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

*JLAB
+ESnet

EPSCI

Jefferson Lab

ESnet

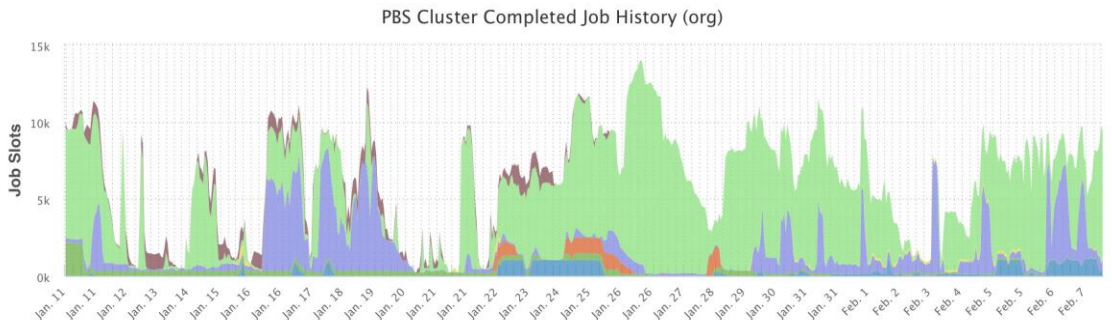
Where Are We Now?

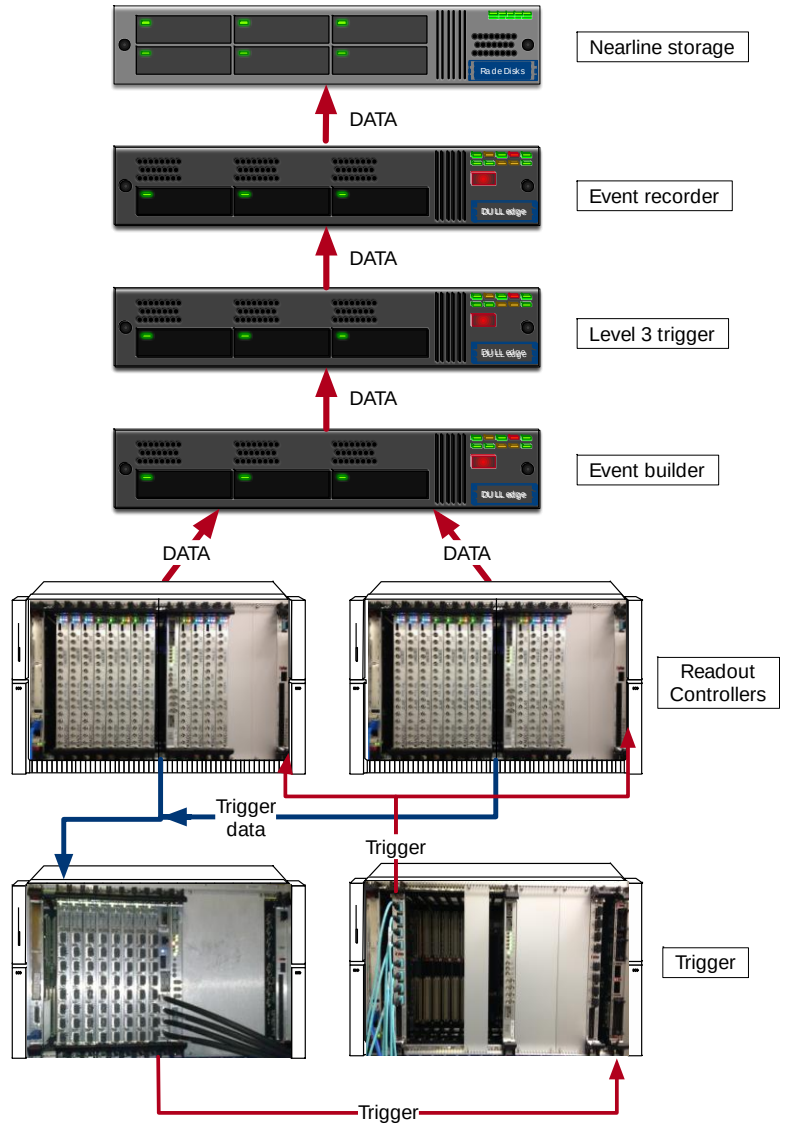
Online:

