FABRIC Overview

Joint Engineering Team (JET) Meeting

January 19, 2021

ESnet
ENERGY SCIENCES NETWORK

CLEMSON
UNIVERSITY

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Overview

- FABRIC Project Overview
- Fabric Across Borders (FAB) Project Overview

FABRIC

What is FABRIC? NSF Mid-Scale Project

intended to enable *new paradigms for distributed applications and Internet protocols:*

- A nation-wide programmable network with compute and storage at each node. Run computationally intensive programs & maintain information in the network.
- GPUs, FPGAs, and network processors (NICs) inside the network
- Quality of service (QoS) - dedicated optical 100Gb
- Interconnects national facilities: HPC, cloud & wireless testbeds, commercial clouds, Internet, and edge
- Design and test applications, protocols and services that run at any node in the network
- Science cases: IoT sensors, Cybersecurity, AI/ML, SDN/P4, Science apps

FABRIC

FABRIC Leadership Team

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

FABRIC

FABRIC Edge

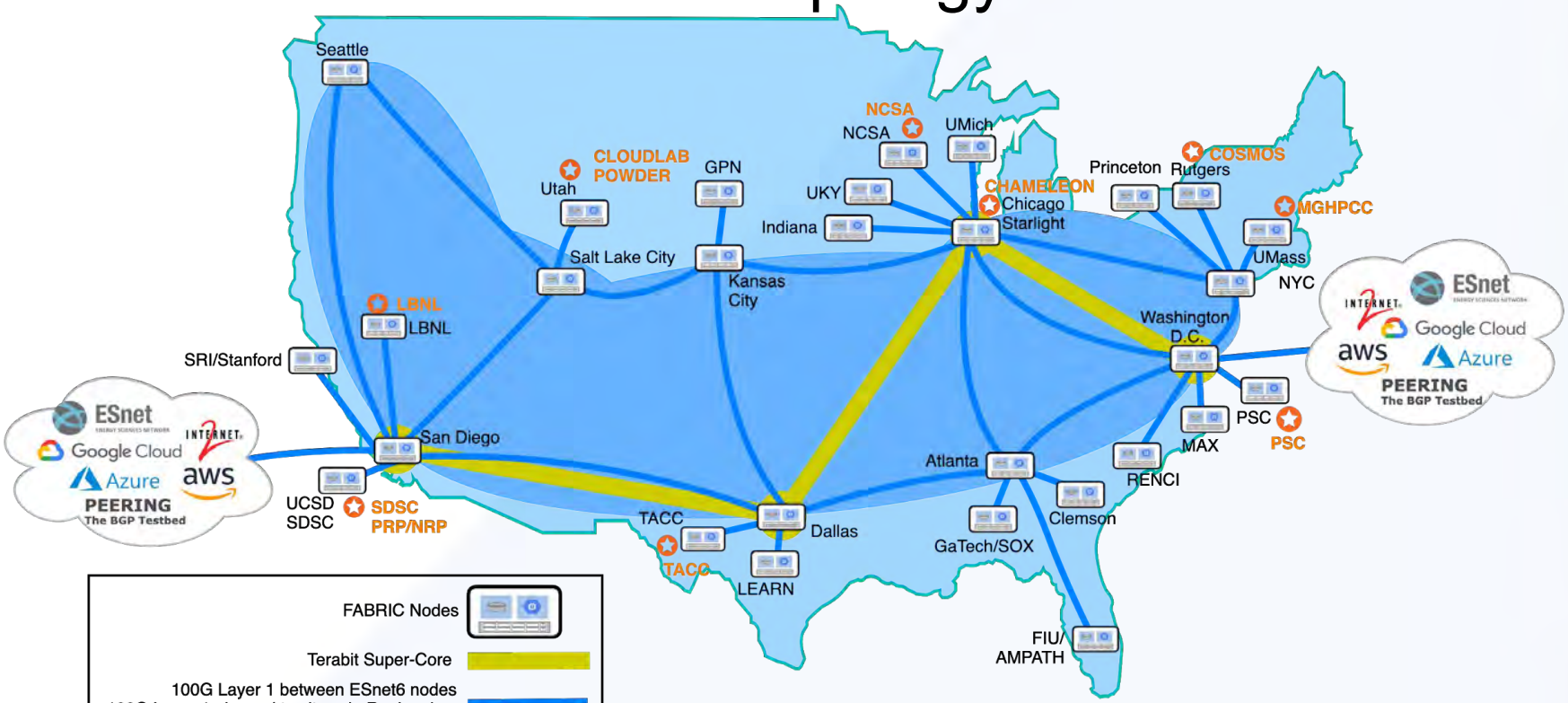

SRI International
Georgia Tech
FIU FLORIDA INTERNATIONAL UNIVERSITY
Colorado State University
UMass Amherst
renci
CLEMSON UNIVERSITY
POWDER
THE UNIVERSITY OF UTAH
INTERNET 2
TACC
COSMOS
LEARN
LONESTAR EDUCATION AND RESEARCH NETWORK
MAX
MID-ATLANTIC CROSSROADS
RUTGERS
3ROX
THREE RIVERS OPTICAL EXCHANGE
PITTSBURGH SUPERCOMPUTING CENTER
CENIC
UK
STARLIGHT SM
The Optical STAR TAP™
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK
PRP PACIFIC RESEARCH PLATFORM
usignite
BERKELEY LAB
UNIVERSITY OF MICHIGAN
UC San Diego
SDSC SAN DIEGO SUPERCOMPUTER CENTER
Chameleon
ILLINOIS NCSA
FABRIC
EASTERN REGIONAL NETWORK
THE UNIVERSITY OF CHICAGO
GPN
GREAT PLAINS NETWORK
PRINCETON UNIVERSITY

Overview

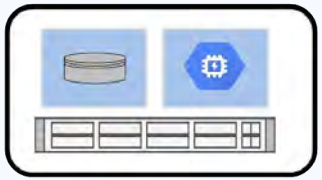
- 29 FABRIC Nodes
  - Development Phase: April 1, 2020 – September 30, 2021: (3 Nodes)
  - Phase 1: July 1, 2020 – September 30, 2021 (16 Nodes)
  - Phase 2: April 1, 2022 – June 30, 2023 (10 Nodes + Supercore)
- 9 nodes co-located at ESnet6 Points of Presence
  - Connected via dedicated 100Gbps DWDM across the new ESnet6 open-line optical system
  - Some sites upgraded to Terabit SuperCore during Phase 2
- 20 other nodes distributed across the R&E community at various regional networks, major cyberinfrastructure facilities, and university hosting sites
  - Working to get as many connected via 100 Gbps Layer 1 as possible

