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

*JLAB
+ESnet

EPSCI

Jefferson Lab

ESnet

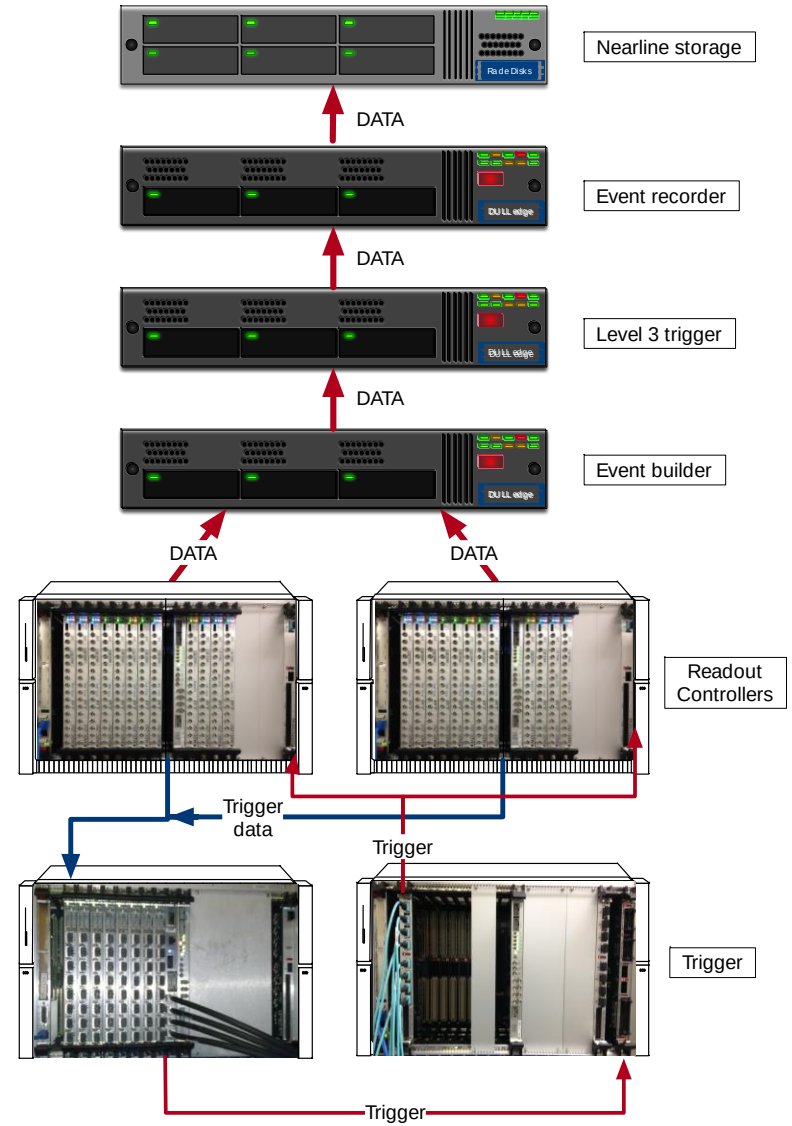
Where Are We Now?

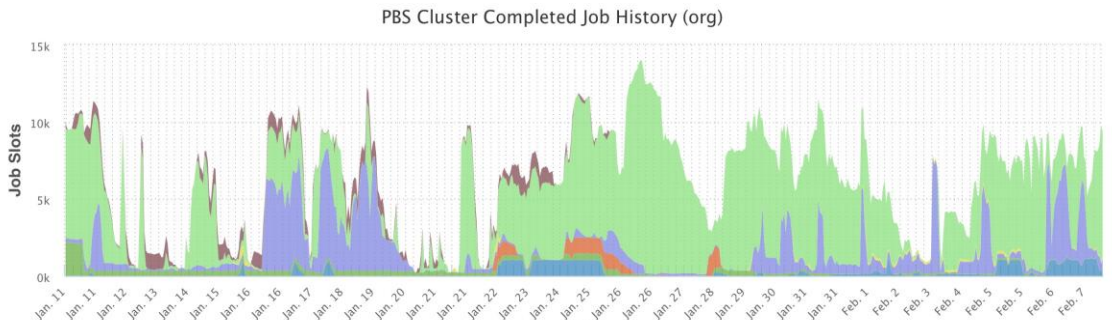
Online:

- Counting House: Custom Electronics, Multi-Level Triggers, Pipelined Readout Systems Build Events Online and Store for Offline Analysis
- Designed To Be Inherently "Stable"
- Stability Often Comes At An Efficiency Cost As The trade Off for Reliable/Acceptable Performance

Offline:

- Events Processed In Steps: Monitoring, Calibration, Decoding, Reconstruction, Analysis.
  - Data Passed Between Stages In Flat Files.
  - Pauses Of Days/Weeks/Months Between Steps.
  - Minimal Automation Between Steps.
  - Analyze with Homogeneous Batch Farms.

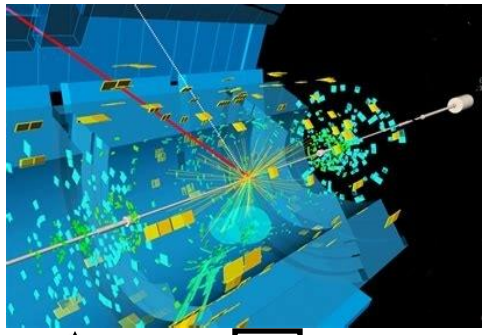




EPSCI

ESnet Jefferson Lab

EJFAT: Line Rate Load Balancer

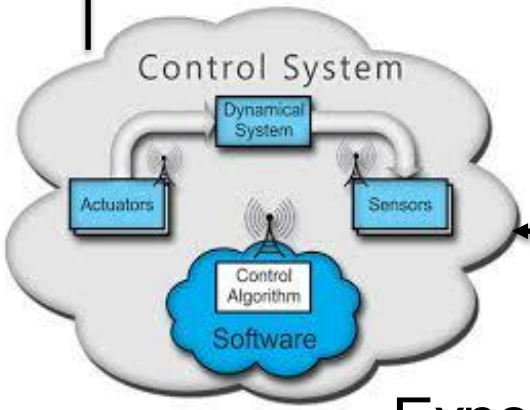


- SRO / Triggerless
- Steady Luminosity
- Hi UpTime

1 Tbps

EJFAT

Higher SNR

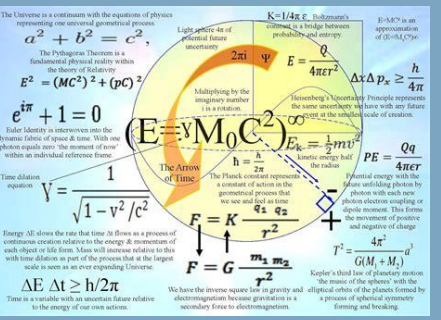


Experiment Steering

Cluster

- **Line Rate**
- Fixed Latency
- Baseline Corrected
- Drift Corrected

Full Physics Model



Near Real-Time Feedback

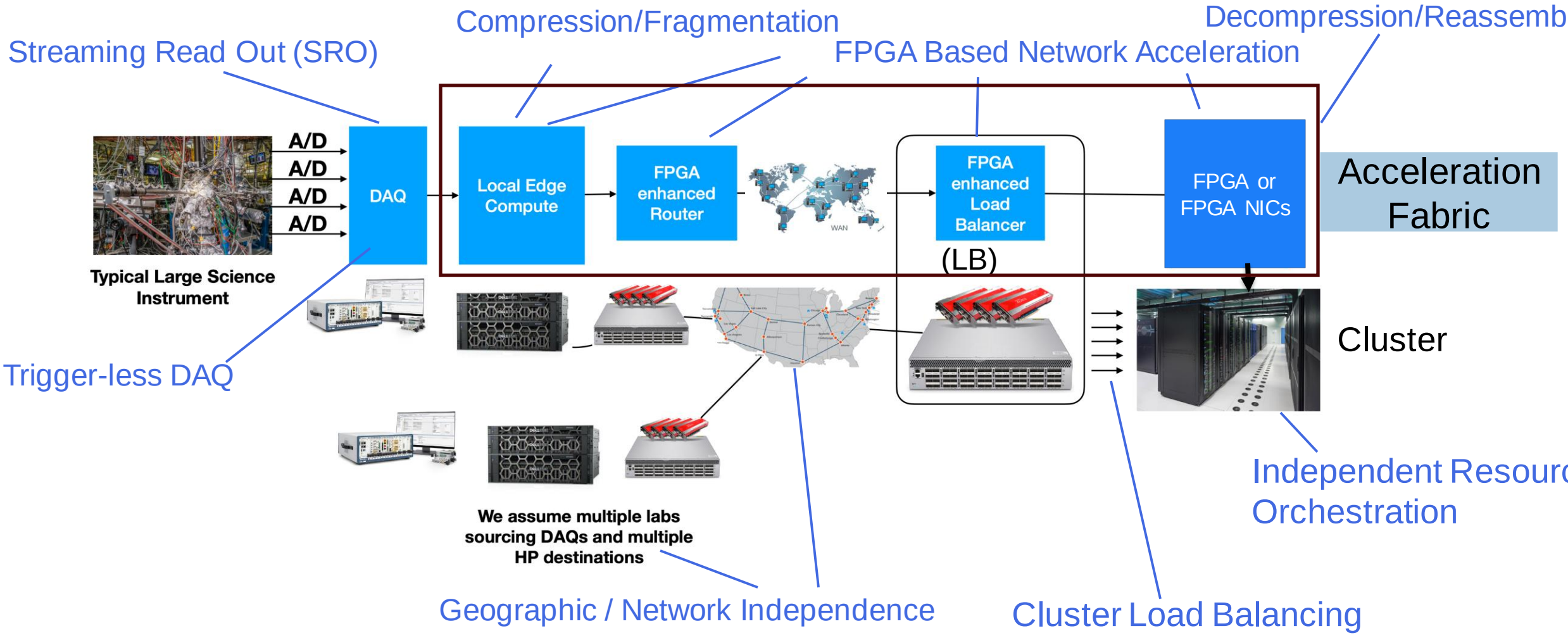
EPSCI

ESnet Jefferson Lab

EJFAT Benefits:

- Move 'Inline/Real-time' Correction Algorithms / Triggers at The Detector to Backend Framework For Greater Flexibility / Power:
  - Easier To Test/Implement/Tweak New Algorithms
  - Algorithms Have Access To Broader Scope of Detector Information