- 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

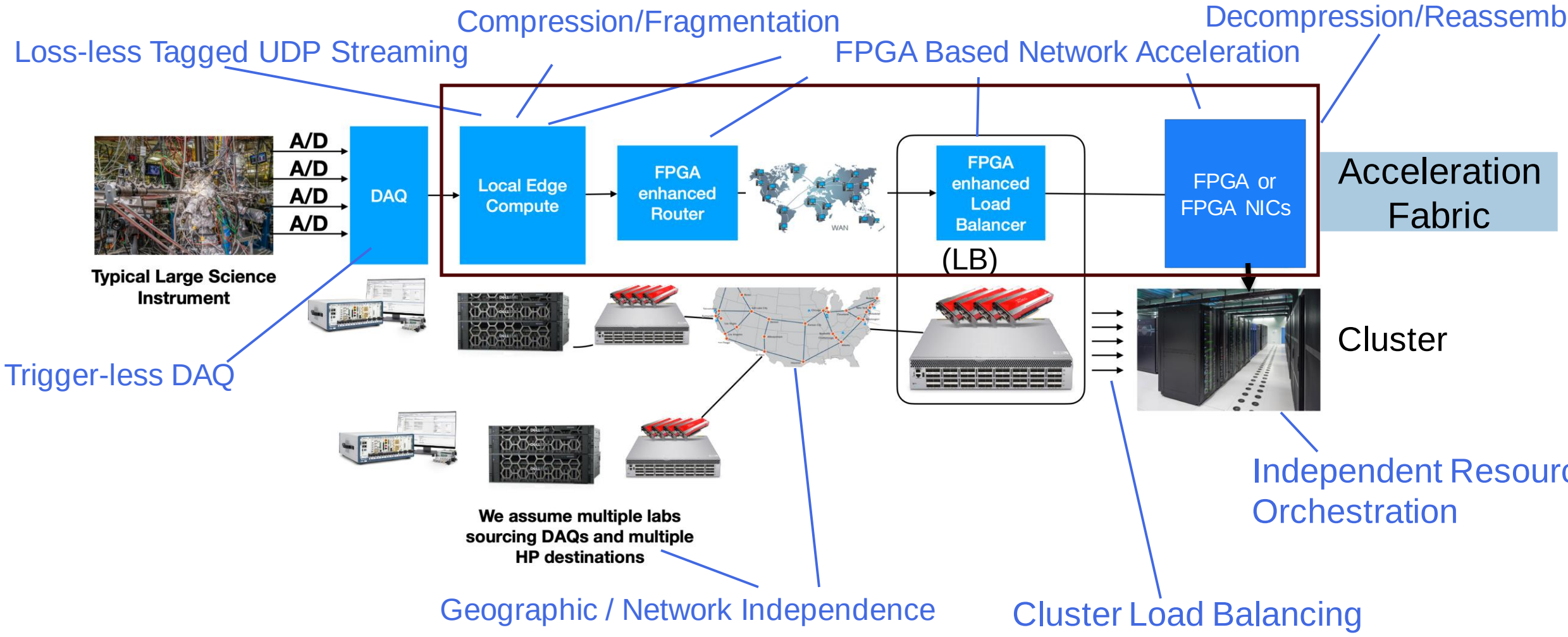
Offline:

- 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.



