FABRIC

(NSF Midscale CyberInfrastructure Project)

U. South Carolina P4 Workshop

Paul Ruth, RENCI - UNC Chapel Hill

August 22, 2023

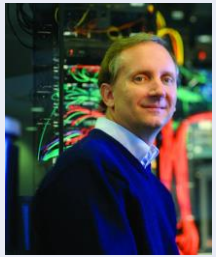
FABRIC

a NSF Mid-scale Research Infrastructure, empowers researchers to **securely prototype** and **validate disruptive designs** in wide-area networks. With its inclusion of advanced hardware uncommon in routers and switches, FABRIC facilitates experiments that **closely emulate real-world production** environments. By prioritizing connectivity with existing research and compute facilities, FABRIC enhances its usability and relevance.

FABRIC

FAB
FABRIC Across Borders

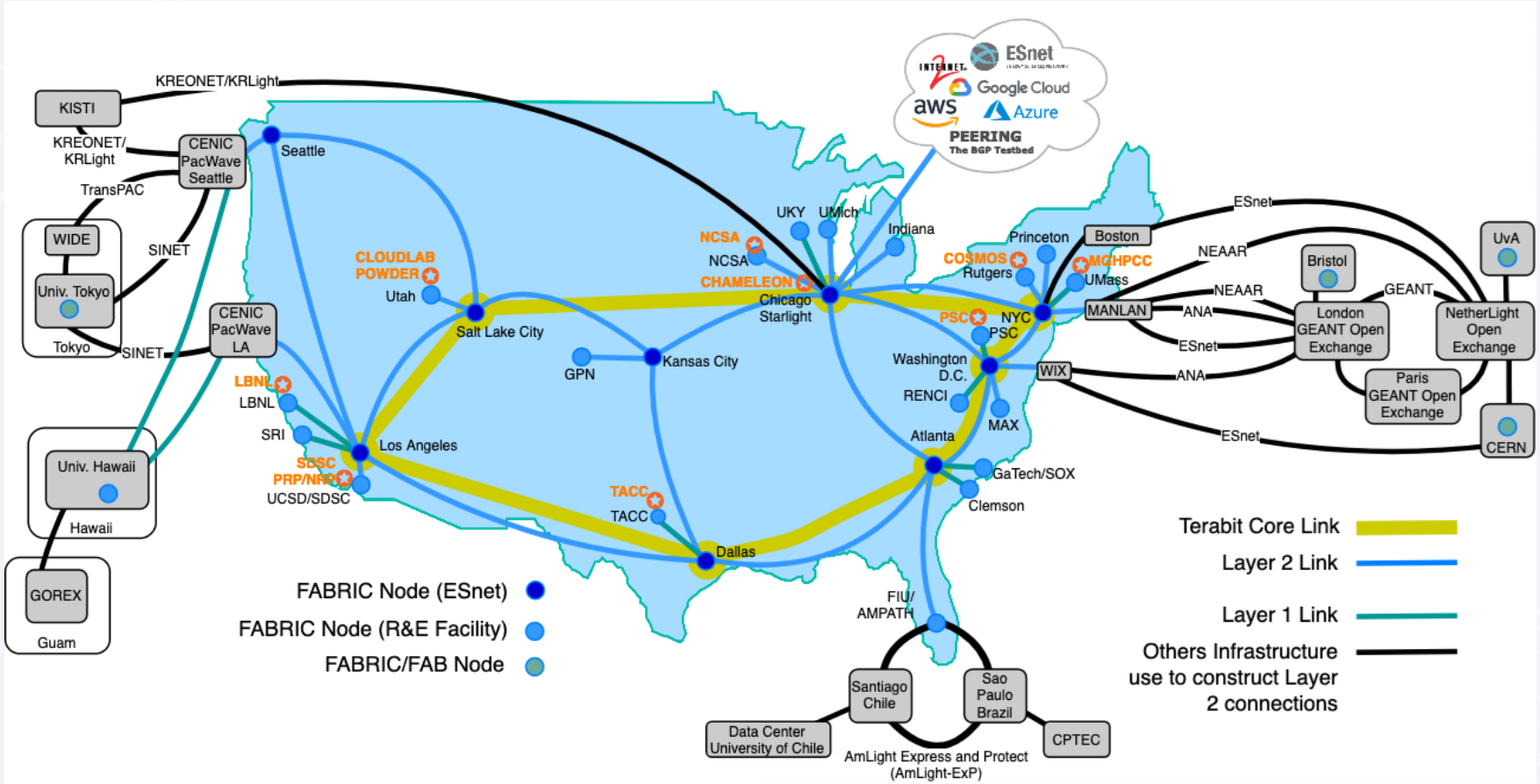
FABRIC
Leadership Team

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

FABRIC

FAB
FABRIC Across Borders

FABRIC: Networking Experiments at Scale



Data Staging & Transfer

- Access to data repositories/instrument streams
- Dissemination of results
- Inter-domain data transfers (cloud, compute centers, campus)

Edge Computing

- Combining edge and Cloud

Distributed AI

- Federated Learning
- Distributed learning/inference

Remote Device Control

- Microscopes, telescopes, medical devices, etc.

Computational Workflows

- Pegasus and many others

Scientific Computing: Across Networks

Facilities

- HPC/RCD Compute Centers
- Public/Private Clouds
- Campuses
- Data Repositories
- Large Instruments (microscopes, etc.)

Infrastructure

- Internet2, ESnet, regional RENs
- Data Transfer Nodes (DTNs)
- Science DMZs

Tools & Features

- Data Transfer tools (Globus)
- Workload Managers (SLURM, etc.)
- Workflow Managers (Pegasus, etc.)

Scientific Computing: Resources and Tools

Applications are Limited to Running at the *Edge*

- Limited by edge-to-edge protocols
- Growing emphasis in Edge-Core-Cloud (+HPC) computing paradigm

End-to-end Performance

- Network performance between data repos, instruments, campuses and compute centers
- Requires Internet2 AL2S, optimized DTNs and/or Science DMZs

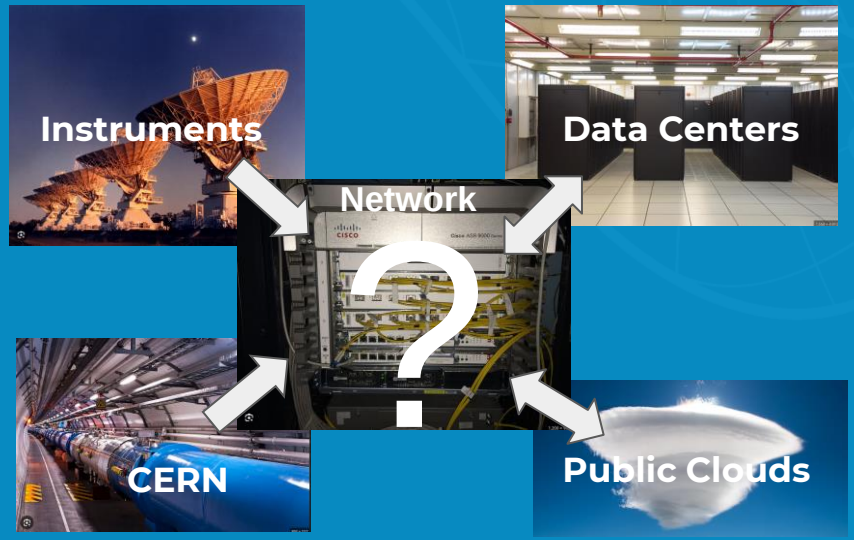
Multi-facility Computation

- Distributed workflows
- Edge-core-cloud paradigm

FABRIC

FAB
FABRIC Across Borders

Realities & Trends



FABRIC: Smart Programmable Cyberinfrastructure

"If I had asked people what they wanted, they would have said faster horses."

– Henry Ford, (maybe)

FABRIC

FAB
FABRIC Across Borders

FABRIC: Smart Programmable Cyberinfrastructure

"If I had asked people what they wanted, they would have said faster Networks."

– Henry Ford, (maybe)

FABRIC

FAB
FABRIC Across Borders

Run Applications Anywhere

- Edge, cloud, or routers in between
- In-network data caching/processing
- Distributed AI learning/inference

Real Facilities

- Experimental core with real edges, clouds, compute centers
- Workflows composing multiple facilities with in-network compute and storage

Production Scale

- Geographic and performance

FABRIC

FAB
FABRIC Across Borders

FABRIC

FAB
FABRIC Across Borders

Sandbox for New Ideas

- Not a replacement for existing infrastructure
- Powerful sandbox for experimentation
- Connected to your real facilities
- Reproducible experiments

Low-level Tools & Protocols

- Dedicated L2/L3 networks
- In-network compute, storage, and accelerators
- Measured bandwidth, latency, jitter

Safe Place to Experiment

- Minimize security exposure
- Rapid test-fix-test cycles

FABRIC

FAB
FABRIC Across Borders

Refine Existing Tools & Apps

- Not a replacement for existing tools
- Powerful sandbox for improving tools

Invent New Tools & Apps

- Powerful sandbox for inventing new tools
- Novel tools and applications using programmable hardware in the network

Impact on Cyberinfrastructure

FABRIC makes the previously unthinkable possible by enabling access to a programmable network core, empowering academic researchers.

FABRIC

FAB
FABRIC Across Borders

Next Generation Cyberinfrastructure

FABRIC

FAB
FABRIC Across Borders

Science Unbound

- Applications natively running across platforms and domains
- Smart in-network processing, caching, and data collection/distribution

New Layers of Software

- Simplified tools for deploying of HPC/RCD applications throughout the edge-core-cloud
- Smarter faster tools for managing scientific workflows and data

Prototype Validation

Full Scale Prototypes

- Production scale
- Connect to real facilities
- Secure Sandbox

Measured Validation

- Justification for expensive/risky systems
- Path for transition to production deployments

FABRIC

FAB
FABRIC Across Borders

Workforce Development & Learning

Educating Next Generation of CI Engineers

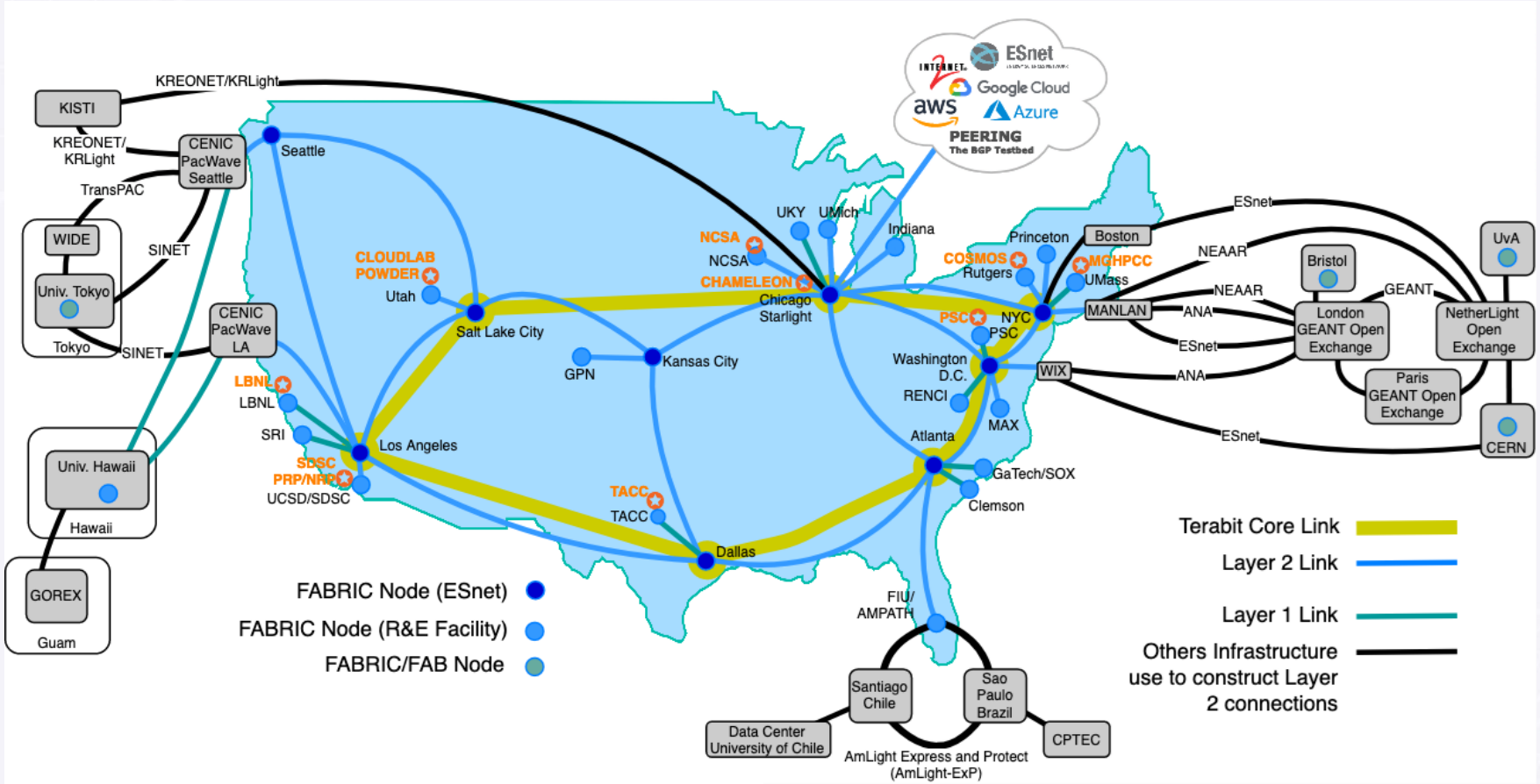
- Hands-on learning
- Experiment in the network core with minimal impact
- Evaluate risky technologies without compromising security of production infrastructure
- Educate with a secure production infrastructure