  - Less Restrictive Near-Realtime Feedback Into Correction Algorithms
  - Backend Processing Relaxes Constraints on Power Consumption, Physical Size, and Radiation hardness Required by Frontend Firmware
- Higher Quality Raw Data To Physics Engine Components Have Compounding Benefits In Speed To Science
  - Less Time Chasing Artifacts Due To Trash In The Input,
  - Near-Realtime Extracted Physics Will Be More Reliable, Etc.

EJFAT: Core of Network-Cluster *Acceleration* Fabric



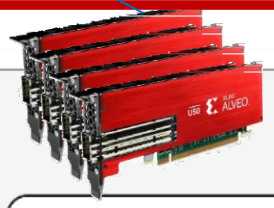
EJFAT LB: Horizontal Scaling

Colors → Events
Shapes → Channels (ROCs)

Multiple Load Balancer FPGAs (Tbps)

Data Plane (DP)

Control Plane (CP)



Load Balancer (LB)

Event Channel Distribution Across CPUs

Experiment edge compute

FPGA - Edge data shaping

FPGA - Data center traffic shaping

Steering

Steering System

Network telemetry

Cluster telemetry

μSec

Event Sequence

Edge Compute

Event Channels

Switch

WAN

Switch

Event Sequence

Node1

Large Compute Resource

NodeN

NodeK

CPU

CPU

CPU

Event Balance Across Nodes (NAT)

Event Aggregation

EPSCI

ESnet

Jefferson Lab

EJFAT LB FPGA Data Plane (DP)

NAT (uSec)

200 Gbps

Xilinx U280

ESnet SmartNIC Open Source:

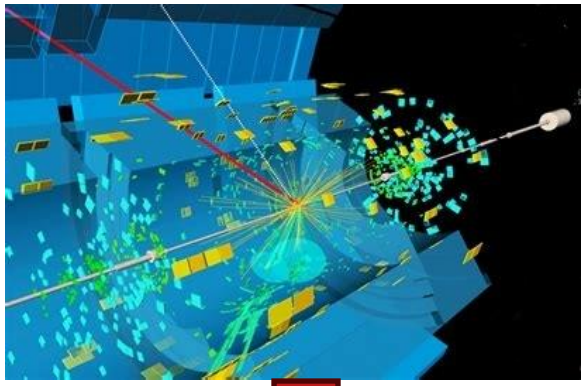
- Ping
- ARP
- Line Rate NAT (EJFAT)
- Some ICMP
- RTL/P4

- Supports Four Virtual DP Pipelines / Separate Experiments
- NAT Look Up Tables Configured by Control Plane
  - Node Network Coordinates
  - Event to Node Dynamic Balancing (1Hz)
  - Destination Ports for Channels

EPSCI

ESnet Jefferson Lab

EJFAT LB Control Plane (CP)

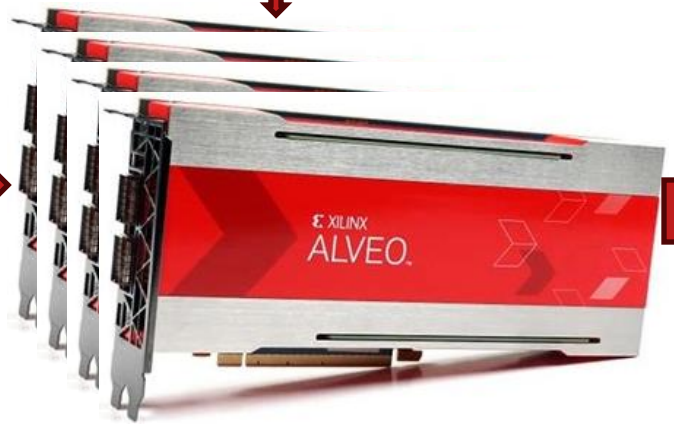


CP

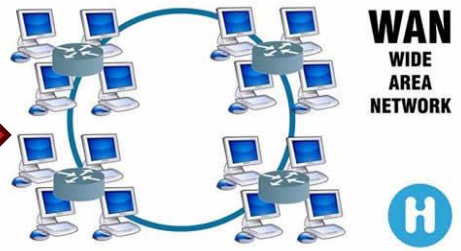
Subscribe
(PID) Feedback (1Hz)

Workload
Balance (1Hz)

Tbps



Tbps
(NAT)