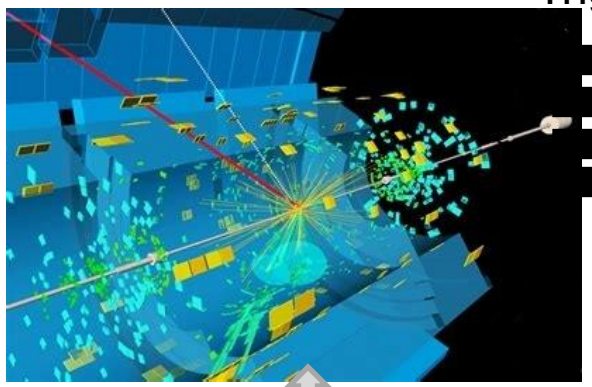


EJFAT: Network-Cluster *Acceleration* Fabric



Full Streaming Readout ➡ Greater Stability, Cluster Resilience/Scaling

Trigger: Legacy ⟷ None

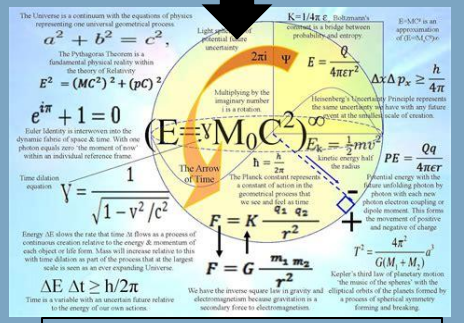


- UnAggregated
- Streaming
- COTS Edge

Network / Cluster Accelerator Fabric:

- Line Rate
  - Channel Aggregation
  - NAT
- Pub-Sub/Load Balance
  - Cluster Scaling
  - Cluster Resiliency
  - Load Mngmnt

Clusters 1,…,N



OffLine: Full Physics Model

- Longer Term Stability
- Higher Quality Data
- Higher Up Time ($)
- Science Rate Acceleration

Control

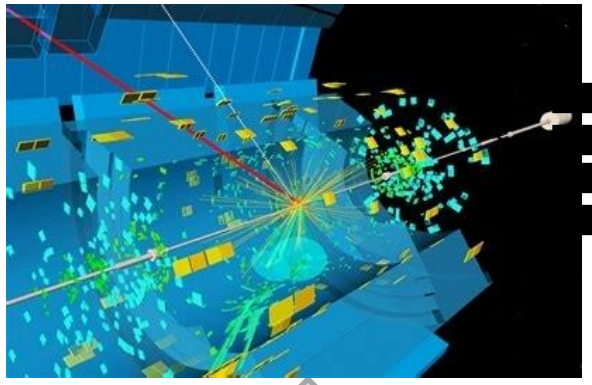
Near Real-Time Feedback
Particle ID, Vertex, etc

Control System
OnLine: Flexible S/W Triggers
Replace fixed H/W, F/W (?)

EPSCI

ESnet Jefferson Lab

Better Detector Feedback ➡ Greater Stability, Better Data



If No Trigger: Full Beam Current

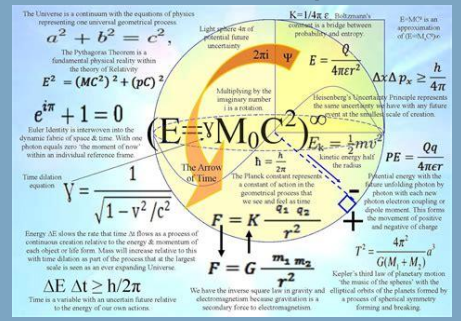
Network / Cluster Acceleration Fabric:

If No Trigger: No Reconfig of Front End DAQ

High-level Analysis Driven Feedback Loops Require Aggregate Data

Clusters 1,…,N

Farm Allows Increasingly Complex Trigger With Increased Flexibility



More Robust Algorithms, Easier Algorithm Monitoring

'Hit By Hit' Drifts O(us)

Suppress Noise And Dramatically Increase The SNR

Baseline Drifts O(s)

Statistics For Fit To Apply Corrections

Control System

Rapid Feedback Based On Low-Level Detector Outputs Provide Rapid Baseline Corrections O(ms)

EPSCI

ESnet Jefferson Lab

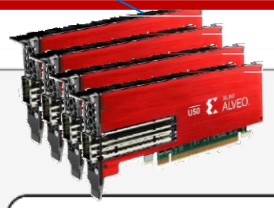
EJFAT LB: Horizontal Scaling

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

Network Device:

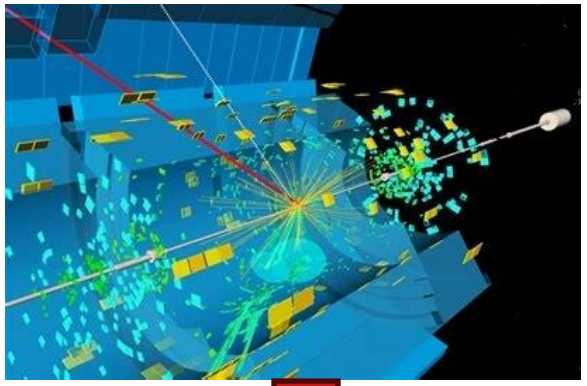
- Ping
- ARP
- Line Rate NAT
- Some ICMP
- RTL/P4

- 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

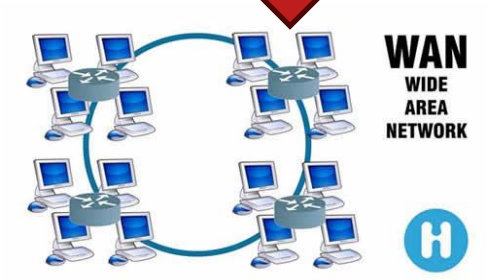
ESnet Jefferson Lab

EJFAT LB Control Plane (CP)

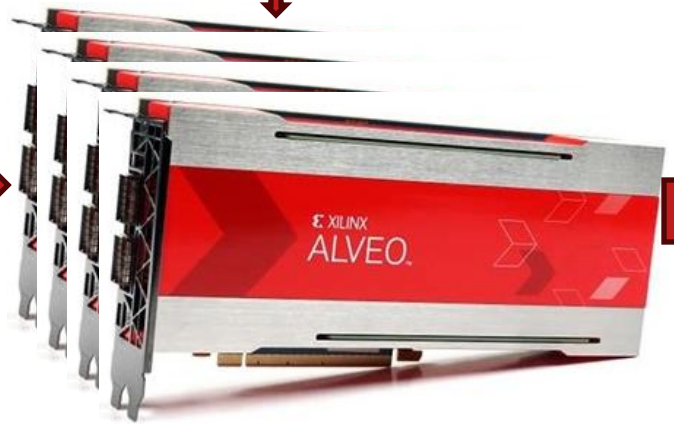


Subscribe
(PID) Feedback (1Hz)

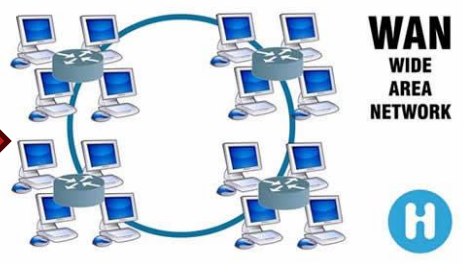
Workload
Balance (1Hz)



Tbps



Tbps
(NAT)



