FABRIC and FAB

Project Overviews and Status

Rob Gardner
Enrico Fermi Institute
University of Chicago

Shawn McKee
Department of Physics and UM ARC-TS
University of Michigan

LHCOPN/LHCONE Virtual Meeting, #47
October 12, 2021

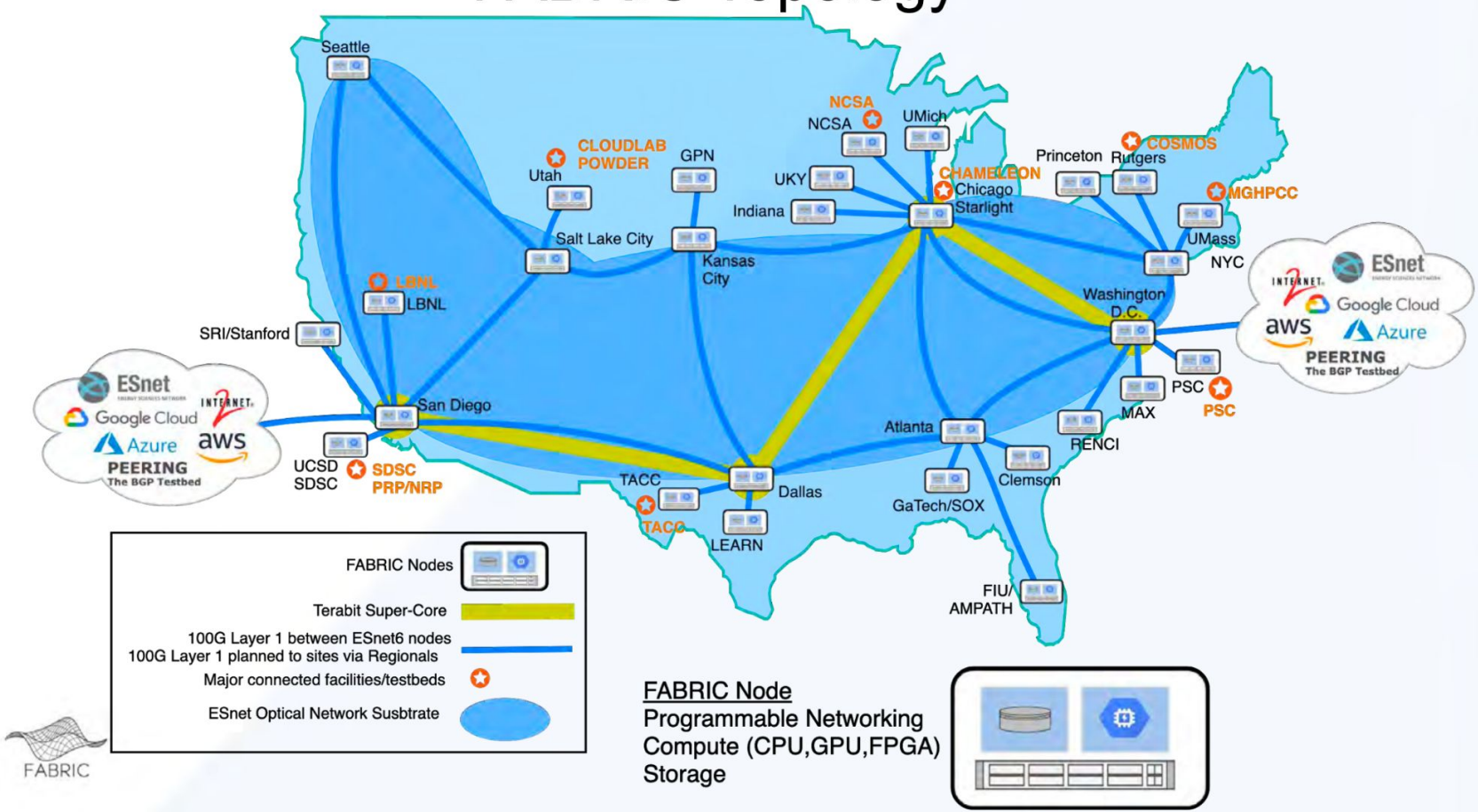
What is FABRIC and FAB?

- **FABRIC** is an NSF R1–midscale project to build a US national scale programmable network with compute and storage at each node.
  - Run computationally intensive programs & maintain information in the network
  - Nodes have 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
- **FAB** (**FABRIC Across Borders**) is a follow–on to FABRIC which is creating an international extension of this testbed to allow at–scale testing for global science.