FABRIC

FABRIC Topology



FABRIC Node
Programmable Networking
Compute (CPU,GPU,FPGA)
Storage

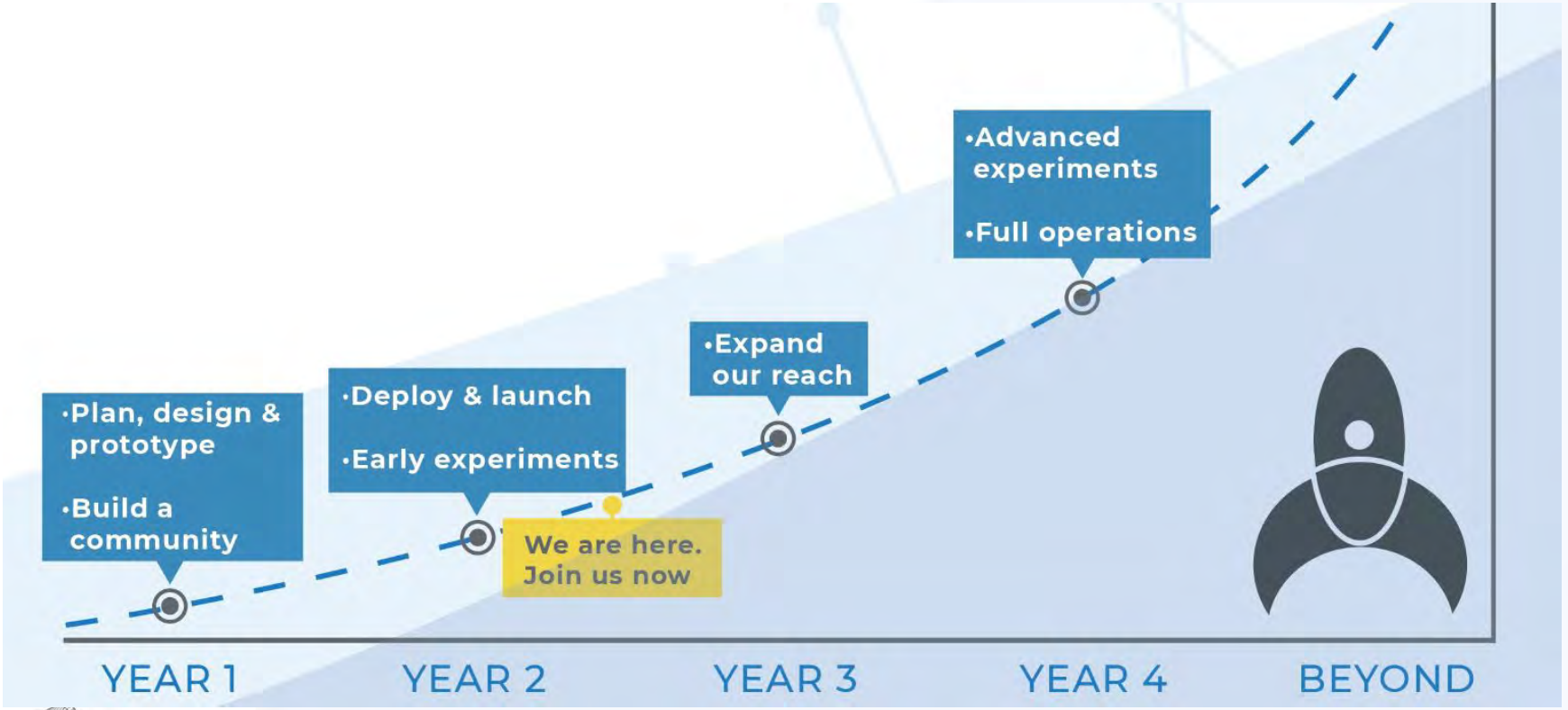


What is a FABRIC node?

- All nodes have compute, storage and programmable networking capabilities
  - Network programming at the level of OpenFlow, P4, eBPF, DPDK
  - GPUs to support ML applications
  - Ability to interpose compute, memory and storage into the path of fast packet flows
  - Processing speeds at 25Gbps, 40Gbps, 100Gbps, Nx100Gbps
  - Experimenters access hardware directly (programmable network cards, GPUs, FPGA cards)
  - Provide sliceable, programmable switching, hierarchical storage and in-network compute
- Node placement and connections
  - 9 *ESnet Core* nodes directly connected to ESnet6 optical substrate at the intersection of multiple high-capacity *dedicated* optical links.
  - 20 *CoreEdge (Layer 1 connected) and Edge (Layer 2 connected)* nodes located on campuses, regional networks, and R&E facilities.

FABRIC

Construction Timeline



FABRIC

FABRIC Deployment Schedule

This is an initial plan which may be modified based ongoing development, engineering, and planning activities

Development Phase April 1, 2020 – September 30, 2021
RENCI
University of Kentucky
LBNL

FABRIC

FABRIC Deployment Schedule

Phase 1 July 1, 2020 – September 30, 2021
StarLight (ESnet)
TACC (LEARN Regional)
Washington (ESnet)
Dallas (ESnet)
Salt Lake (ESnet)
UCSD (CENIC Regional)
FIU/AMPATH
GPN (Regional)

Phase 1 July 1, 2020 – September 30, 2021
MAX (Regional)
University of Michigan (MERIT Regional)
University of Utah (UEN Regional)
UMASS/MGHPCC (UMASSNET, NEREN)
NCSA (ICCN Regional)
Clemson
Georgia Tech/SoX

FABRIC

FABRIC Deployment Schedule

Phase 2 April 1, 2022 – June 30, 2023
Kansas City (ESnet)
New York (ESnet)
Atlanta (ESnet)
San Diego (ESnet)
Seattle (ESnet)

Phase 2 April 1, 2022 – June 30, 2023
Princeton
University of Indiana
PSC
LEARN
Rutgers
SRI/Stanford

FABRIC

FABRIC Design Information

FABRIC

FABRIC Node Design: Network + Compute

- High-performance servers equipped with:
  - GPUs (RTX 6000 or T4), FPGA (Xilinx) compute accelerators
  - Storage: User-provisionable short term & shared high volume. Not meant to be persistent.
  - All ports interconnected by a 100G+ switch programmable through testbed control software
  - VM/Containers sized to support full-rate DPDK for access to Programmable NICs, FPGA, and GPU resources
- Reconfigurable Network Interface Cards (FPGA and P4/network processors)
  - Mellanox ConnectX, Xilinx FPGA based, Multiple interface speeds (10G, 25G, 40G, 100G, 200G+(future)
- Network Dataplane Switch/Router (Cisco NCS 5500 Shadow Tower, IOS XR, IP, MPLS-SR, Layer 2)

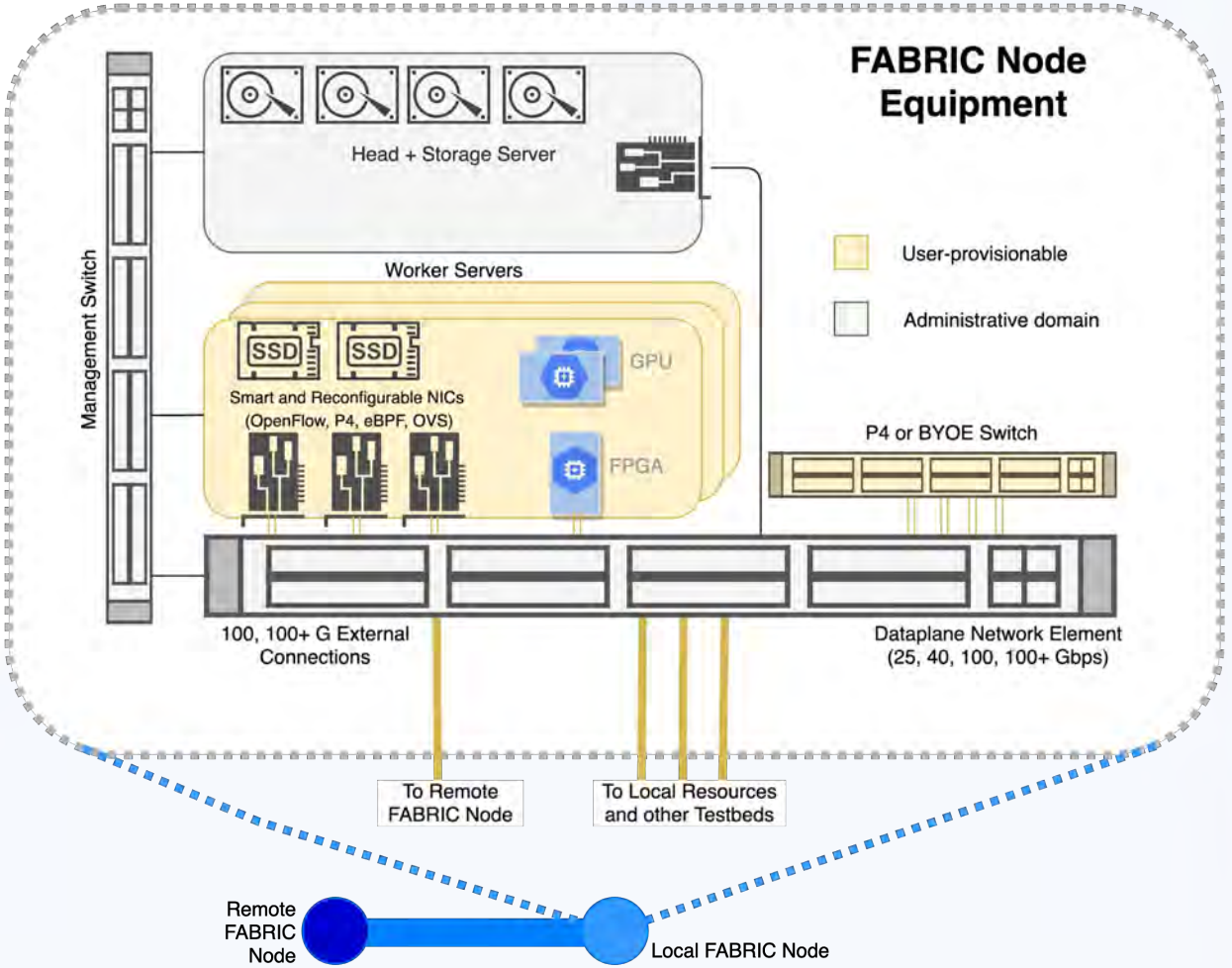
FABRIC

FABRIC Node Design: Storage

- Multiple storage options:
- OS/base filesystem
  - With each bare metal, VM or container
- User-provisionable
    - NVMe drives in servers
    - Shared allocatable block storage in each site
- Large-scale dataset storage
  - 300+ TB of NAS storage with each rack

