Michael Goodrich*, Vardan Gyurjyan*, Graham Heyes*,
Derek Howard+, Yatish Kumar+,
David Lawrence*, Stacey Sheldon+ ,Carl Timmer*

*JLAB
+ESnet

EPSCI

Jefferson Lab

ESnet

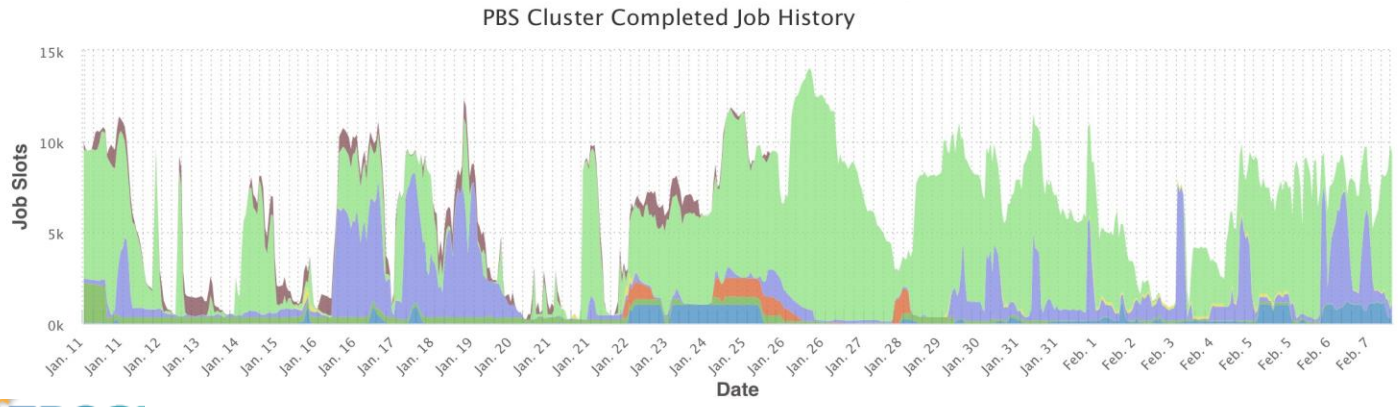
Where Are We Now?

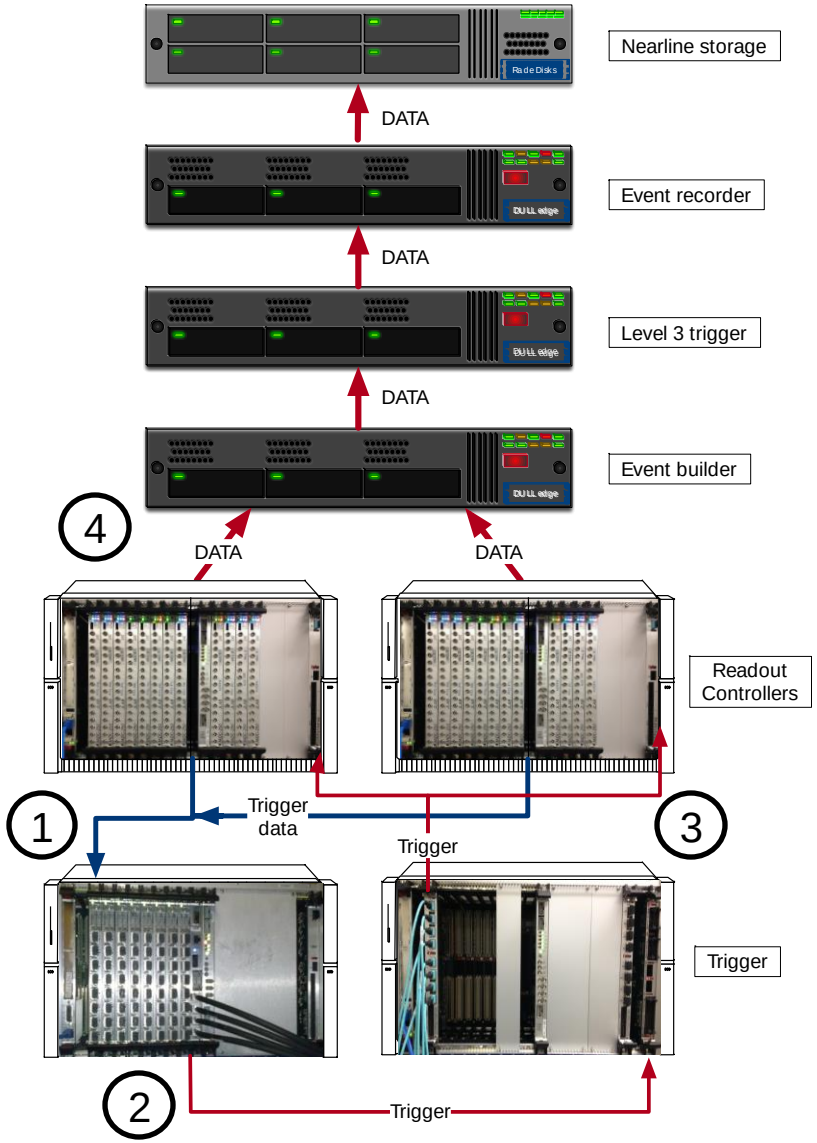
Online:

- Counting House: Custom Electronics, Multi-Level Triggers, Pipelined Readout Systems Build Events Online and Store for Offline Analysis

Offline:

- Events Processed In Steps: Monitoring, Calibration, Decoding, Reconstruction, Analysis.
  - Data Passed Between Stages In Flat Files.
  - Pauses Of Days/Weeks/Months Between Steps.
  - Very Little Integration Between The Various Steps.
  - Analize with Homogeneous Batch Farms.



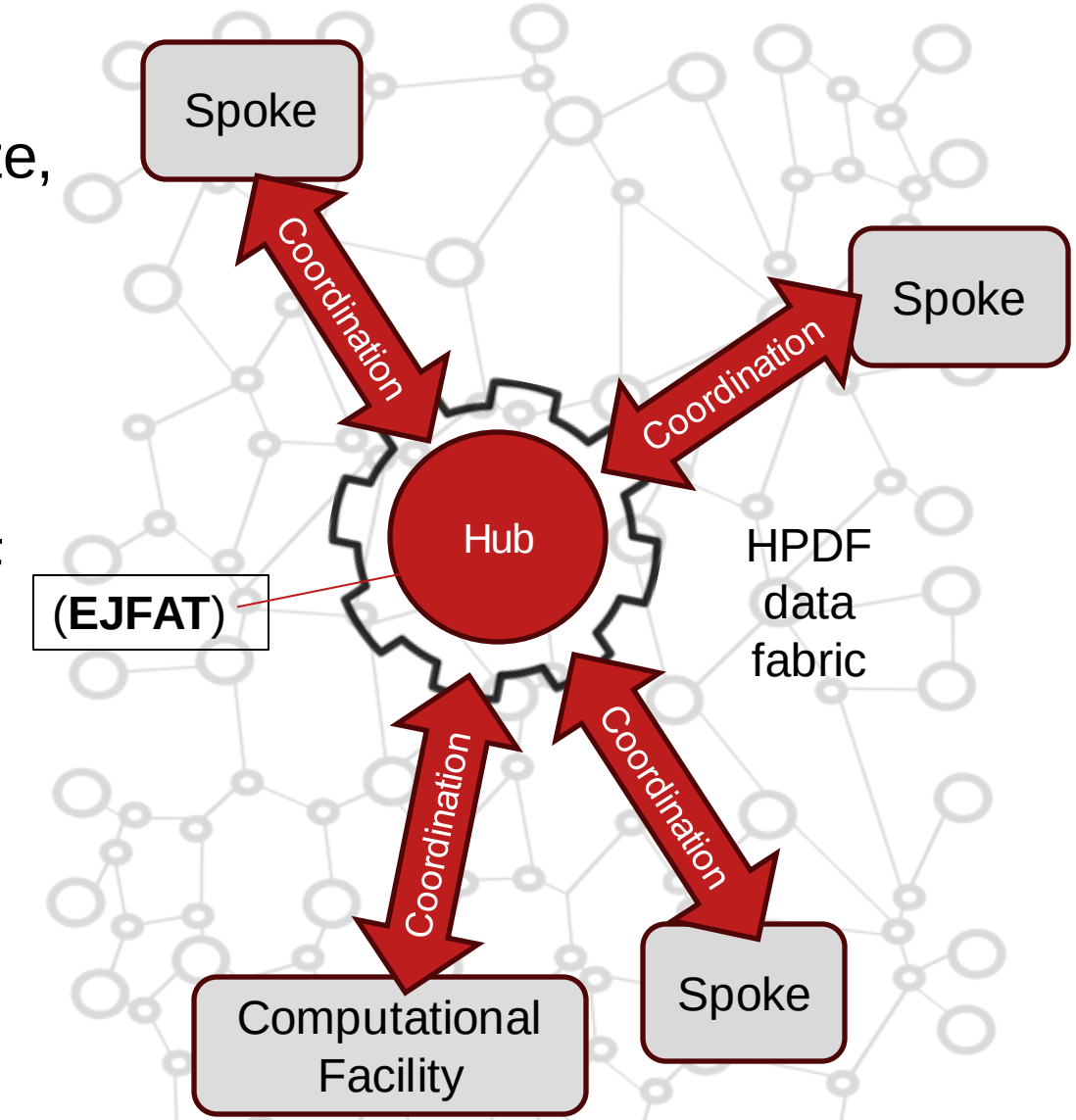


EPSCI

ESnet Jefferson Lab

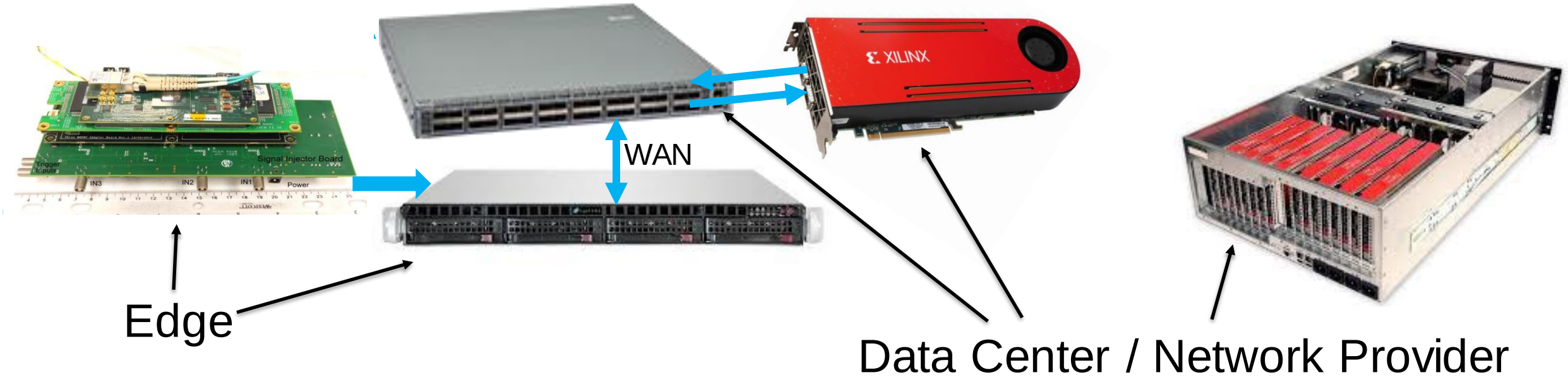
Where Are We Going? - Global HPDF Concept

- Distributed Facility - Hub And Spoke
- Hub Encompasses HW, SW, Staff To Organize, Orchestrate, And Connect Resources
- Technical Approach:
  - Streaming Transport
  - Distributed And Local Data Storage
  - Federated Data Cataloging Across HPDF
  - Orchestration Services Global To HPDF