Shared Discoveries

- Platform for sharing new tools and technologies
- Reproducible experimental artifacts
- Publishable experiments

Design and Architecture

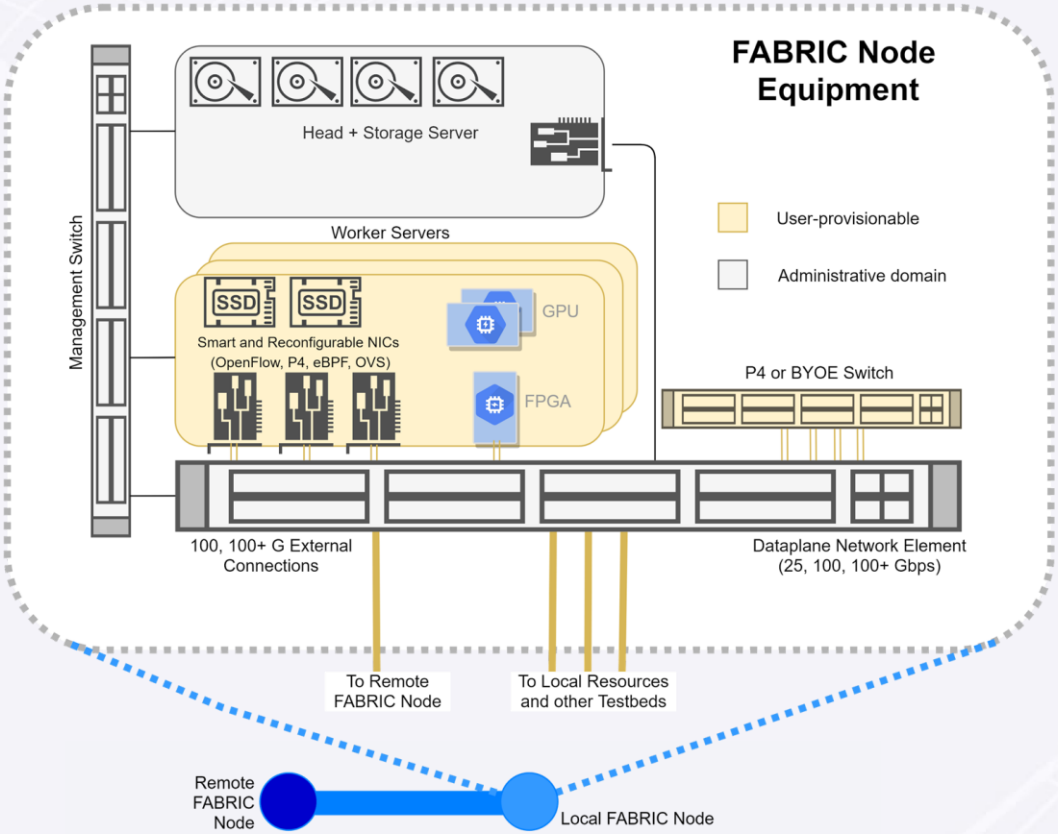
FABRIC: Networking Experiments at Scale



Hardware

- Rack of high-performance servers (Dell 7525) with:
  - 2x32-core AMD Rome and Milan with 512G RAM
  - GPUs (NVIDIA RTX 6000, T4, A30), 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
  - Tofino-based P4 switches (4 sites)
- Reconfigurable Network Interface Cards
  - FPGAs (U280 XILINX with P4 support)
  - Mellanox ConnectX-5 and ConnectX-6 with hardware off-load
  - Multiple interface speeds (25G, 100G, 200G+(future)
- Kernel Bypass/Hardware Offload
  - VMs sized to support full-rate DPDK for access to Programmable NICs, FPGA, and GPU resources via PCI pass-through

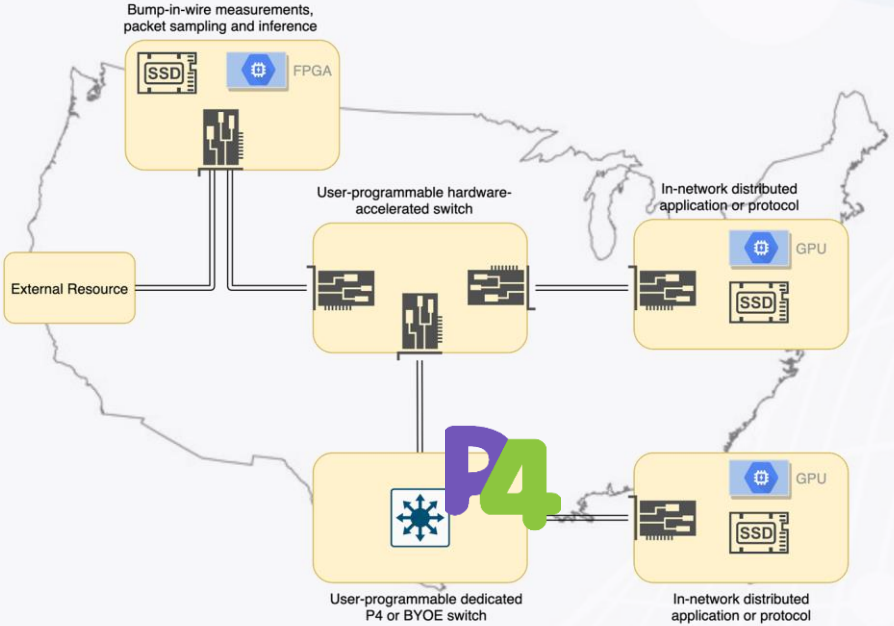
FABRIC “Hank”



Hank: a measured unit of coiled or wrapped yarn or twine

Features and Capabilities

- **Networking**
  - L2: Local bridges and wide-area circuits
  - L3: FABnet IPv4/IPv6
  - Dedicated smart NICs (ConnectX-5/6)
  - Tofino P4 Switches (coming soon)
- **Storage**
  - Local disk
  - Dedicated NVMe PCI devices
  - Persistent network storage volumes
- **GPUs**
  - NVIDIA RTX 6000
  - NVIDIA Tesla T4
  - NVIDIA A30
- **FPGAs**
  - Xilinx U280 (2x100 Gbps Network Ports)
- **Facility Ports**
  - L2 connections to external facilities
  - Chameleon, Cloudlab, and more.



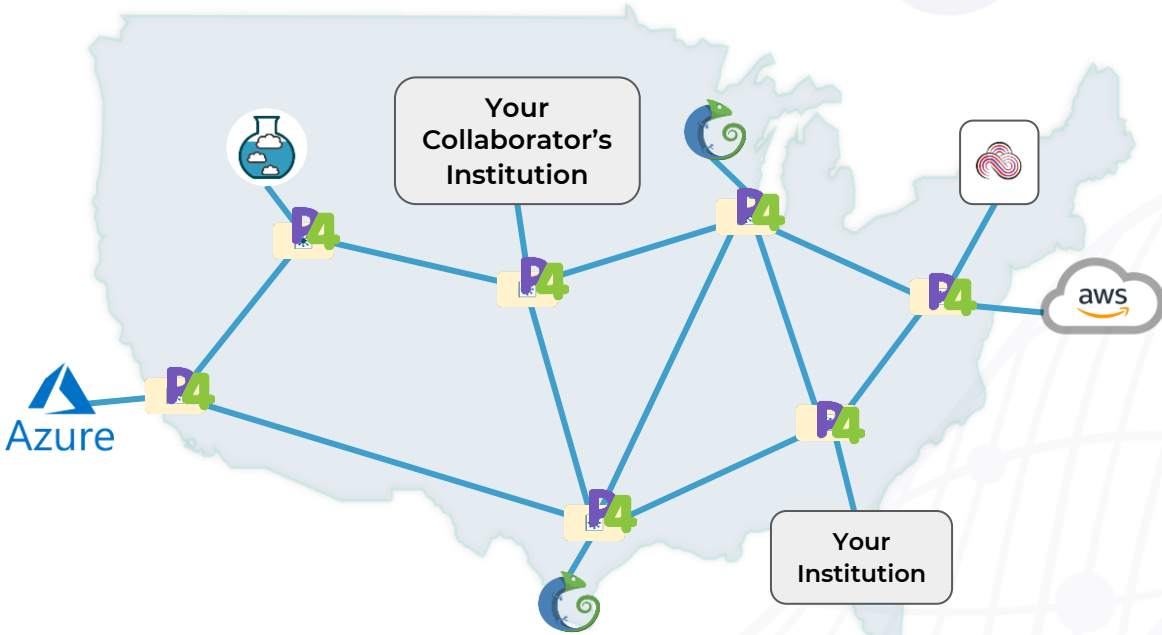
FABRIC Experiments

Programmable Internet core

- Smart routing and switching
- In-network processing
- In-network caching

Realistic experiments

- At-scale geographic distribution
- At-scale performance
- Connections to real external facilities
- Connections to other testbeds you can use



FABRIC

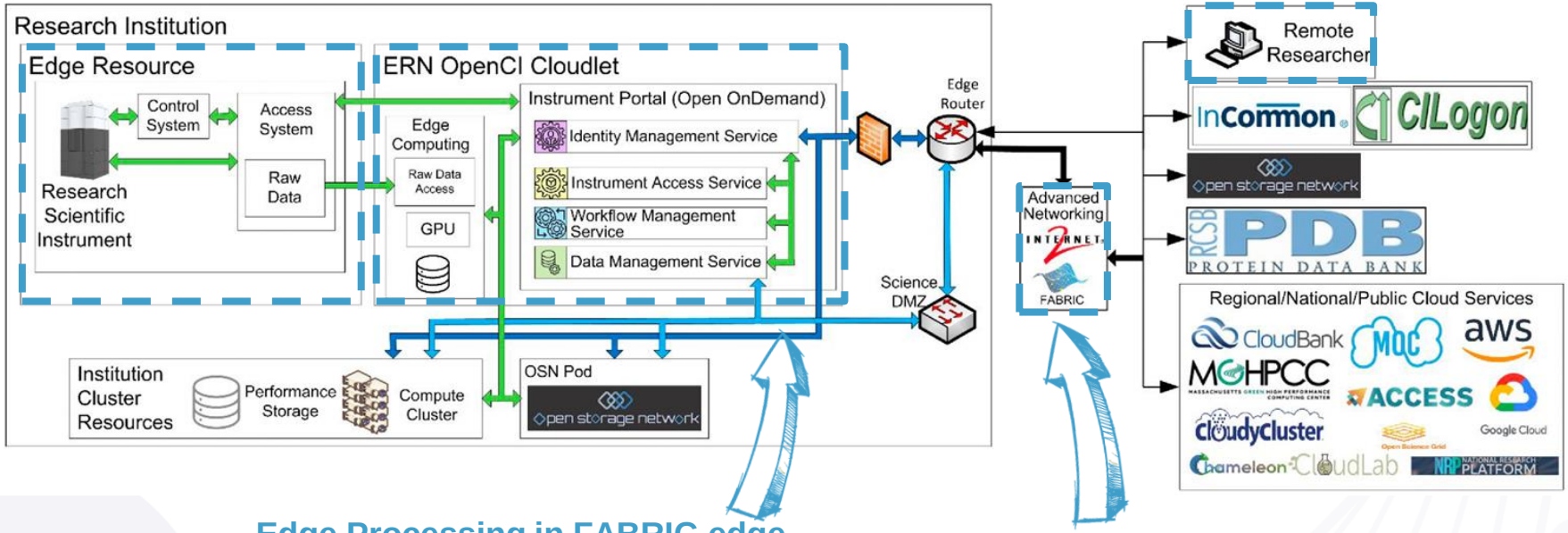
FAB
FABRIC Across Borders

Example Use Cases

FABRIC

FAB
FABRIC Across Borders

Remote Instrument Collaboration: Cryo-EM



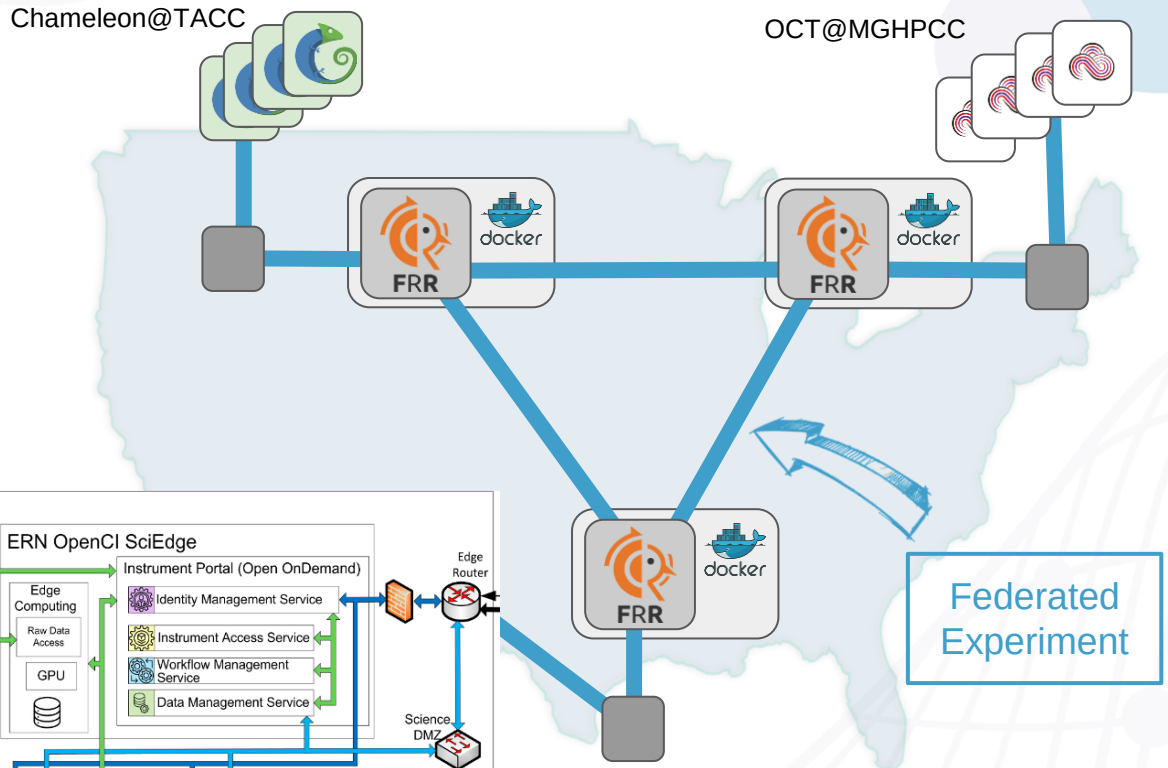
Edge Processing in FABRIC edge (or instrument's cluster)