FABRIC

FABRIC Node ‘Hank’ Overview

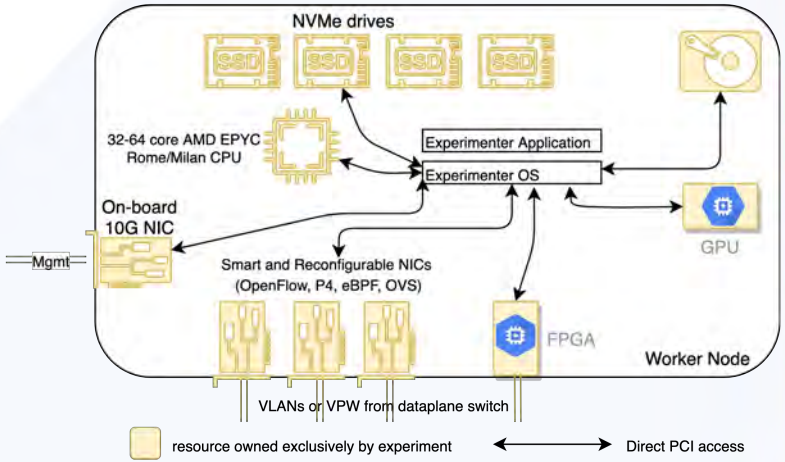


FABRIC

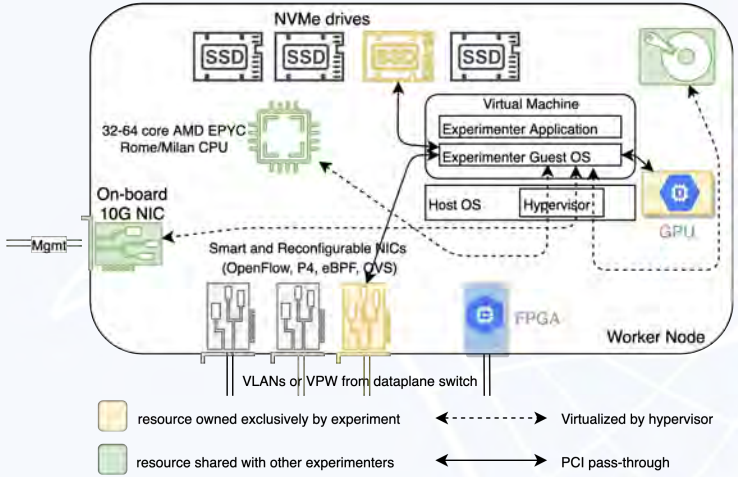
Node Level Programmability Abstractions

- Main capabilities are various PCI cards in individual servers
  - NICs, GPUs, FPGAs
- Additional switches and BYOE hardware
- Depending on experimenter request can be provided as part of a baremetal server or via PCI pass-through for VMs and containers

FABRIC experiment using a bare-metal server



FABRIC experiment using a virtual machine



FABRIC

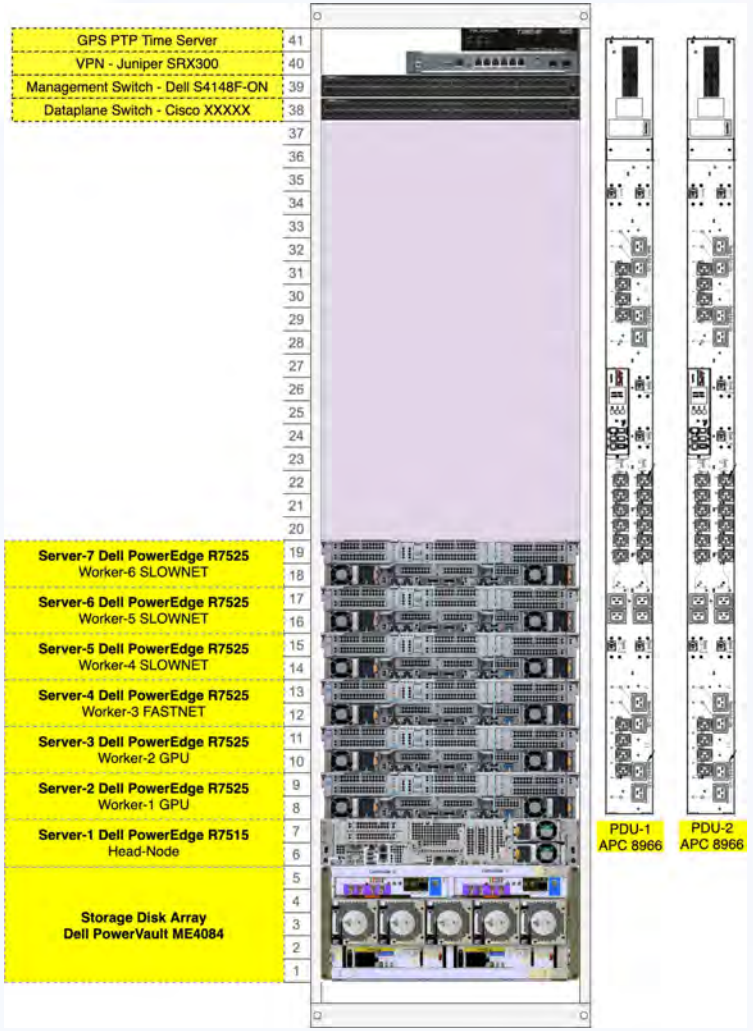
FABRIC Node Design: Measurement Hardware

- GPS-disciplined clock source at most sites
  - 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 Rack Configuration

This is an example FABRIC Rack Configuration. There are multiple configurations which vary the number and type of compute and storage elements.