Tbps

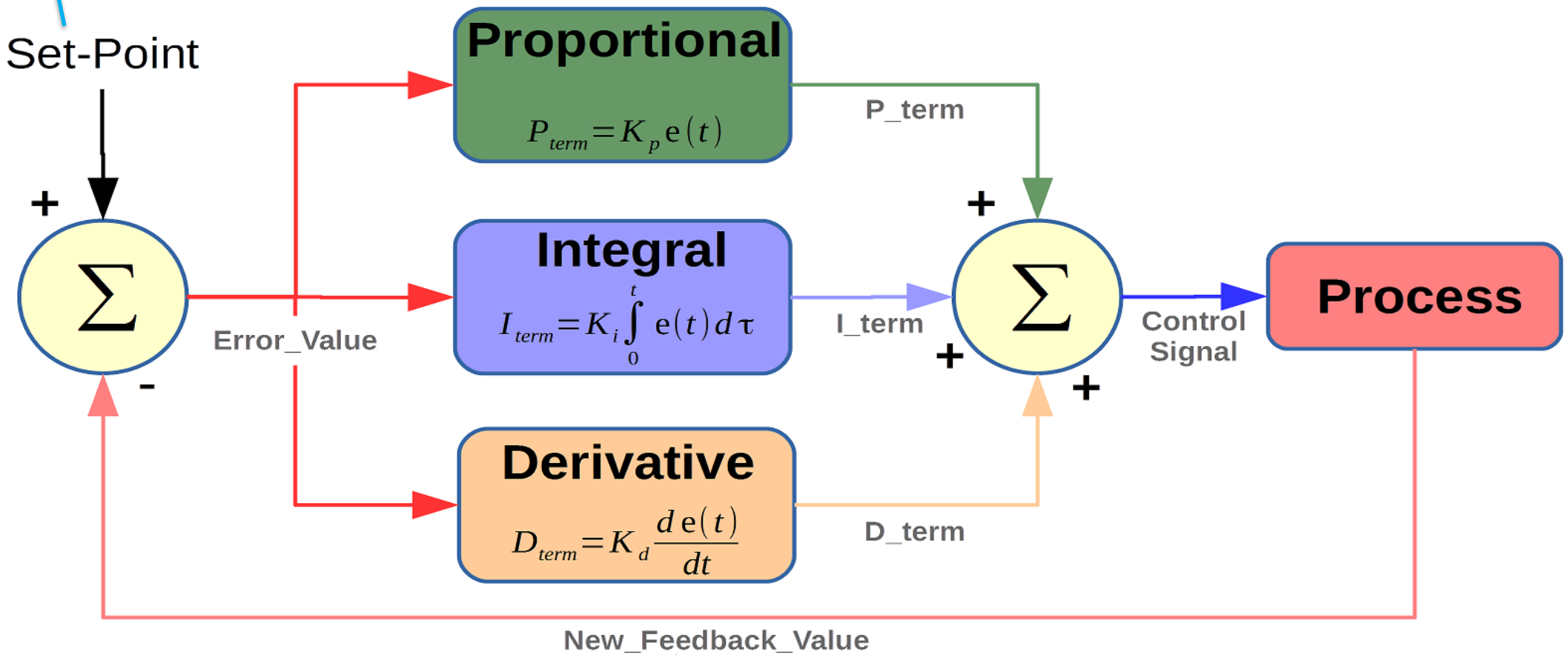
USec Latency

EPSCI

ESnet Jefferson Lab

CP Load Balancing: PID Control

FIFO Full = X%

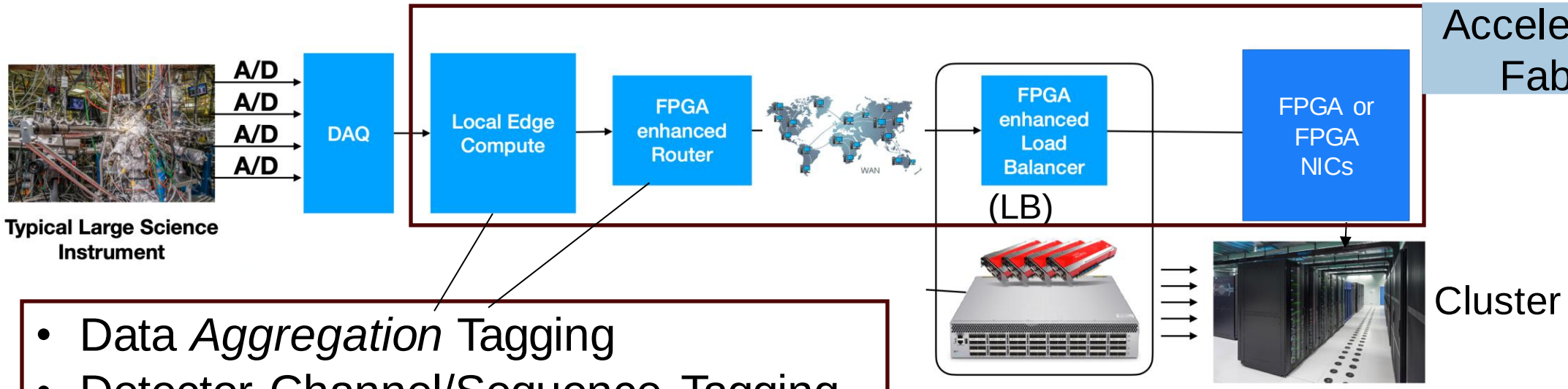


Node Event Rate = DP Schedule Occupancy

EPSCI

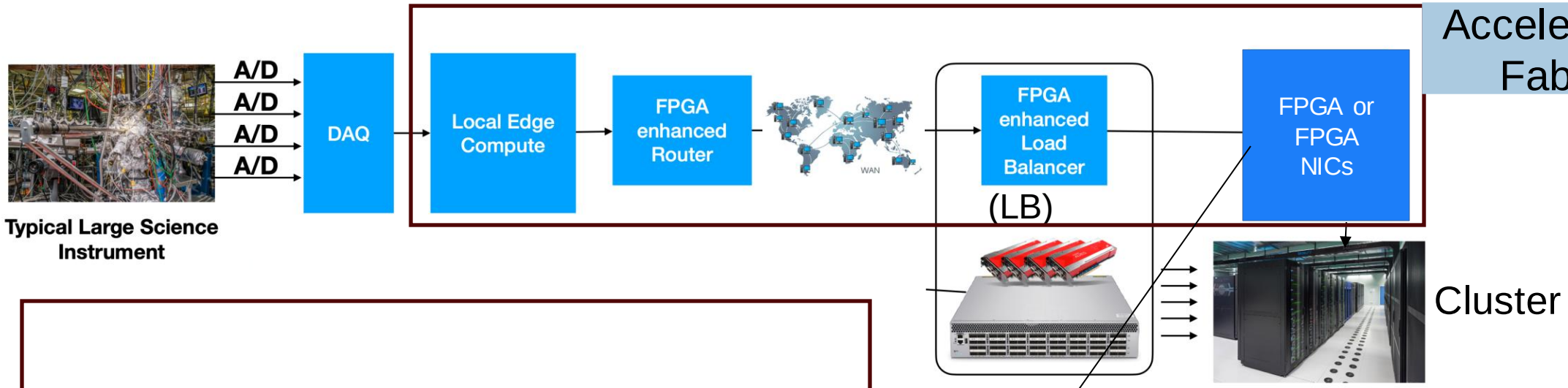
ESnet Jefferson Lab

EJFAT: Data *Producer* Acceleration



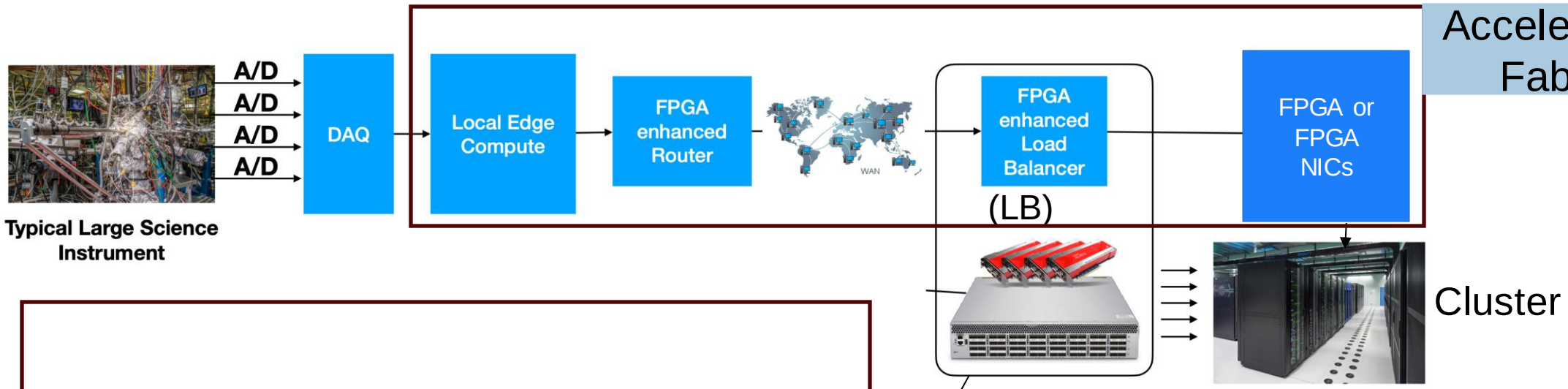
- Data *Aggregation* Tagging
- Detector Channel/Sequence Tagging
- UDP Fragmentation
- Compression

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

Status / Future

- EJFAT LB Data Plane, Control Plane 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?