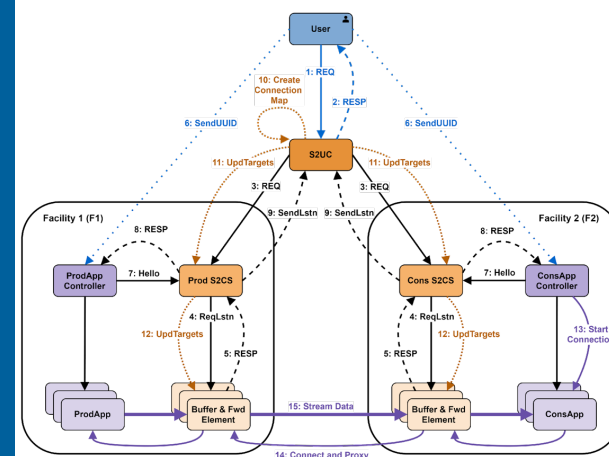


5GRP Workshop 2024



SciStream: Enabling Applications to Stream Data between Science Instruments and HPC

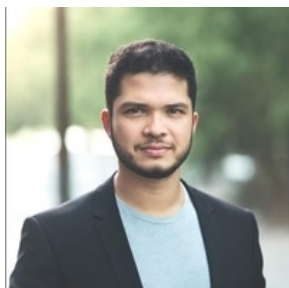


Rajkumar Kettimuthu

Argonne National Laboratory and The University of Chicago, USA

Sep. 17, 2024

Team



Flavio Castro



Joaquin Chung



Tekin Bicer



Rohan Vardekar

Outline

Motivation

Overview of SciStream

Releases and Demonstrations

Application Integration Activities

- Recently accomplished

- Ongoing

Deployment Activities

Upcoming Demos and Tutorials

Motivation

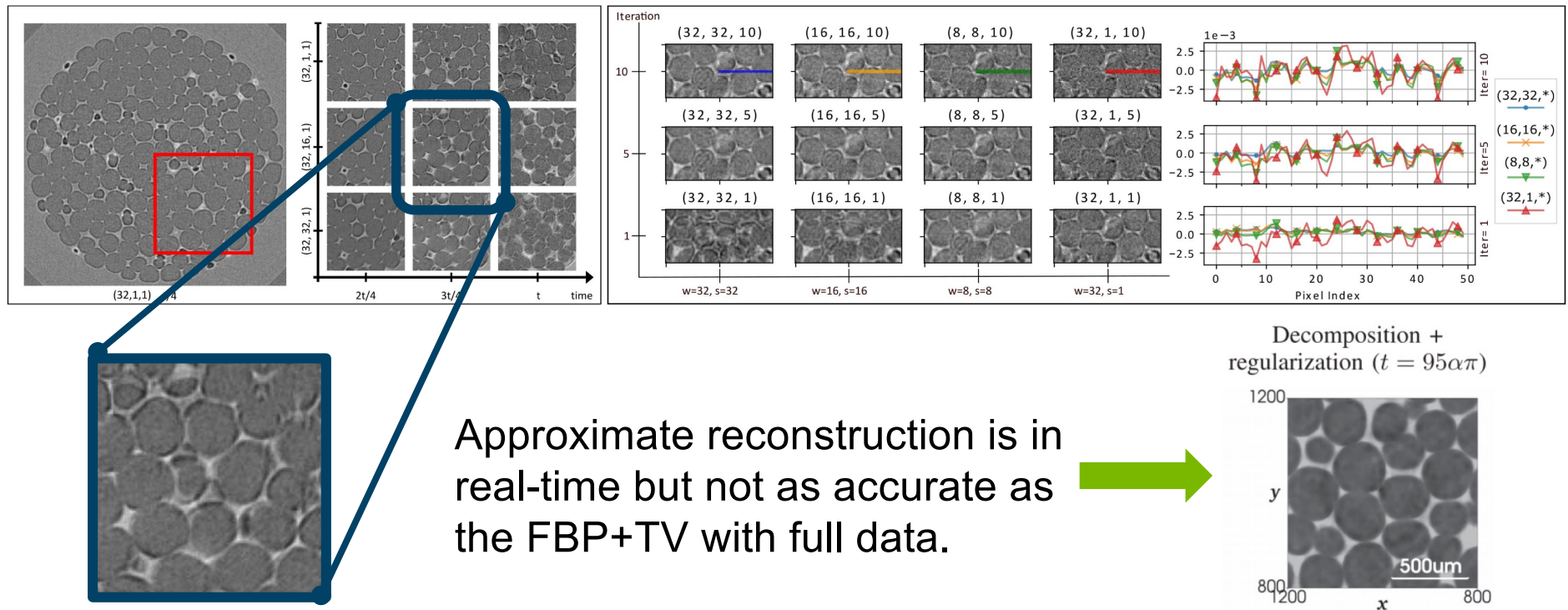
Stream Processing as a Science Enabler



- Foam dataset with increasing temperature
- Initial stage does not require fast data acq.
 - Time (secs): [0, 17]
- The faster data acquisition is desired when features change rapidly
 - Time (secs): (17,-)
- Ideal experimentation with stream (processing)
- Real-time processing with rapid feature detection, adjusting data acquisition
- Other potential science use cases
- Region of Interest, on-the-fly data reduction...

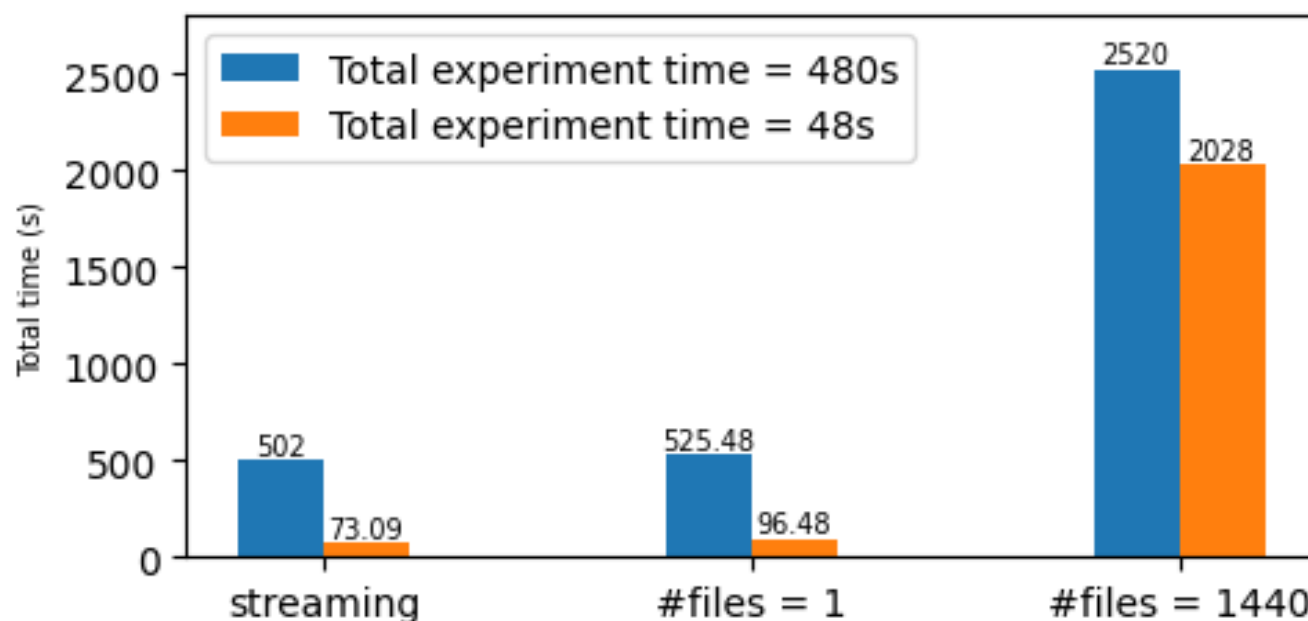
Motivation

Complementary Data Analysis Techniques: Offline and Online



Motivation

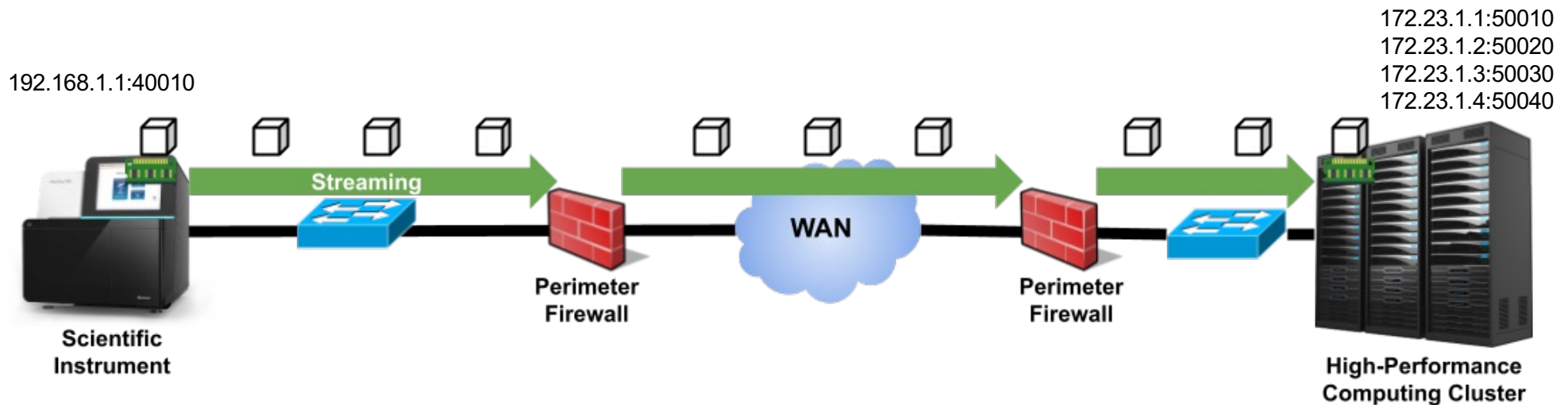
Memory-to-Memory Streaming vs File-based Transfers



Total data generated:
1440 frames with 12 GB

Instrument and HPC are in Different Security Domains

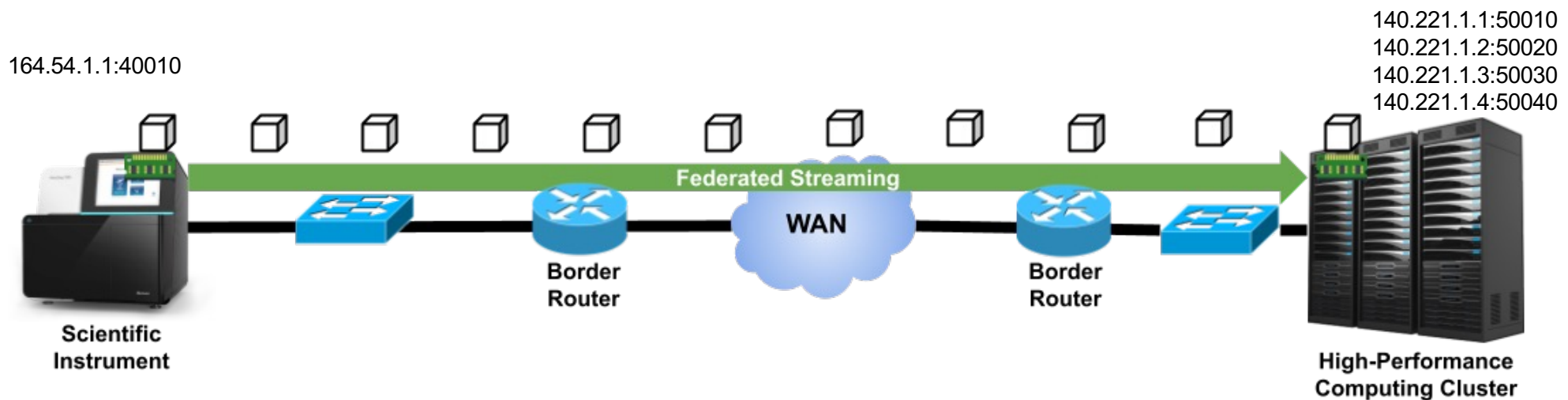
On private networks and not directly visible from outside



- ✗ Scientists need to negotiate with each institution opening TCP ports on their firewalls to enable streaming
- ✗ Scientists also need to implement the mechanisms to comply with the security policies of each institution

Instrument and HPC are in Different Security Domains

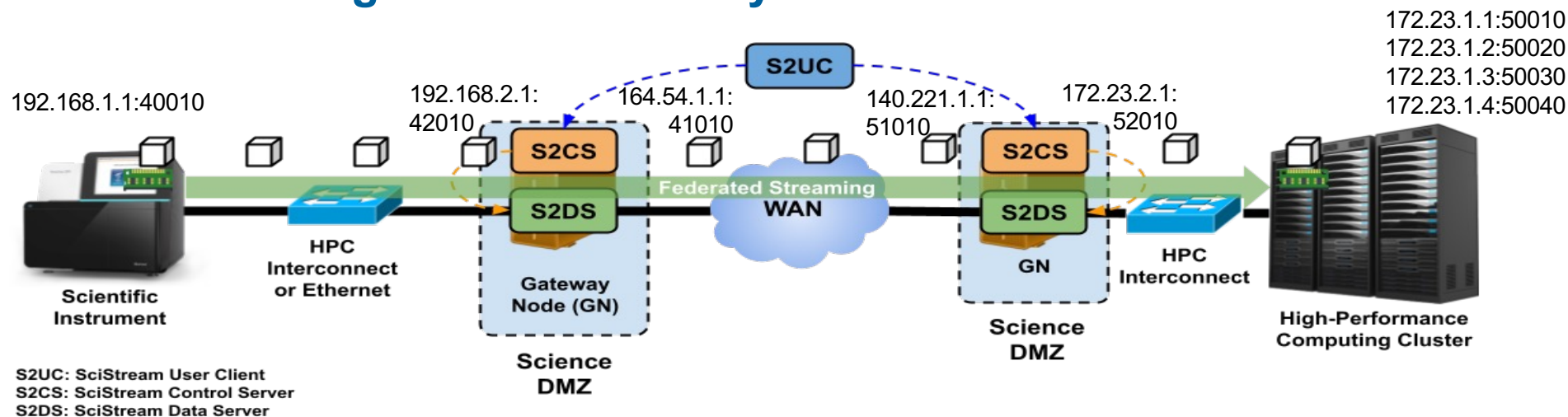
Public IPs a possibility but issues remain



- Scientists still need to worry about the TCP ports needed for streaming data from the data producer to the data consumer
- As well as the authentication of the connections

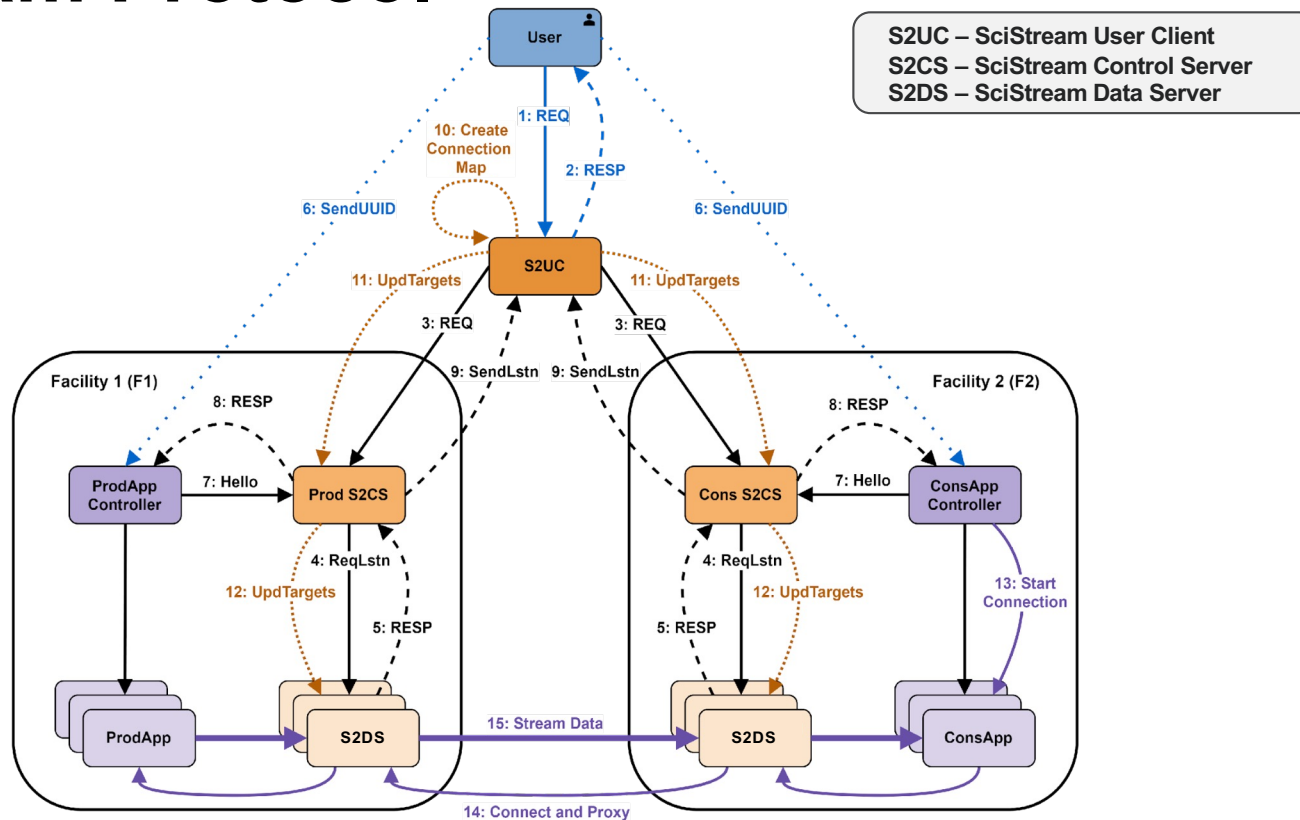
The Future with SciStream

Secure streaming between security domains



- ✓ Automates the process of deploying secure bridges between security domain
- ✓ Keeps streaming app semantics intact
 - ✓ One end listens on one or more [IP:Port]
 - ✓ Other end connects to given [IP:Port]
- ✓ Requires only minimal changes to app to authenticate itself to SciStream
- ✓ Supports a lot of streaming applications
 - ✓ Operate at the transport protocol layer
 - ✓ One-time deployment and vetting for institutions
- ✓ Uses Globus Auth for federated id & securing data & control channel
- ✓ Supports 3rd party control for hosted orchestration and monitoring

SciStream Protocol



J. Chung, W. Zacherek, A. Wisniewski, Z. Liu, T. Bicer, R. Kettimuthu, and I. Foster, "SciStream: Architecture and toolkit for data streaming between federated science instruments," in Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing, HPDC '22, (New York, NY, USA), p. 185–198, Association for Computing Machinery, 2022.

Outline

Motivation

Overview of SciStream

Releases and Demonstrations

Application Integration Activities

- Recently accomplished

- Ongoing

Deployment Activities

Upcoming Demos and Tutorials

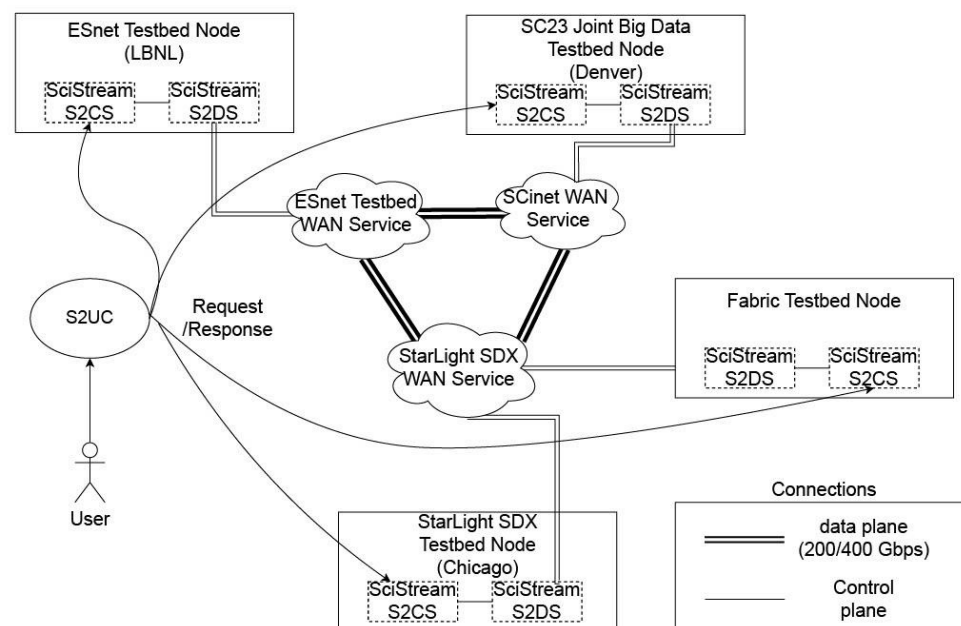
SciStream Demo at SC23 in collaboration with ESnet

SciStream achieved up to 87 Gbps with 32 concurrent iperf streams

- SSH tunnels achieved 2.8 Gbps with similar concurrency

SciStream achieved up to 10 Gbps with 1 single iperf stream

- SSH tunnels achieved mere 300 Mbps



Recent Releases

Releases 1.0 and 1.1

- Integrated HAProxy, Nginx and Stunnel for production-grade forwarding quality
- Integrated with Globus Auth
- Data-channel authentication allowing ports to be open securely
- Implemented encryption for control and data-channel
- TLS protocol and PSK authentication
- Open Source Code <https://github.com/scistream/scistream-proto>
- Documentation <https://scistream.readthedocs.io/en/latest/scistream/>

Outline

Motivation

Overview of SciStream

Releases and Demonstrations

Application Integration Activities

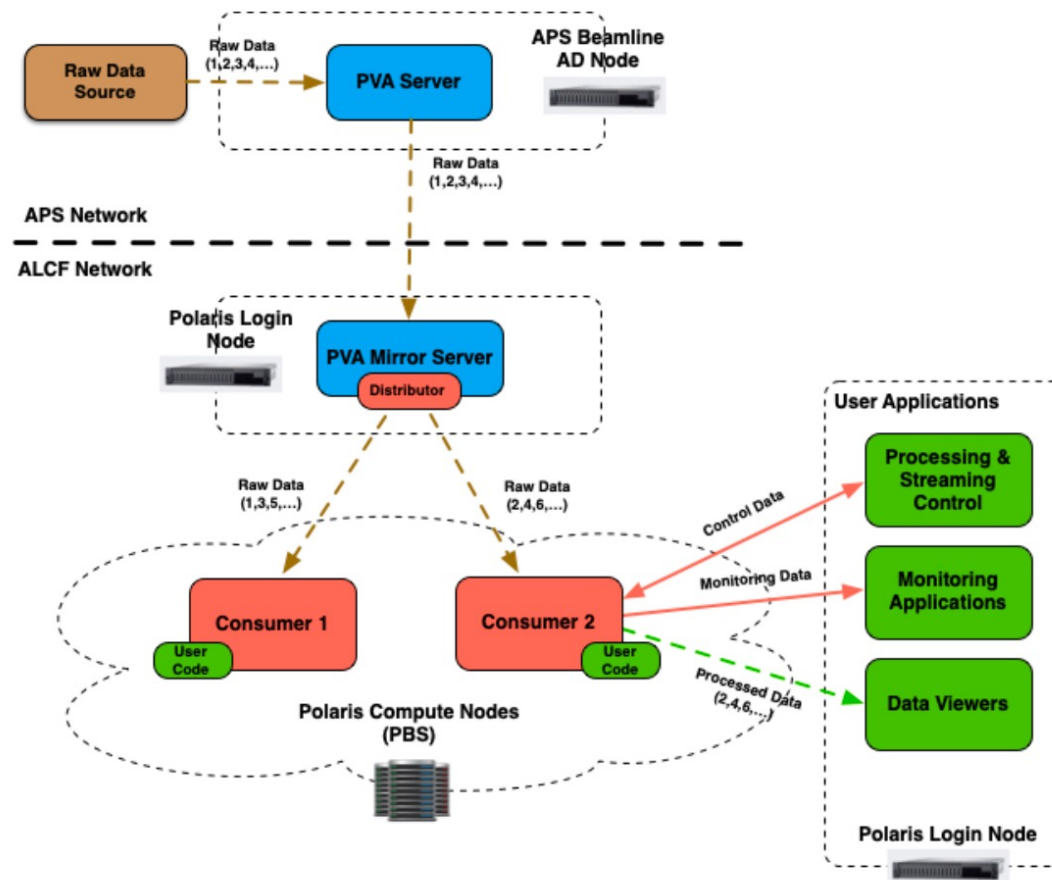
Recently accomplished

Ongoing

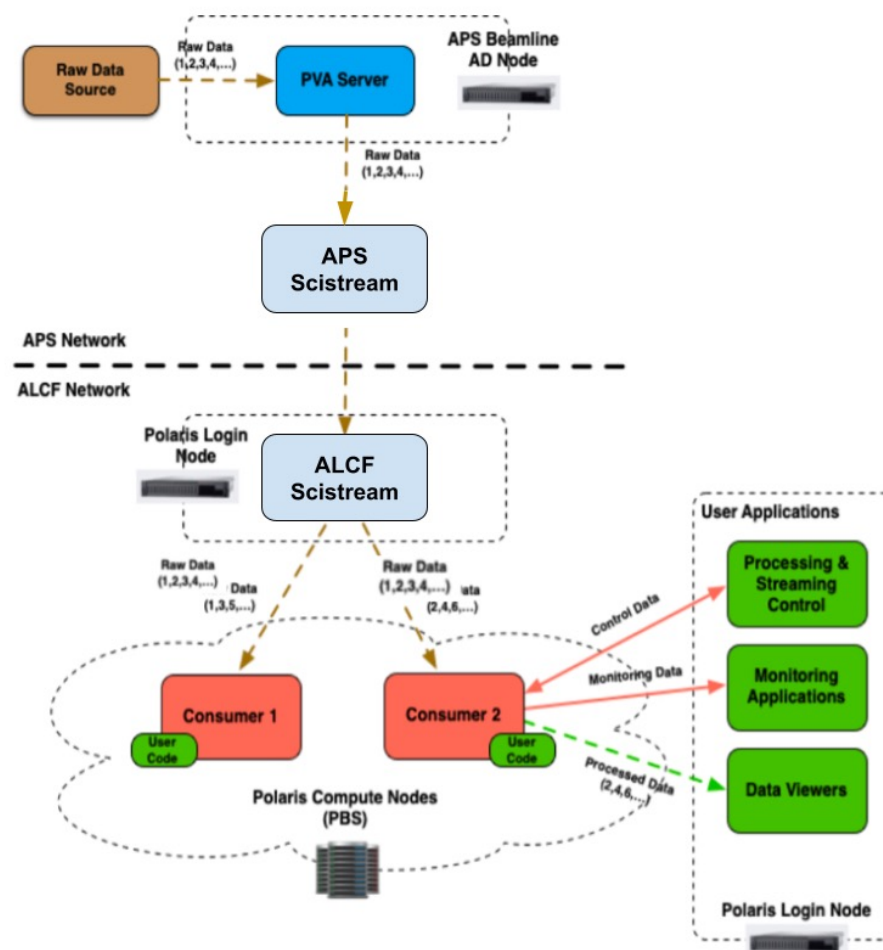
Deployment Activities

Upcoming Demos and Tutorials

PvaPy Data Streaming from APS to ALCF (Before SciStream)

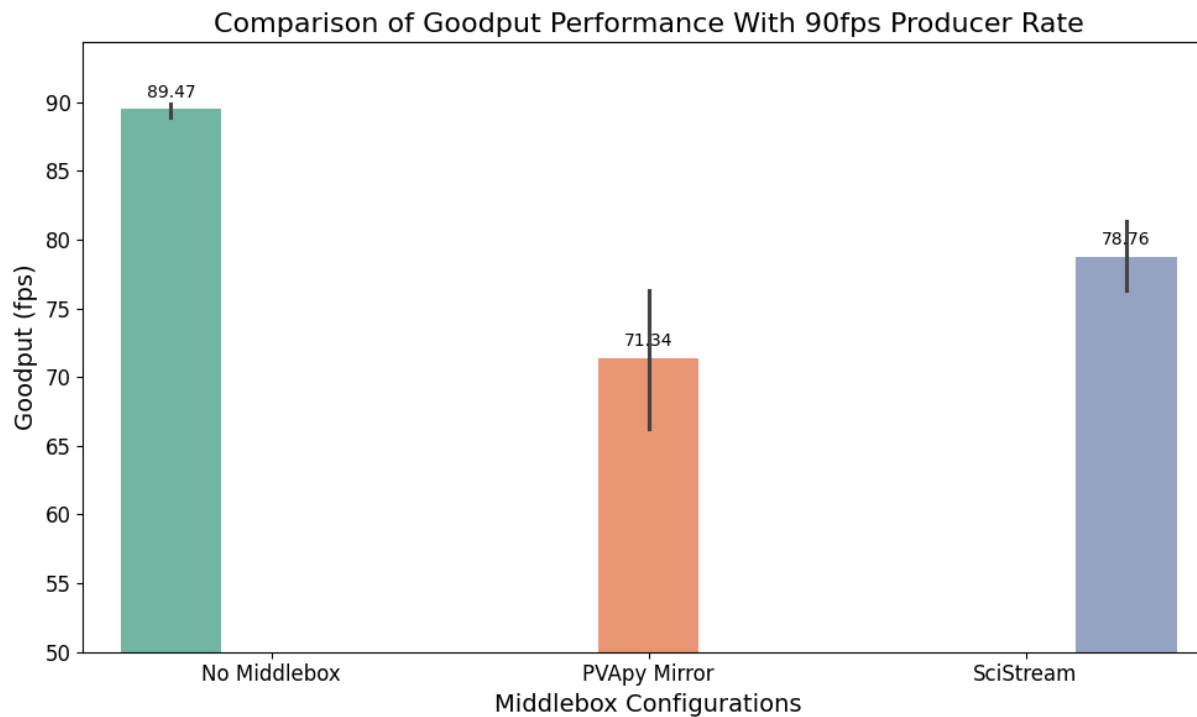


PvaPy Data Streaming from APS to ALCF with SciStream



PvaPy Data Streaming from APS to ALCF with SciStream

Preliminary results



Deleria workflow for GRETA

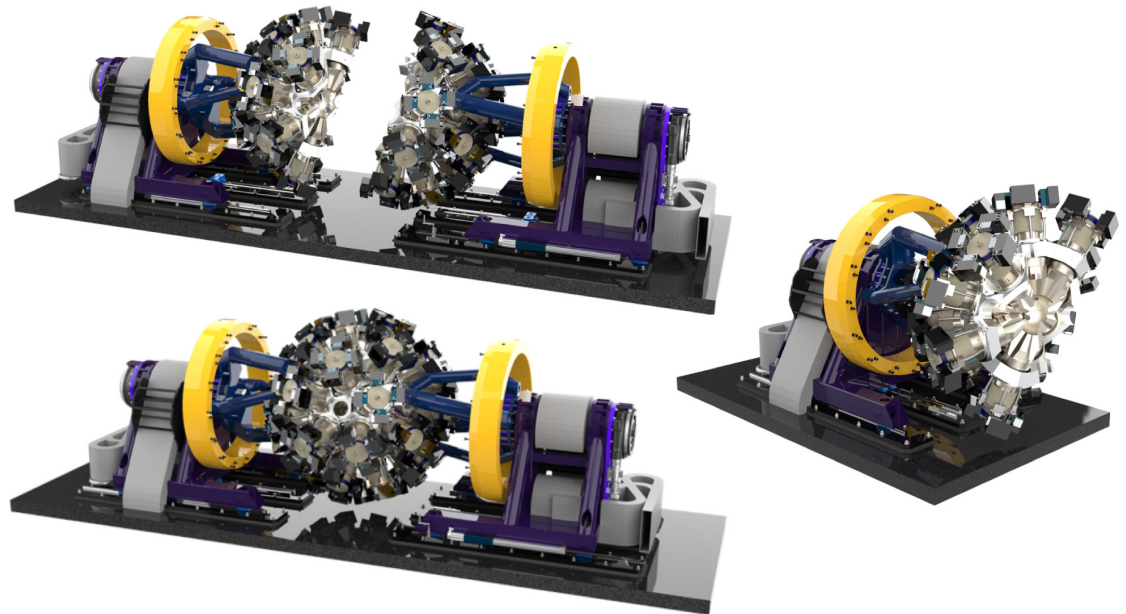
Scientific instrument at
Michigan State

Real-time study of the energy
and 3D position of gamma
rays within atomic nuclei

Deleria continuous streaming
of live measurement data to
compute nodes

Uses nanomsg for network
communication

SciStream Deleria integration is underway for a potential demo at SC24



DIII-D

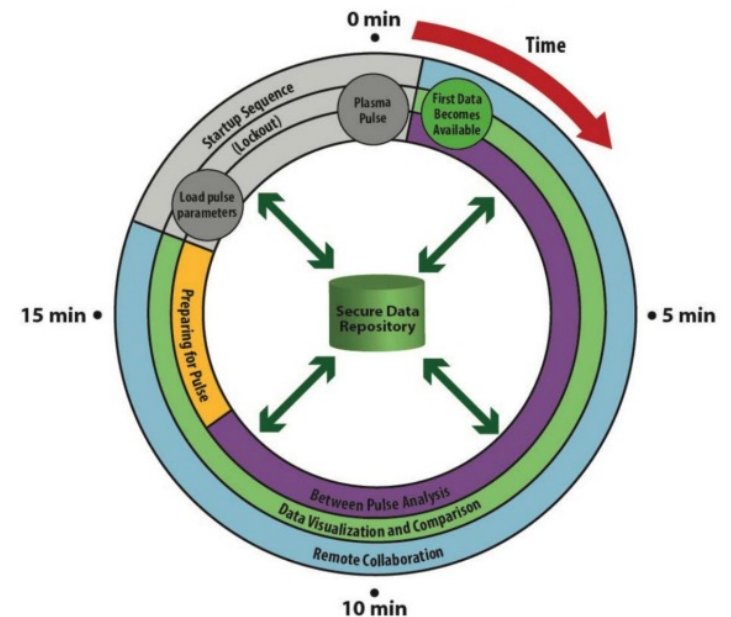
National Fusion Facility, operated by GA for DOE since late 1980s

Substantial influence on ITER (The Way), largest fusion experiment

Make decisions about how to tune their experiment parameters between plasma shots

Computing model for ITER, which will produce 1PB/day

SciStream to enable data streaming between GA and ALCF



Outline

Motivation

Overview of SciStream

Releases and Demonstrations

Application Integration Activities

- Recently accomplished

- Ongoing

Deployment Activities

Upcoming Demos and Tutorials

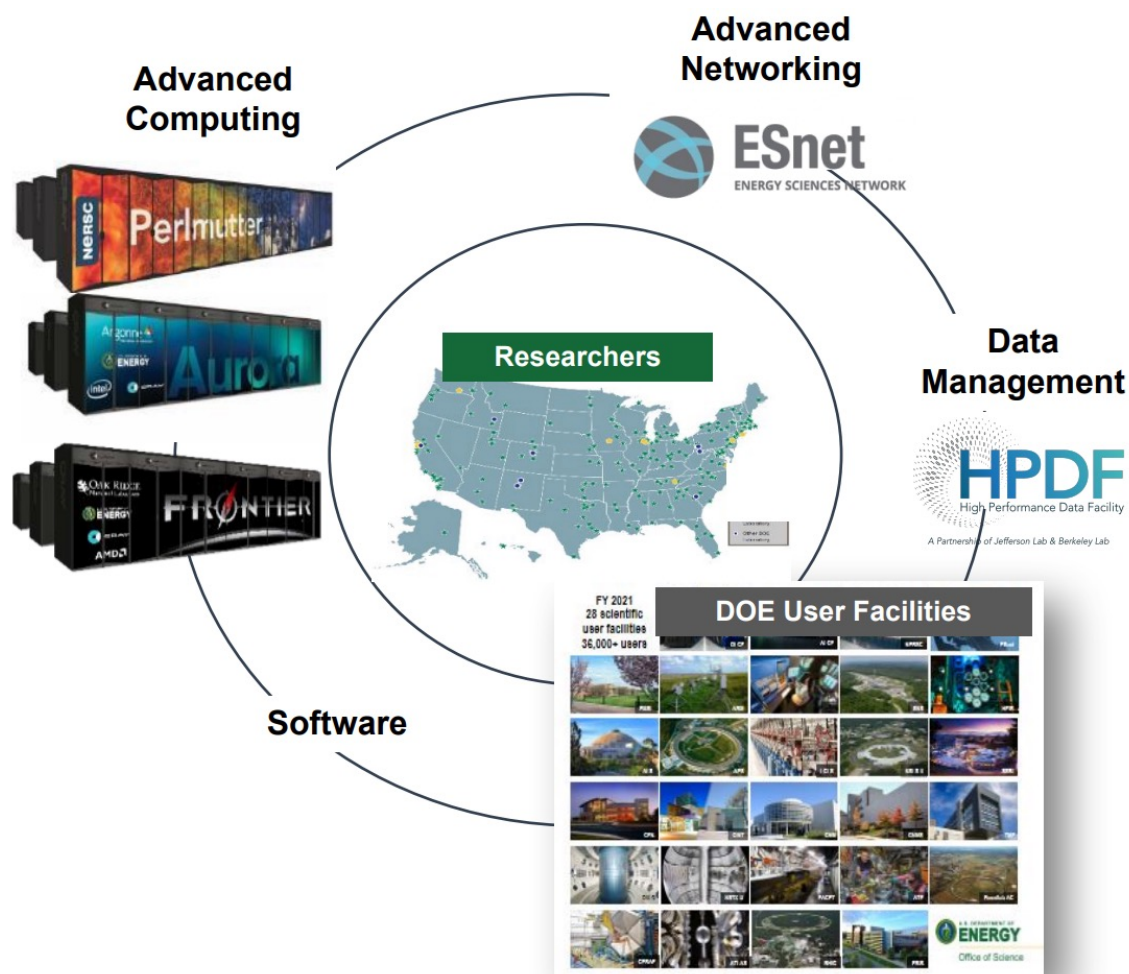
Integrated Research Infrastructure

Seamless and dynamic integration

- Experiment, theory, observation, modeling, simulation, viz, machine learning, AI & analysis

Multiple research facilities

Ongoing prototyping activities



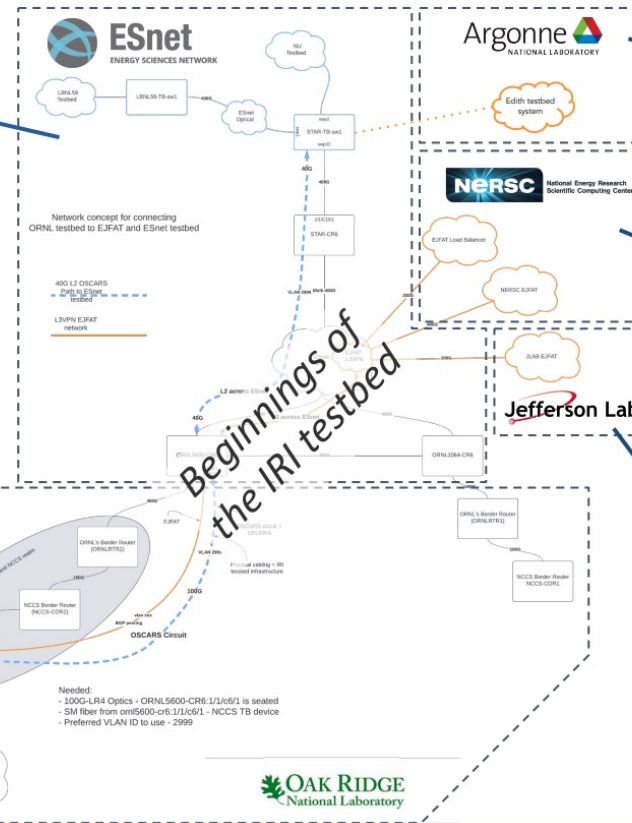
IRI Testbed

SciStream

ESnet Testbed-NextGen upgrades driven by IRI testbed requirements

OLCF ACE testbed effort to explore gaps and develop novel capabilities for OLCF-6 and OLCF-7

SciStream



SciStream

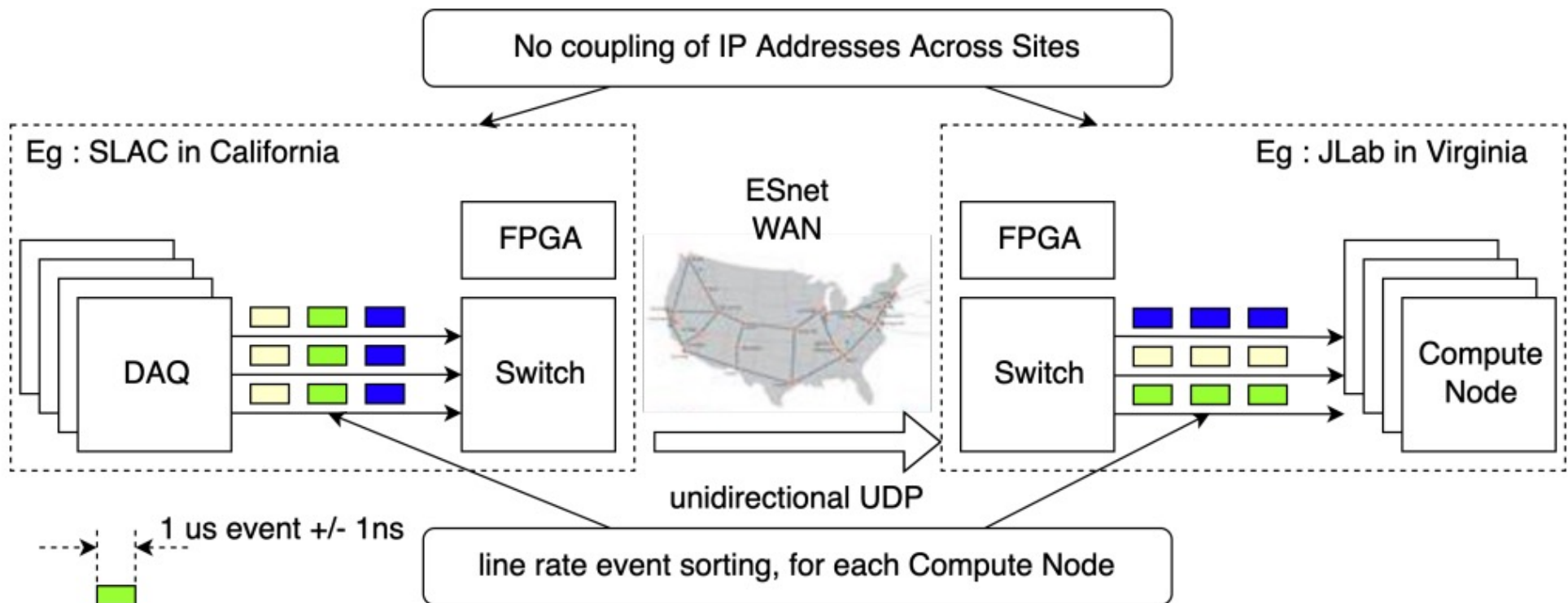
Edith IRI testbed resource is available for exploratory IRI use cases

The "POD" testbed is used for IRI experiments

HPDF will have a phased approach, starting with a testbed, followed by a small scale production facility that will be extended into the full scale HPDF

EJFAT: ESnet JLab FPGA Accelerated Transport Load Balancer

SciStream EJFAT integration is planned



Upcoming Demos and Tutorials

Demo: SciStream

Time: Wednesday, 25 September 2024, 5-7pm CDT

Location: Kansas City, MO, USA

Tutorial: SciStream: Enabling Data Streaming
Between Science Instruments and HPC Nodes

Time: Monday, 18 November 2024, 8:30-12 EST

Location: Atlanta, GA, USA

Tutorial: SciStream: Enabling Data Streaming
between Science Instruments and HPC Nodes

Time: Monday, 9 December 2024, 8-11:30am EST

Location: Boston, MA, USA

MERIF24



Thanks!

Acknowledgments