Dealing with BYOE hardware

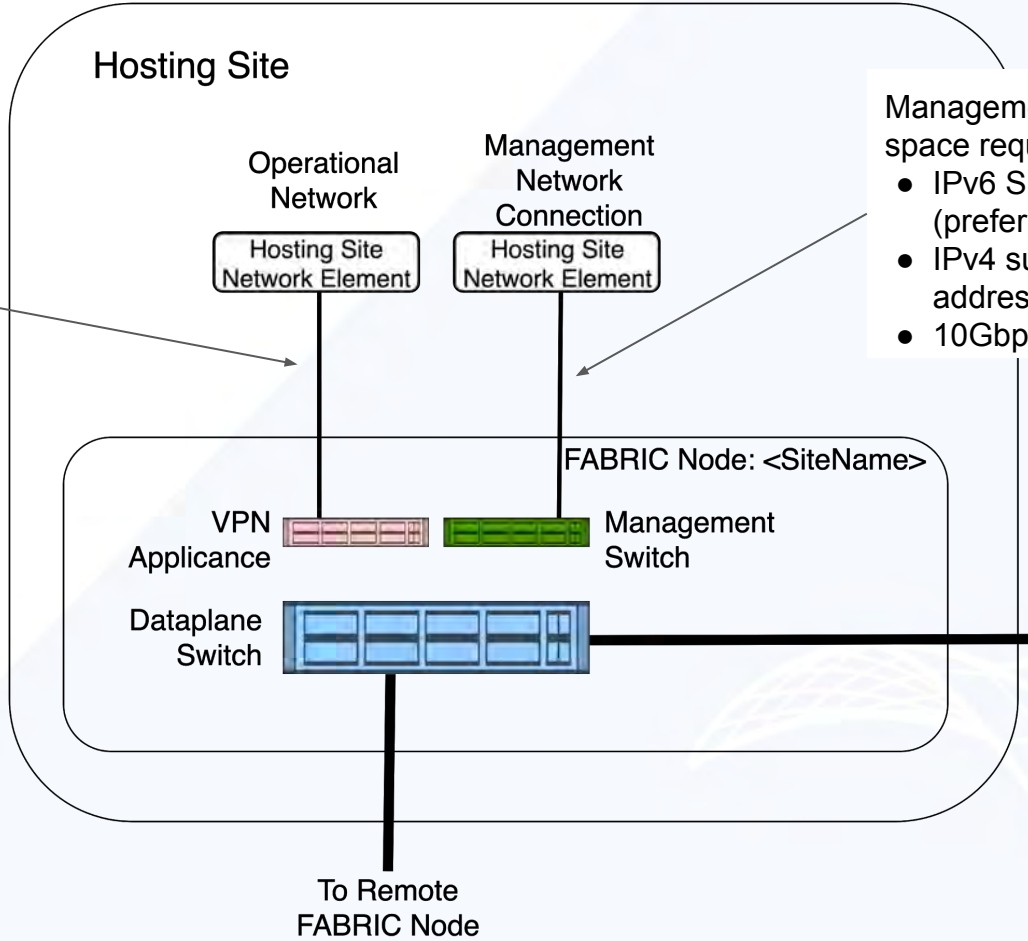
- 3 possible ‘archetypes’ of BYOE
  - PCI card (integrates into a FABRIC worker node)
  - Standalone server
  - Standalone switch
- Can be placed in FABRIC locations subject to constraints on
  - Power/space
  - Available management and dataplane ports
  - Available PCI slots in servers
- BYOE standalone can be integrated with Control and Measurement frameworks
  - Requires better understanding of capabilities, programmability model and APIs
  - Authorization for BYOE can be tuned to allow priority/exclusive access to contributing experimenters.

FABRIC

FABRIC Management/Control Plane Connections

Operations Net IP Address space request:

- single IPv4 address (preferred)
- single IPv6 address (optional)
- different subnet from the Management network
- 1Gbps connection

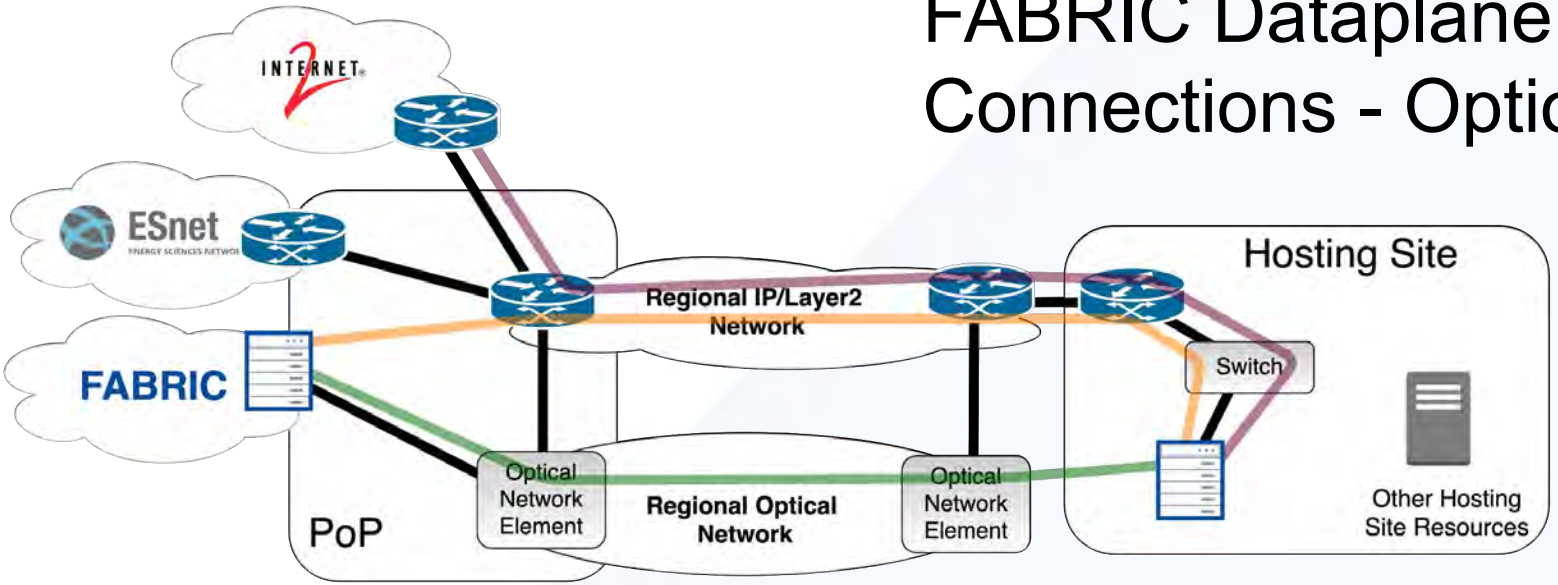


Management Network IP Address space request:

- IPv6 Subnet (standard /64) (preferred)
- IPv4 subnet (/26 or more address) (optional)
- 10Gbps connection

FABRIC

FABRIC Dataplane Connections - Options



FABRIC Rack

Option 1 - Layer1 DWDM Connection across Regional Optical Network to another local FABRIC Node

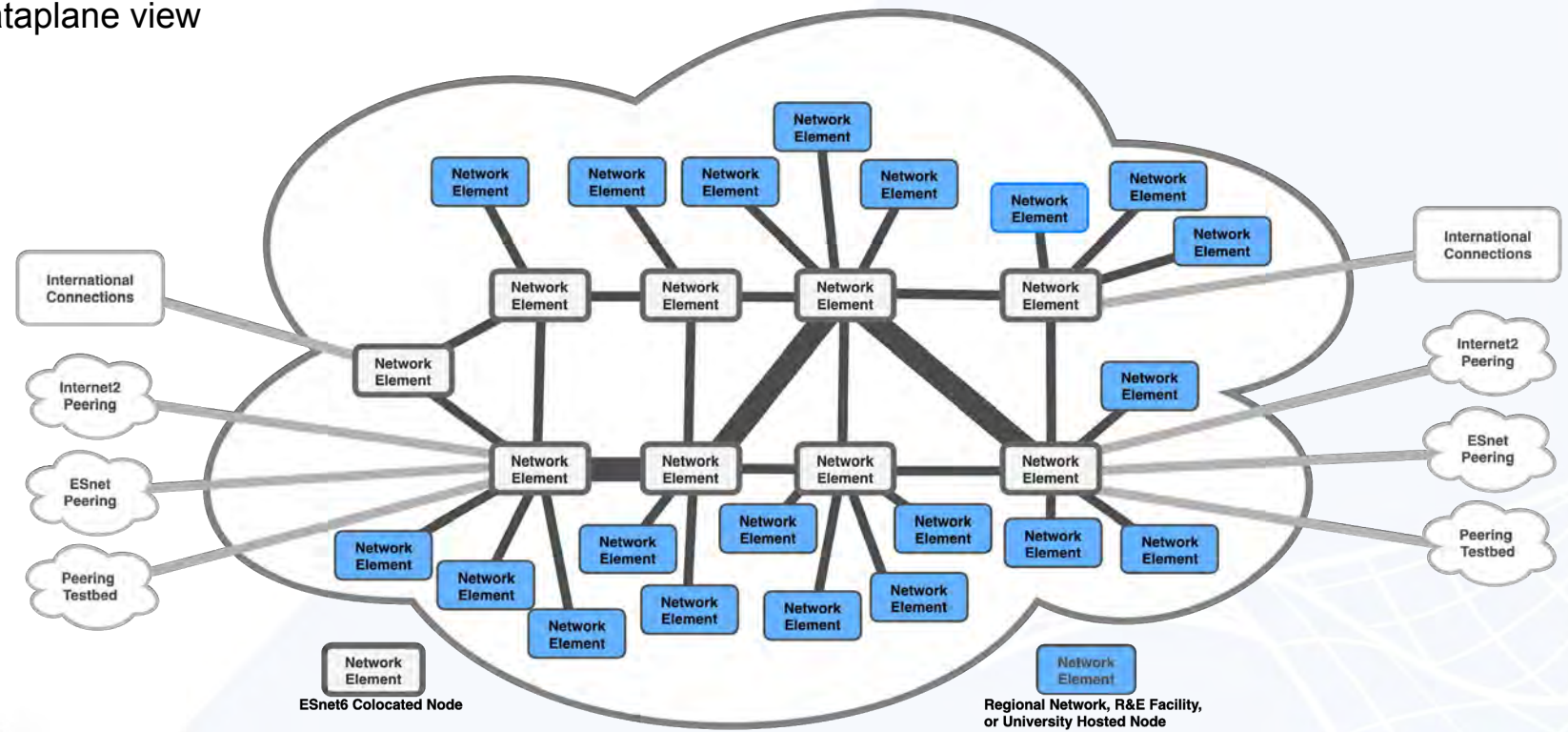
Option 2 - Layer2 Ethernet Connection across Regional IP/Layer2 Network to another local FABRIC Node