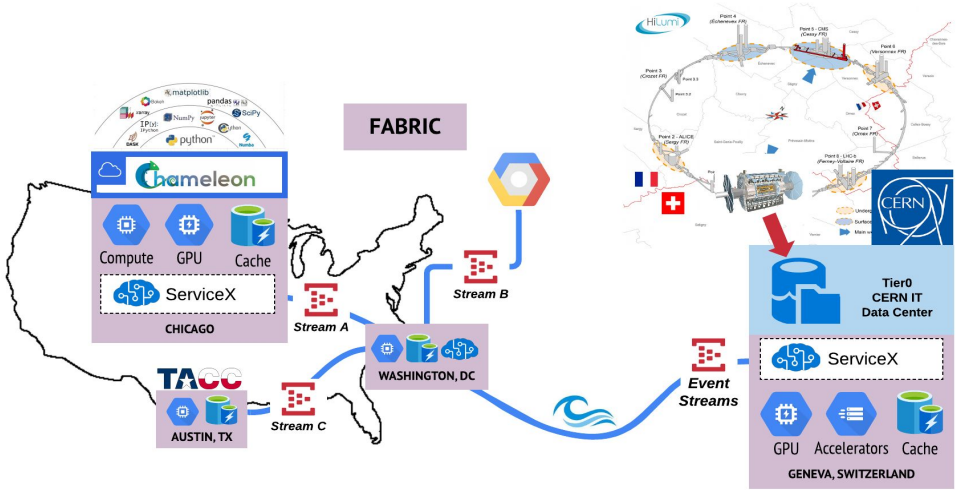
FABRIC Overview https://fabric-testbed.net/

- 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 Topology



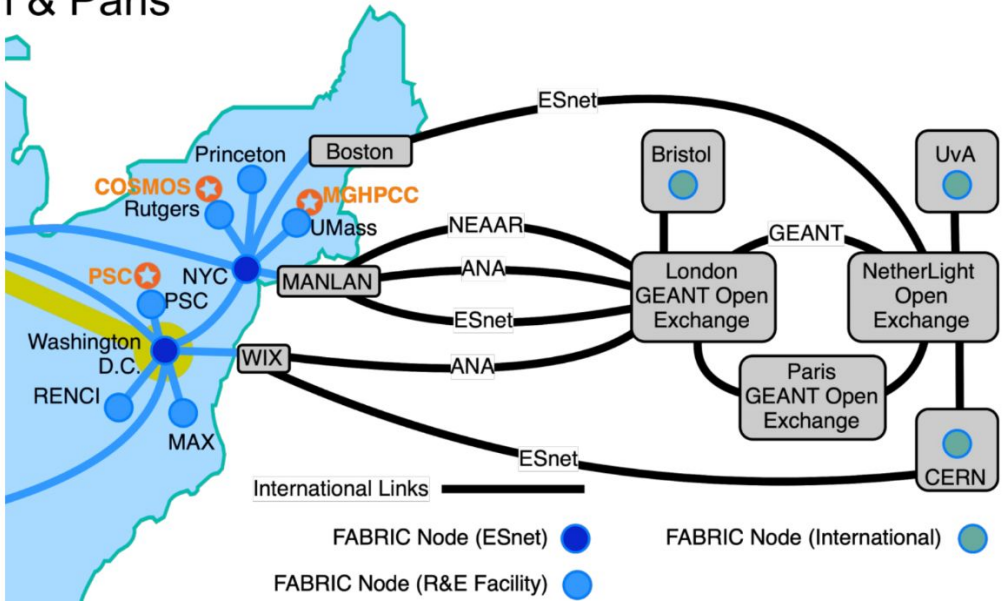
FAB Proposal Details



- Deployment of a **FABRIC** node at **CERN** to enable networking R&D
- Explore new capabilities in the network
- Some early use case ideas
- Optimizing transatlantic data transfers (packet marking, traffic shaping, orchestration)
  - Caching in the WAN
  - Accelerated data delivery
- Others are welcome!