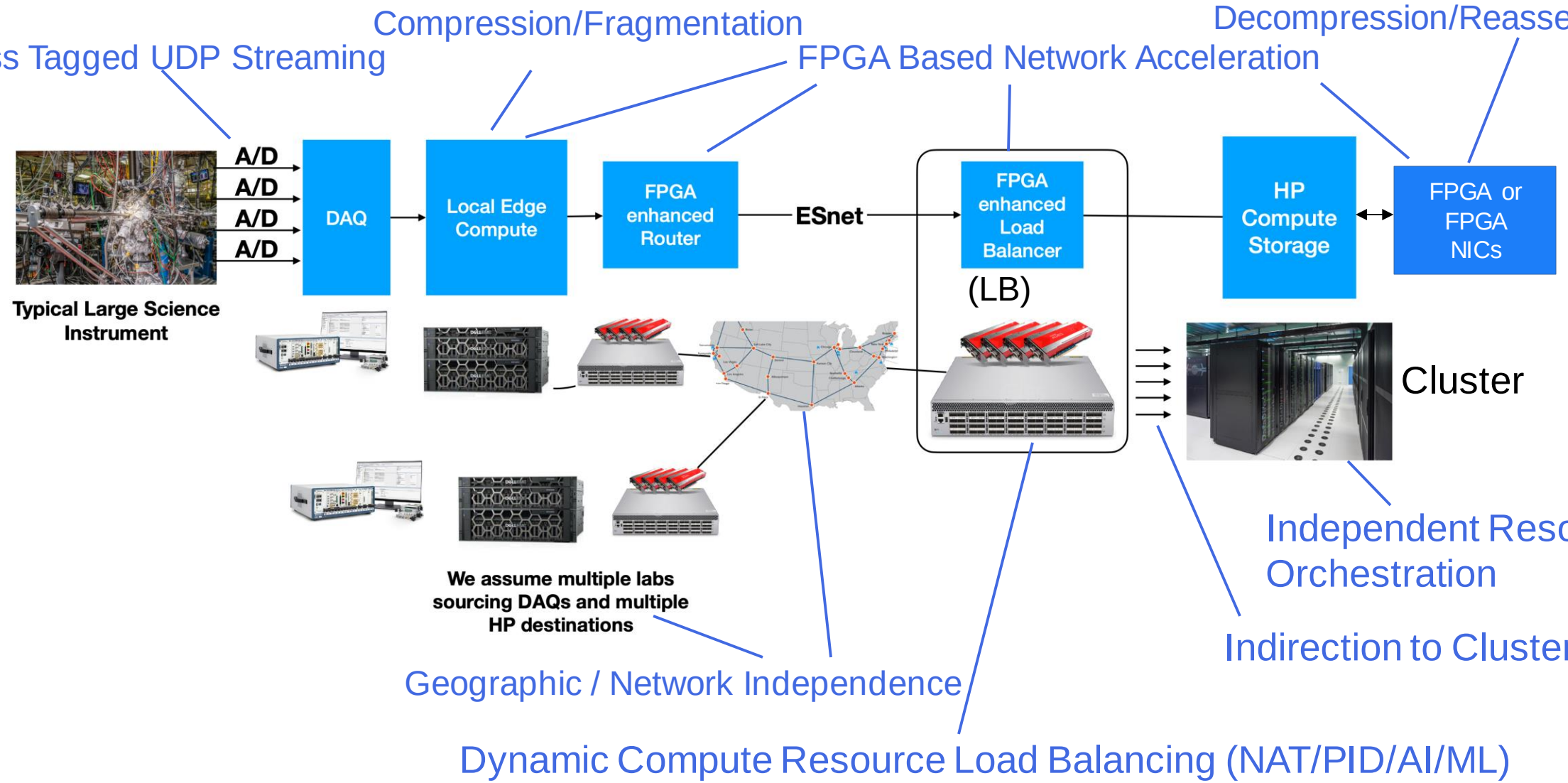


DAQ Goal: Stream Data Through Commercial Hardware

- Readout: Replace Complex Triggering with COTS Streaming.
- Route Edge Data Over WAN to Distributed Compute Centers.
- Replace Custom Edge HW / FW With Generic FPGAs In PCIe.



EJFAT: Accelerated Edge to Core Data Steering

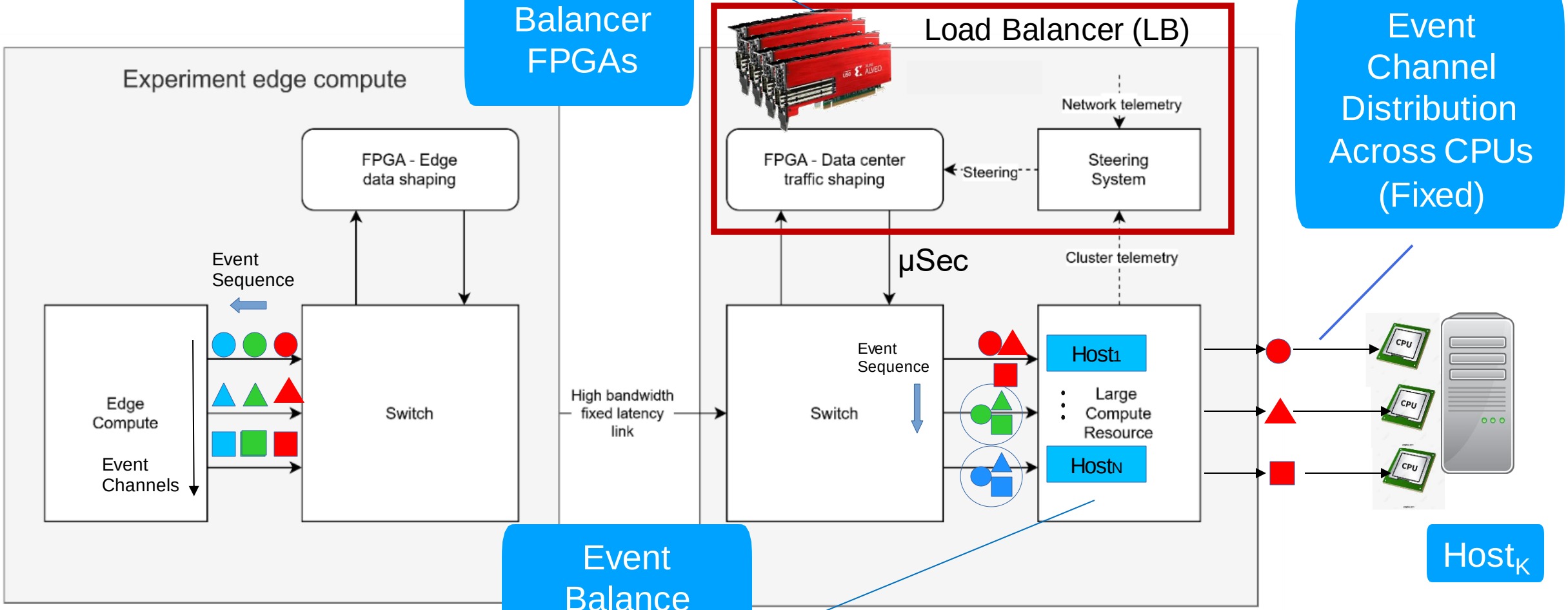


Horizontal Scaling:

Colors → Events
Shapes → Channels (ROCs)

Data Plane (D

Multiple Load Balancer FPGAs

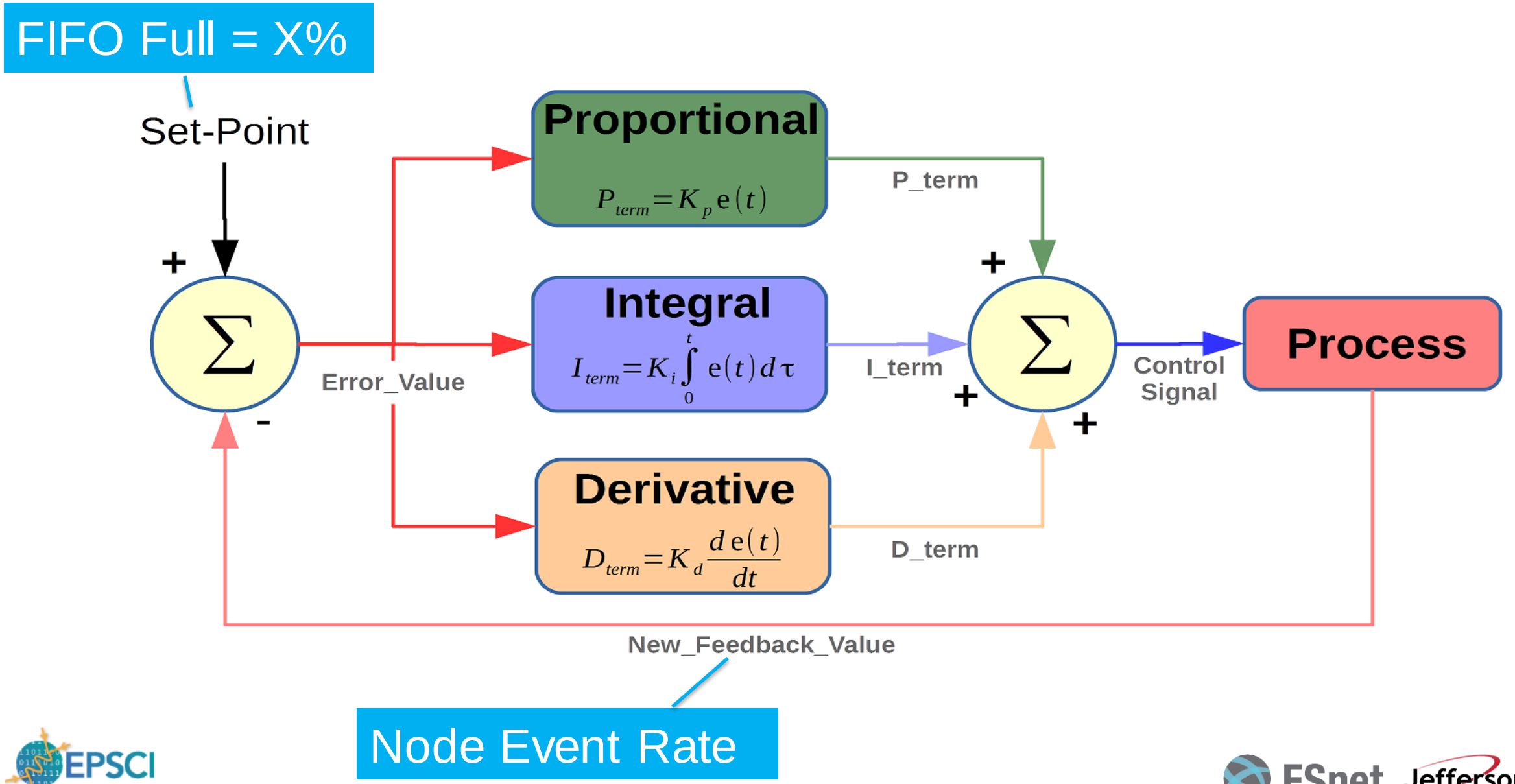


Event Channel Distribution Across CPUs (Fixed)

Event Balance Across Hosts (NAT)

$Host_K$

CP Load Balancing: PID Control



EPSCI

ESnet Jefferson Lab

EJFAT Design Principles

- Data Producer Responsibilities
  - Identify Data *Events*
  - UDP Fragmentation
  - Event, Channel Sequence Packet Tagging
  - Send UDP to LB-DP
- Data Consumer Responsibilities
  - Register with LB-CP
  - Channel Reassembly / Aggregation into Event
  - Post Reassembly Processing for Use Case
- LB Responsibilities
  - Predict Arrival of Future Event Tags (Load Weighting Revision)
  - Load Balance Events Across Registered Nodes
  - Dynamically Weave New Registrants into Load Balance
  - Dynamically Evict Retiring Nodes from Load Balance

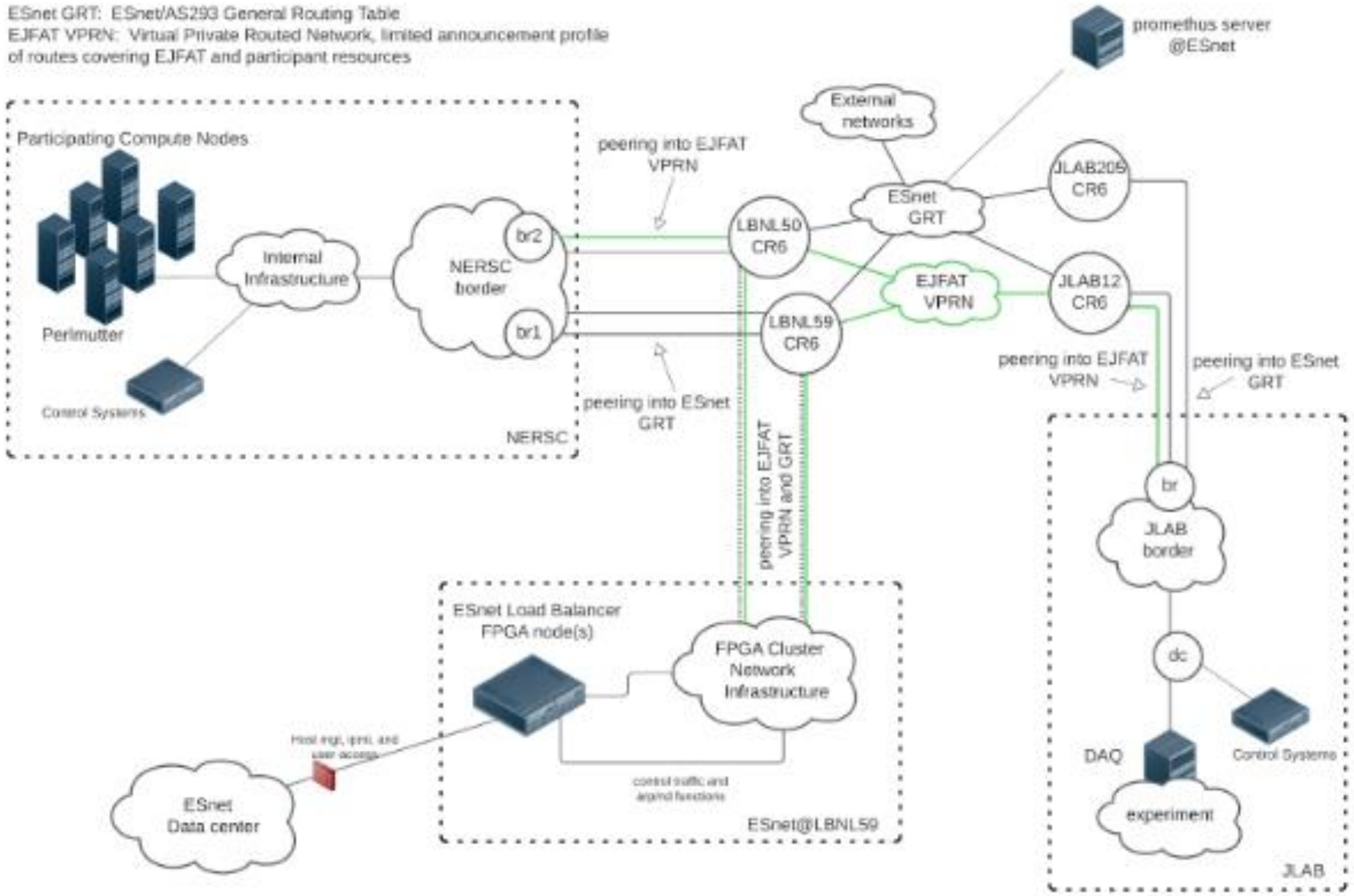
EJFAT LB FPGA Data Plane (DP)

- Network Address Translation (NAT)
- NAT Look Up Tables Configured by Control Plane
  - Network Coordinates for Subscribers (IPv4,6 / MAC)
  - Destination Ports for Channels
  - Data Event to Subscriber
  - Subscriber Workload Weighting/Balancing
- FPGA Supports Four Virtual DP Pipelines
- Network Device
  - Ping
  - ARP

EJFAT LB Control Plane (CP)

- Publish / Subscribe
- Receives
  - (PID) Feedback From The Cluster Nodes
  - Event Sync Messages From Data Source.
- Dynamically Controls DP (FPGA) Distribution weighting
  - Nodes Overworked/Underworked
  - Nodes Added Or Removed
  - Node Data Event Rate Adjustments
- Controls Multiple FPGAs Simultaneously
- Facilitates Tbps Throughput (!)

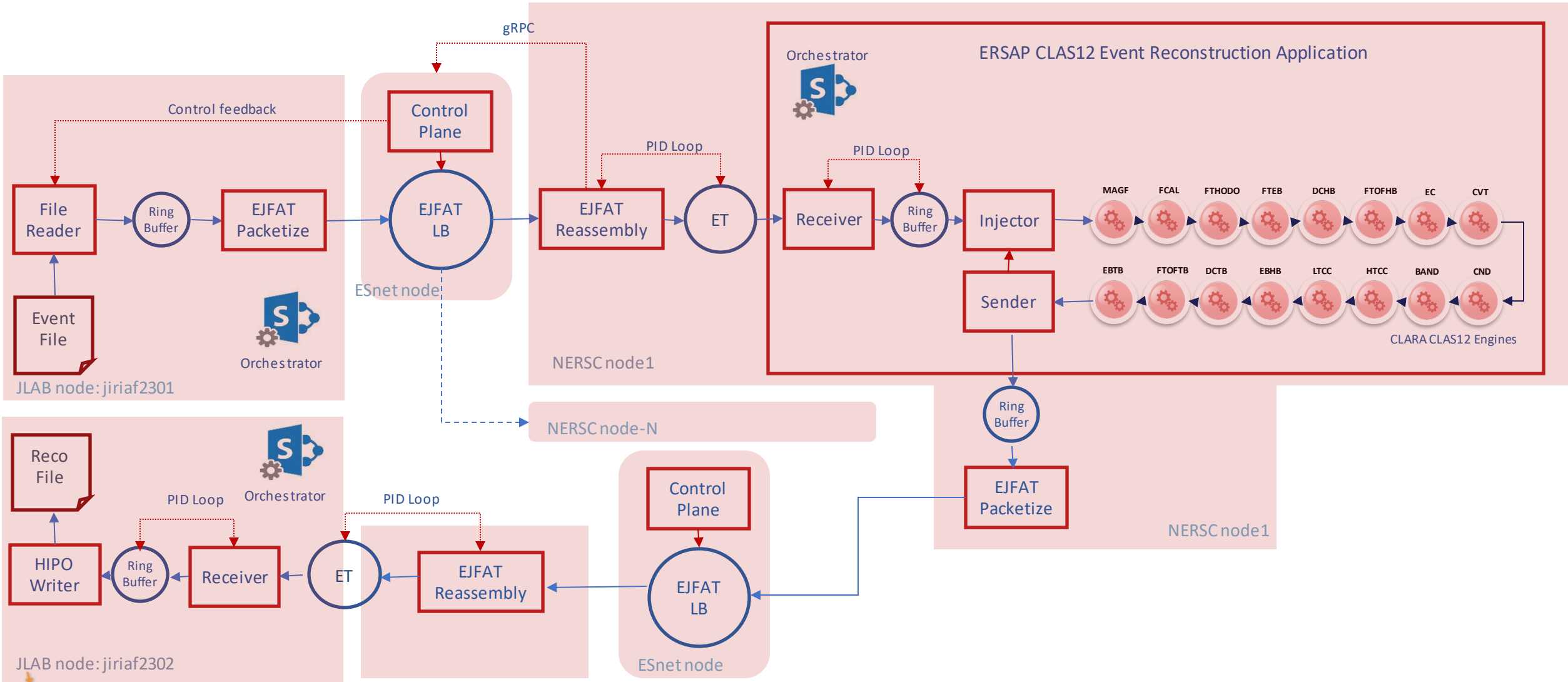
Concept Validation Experiment – CLAS12 Event Streaming



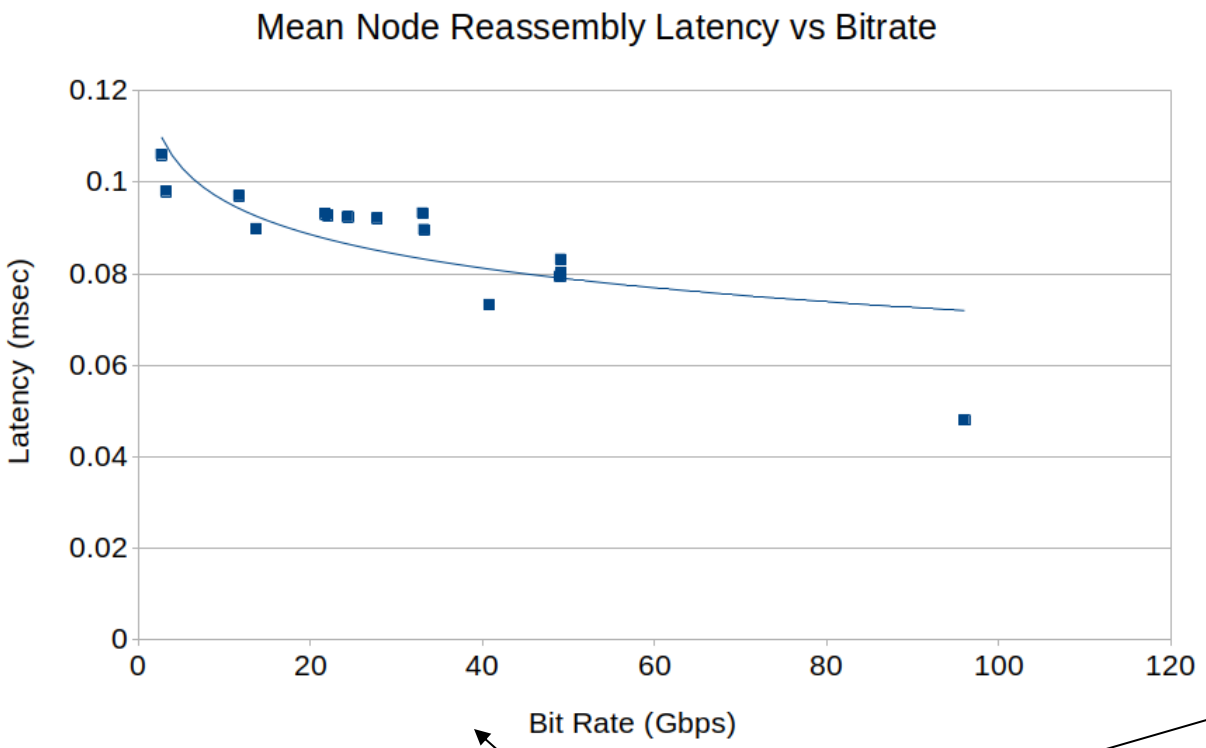
EPSCI

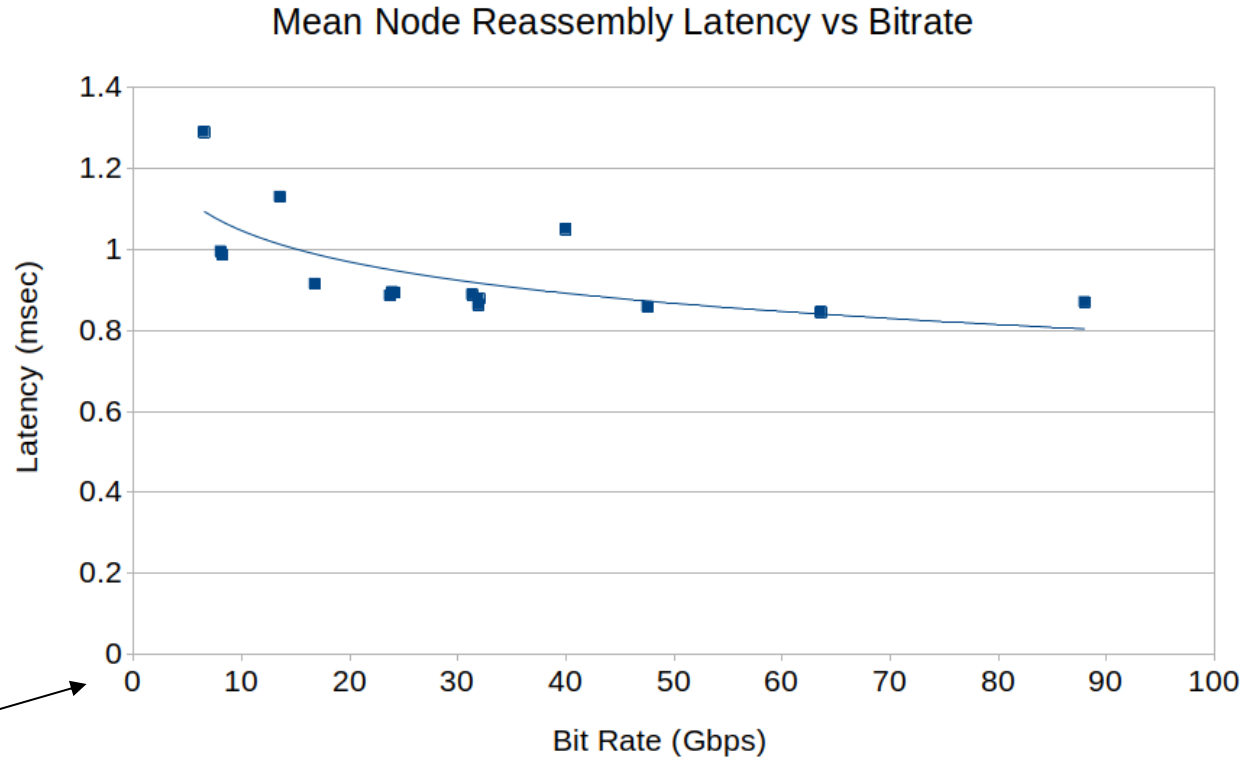
ESnet Jefferson Lab

CLAS12 Online Data-stream Processing JLAB-NERSC. Full Cycle.



Reassembly Latency:



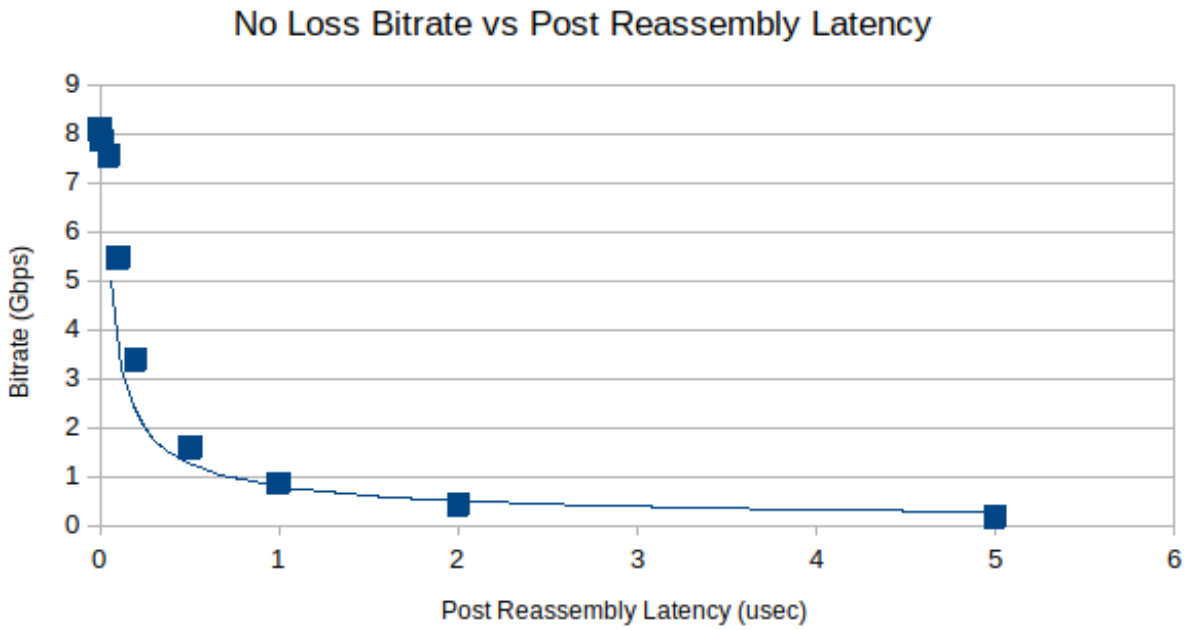


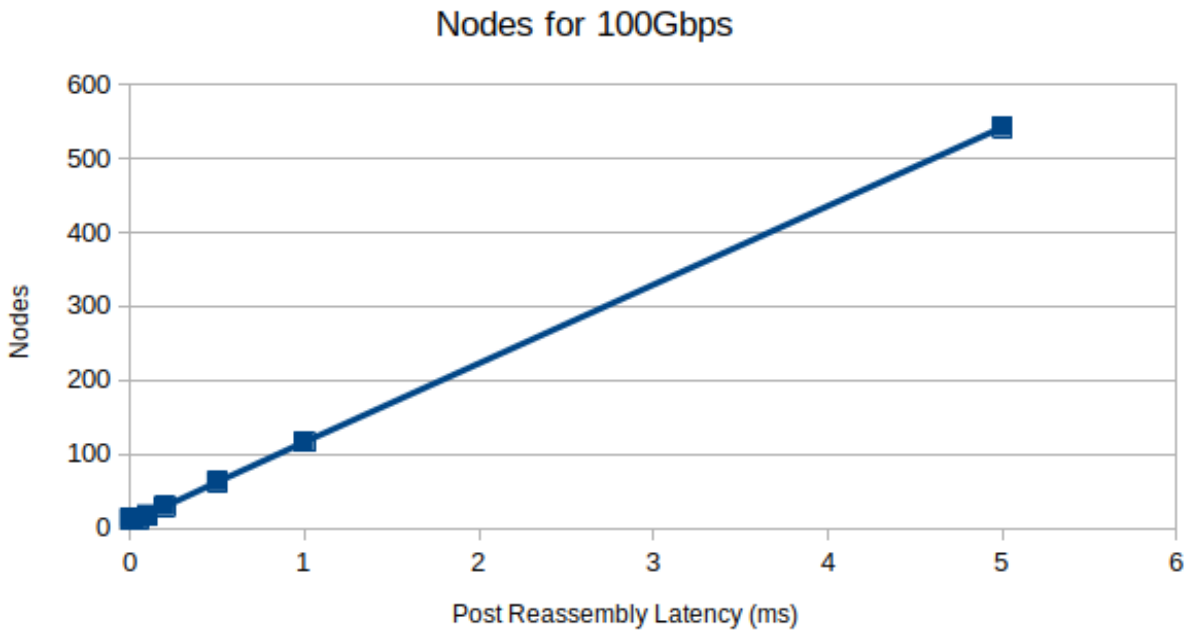
- 100kB Event vs 1 MB Event: Linear Scaling
- Many Factors Determine Latency : OS, UDP/IP Stack, Data Source, Data Rate, Event Size.
- Reassembly CPUs Used In NUMA Domain Of NIC With Realtime Priority
- Latency Decline/Bitrate Believed Due To Efficiency Gains

EPSCI

ESnet Jefferson Lab

Post Reassembly Latency Implications:



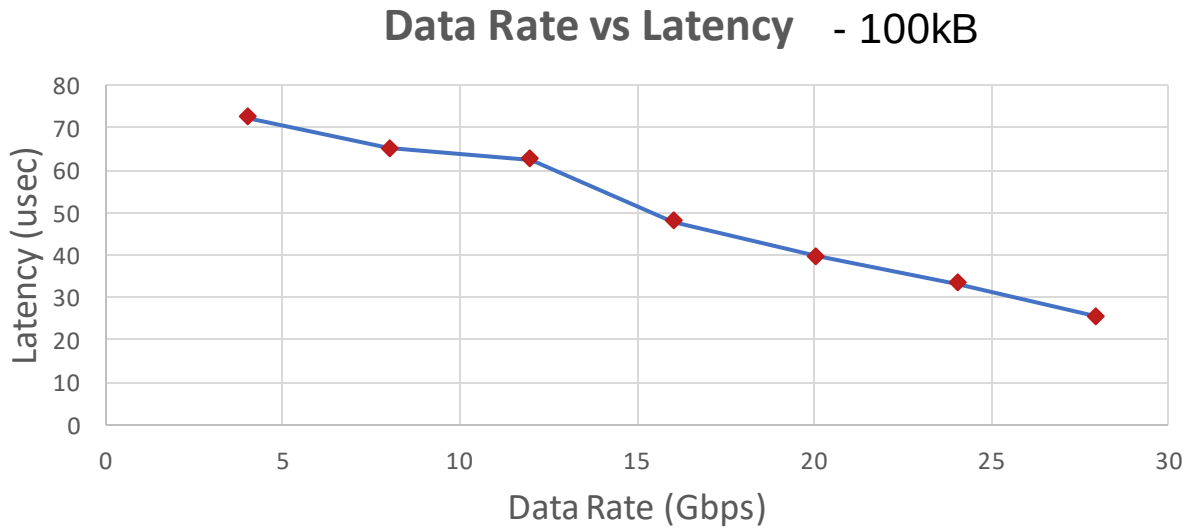


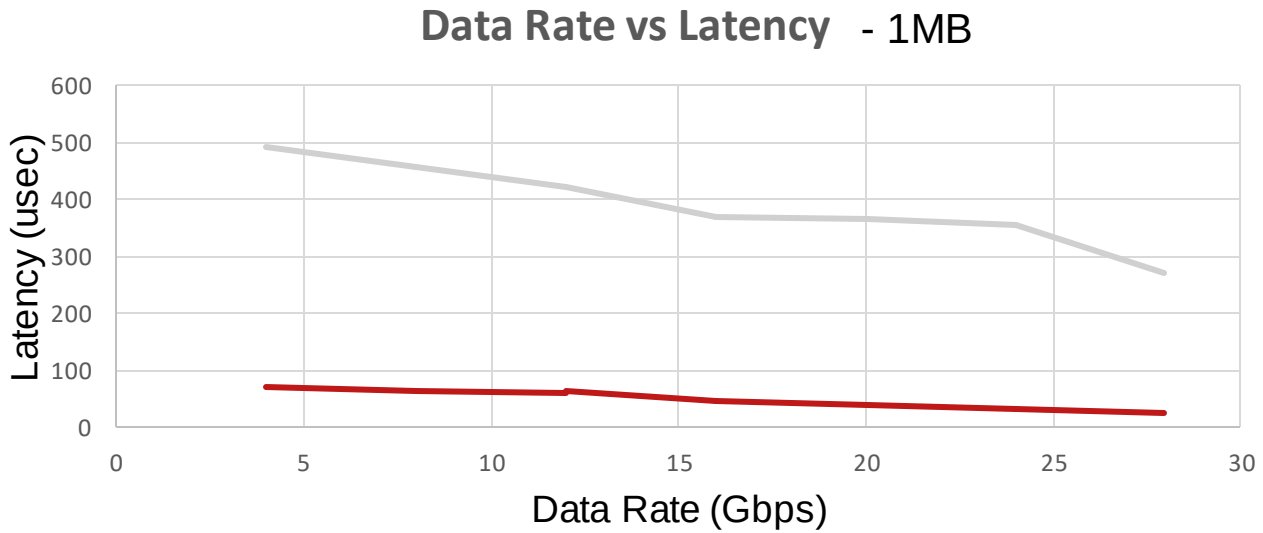
- Measured Max Bit Rate For No Drops
- Max No Drop Bitrate Used To Estimate Nodes For 100gbps
- Measured Mean Reconstruction Latency ERSAP/CLAS12 = 25 Ms / Reass/Recon Suite
- Box with 128 Cores Hosts 4 Nodes -> Measured 5 ms Mean Reass/Recon Latency / Box
- Total Data Latency of EJFAT = Flight Time to LB + Flight Time LB to Node + Reassembly Latency
- E.g., 1MB Event @ 100 Gbps = 100 usec + 100 usec + 1msec = 1.2 msec

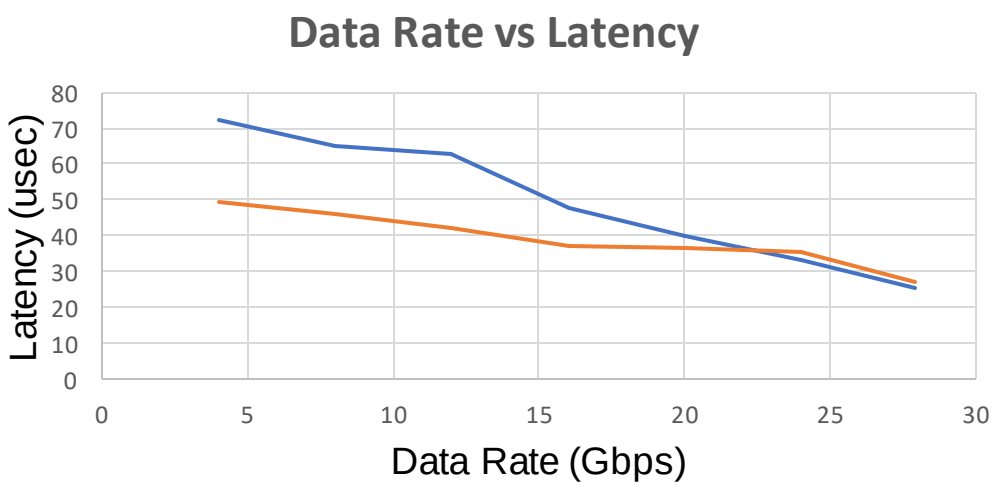
EPSCI

ESnet Jefferson Lab

Reassembly Latency Scaling







- EJFAT Latency
  - Latencies Many Factors: OS, UDP/IP Stack, Data Source/Rate.
  - EJFAT Latency = Flight Time to LB + Flight Time LB to Node + Reassembly Latency
  - E.g., 1MB Event @ 100 Gbps = 100 usec + 100 usec + 1msec = 1.2 msec
  - Latency Improves with Rate
  - Latency Improves With Event Size

EPSCI

ESnet Jefferson Lab

Summary – Big Wins

- EJFAT Streaming Simplifies Experimental Data Collection/Migration Logistics
- Eliminates/Simplifies Counting House Custom Electronics Engineering
- Eliminates HW Triggers/Bias
- Lossless UDP Streaming Enables Global Connections b/n Producer/Consumer
- Decouples Data Source/Consumer
  - Networking
  - Administration
  - Orchestration
- 100 Gbps EJFAT Latency Dominated by Reassembly SW Latency
- Event Latency Dominated by Reconstruction SW Latency