Option 3 - Layer2 Ethernet Connection across Regional IP/Layer2 Network to Internet2 AL2S or ESnet OSCARS (will then transport to another FABRIC Node)

FABRIC

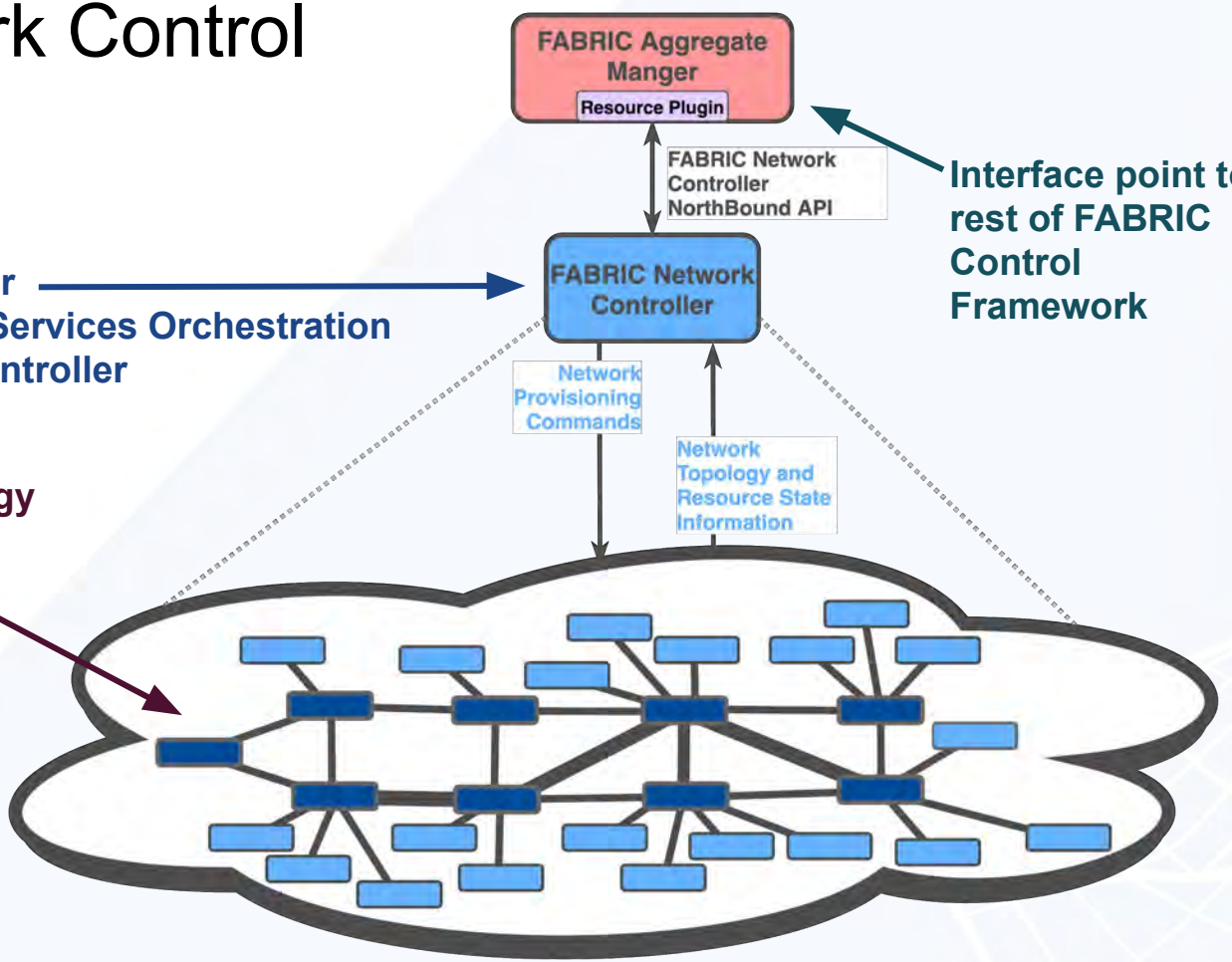
FABRIC Network Dataplane

Dataplane view



FABRIC

FABRIC Network Control Plane



FABRIC

FABRIC Network Services

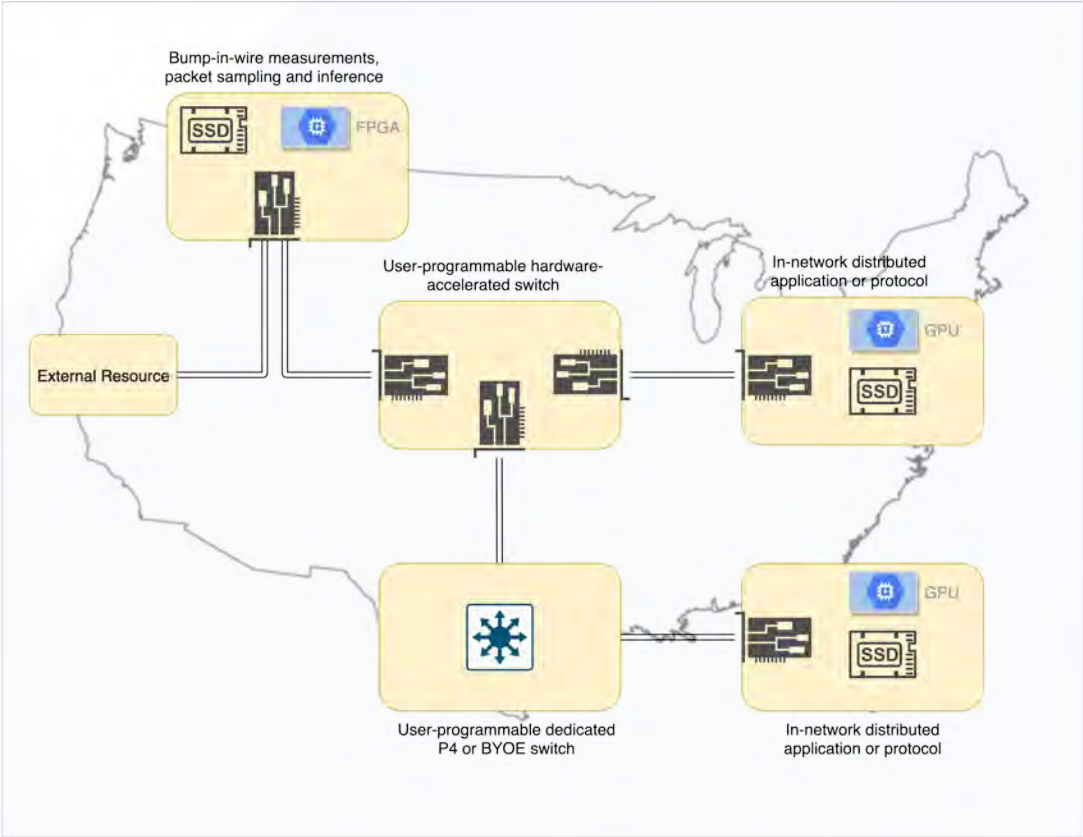
- Layer 3 Routing
- Layer 2 Connections
- Layer 3 Virtual Private Network (VPN)

