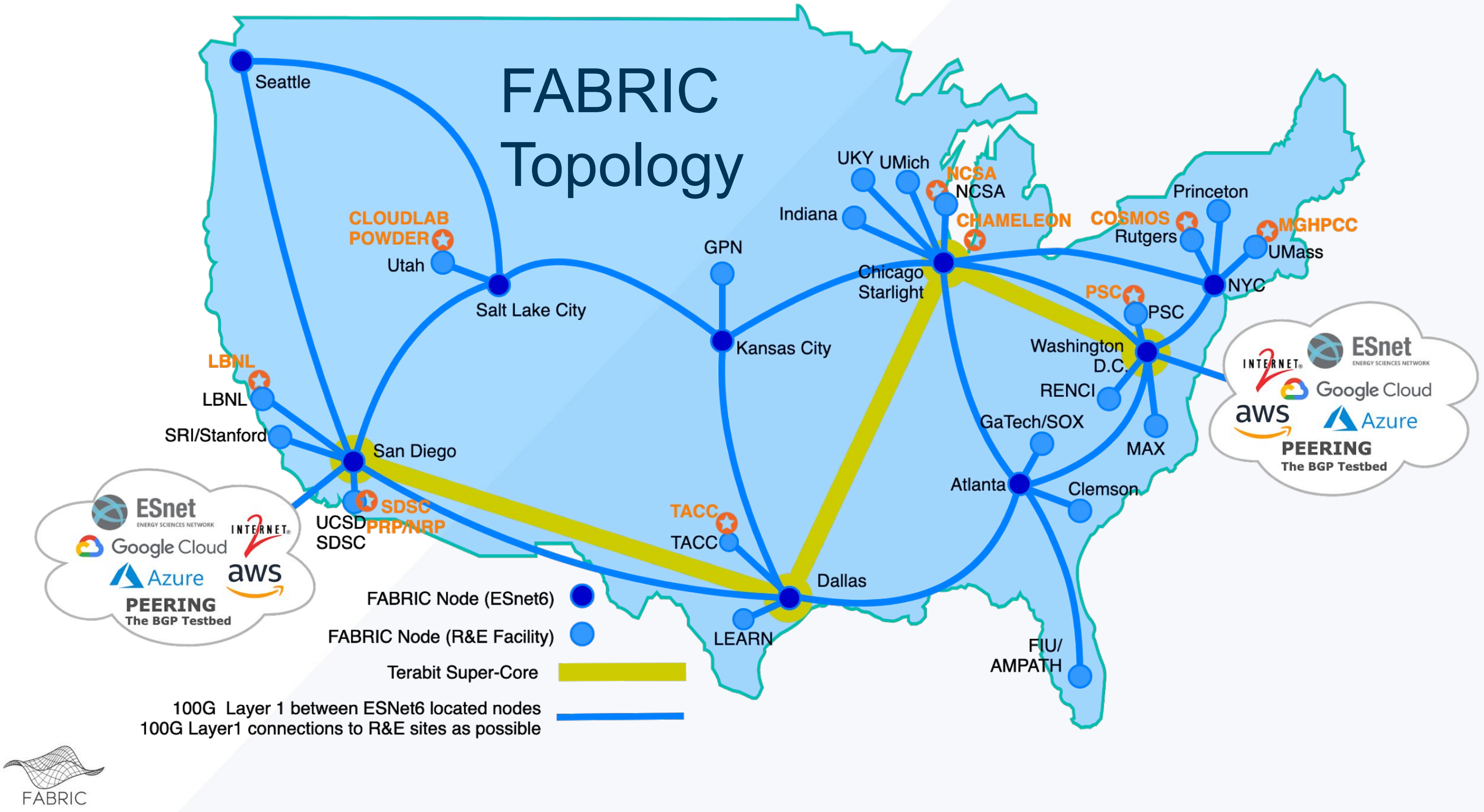
FABRIC

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
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
SDSC
UCSD
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

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

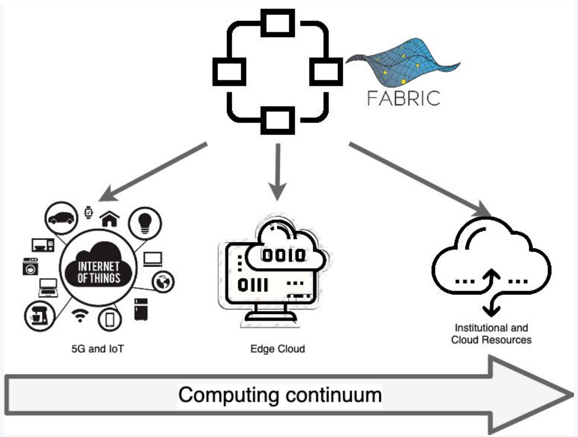
FABRIC Integrates Machine Learning & Artificial Intelligence

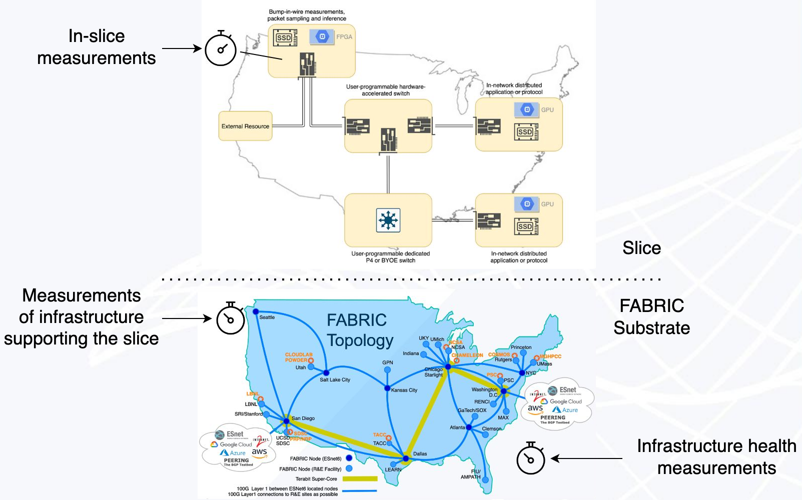
- Support for in-network GPU-accelerated data analysis and control

FABRIC helps train the next generation of computer science researchers

Key FABRIC features

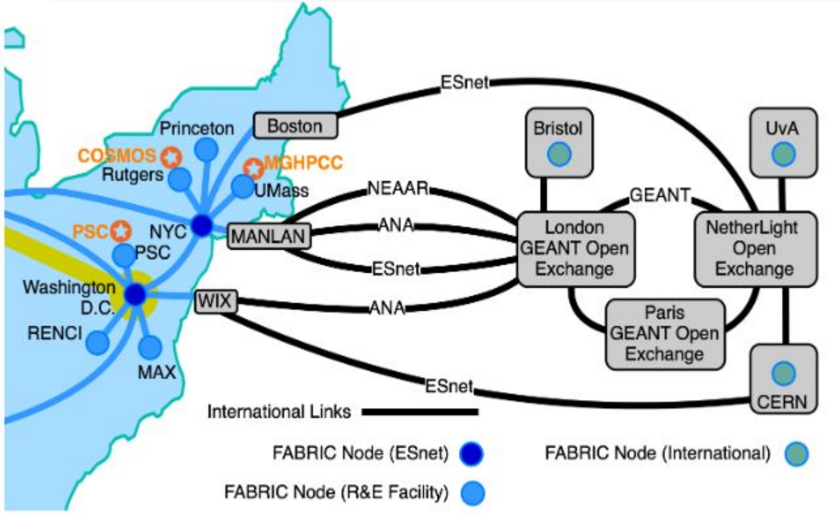
- Network as part of computing continuum
  - ‘Everywhere-programmable’ using different abstractions (P4, OpenFlow, others)
  - Diverse compute, storage capabilities in places where routers typically reside today
  - Dedicated 100G optical links between many sites
  - Support new paradigms in network aware applications and protocols
  - Ability to peer with Internet
- Network as a scientific instrument
  - Pervasive measurement collection capabilities in- and outside the slice available to researchers
  - GPS-disciplined PTP clock sources at every site
- Serve a broad range of scientific domains and applications
  - Concerned with data transport for big-data science, cyber-security, terrestrial and 5G hybrid network architectures, federated ML/AI, Internet measurements and many more





FAB (FABRIC Across Borders): Global Expansion

- Japan (University of Tokyo)
- UK (University of Bristol)
- EU (University of Amsterdam)
- CERN



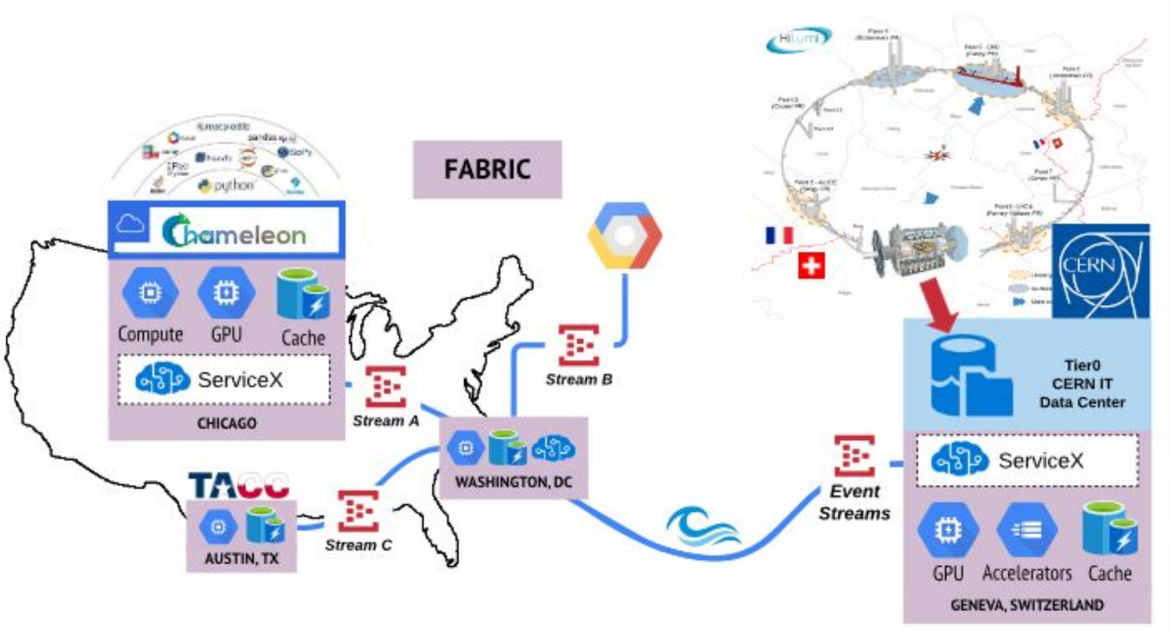
FAB's EU connectivity

- New Use-cases & Partners
  - Astronomy/Cosmology (CMB-S4, LSST)
  - Weather (UMiami)
  - High-Energy Physics (CERN)
  - Urban Sensing/IoT/AI at Edge (UBristol)
  - Computer Science: 5G across borders, P4/SDN, Cybersecurity/Censorship Evasion

FABRIC

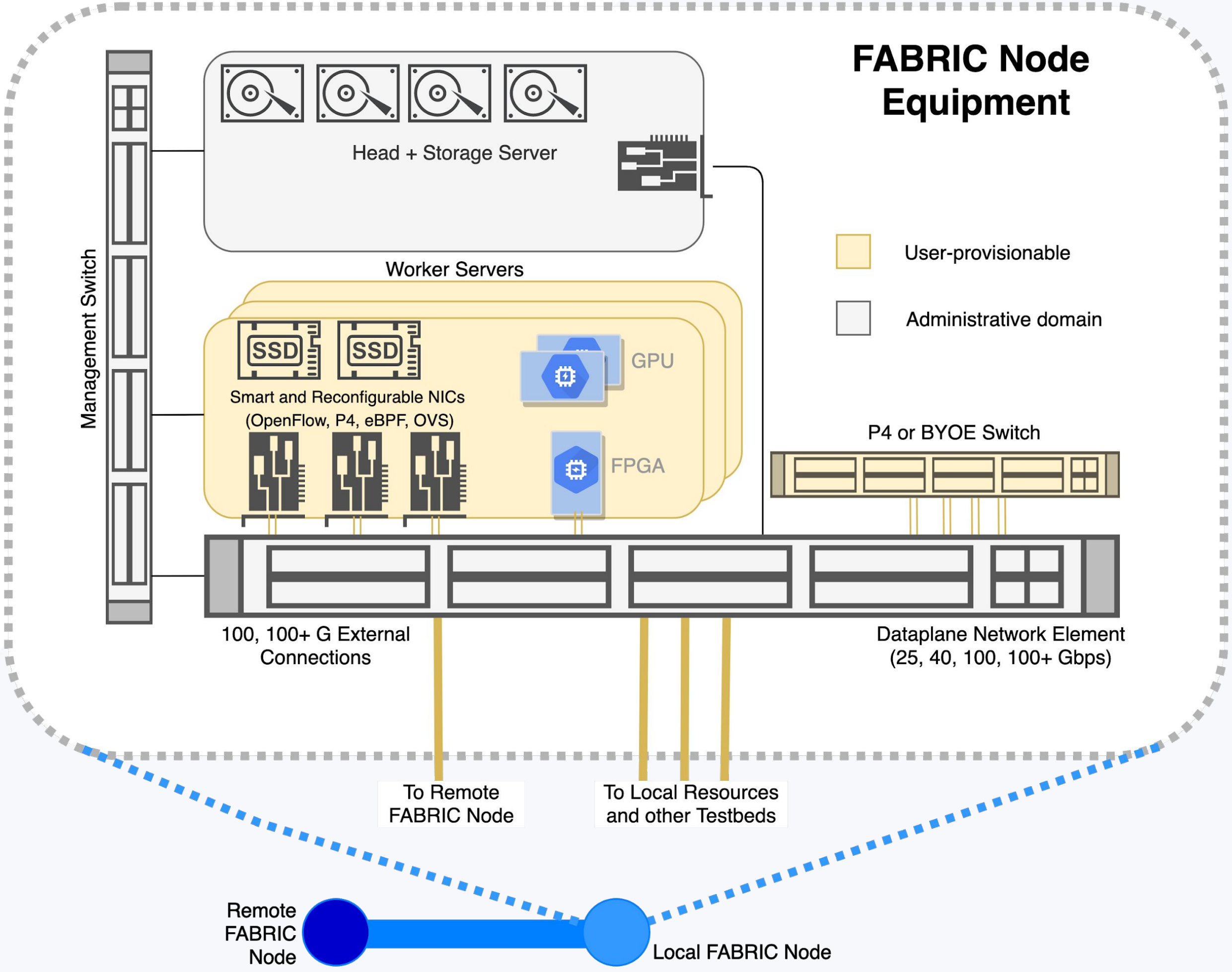
Example: Testing HEP data analysis approaches

- Real-time filtering & accelerated HEP data delivery
- Develop and test ML algorithms - inferencing within FABRIC nodes for real-time data processing



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

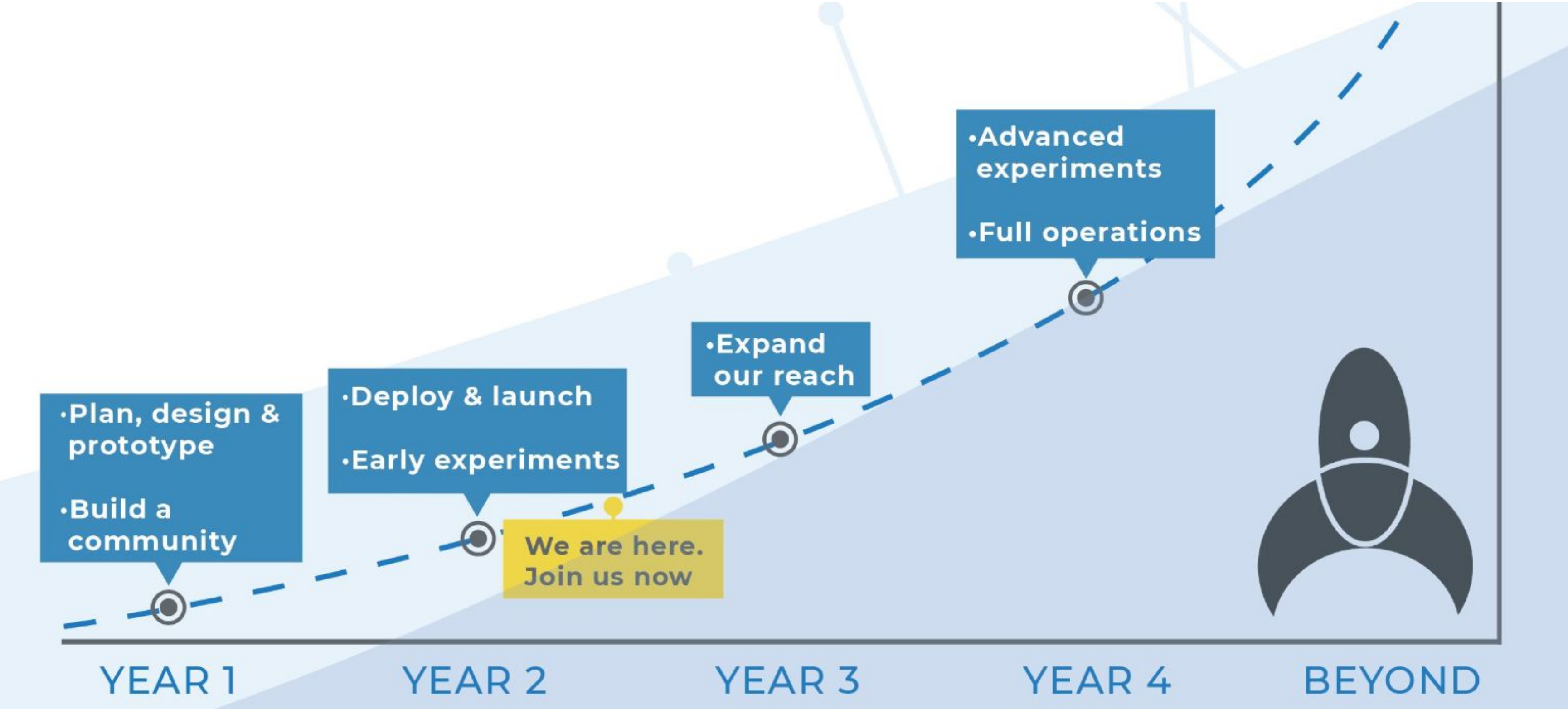
FABRIC

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)

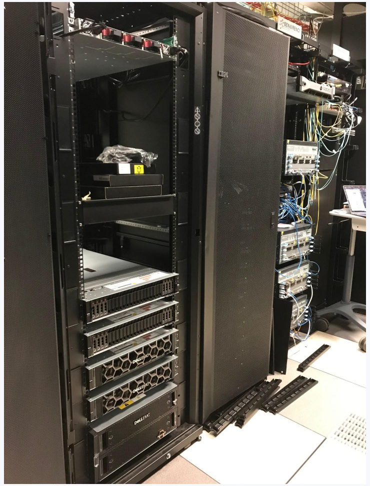
FABRIC

Construction Timeline



Where we are today

- Year 1 completed
- First 3 'dev' sites are being integrated: RENCI, UKY, LBNL
- First production site (StarLight) being assembled by the integrator
- Software:
  - Control framework, network control plane, measurement framework, portal, system services all being implemented
- Expect to have early experimenters summer 2021



FABRIC


FABRIC
Scientific Advisory Committee
Sujata Banerjee
Terry Benzel
Kaushik De
Cees de Laat
Phillipa Gill
Abraham Matta
Craig Partridge
Jennifer Rexford
Scott Shenker
Frank Wuerthwein

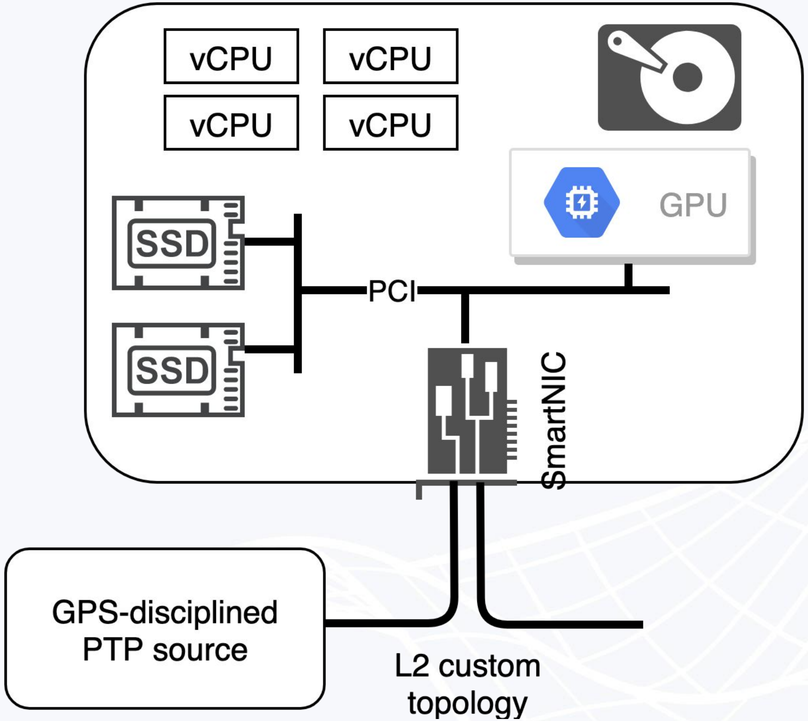
FABRIC Experiment building blocks

- 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

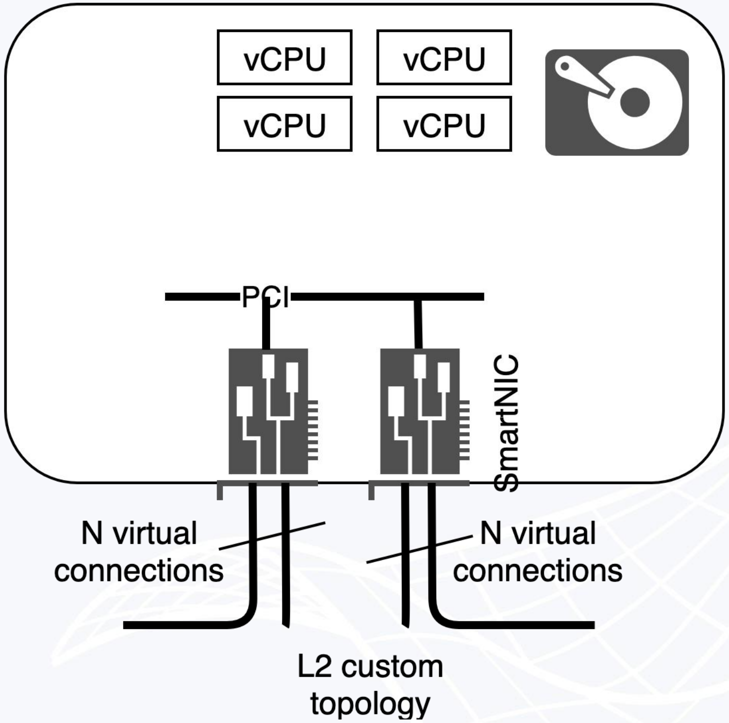
- Useful for collecting and analysing high-volume packet traces
  - Rely on NVMe drive for high-throughput local storage
  - Use GPU to assist in analysis
- Can optionally use a local GPS-disciplined PTP source to achieve millisecond-level accuracy for measurements
  - Multiple 'bumps-in-wire' in a slice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



FABRIC

SmartNIC router sliver

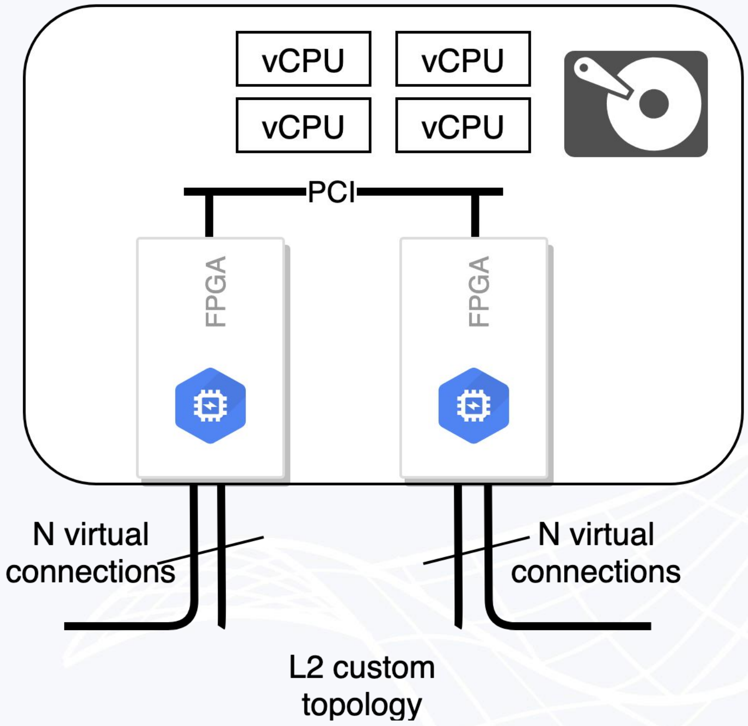
- Can create an small-port-count OpenFlow router with hardware acceleration via Mellanox ConnectX-[5,6] cards
    - Direct access to PCI allows to bypass CPU in many cases.



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
-



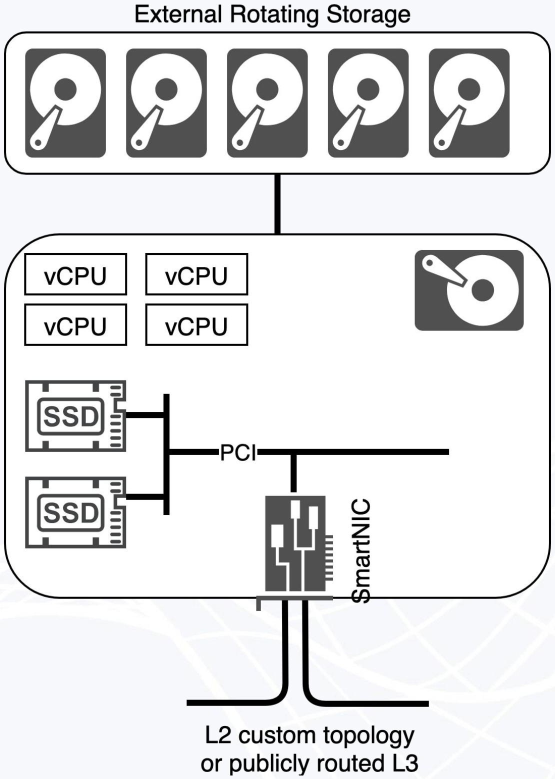
FABRIC

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

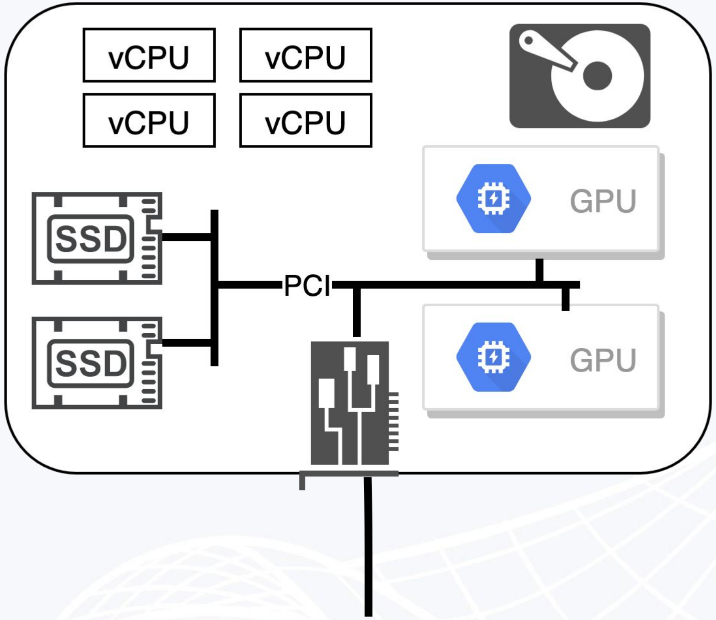
Speed

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

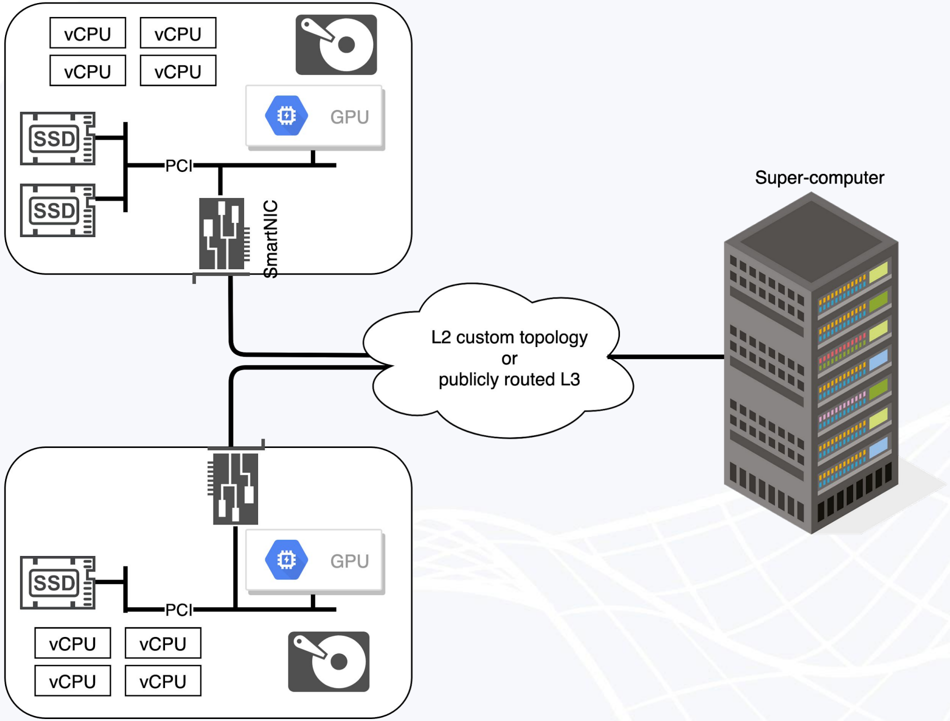


Publicly routed L3
FABRIC

FABRIC

Attaching external facilities

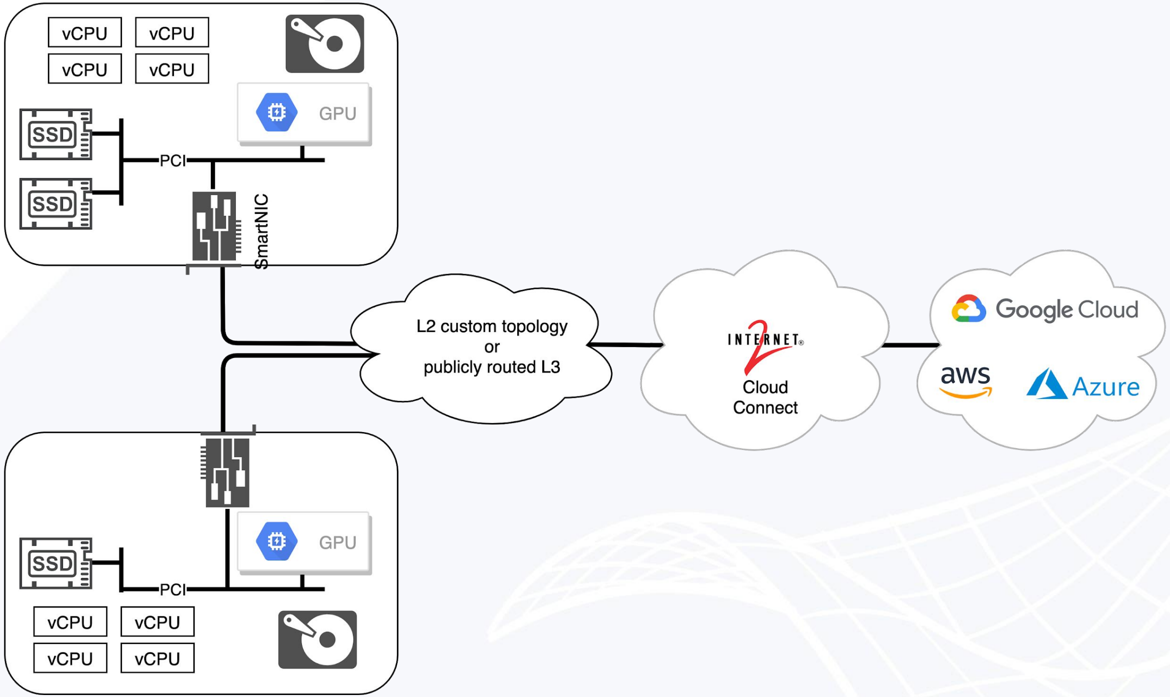
- The US NSF has made significant investments in scientific CI
- Future networks must better support domain science needs
- FABRIC connects to a number of facilities and testbeds to enrich the set of resources that can be used in experiments
  - Supercomputing centers (PSC, NCSA, SDSC, TACC, MGHPCC)
  - 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

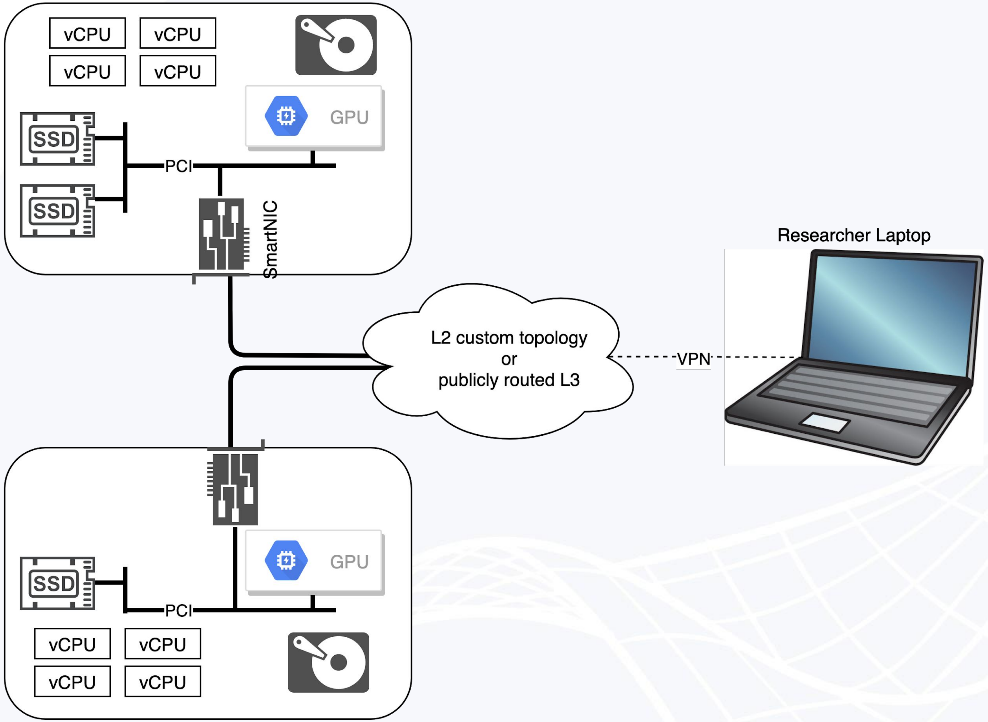
- Future networks will connect clouds and their customers
- 5G+Cloud experiments
- Through partnership with Internet 2 FABRIC will provide connectivity to commercial clouds
    - Utilize I2 CloudConnect system



FABRIC

Adding experimenter-owned resources

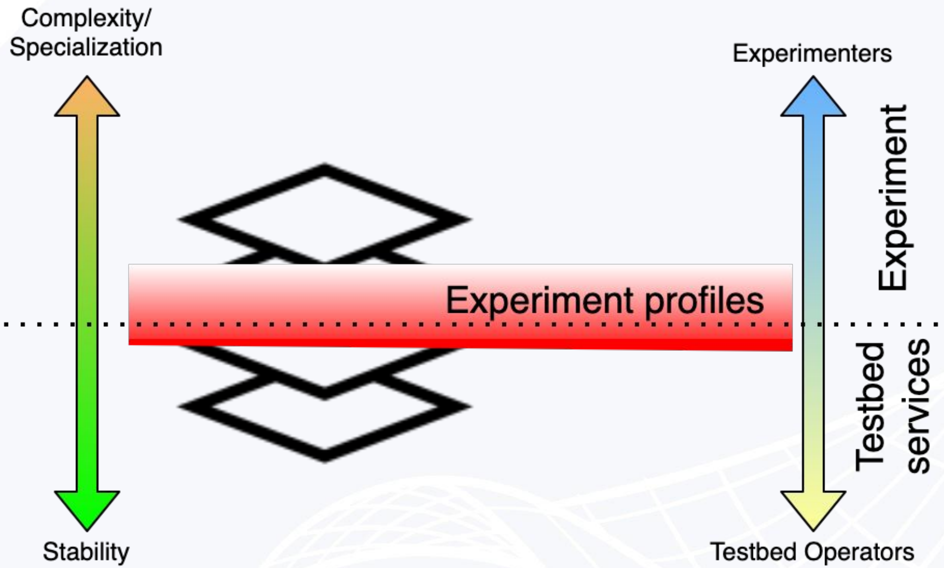
- Many experimenters may be interested in connecting their own resources to their slice topologies
  - FABRIC may not be able to reach every campus with a dedicated connection
- VPN/VPW options will be available to support these cases.
  - Allow experimenters to offer services to others from their slices



FABRIC

FABRIC Testbed Services

- Central to FABRIC are ideas of ‘testbed services’ and ‘experiment profiles’
- Testbed services are provided by the testbed
  - A relatively small number of abstractions
  - Reliable
  - Well-understood
- Experiment profiles can be created by testbed operators or experimenters and shared with others
  - Contain additional configuration, reproducible building blocks to help build experiments faster



FABRIC

FABRIC Network Services

- FABRIC is *not a network,* rather a testbed that provides network services
- Provides a variety of options to connect compute slivers into topologies
- L2 point-to-point (GENI-like, but richer)
  - Built on top of MPLS-SR
  - Support for QoS for individual services
  - Port-to-port and site-to-site
  - Does not assume the use of IP
  - Any routing must be built into the experiment via experiment profiles (e.g. OSPF instances or NDN forwarded instances) or built by experimenter by hand
  - Dedicated to individual experiments
- L3 routed
  - Relies on FABRIC's allocation of public IPv6 addresses
  - Provides high-performance routing using FABRIC's hardware routers
  - Peers with production networks
  - Shared between multiple experiments

FABRIC

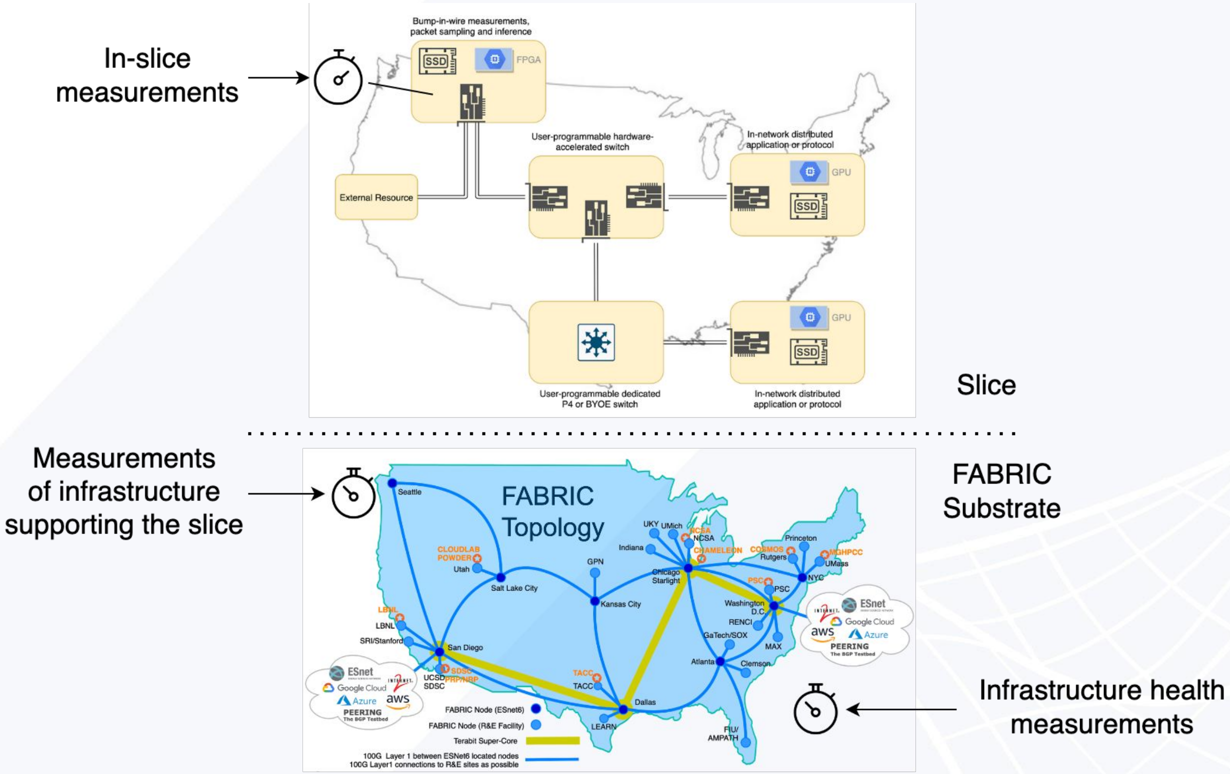
FABRIC Peering

- A combination of testbed services and experiment profiles
- Provides connectivity to other L2 and L3 networks and public clouds
- Peering of the publicly-routable L3 topology
  - Using FABRIC-managed BGP instances
- Programmatic peering with production networks of L2 topologies
  - Using experiment profiles with BGP instances to e.g. peer with commercial cloud VPCs
- Multiple peering points using ESnet and Internet2 infrastructures

FABRIC

FABRIC Measurement Capabilities

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



FABRIC

FABRIC Near-term use-cases

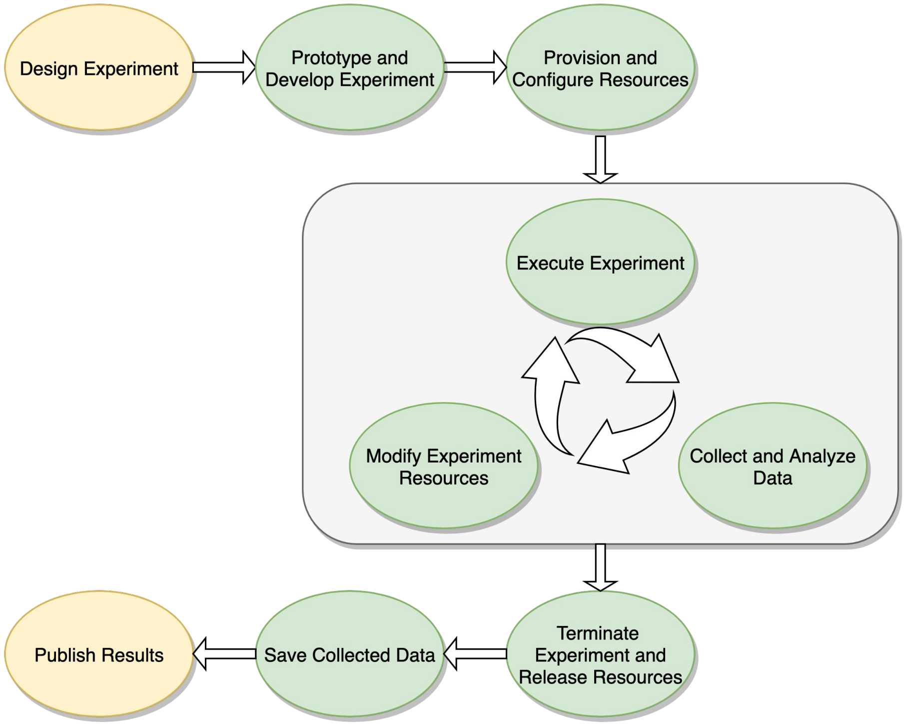
- FABRIC originally proposed several ‘Science Design Drivers’:
    - SRI - Network Security
    - Georgia Tech - 5G/IoT/Network Resilience
    - University of Virginia - ML/Autonomous Network Management
    - FIU - Named Data Networking and AR/5G
- The goal of the design drivers is to help hone requirements and test early capabilities
    - They are meant to be a diverse sampling of the possible experiment space
- FAB brings in several more domain-oriented use-cases
    - Efficient distribution and in-network fusion of astronomical event data
        - LSST/Vera Rubin and CMB-S4
    - Urban Sensing
        - Connecting COSMOS and University of Bristol testbeds
    - Weather science
        - Efficiently distributing data on weather events
    - Computer Science
        - Censorship evasion
        - Private 5G across borders
        - SDX policy negotiation

FABRIC

FABRIC Experiment Workflow

Experiment Phases:

- Design - an experiment is imagined and defined
- Prototyping and development - experiment software is written and prototyped (in-house, using FABRIC or other testbed hardware)
- Provision resources - FABRIC and other resources are acquired and configured via APIs or portal
- Experiment is run:
    - Multiple experiment runs include collecting data and modifying resources
- Termination - experiment ends, all resources released
- Saving data - collected data is retrieved from FABRIC storage
- Publish - paper citing FABRIC is prepared, submitted and published



Thank You!

Questions?

Visit https://whatisfabric.net

Ask info@fabric-testbed.net

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

This work is funded by NSF grants CNS-1935966, CNS-2029261, CNS-2029235, CNS-2029200, CNS-2029261, CNS-2029260