Tbps

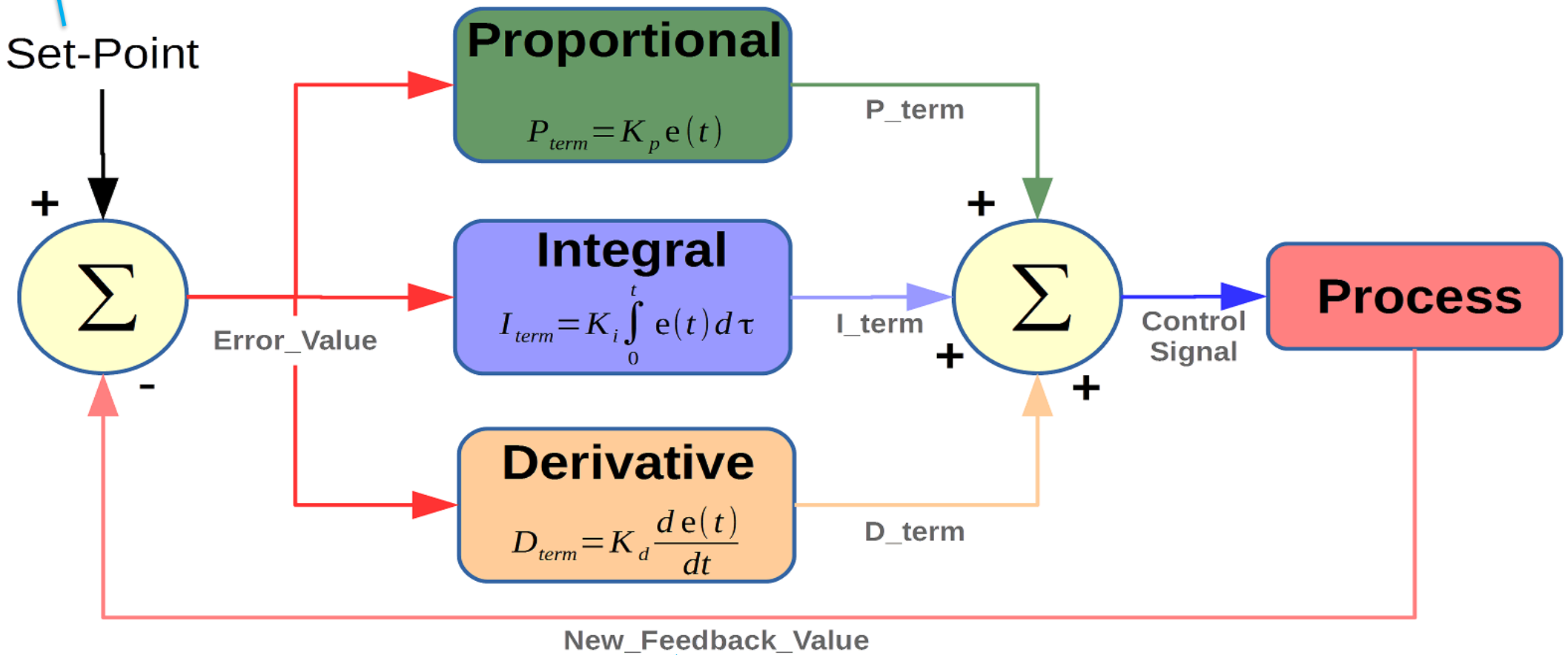
USec Latency

EPSCI

ESnet Jefferson Lab

CP Load Balancing: PID Control

FIFO Full = X%

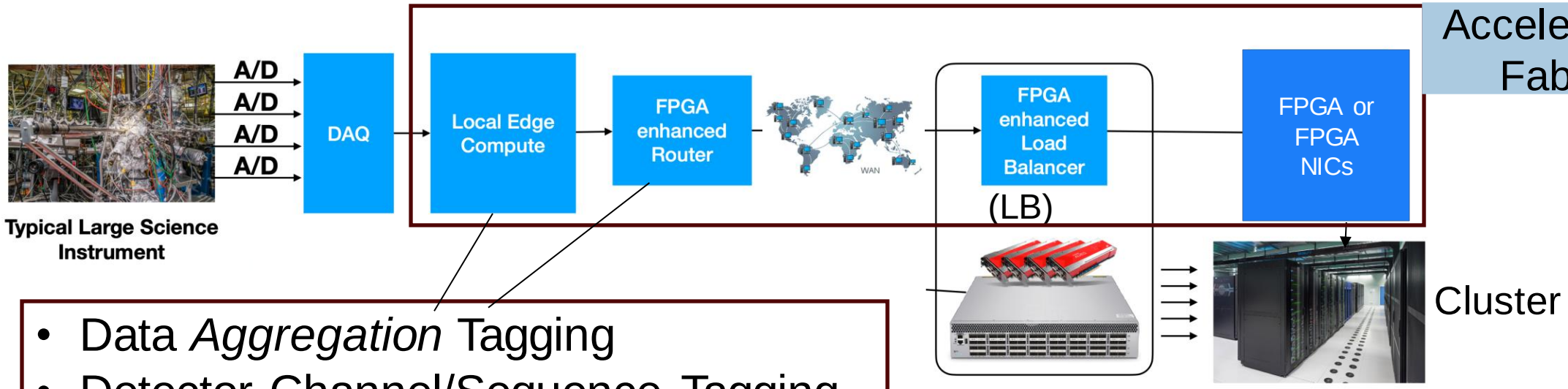


Node Event Rate = DP Schedule Occupancy

EPSCI

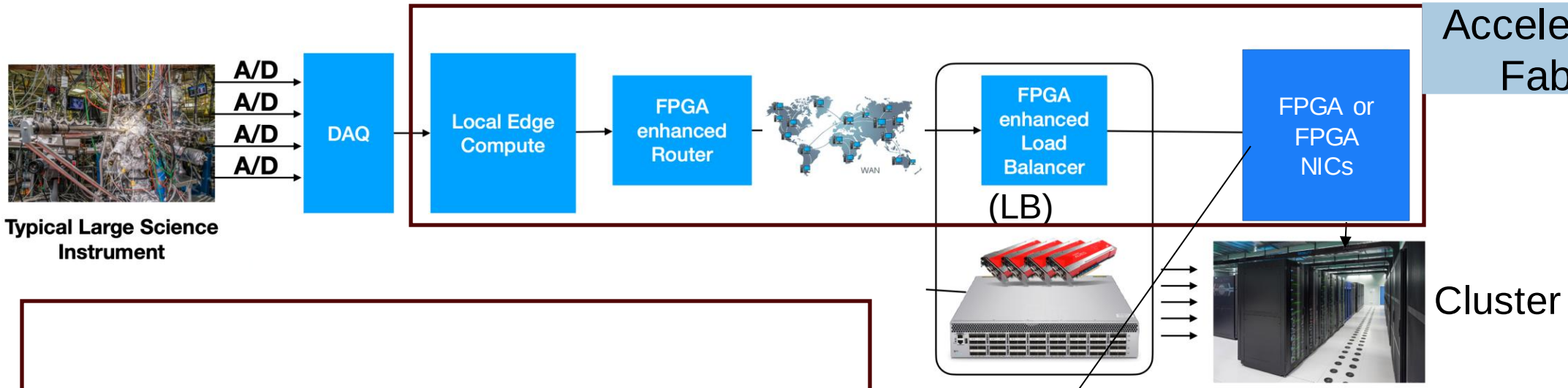
ESnet Jefferson Lab

EJFAT: Data *Producer* Acceleration



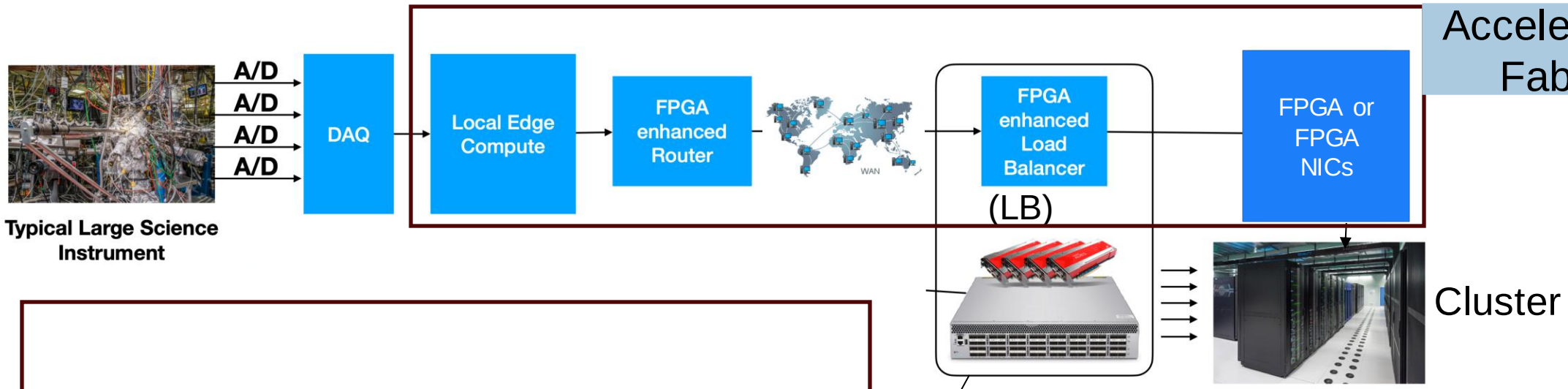
- Data *Aggregation* Tagging
- Detector Channel/Sequence Tagging
- UDP Fragmentation
- Compression

EJFAT: Data *Consumer* Acceleration



- Elastic LB-CP Subscription
- DeCompression
- Detector Channel Reassembly
- Event Aggregation

EJFAT: LB Acceleration



- Dynamic / Elastic Data Publishing
- Line Rate Network Address Translation
- Dynamic Load Balancing

EJFAT: Status / Future

- EJFAT LB DP, CP Developed and Deployed
- Alpha Testing:
  - Jlab Based Data Fabric Research Efforts, LDRDs
  - Jlab Data Source, ESnet based EJFAT LB, LBNL based Cluster (Perlmutter)
  - Jlab Data Source, ESnet based EJFAT LB, ORNL based Cluster (soon)
- Beta Testing:
  - Advanced Light Source (ALS) / LBNL (summer 2024)

Questions?