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



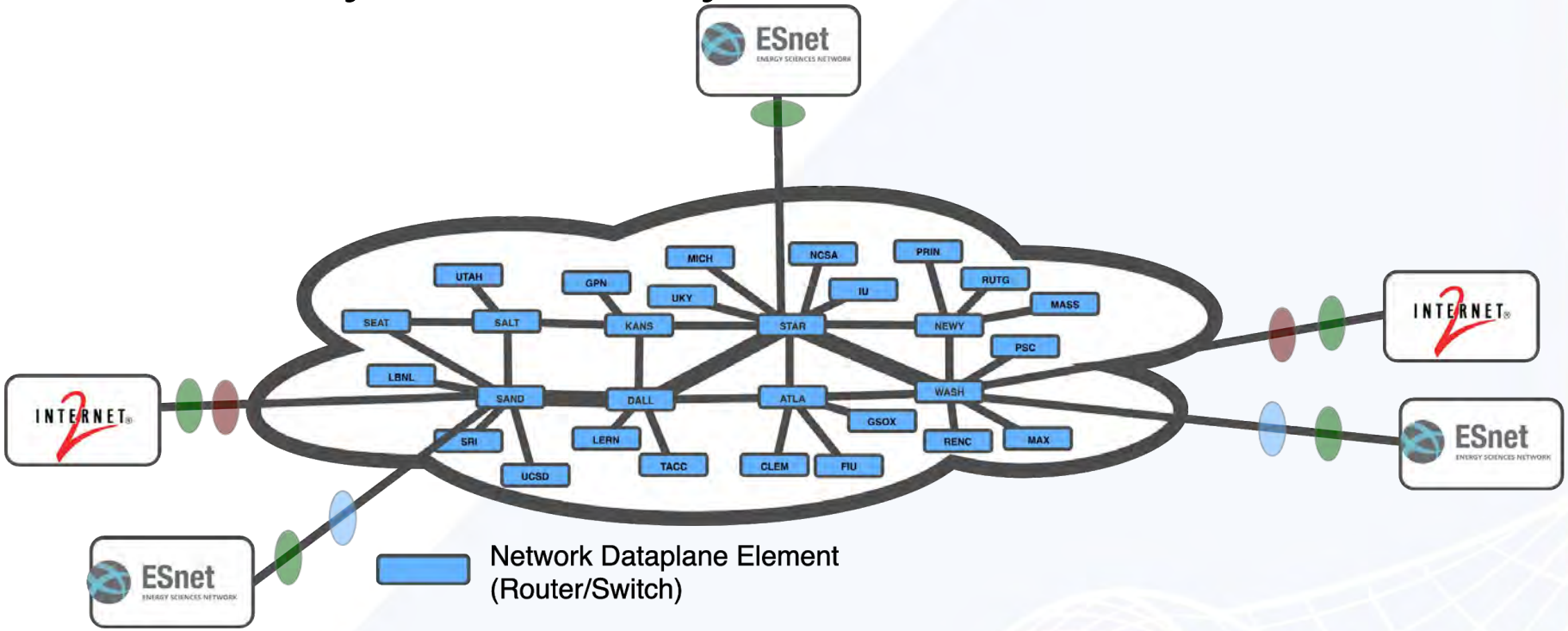
FABRIC

FABRIC External Connections Overview

- FABRIC experiments (slices) can run in an isolated manner within FABRIC Infrastructure, and isolated from external networks.
- Slices can also utilize FABRIC's external connections and peerings to access a variety of external experimental and production resources.
- These external connections and peerings are organized as follows:
  - Layer 3 Peering
  - Layer 2 Services Peering
  - Public Cloud Connections
  - PEERING (The BGP Testbed)

FABRIC

FABRIC Layer 2 and Layer 3 Connections Overview



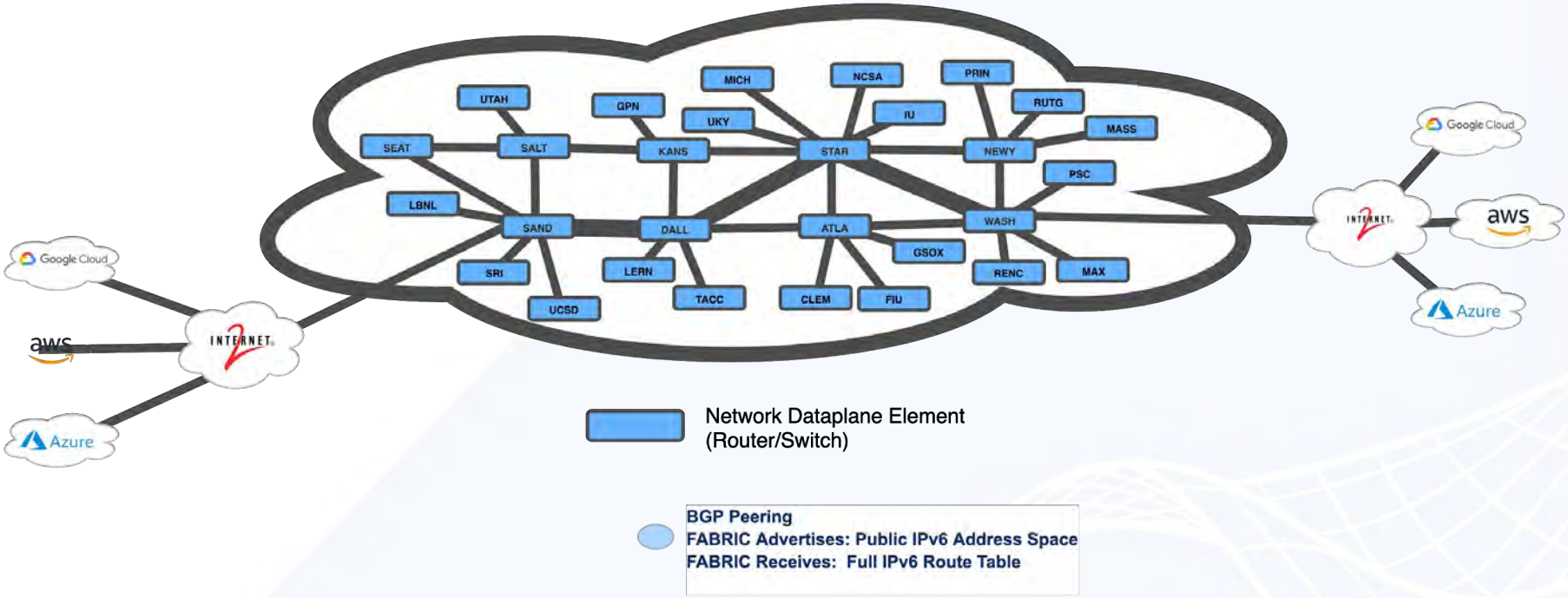
Layer 2 Service Peering:
Internet2 AL2S (OESS)

Layer 2 Service Peering:
ESnet OSCARS

Layer 3 Peering: BGP Peers
FABRIC Advertises: Public IPv6 Address Space
FABRIC Receives: Full IPv6 Route Table

FABRIC

FABRIC Public Cloud Connections Overview



FABRIC

FABRIC Hosting Site Resource Connections

- After the basic FABRIC Node Deployment (as described earlier), the FABRIC node could do one of the following with respect to dataplane connections:
  - FABRIC Dataplane is connected to other remote FABRIC Nodes
    - FABRIC dataplane switch is not connected to any Hosting Site Resources
    - the local management/control plane connections will be always be needed
  - FABRIC Dataplane is connected to other remote FABRIC Nodes and some Hosting Site Resources
    - This allow specific and controlled connections between some Hosting Site resources for the purposes of slice/experiments
- For the Hosting Site Connected, a slice may include a port/vlan combination which faces a Hosting Site resource. That port/vlan combination may be attached to any one of the FABRIC Network Services (as described earlier).
- The FABRIC slice that is connecting to the Hosting Site resource in this manner, may also be connected to other resources on FABRIC, and external resources, all depending on how that slice was requested by the user.

FABRIC

FABRIC Security Policy, Plans, and Best Practices

- Specific Security Policies and Best Practices are being developed
- Section 7 of the FABRIC Network Services and Peering Design document contains more details and thinking on these topics
- The default approach describe at this time is generally:
  - Experiment slices are isolate by default
  - Specific approvals are need to enable automated external access connections. The granularity of this access is to be defined.