FAB Network & Facility Partners: EU

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



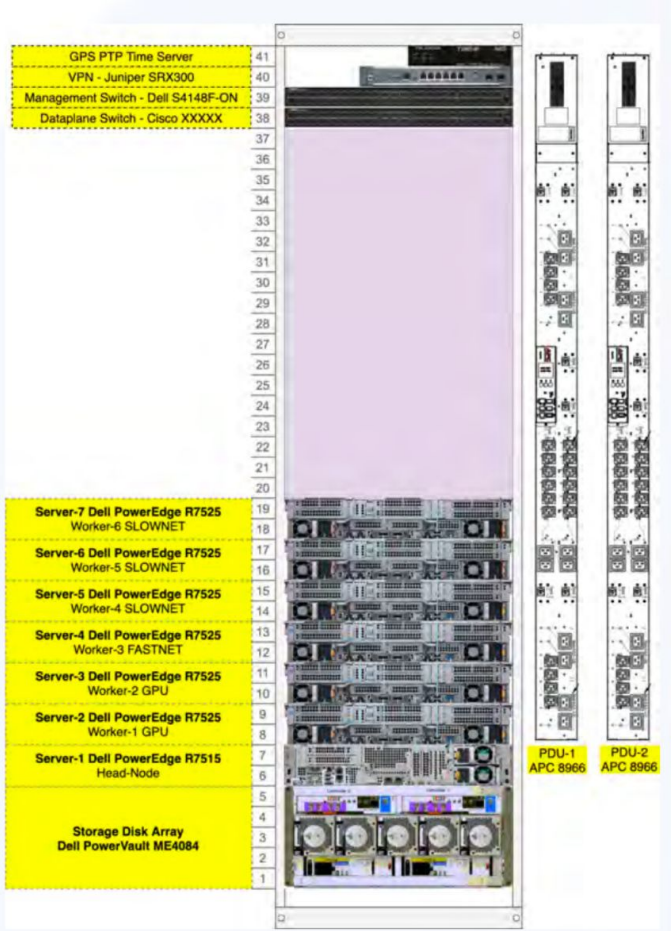
FAB
FABRIC Across Borders

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 Rack Configuration

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



Deployment Status

- **FABRIC** in initial deployment phase
- Relevant node deployments ongoing – UMich, TACC, Starlight (Chicago)
  - and others – MGHPCC, NCSA
- Tried for CERN node by September
  - Initial approval for 3 racks in networking room
  - Design of node deployment relevant to our use cases
  - Delayed (hardware availability) till January 2022.

FAB Installation Plans at CERN

Discussions with Edoardo Martelli/CERN identified 2 racks (@5-6kW/rack) for FAB

- CERN prefers to install equipment to follow their best practices
- Equipment deliver/installation now scheduled for January 2022.

If FAB is to interconnect to the Tier-0, we need to submit plans to security for validation. Two options

1. **FAB connects as a remote site to CERN LHCOPN/ONE border routers. FAB will be local and will get remote connectivity via direct connections to the ESnet routers at CERN (preferred)**
2. FAB connects to the data centre network with a CERN IP addresses and uses the normal LHCOPN/ONE border routers as any other server at CERN. Need special ESnet(?) link for WAN

Consideration: *How to use FAB to validate if net services help production?* We need to be able map production flows between sites over FAB/FABRIC and would like to plan for that from the start.