Network Optimized for Interactive Control (Low Latency)

https://www.ernrp.org/cryoem-remote-instrument/

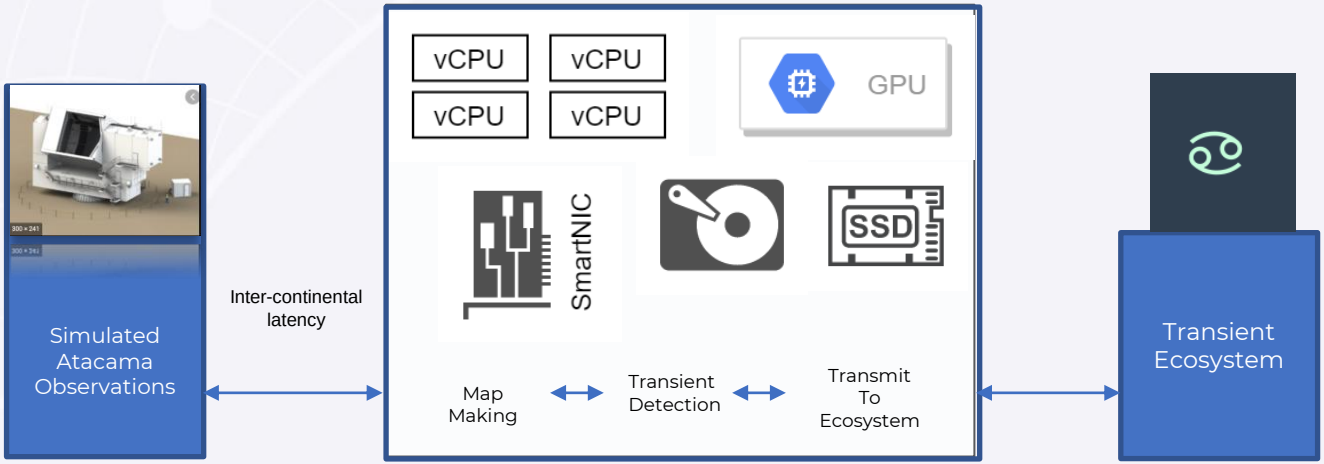
Remote Instrument Collaboration: Cryo-EM

- Packaged for Jupyter
  - Notebooks plus tools
- FRRouters
  - Dockerized
  - Three sites
  - Gateways for local network
- Protocol: OSPF
  - Many other protocols available
- Local networks
  - Cryo-EM @ Rutgers
  - MGHPCC
  - Chameleon



https://www.ernrp.org/cryoem-remote-instrument/

Cosmology

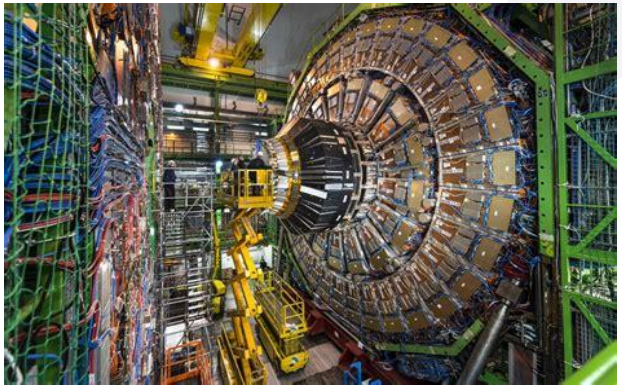


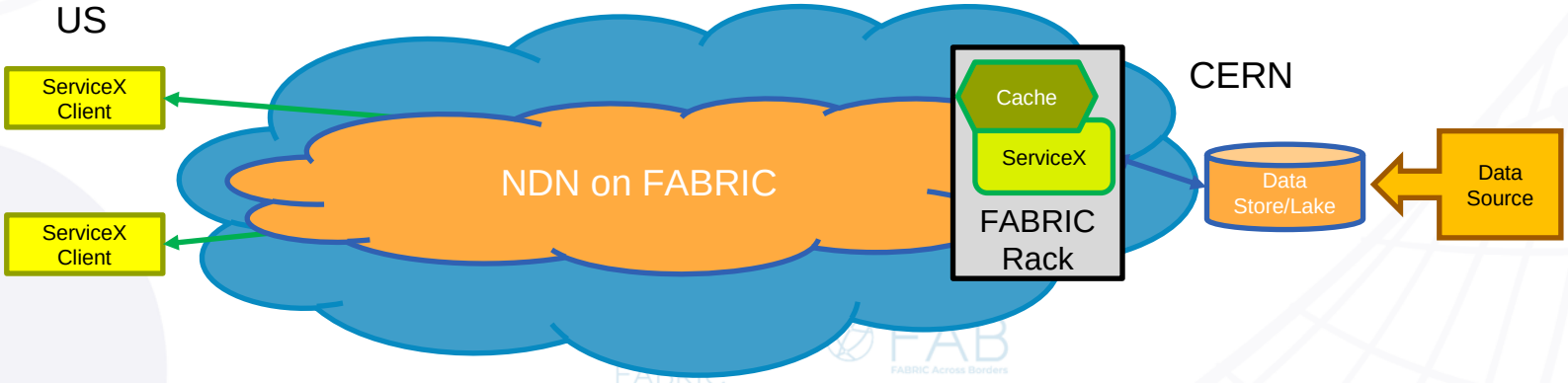
- Don Petravick, Gregory Daues (UIUC/NCSA)
- Designed/deployed CMB-S4 experiment(s) on FABRIC
- Simulated observatory source at FIU (projected actual path)
- In-network data processing
- Implemented a shell on top of FABlib to control their experiment

ServiceX: LHC Atlas

ServiceX over NDN on FABRIC

- Incorporates in-network caching
- Extend the FABRIC ServiceX deployment by replacing TCP/IP flows with NDN data transfers that cache results at every hop in the network





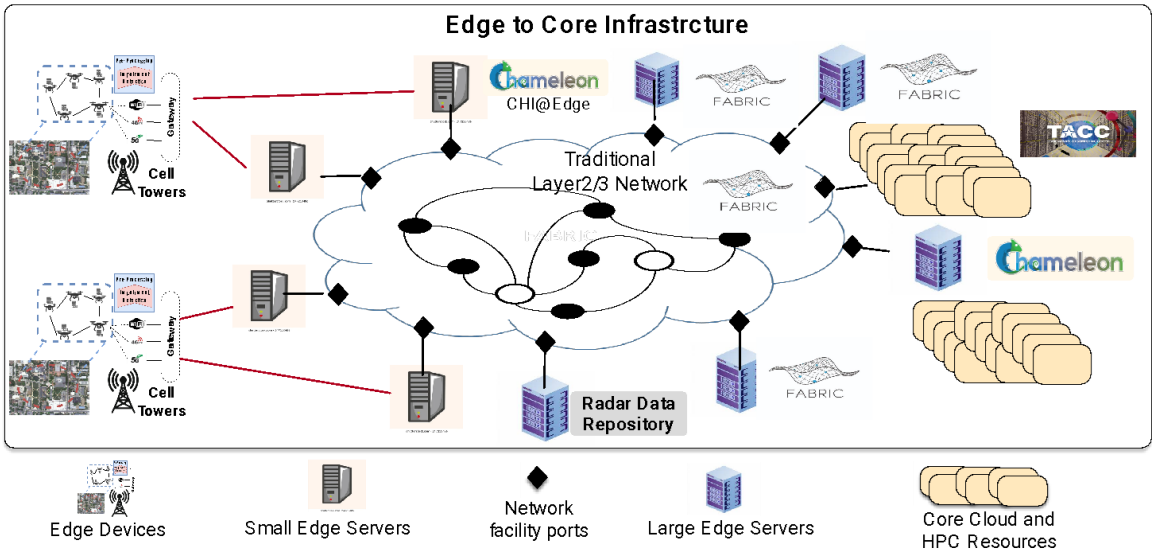
FlyNet

Michael Zink, University of Amherst, PI; contact: zink@ecs.umass.edu

Funded by the National Science Foundation Grant #OAC-2018074

CASA Wind Workflow

Orcasound Workflow



Scientific workflows deployed by the Pegasus workflow management system in the edge-to-cloud continuum leveraging the FlyNet architecture[1];
Using resources from Chameleon and/or FABRIC and utilizing the FABRIC network to move data

1. R. Tanaka, G. Papadimitriou, S. Charan, C. Wang, E. Lyons, K. Thareja, C. Qu, A. Esquivel, E. Deelman, A. Mandal, P. Calyam, and M. Zink, "Automating Edge-to-cloud Workflows for Scien Traversing the Edge-to-cloud Continuum with Pegasus", in *2nd ACM/IEEE International workshop on Cloud-to-Things continuum: towards the convergence of IoT, Edge and Cloud Computing (Cloud2Things 2022)*, Taormina, Messina Italy, May 2022, pp. 826-833, doi: 10.1109/CCGrid54584.2022.00098.

#OAC-2018074

renci

USC Viterbi School of Engineering Information Sciences Institute

UMass Amherst

College of Engineering University of Missouri

Education: P4

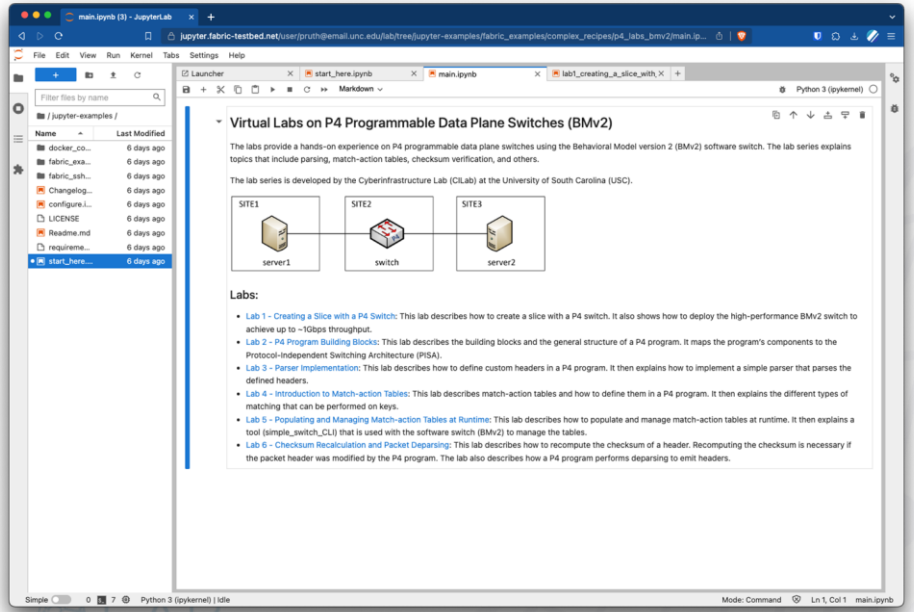
UNIVERSITY OF
South Carolina

Educational Laboratory

- Cyberinfrastructure Lab (CILab)
- University of South Carolina
- Available in FABRIC's JupyterHub

P4 Experiments

- Software P4 BMv2
- Learn match action tables
- Deployed across real distributed infrastructure
- Tofino P4 Switches (coming soon)



The Community: FABRIC & Workshops

Users

- 500+ Total
- ~200 active users in the last 6 weeks

Projects

- 74+ Total
- 57+ Public
- 17+ Private

Organizations

- 108+ distinct organizations

KNIT Workshops

- ~150 virtual attendees
- 80+ in-person attendees
- 6+ countries represented

Join the Community!

FABRIC Account

https://portal.fabric-testbed.net/

Newsletter Signup

bit.ly/FABRICnewsletter

KNIT7 Workshop

September 27-29, 2023 in Columbus, OH
Co-located with NSF CC* PI workshop
Consider attending and inviting others who may be interested.

Ambassadors Program

Coming soon! Join our newsletter for updates.

FABRIC

FAB
FABRIC Across Borders

Thank You!

Questions?

Visit https://whatisfabric.net

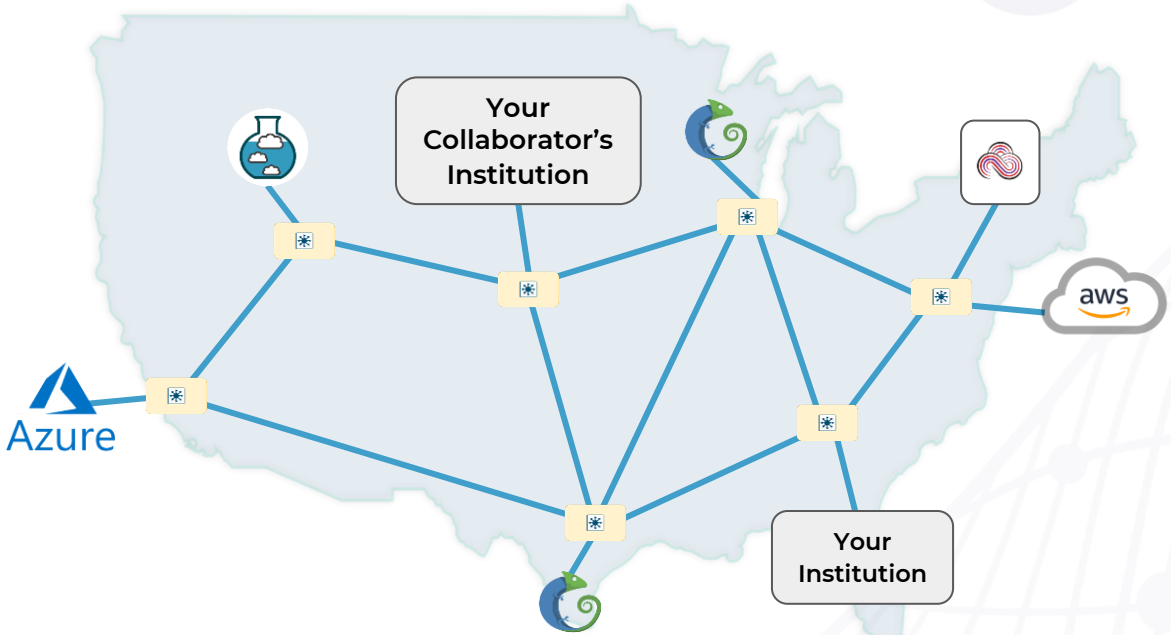
This work is funded by NSF grants CNS-1935966, CNS-2029261, CNS-2029235, CNS-2029200, CNS-2029261, CNS-2029260

FABRIC

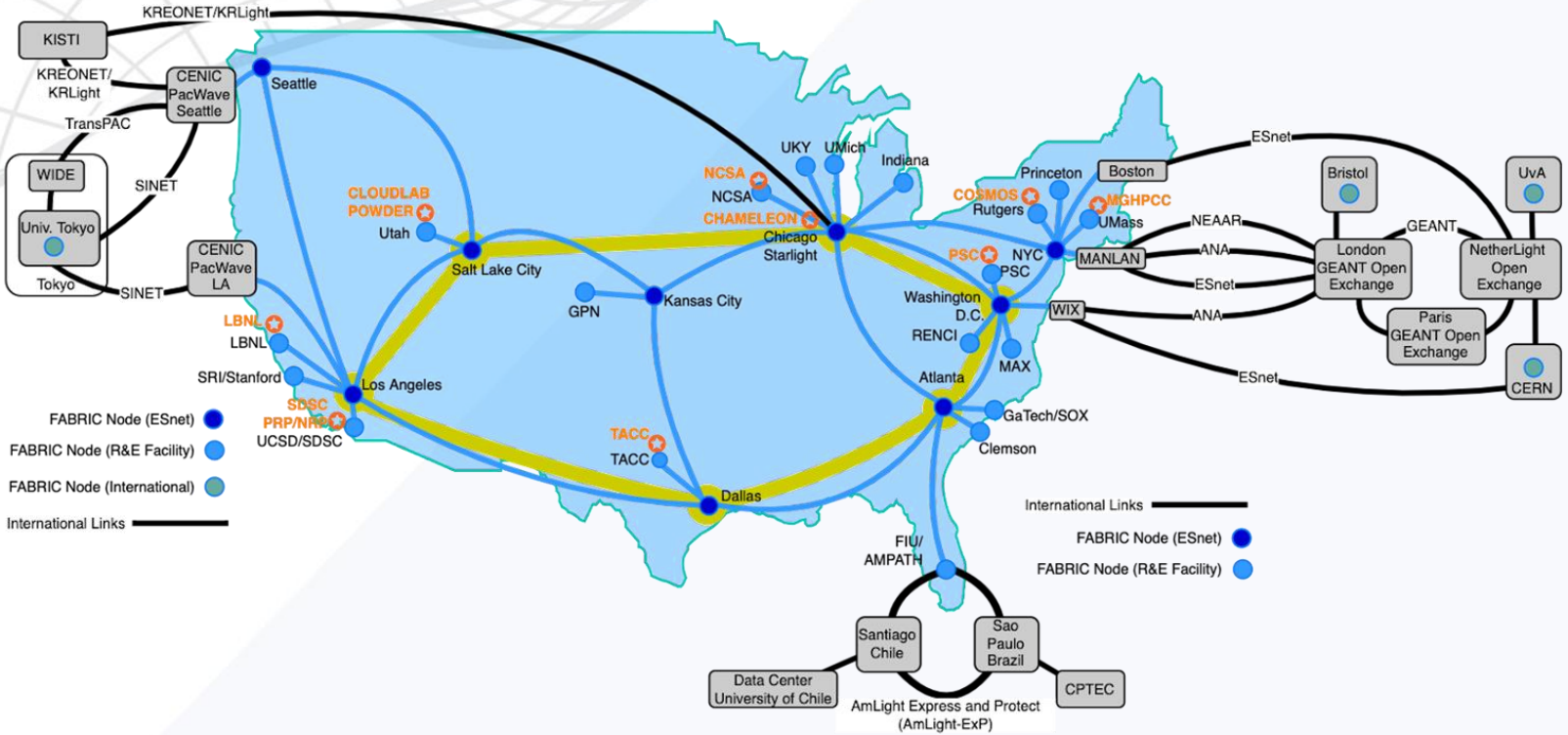
EXTRA SLIDES

FABRIC Experiments

- Programmable Internet core
  - Smart routing and switching
  - In-network processing
  - In-network caching
- Realistic experiments
  - At-scale geographic distribution
  - At-scale performance
  - Connections to real external facilities
  - Connections to other testbeds you can use



FABRIC + FAB

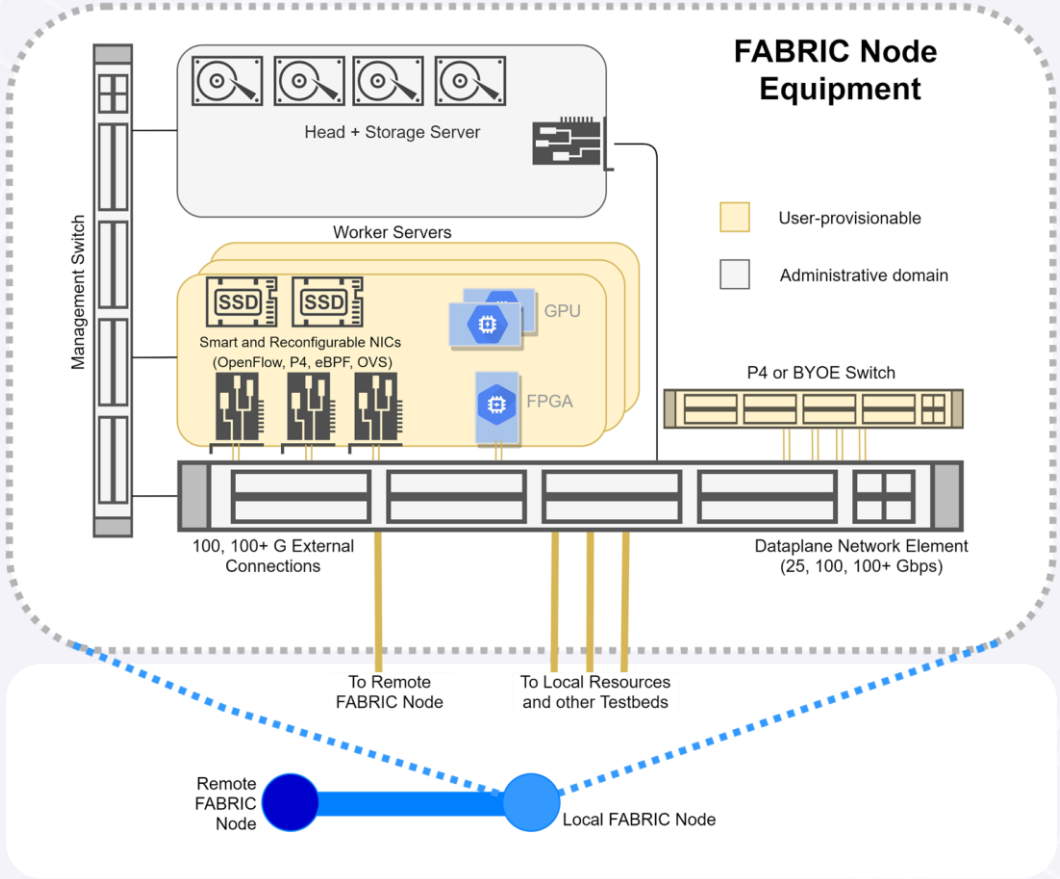


FABRIC

FAB
FABRIC Across Borders

FABRIC Topology - with FAB Sites

FABRIC Node (Rack)



Access Integration (Ilya/Jim)

- CI Logon
- Allocations
- Access Support
- OMNISOC
- Cloudbank

Hardware

- **Rack of high-performance servers (Dell 7525) with:**
  - 2x32-core AMD Rome and Milan with 512G RAM
  - GPUs (NVIDIA RTX 6000, T4, A30), 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
  - Tofino-based P4 switches (4 sites)
- **Reconfigurable Network Interface Cards**
  - FPGAs (U280 XILINX with P4 support)
  - Mellanox ConnectX-5 and ConnectX-6 with hardware off-load
  - Multiple interface speeds (25G, 100G, 200G+(future)
- **Kernel Bypass/Hardware Offload**
  - VMs sized to support full-rate DPDK for access to Programmable NICs, FPGA, and GPU resources via PCI pass-through

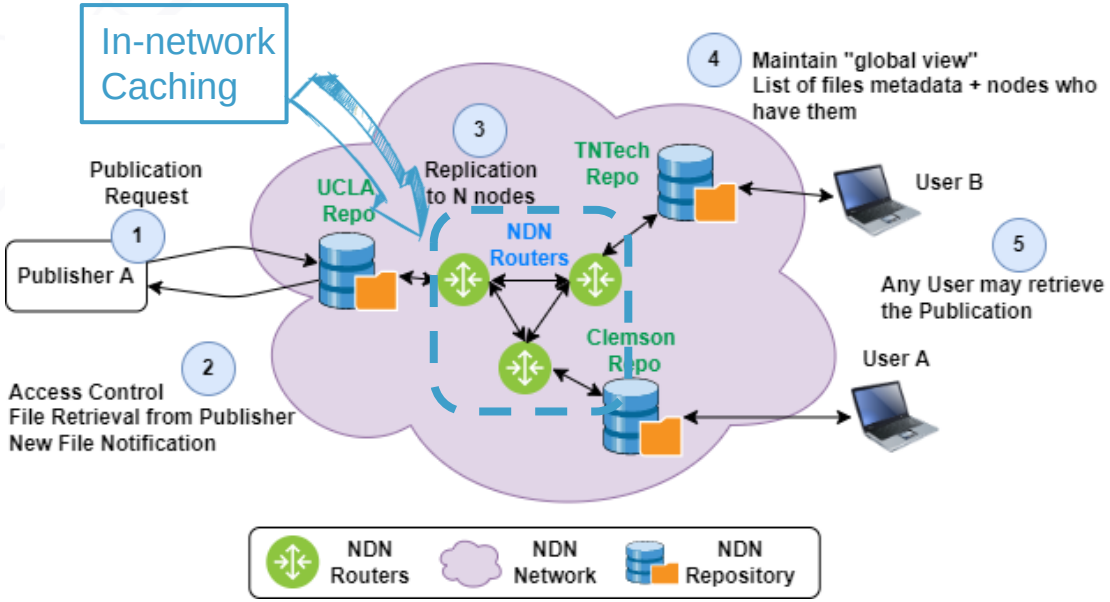
What is FABRIC

FABRIC is a scientific instrument and laboratory that enables a *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.

Hydra: Secure Distributed Storage Framework

(A Federated Data Repository for Big Data)



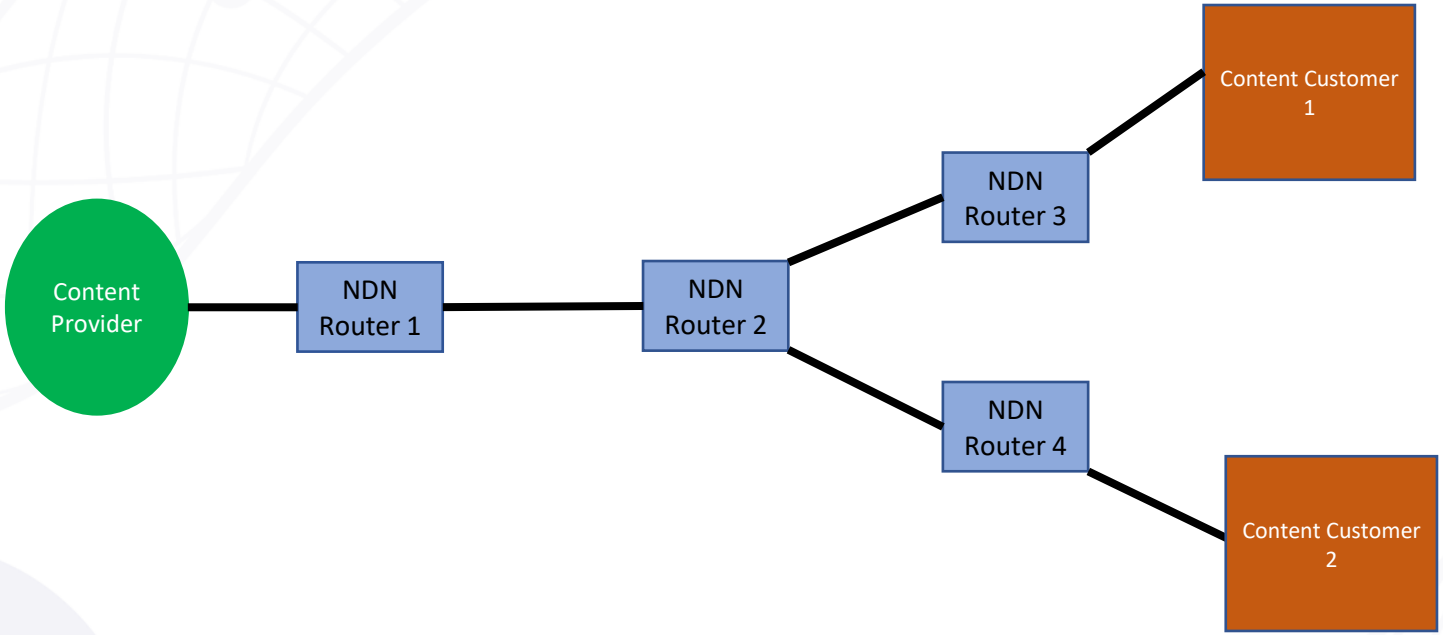
Hydra is completely content centric and does not establish any host-to-host connectivity, allowing location independent data movement and replication.

Source: KNIT5 FABRIC Community Workshop (9/21/22), "Hydra - A Federated Data Repository for Big Science". See also www.hydra-repo.io/

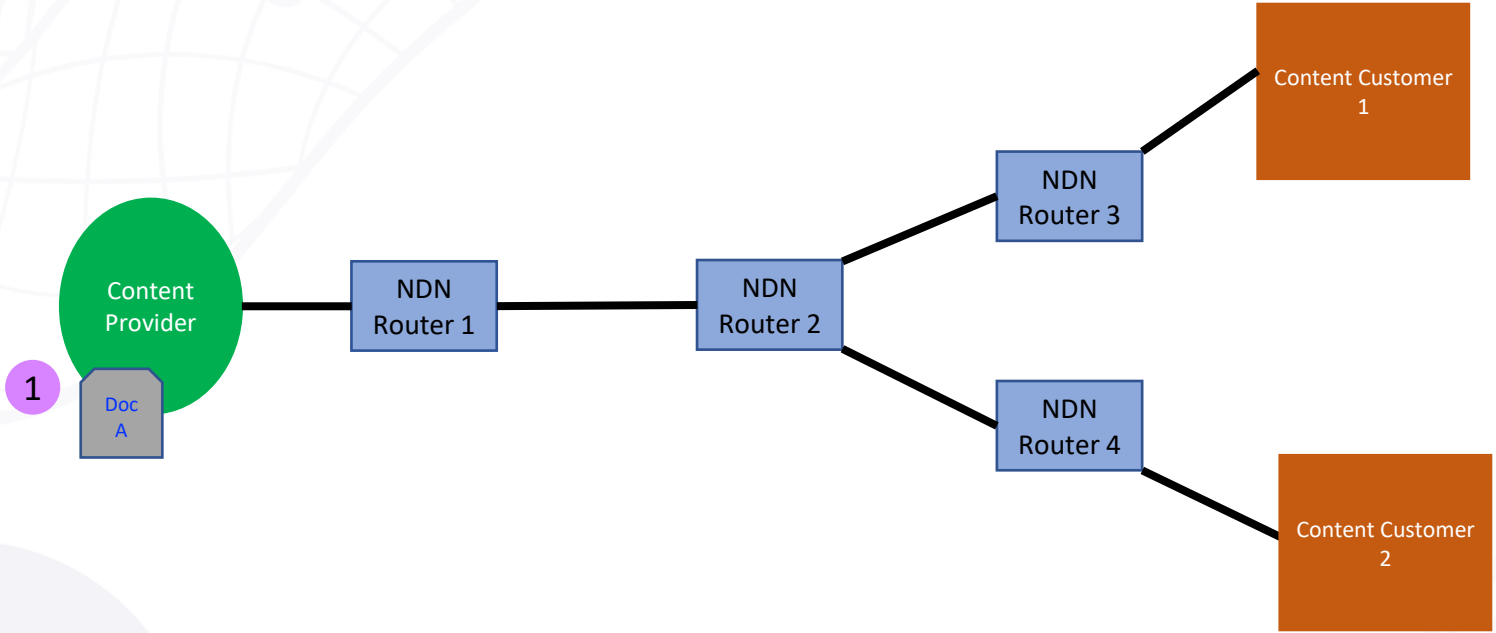
Named Data Networking with In-Network Caching

Additional information: https://named-data.net/

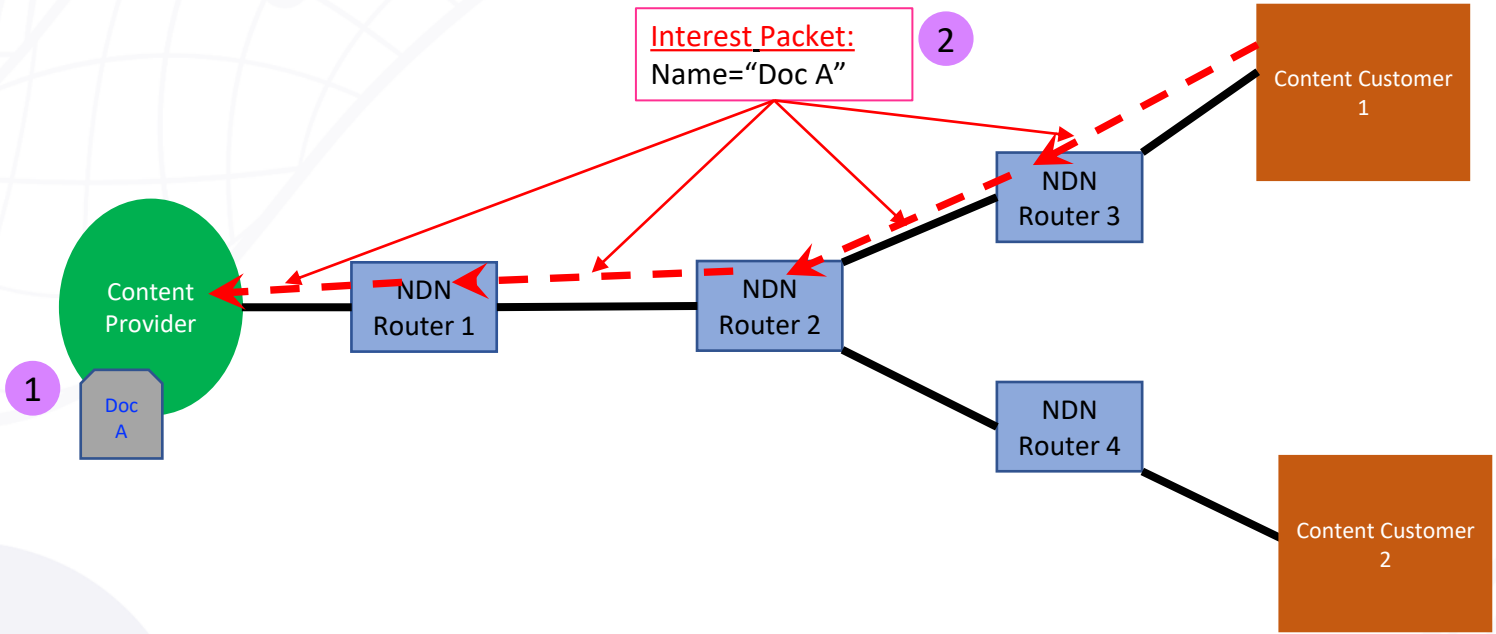
Named Data Networking



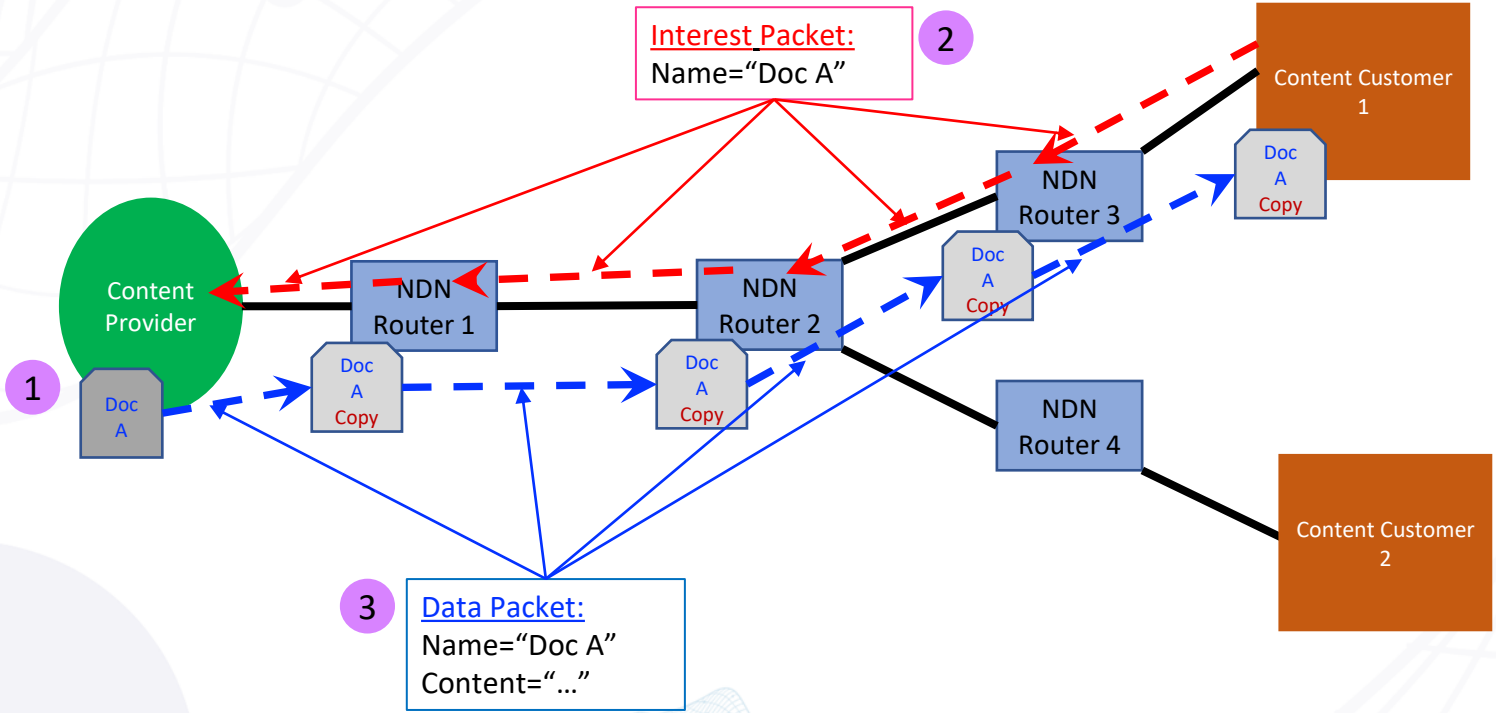
Named Data Networking



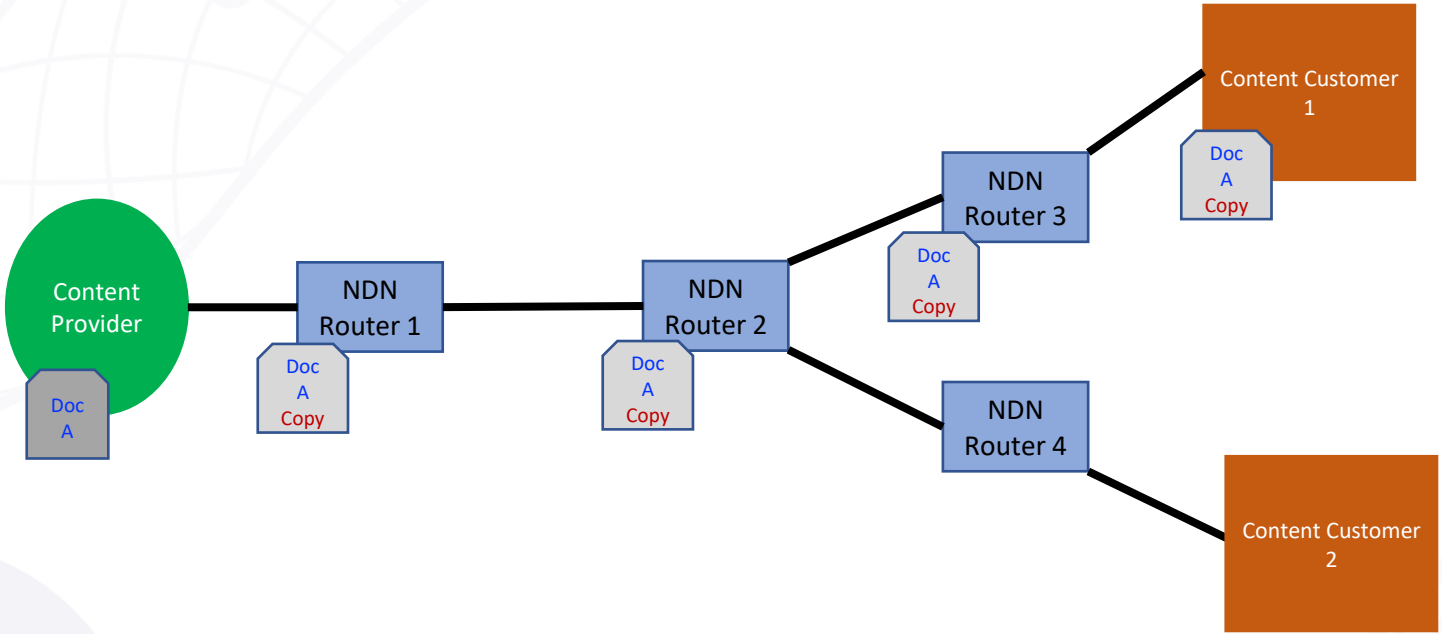
Named Data Networking



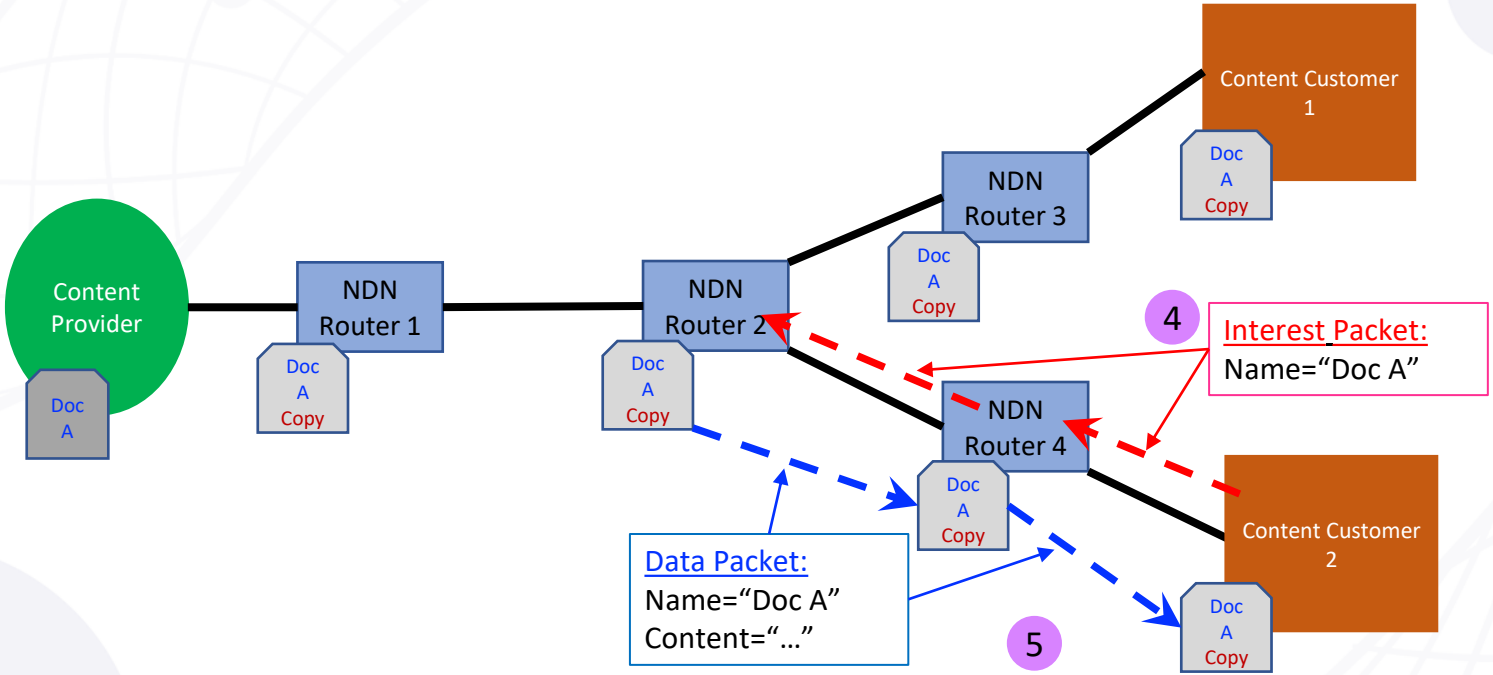
Named Data Networking



Named Data Networking



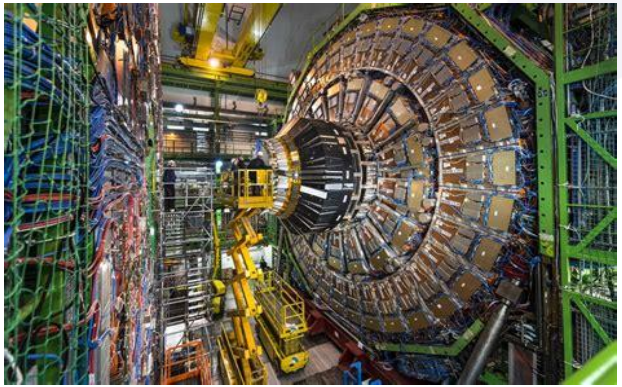
Named Data Networking

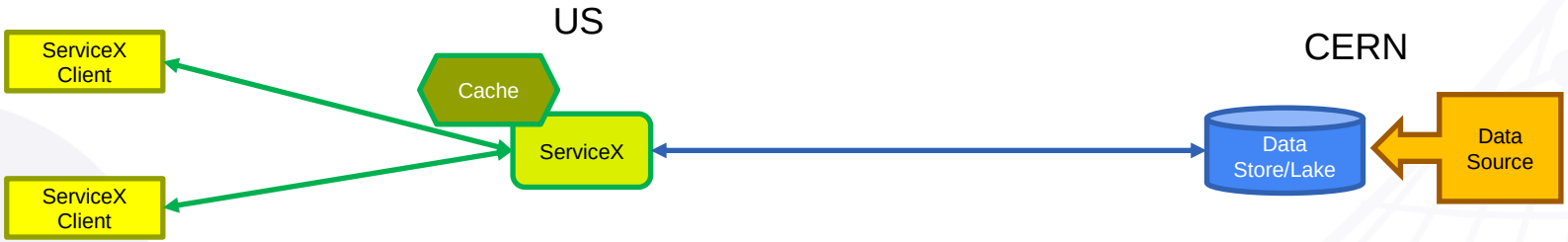


ServiceX: LHC Atlas

ServiceX [1]

- HL-LHC produces exabytes of data
- ServiceX (backend) supports filtering, transformation, processing, and caching of data
- Provides a distributed, caching, columnar data delivery service [2]
- Accessed via ServiceX frontend services



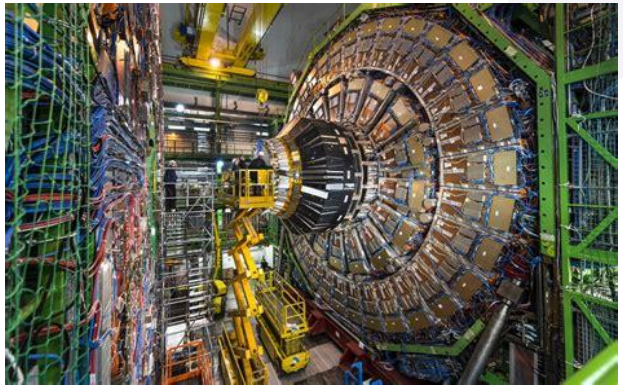


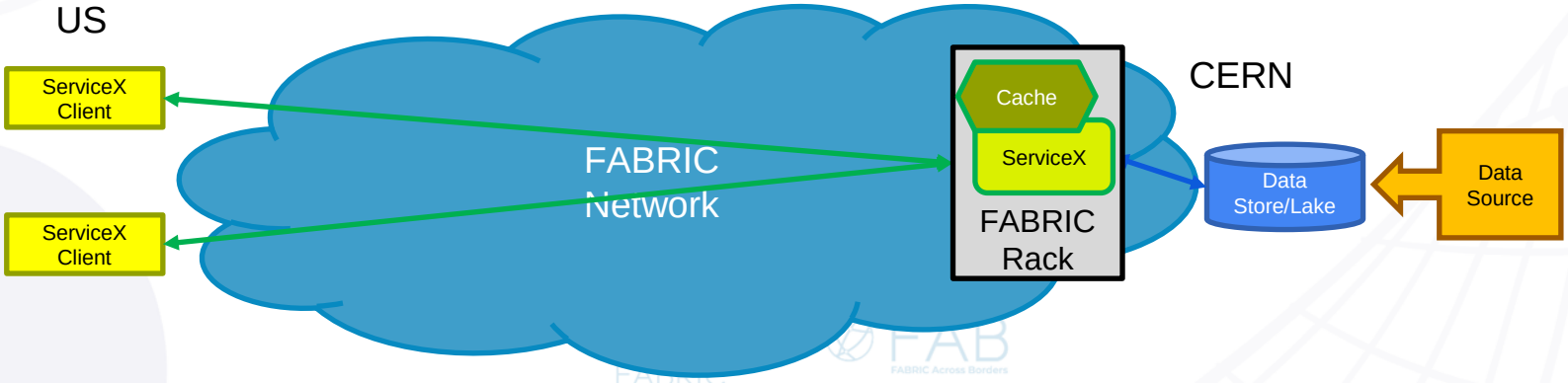
[1] https://servicex.readthedocs.io/en/latest/introduction/
[2] https://indico.cern.ch/event/890991/contributions/3778330/attachments/2007255/3352709/weinbergServiceX200323.pdf

ServiceX: LHC Atlas

FABRIC ServiceX Deployment

- Deploy ServiceX in FABRIC to achieve filtering/transformation near the data source
- Efficiently move data from CERN to UChicago

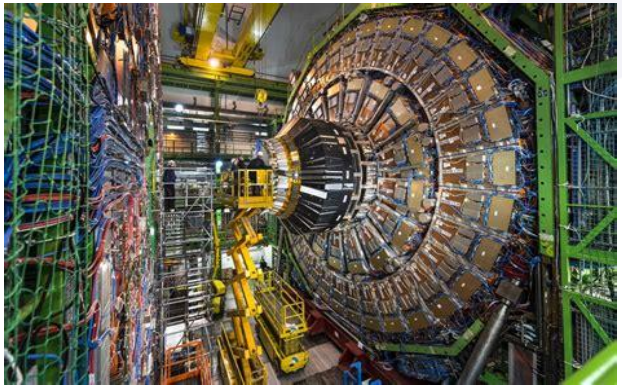




ServiceX: LHC Atlas

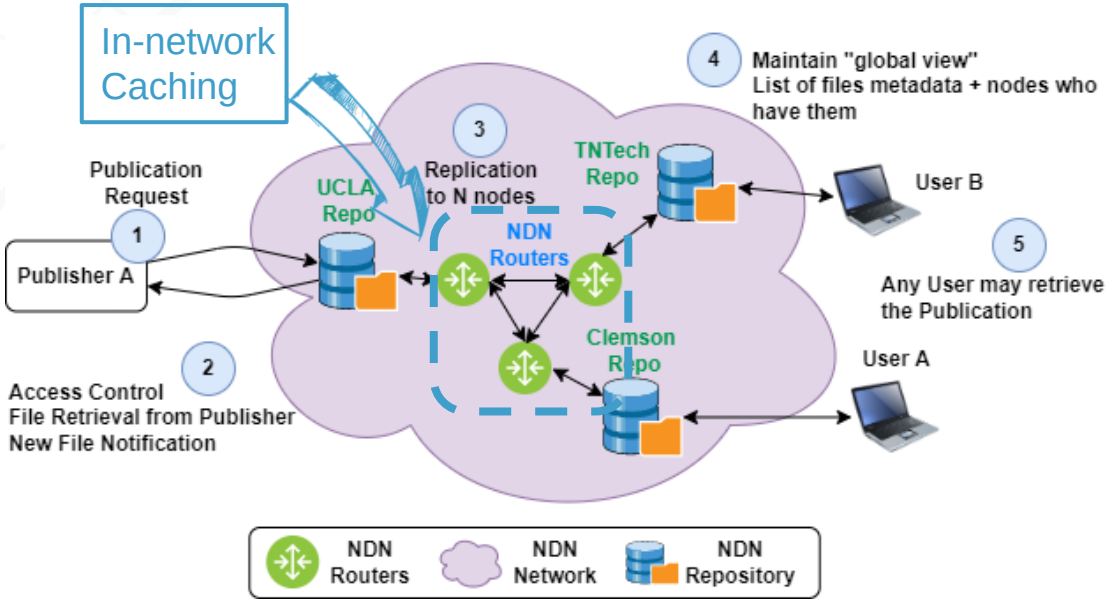
ServiceX over NDN on FABRIC

- Incorporates in-network caching
- Extend the FABRIC ServiceX deployment by replacing TCP/IP flows with NDN data transfers that cache results at every hop in the network



Hydra: Secure Distributed Storage Framework

(A Federated Data Repository for Big Data)



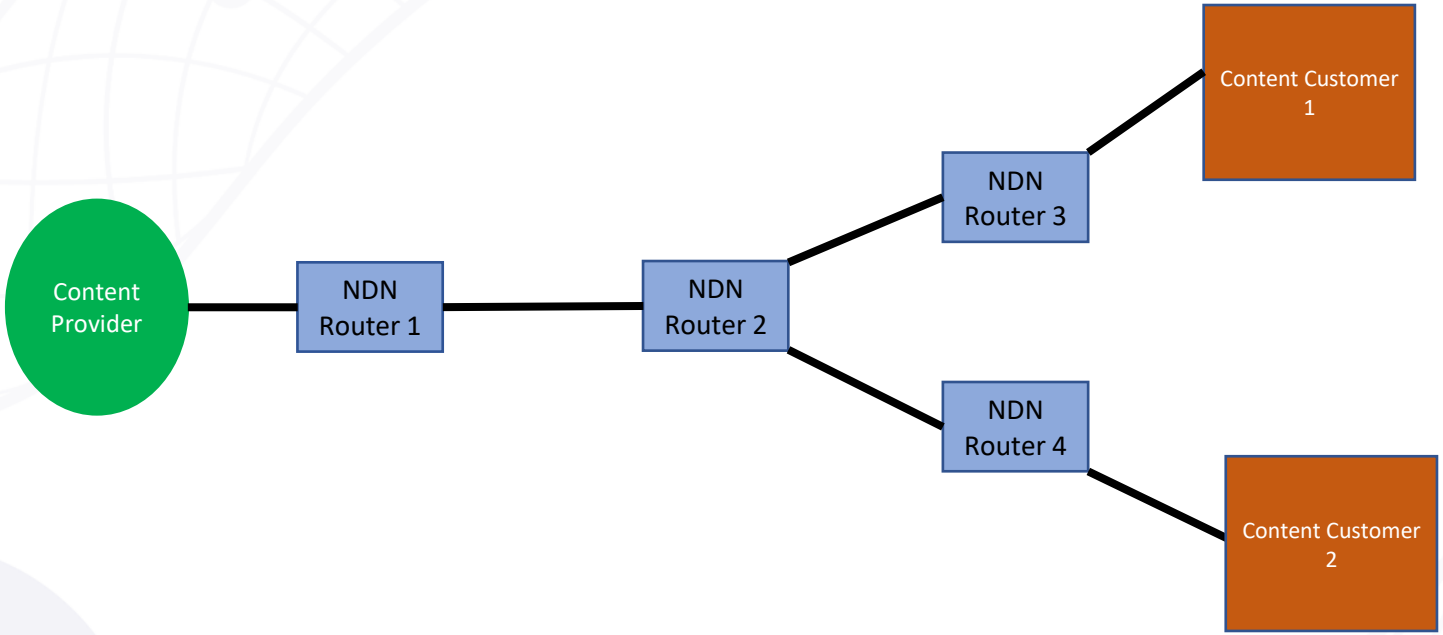
Hydra is completely content centric and does not establish any host-to-host connectivity, allowing location independent data movement and replication.

Source: KNIT5 FABRIC Community Workshop (9/21/22), "Hydra - A Federated Data Repository for Big Science". See also www.hydra-repo.io/

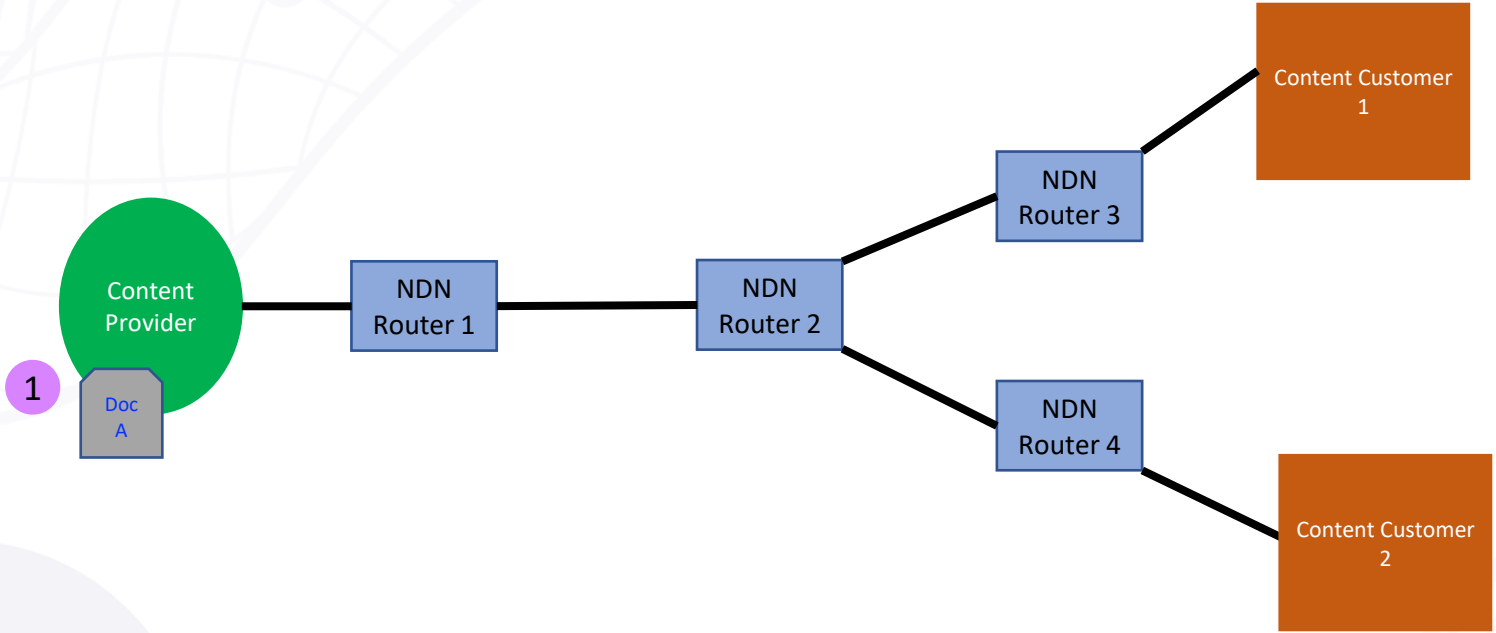
Named Data Networking with In-Network Caching

Additional information: https://named-data.net/

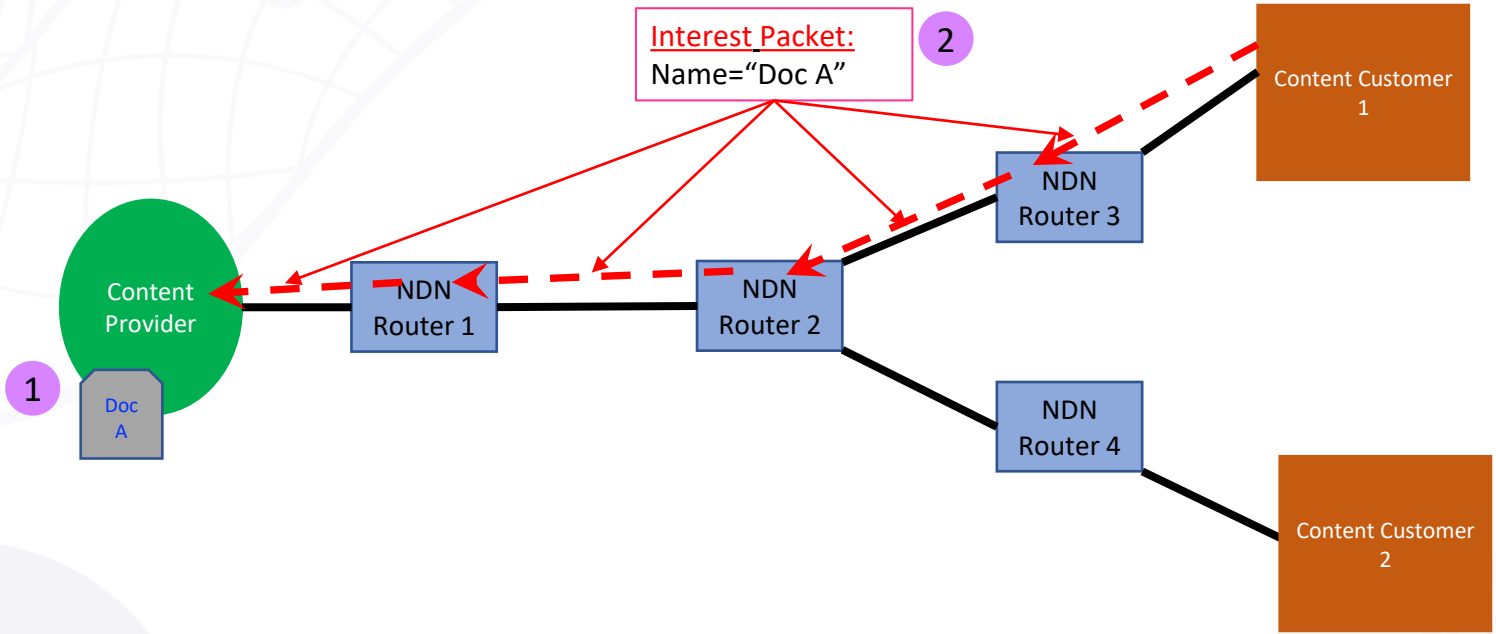
Named Data Networking



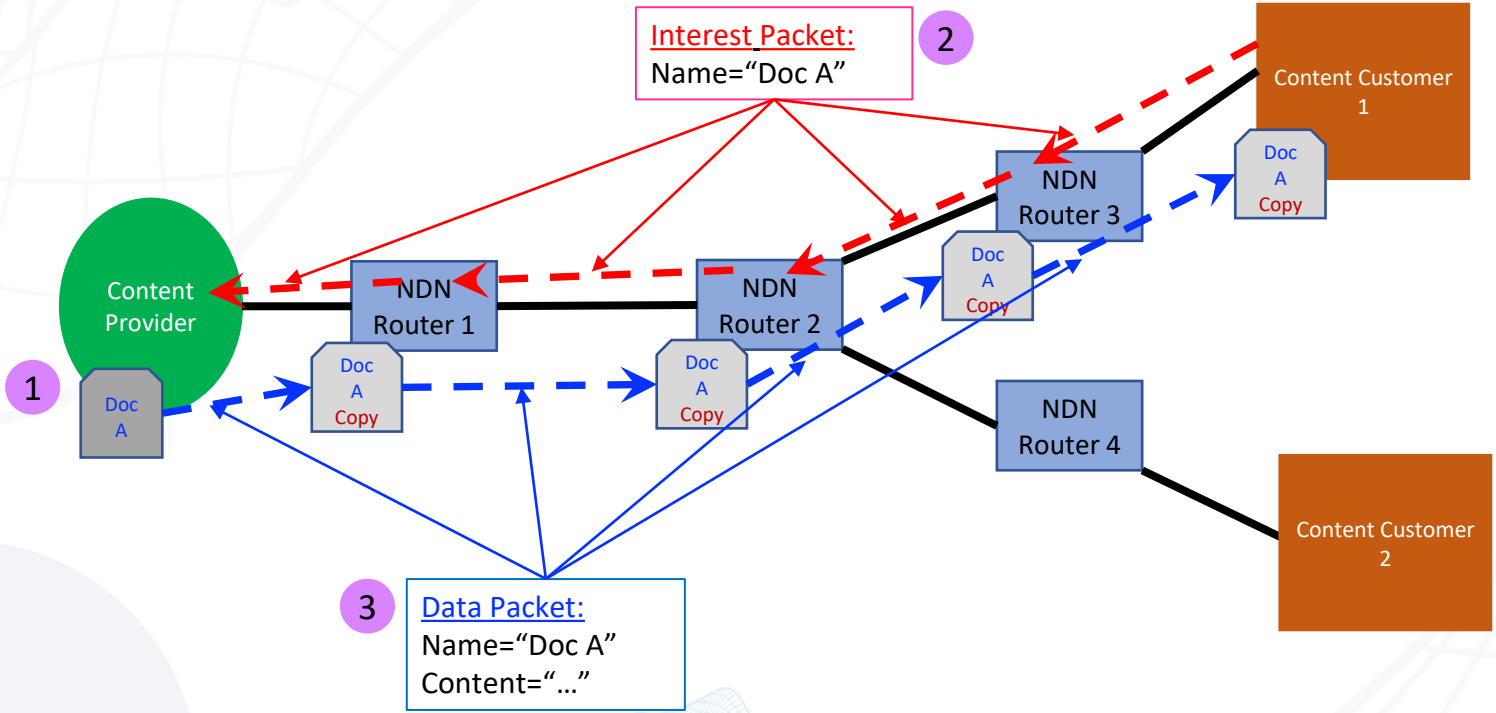
Named Data Networking



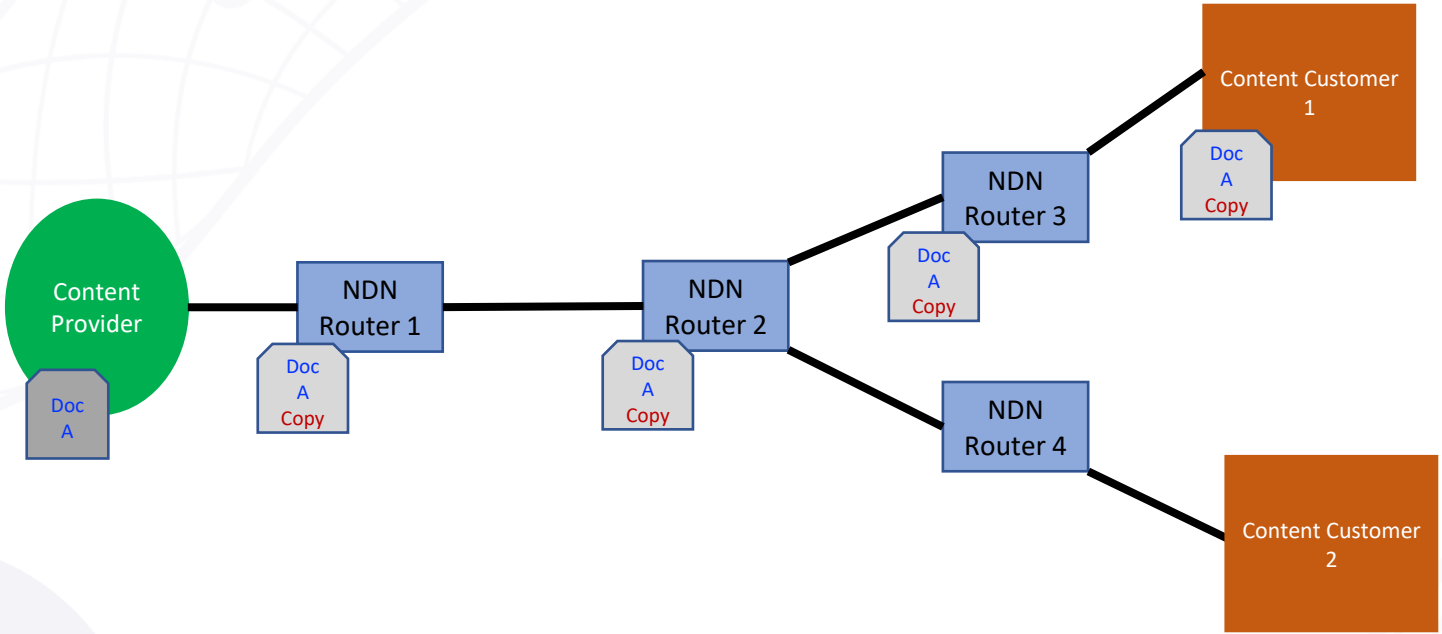
Named Data Networking



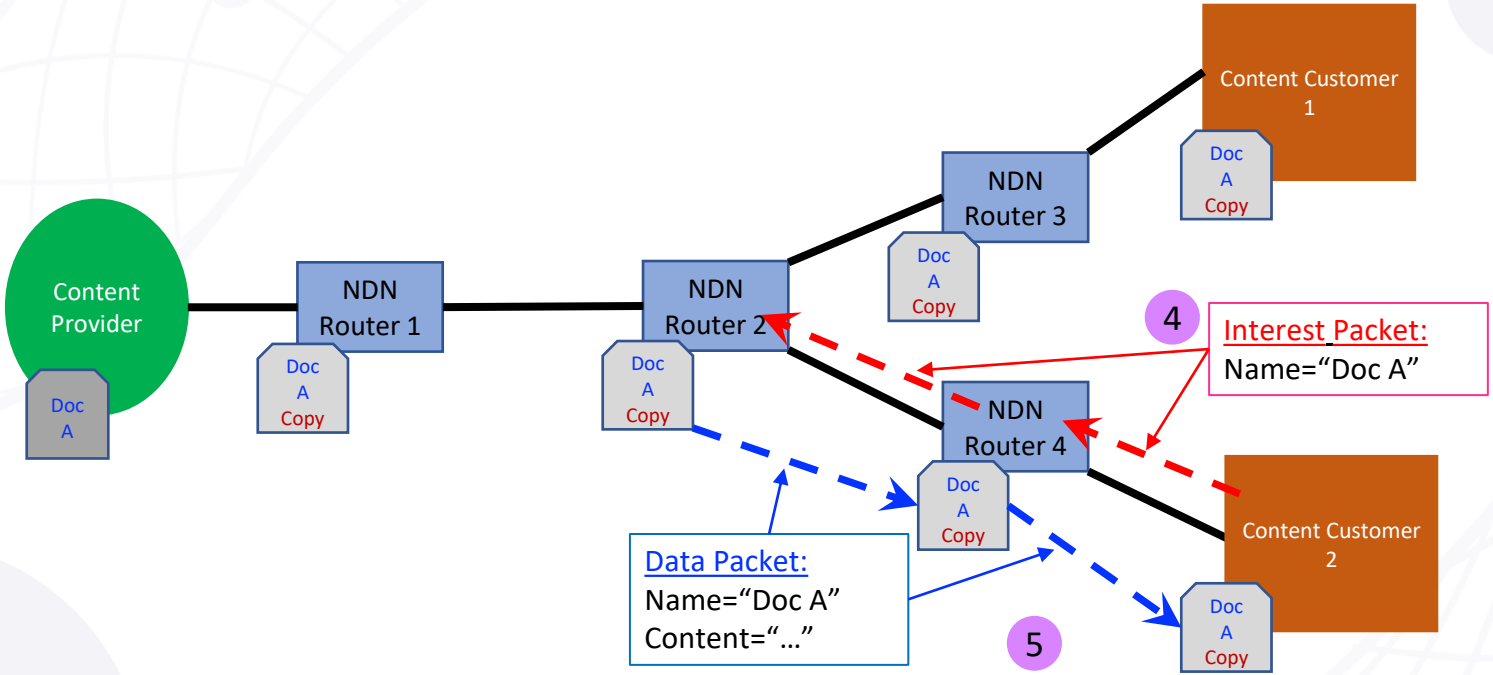
Named Data Networking



Named Data Networking



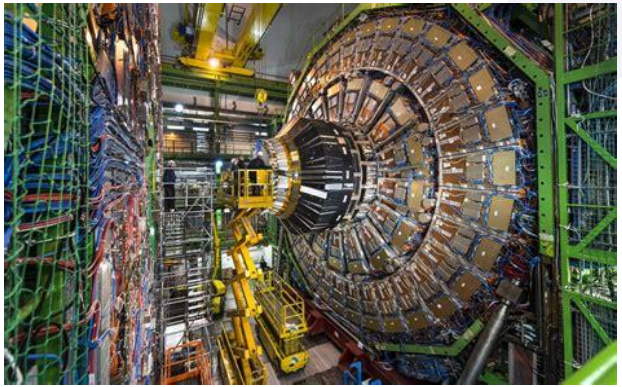
Named Data Networking

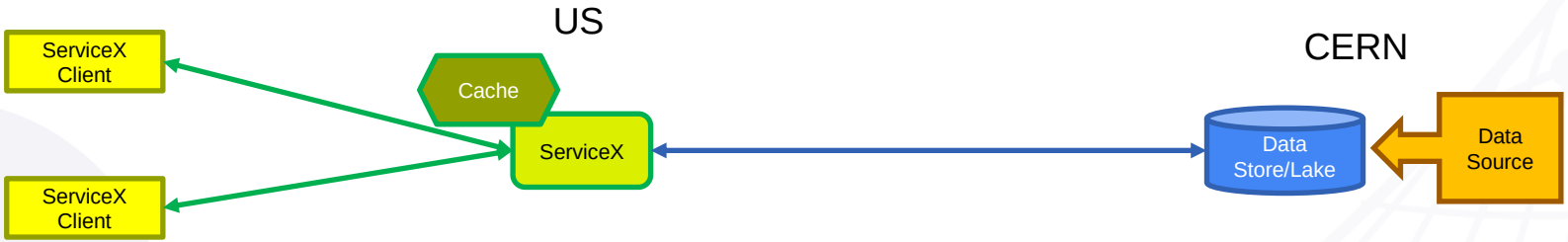


ServiceX: LHC Atlas

ServiceX [1]

- HL-LHC produces exabytes of data
- ServiceX (backend) supports filtering, transformation, processing, and caching of data
- Provides a distributed, caching, columnar data delivery service [2]
- Accessed via ServiceX frontend services



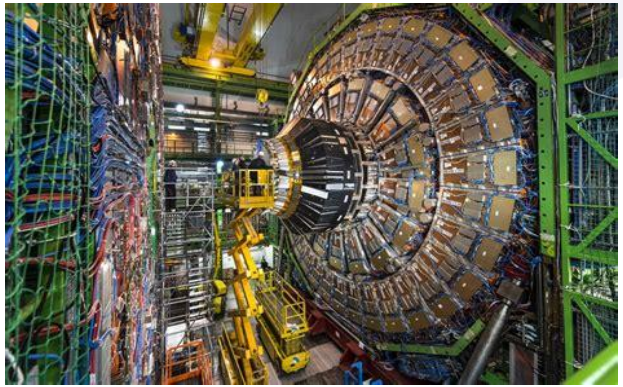


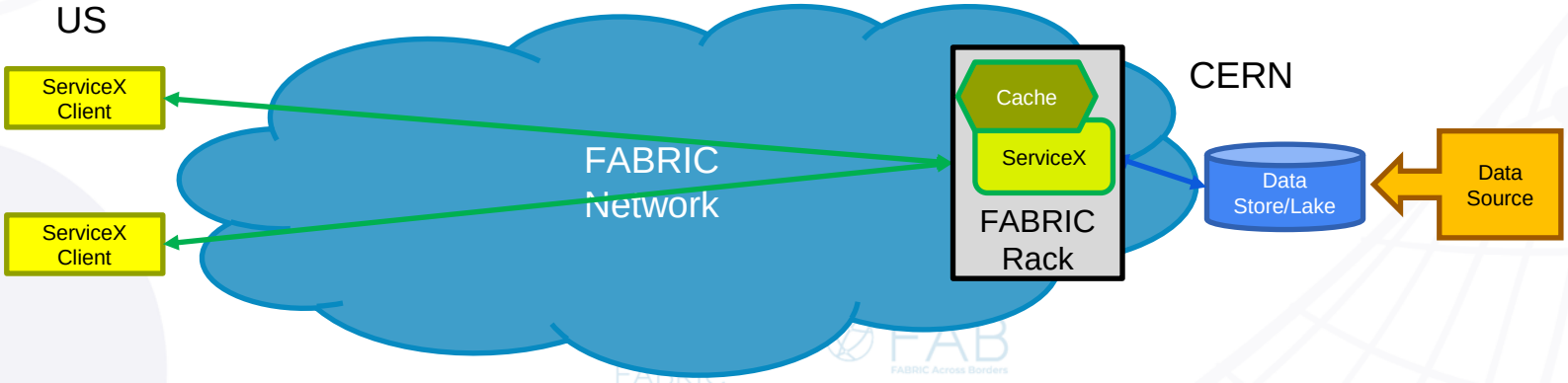
[1] https://servicex.readthedocs.io/en/latest/introduction/
[2] https://indico.cern.ch/event/890991/contributions/3778330/attachments/2007255/3352709/weinbergServiceX200323.pdf

ServiceX: LHC Atlas

FABRIC ServiceX Deployment

- Deploy ServiceX in FABRIC to achieve filtering/transformation near the data source
- Efficiently move data from CERN to UChicago

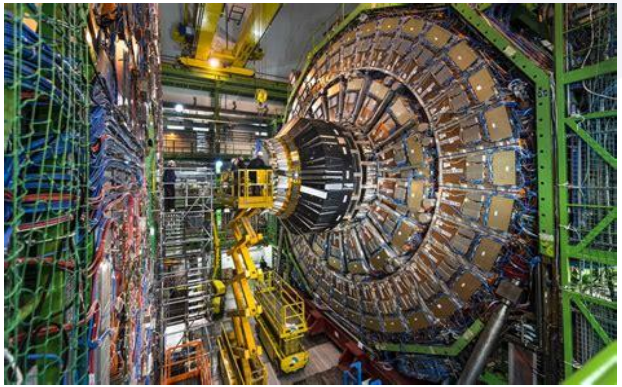




ServiceX: LHC Atlas

ServiceX over NDN on FABRIC

- Incorporates in-network caching
- Extend the FABRIC ServiceX deployment by replacing TCP/IP flows with NDN data transfers that cache results at every hop in the network



Processing Continuum