FABRIC

FABRIC Design Documents

- FABRIC Design Documents
  - https://fabric-testbed.net/resources/design-documents
- The following documents are most relevant to this presentation
  - Topology Design
  - Network Services and Peering Design
  - Site Questionnaire (Deployment and Hosting Plan)
  - Partner Security Questionnaire

FABRIC

Questions?

Ask info@fabric-testbed.net

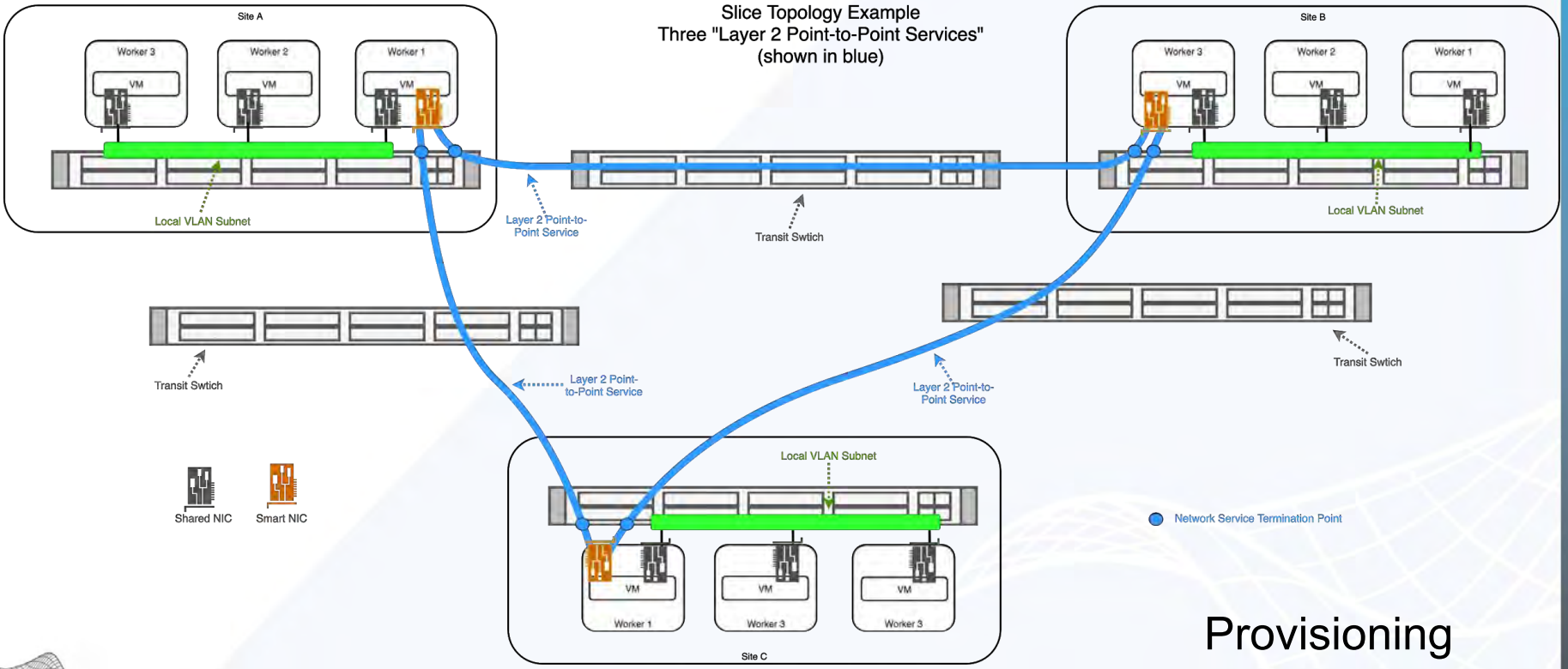
Website: https://fabric-testbed.net

This work is funded by
NSF grant CNS-1935966

FABRIC

Extra Slides

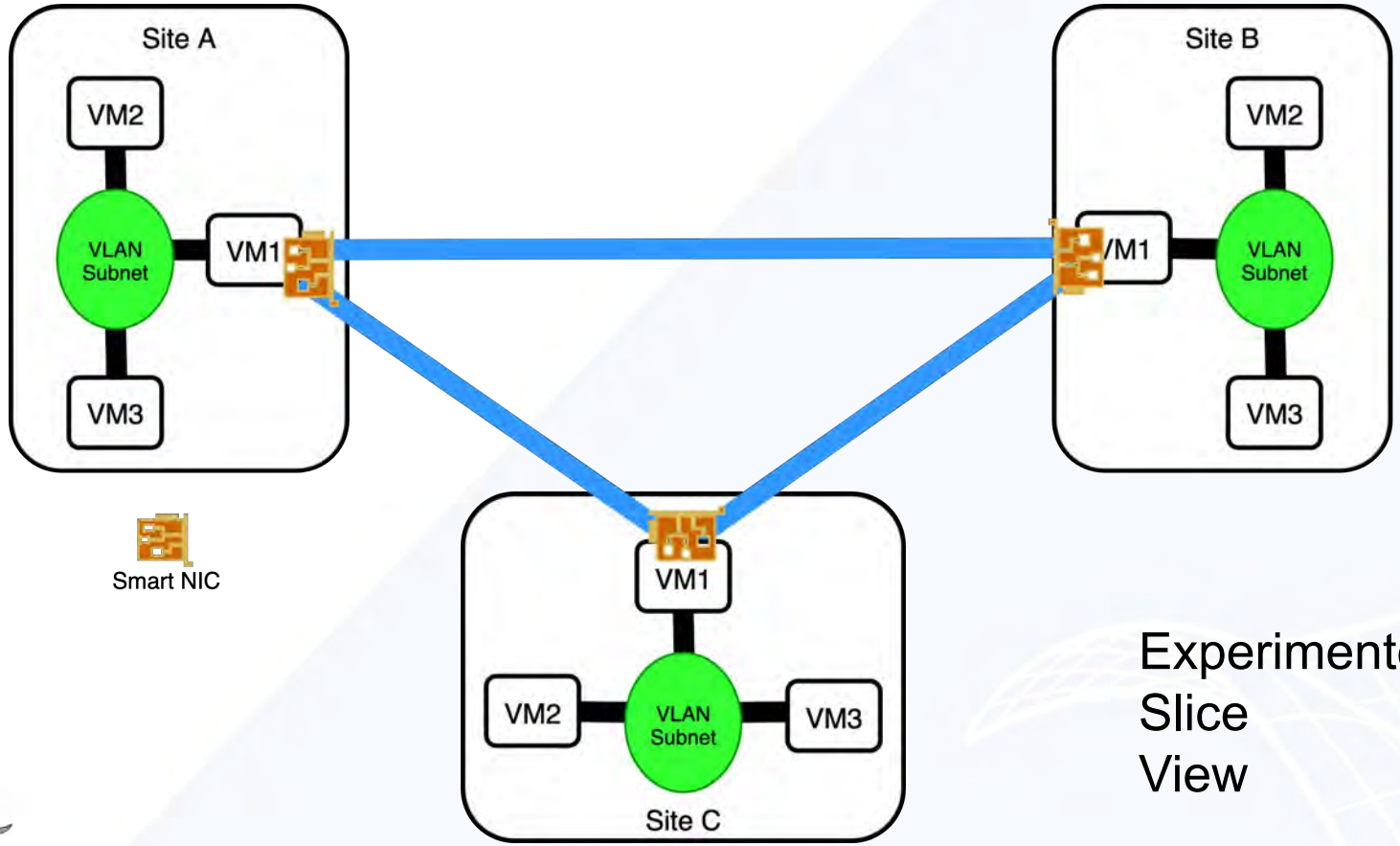
FABRIC Layer 2 Point to Point Service



Provisioning View

FABRIC

FABRIC Layer 2 Point to Point Service



Experimenter
Slice
View

FABRIC

FABRIC Network Services - Quality of Service

- **Network Control (Highest Priority)** - critical network traffic such as routing protocols, time synchronization messages, and control system communications.
- **Deterministic Service** - experiment slices which request guaranteed bandwidth services.
  - Soft-capped: guaranteed bandwidth in Deterministic Service Queue with oversubscribed traffic placed in Best Effort queue.
  - Hard-capped: guaranteed bandwidth in Deterministic Service Queue with oversubscribed traffic dropped
- **Best Effort Service (Lowest Priority)** - All other traffic is placed in the Best Effort queue.