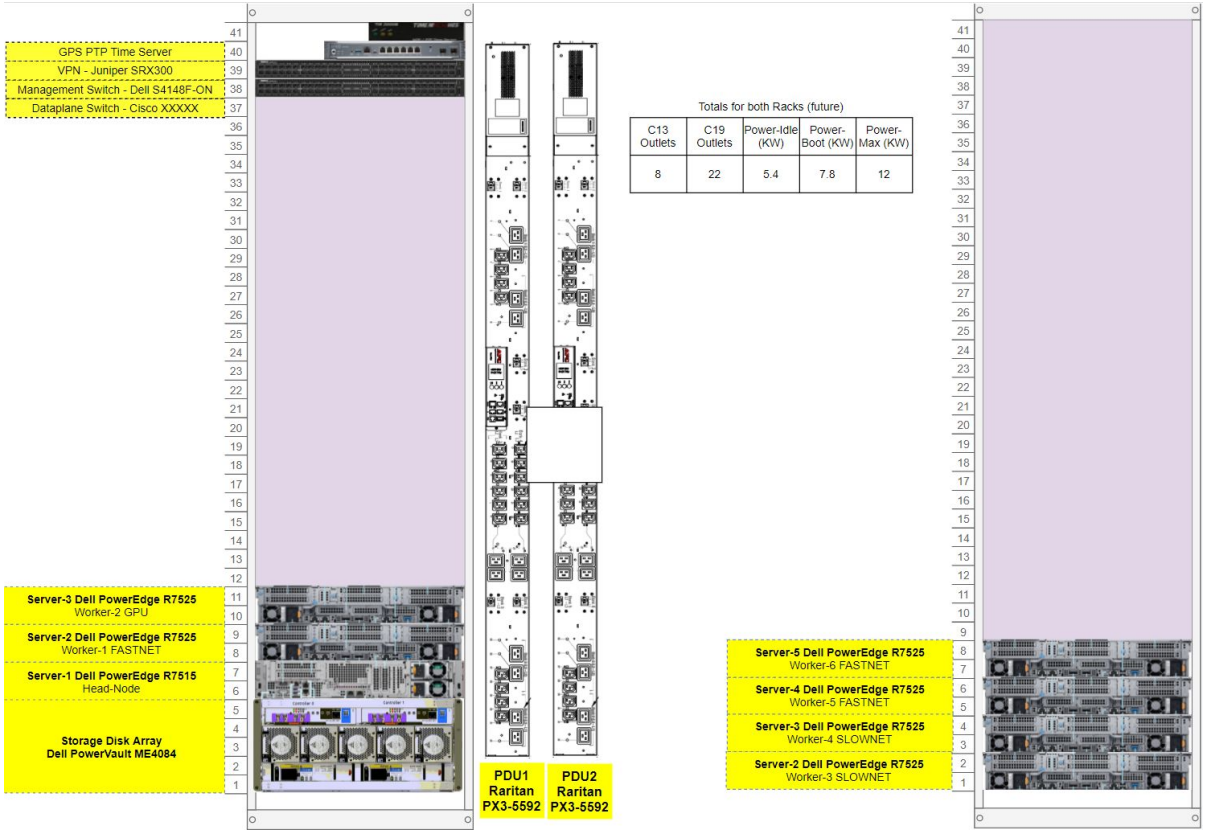
CERN FAB Install Details

Rack 1

GPS PTP Time Server
VPN - Juniper SRX300
Management Switch - Dell 4148S-ON
Dataplane Switch/Router - Cisco NCS 5700
SLOWNET - PowerEdge R7525
SLOWNET - PowerEdge R7525
SLOWNET - PowerEdge R7525
FASTNET - PowerEdge R7525
FASTNET - PowerEdge R7525
HeadNode - PowerEdge R7515

Rack 2

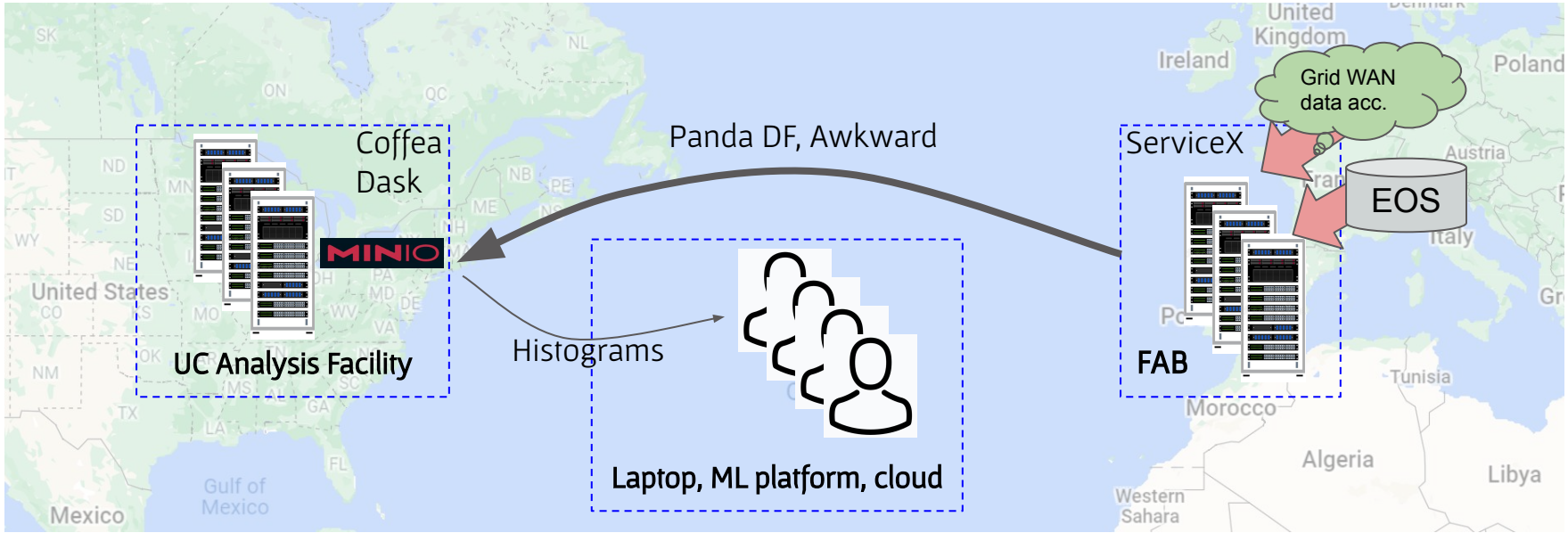
SLOWNET - PowerEdge R7525
SLOWNET - PowerEdge R7525
SLOWNET - PowerEdge R7525
FASTNET - PowerEdge R7525
FASTNET - PowerEdge R7525
GPU - PowerEdge R7525



FAB Use Case: Accelerated Data Delivery Demonstrator

- Read ATLAS data from the Tier0
- Cache locally (CERN FABRIC node)
- Transform locally to columnar format with **ServiceX**
- Write output to MinIO database to analysis facilities in the US
  - FABRIC-peered: Chameleon, UMich / AGLT2
  - Others: IRIS-HEP SSL cluster / new US ATLAS Analysis Facility (UChicago)
- Analyze in Jupyter Notebooks using Coffea & Dask
  - TRExFitter uses ServiceX for ATLAS analysis

ServiceX @ FAB

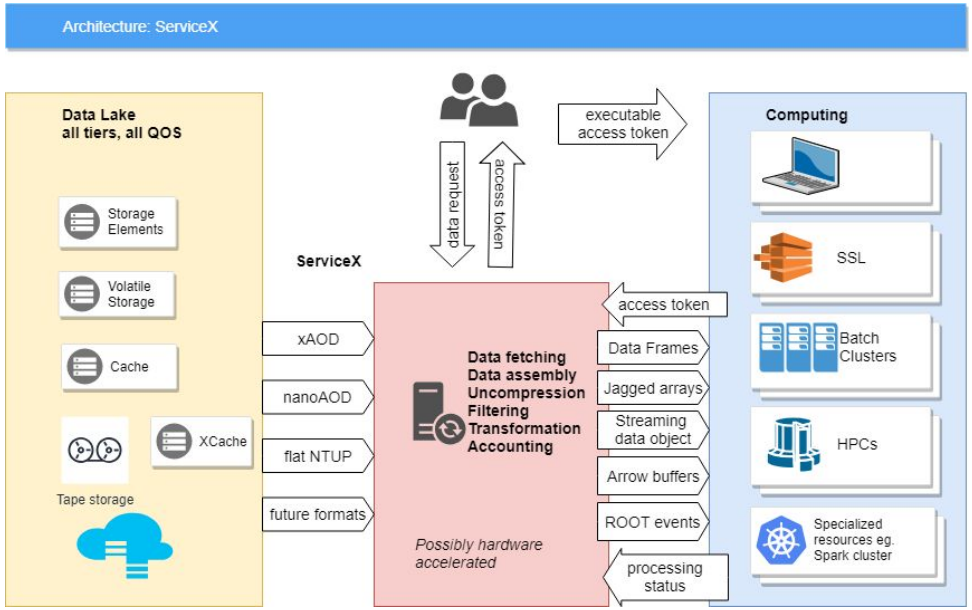


ServiceX – big picture

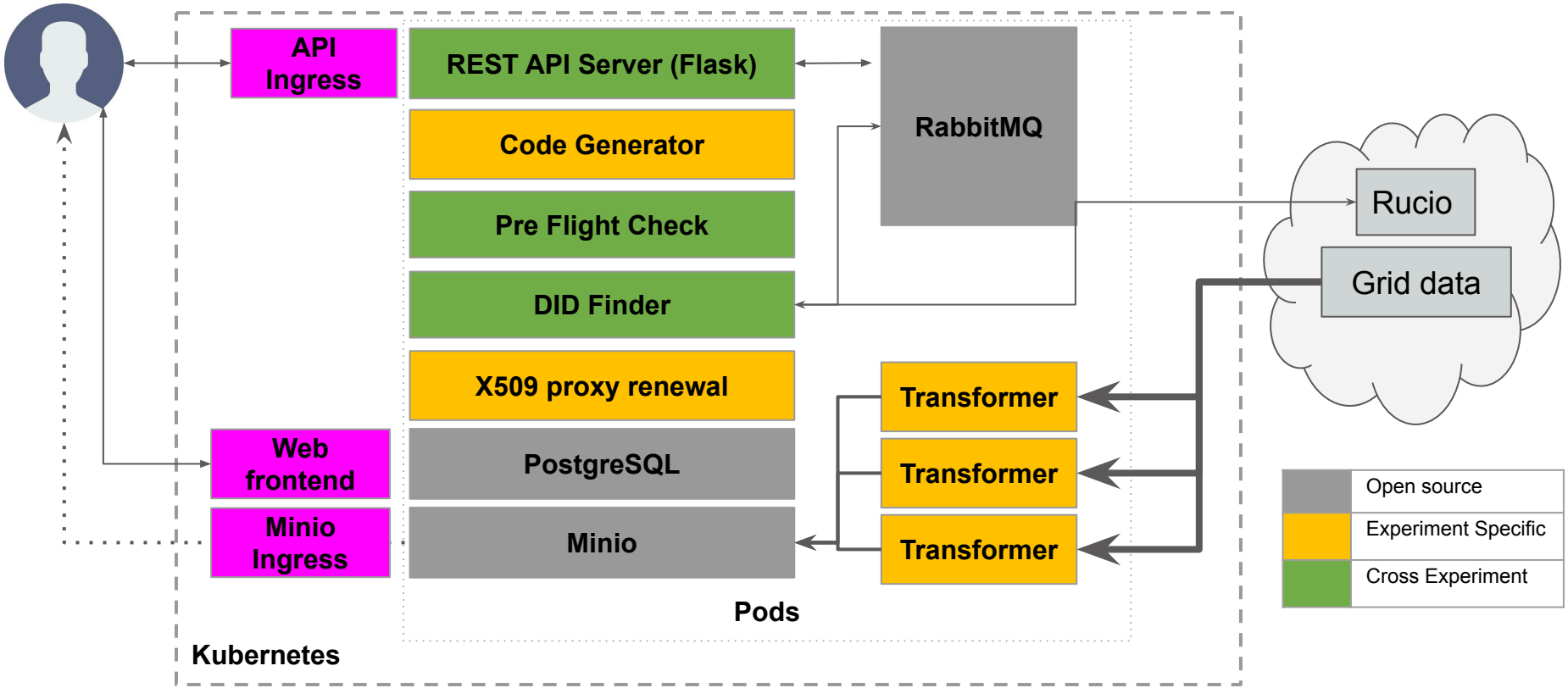
Tailored for **nearly-interactive,** high-performance **array-based analyses**

Performs On-the-fly data access, filtering, derivation, delivery into variety of formats

Project Page



ServiceX Internals



ServiceX Requirements

- Kubernetes
- As much **CPU** as the racks can support
- Doesn't require **GPU**

XCache Requirements

- NVMe disk
  - 20–30 TB would be sufficient
- As much **disk** as we can afford in the FABRIC node
- Doesn't require **GPU**
- As good connection to **EOS** and **to WAN** as we can get.

Network Prototyping Use-Cases

- Packet marking
  - P4 might be useful for both marking packets and accounting on packets
  - Can help prototype this functionality for WCLG
- Can we use **SENSE**?
  - Are network orchestration components in FAB/FABRIC compatible with SENSE?
  - Will SENSE be a standard FAB/FABRIC service?

Putting it all together

- We are still discussing an optimal mix of three resources: fast disk, CPU (bus on the node), and network to build an impactful accelerated data delivery demonstrator
- Note – our "use case" catalog is flexible – **there are multiple configurations worth exploring with FABRIC and FAB**

Questions Being Resolved...

- How will CERN FAB node connect back to FABRIC?
  - Physical path options and available bandwidth?
- Power is the obvious concern (5-6KW/rack; 2 racks with 2 adjacent empty racks) but should be sufficient for planned deploy.
- Can CERN provide a GPS antenna? (PTP)
- What additional prototyping and use-cases should we explore?

Summary and Conclusion

We are working to complete and utilize an at-scale network testbed to demonstrate the benefits and capabilities of new services and applications able to utilize high performance global networks.

We welcome suggestions and participants!

Questions or Comments?

Acknowledgements

Many thanks to the FABRIC and FAB project teams for their work, as well as to Ilija Vukotic for ServiceX details.

This work was funded by NSF grants FABRIC (NSF #1935966) and FAB (NSF #2029260)