Key Differences from GENI

- FABRIC has a programmable core network infrastructure
- FABRIC provides guaranteed quality of service by utilizing its own dedicated optical 100G infrastructure or relying on dedicated L2 capacity wherever possible to create QoS-guaranteed connections.
- FABRIC provides access to a variety of programmable PCI devices
  - Network cards, GPUs, FPGAs
- FABRIC interconnects a large number of existing scientific, computational and experimental facilities
- FABRIC experimenter network topologies can peer with the Internet, R&E Networks for access other facilities and public clouds on-demand

FABRIC

FAB (FABRIC Across Borders)

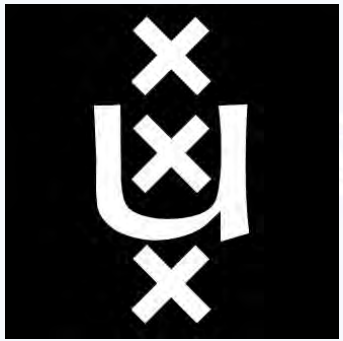
Enabling Global Experimentation

NSF IRNC Awards: 2029261, 2029235, 2029200, 2029261, 2029260

FABRIC

Global FABRIC Nodes

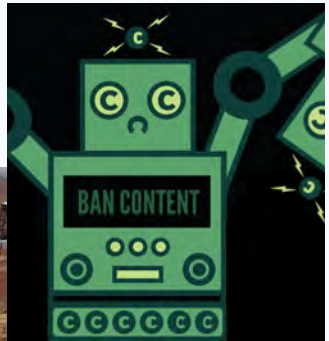
- Japan (University of Tokyo)
- UK (University of Bristol)
- EU (University of Amsterdam)
- EU (CERN)



FABRIC

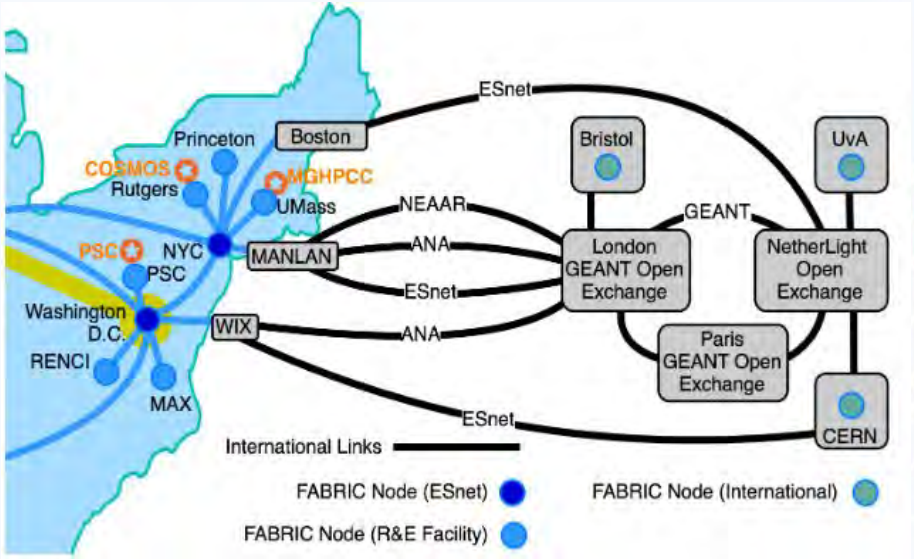
Additional Science Use Cases & Partners

- Astronomy (Vera Rubin Observatory/LSST, Chile)
- Cosmology (CMB-S4)
- Weather (UMiami & CPTEC, Brazil) - Ben Kirtman, Atmospheric Science & Paolo Nobre
- High-Energy Physics (CERN) - Rob Gardner, **FAB Co-PI**, Physicist
- Urban Sensing/IoT/AI at Edge (UBristol) - Dimitra Simeonidou, Prof. of Networking
- 5G across borders, P4/SDN - (UTokyo) Aki Nakao, Prof. of CS; KISTI (Korea Institute of Science and Technology Information)
- Censorship Evasion - Richard Brooks, Prof. of ECE



FAB Network & Facility Partners: EU

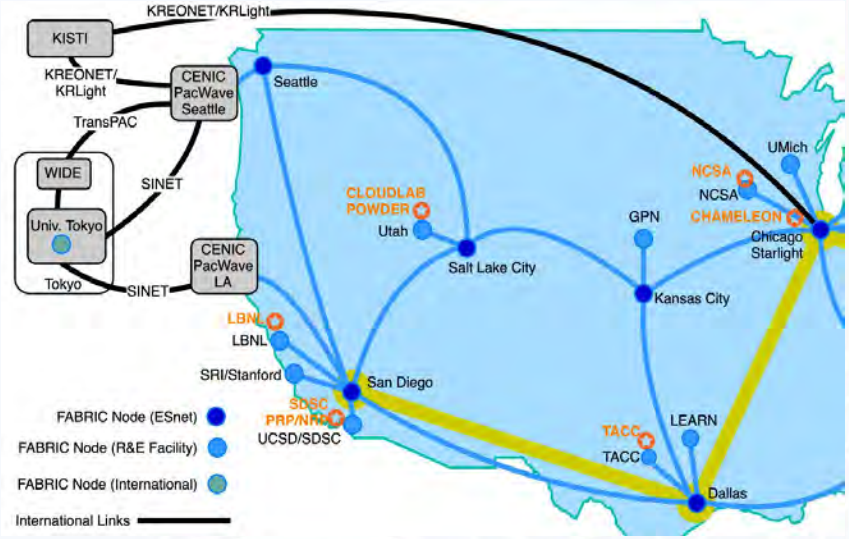
- NEAAR (Networks for European, American, and African Research)
- ANA (Advanced North Atlantic)
- ESnet
- CERN
- GEANT Open Exchange London & Paris
- NetherLight Open Exchange
- SURFnet
- University of Bristol
- University of Amsterdam
- University of Antwerp
- SAGE (MidScale project)



FABRIC

FAB Network & Facility Partners: Asia-Pacific

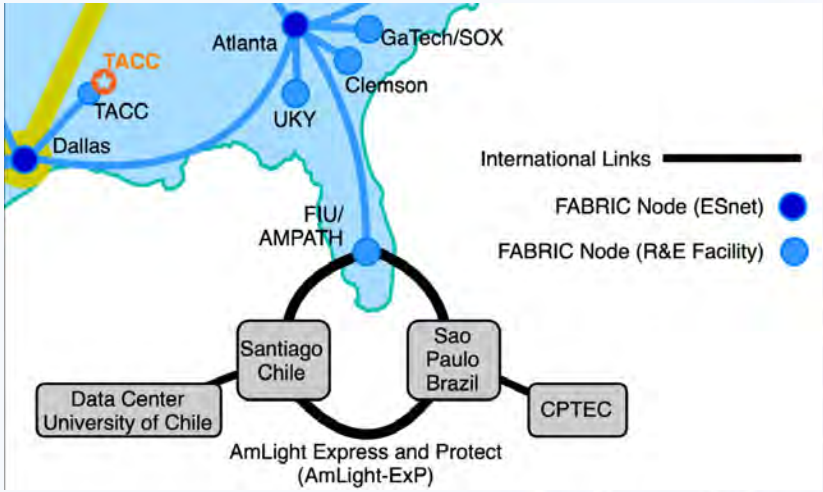
- TransPac
- WIDE (to TransPac)
- University of Tokyo
- Japanese Science Information Network (SINET)
- Korean Institute of Science and Technology Information (KISTI)
- Korea Research Environment Open Network (KREONET)
- StarLight International/National Communications Exchange Facility



Future: Hawaii (?) and/or Guam (?)

FAB Network & Facility Partners: South America

- AmLight International Exchange Point
- FIU
- University of Miami
- Center for Weather Forecast and Climatic Studies (CPTEC)
- AmLight Express and Protect
- Academic Network at Sao Paolo (ANSP)
- RedCLARA



FABRIC

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov