

EJFAT

ESnet-Jefferson Lab FPGA Accelerated Transport

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

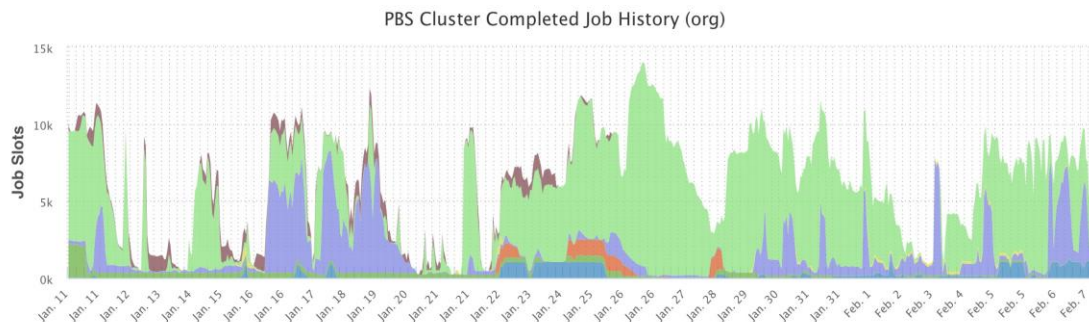
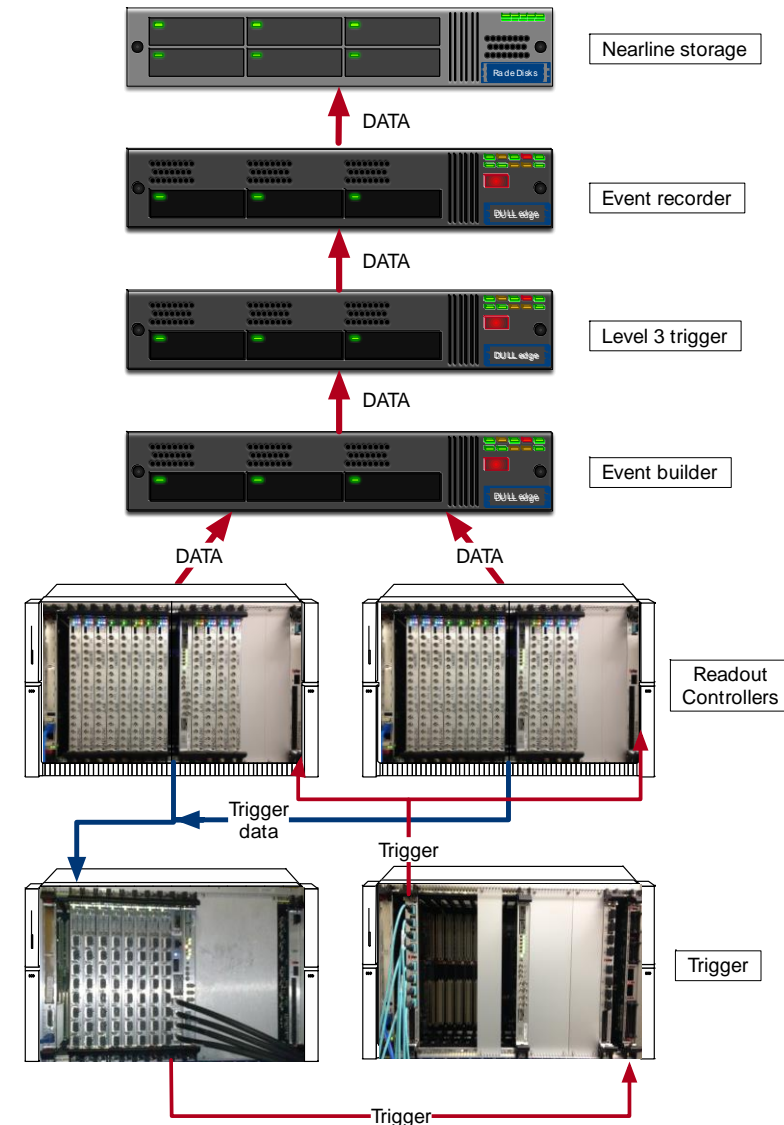
Where Are We Now?

Online:

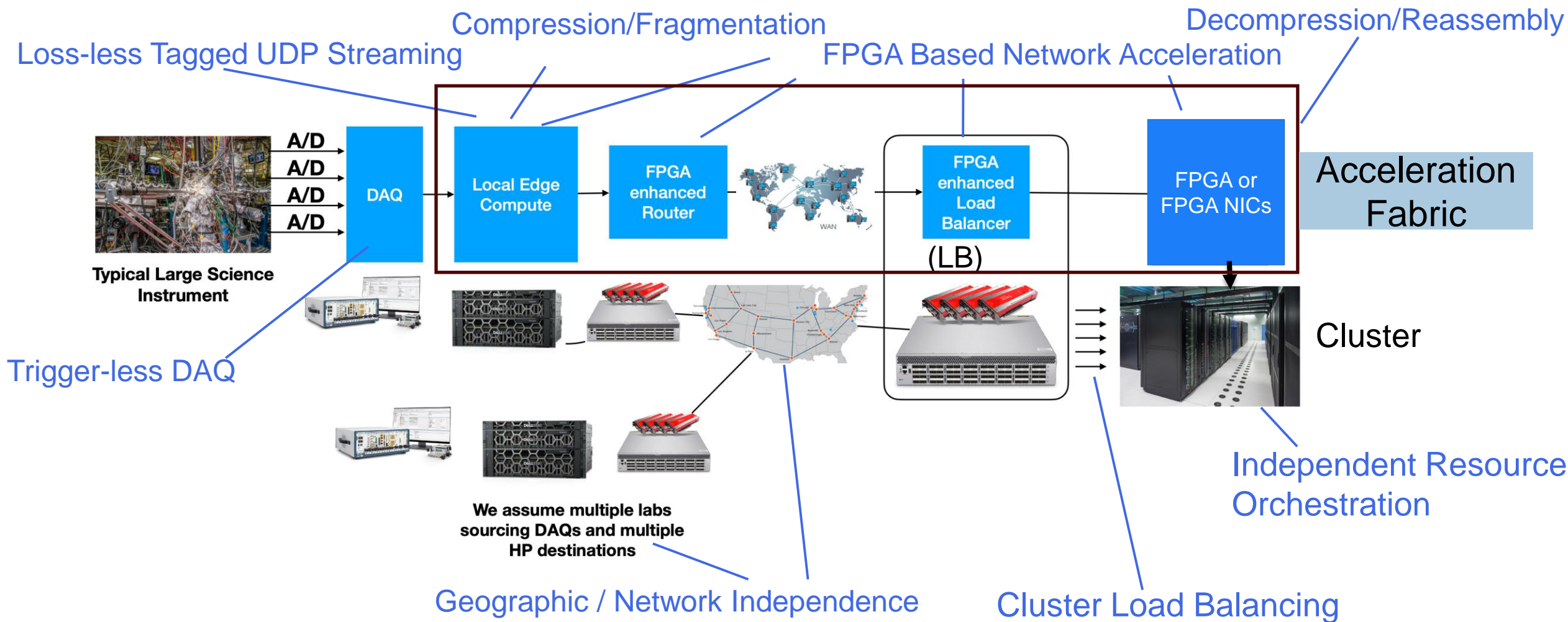
- Counting House: Custom Electronics, Multi-Level Triggers, Pipelined Readout Systems Build Events Online and Store for Offline Analysis
- Designed To Be Inherently "Stable"
- Stability Often Comes At An Efficiency Cost As The trade Off for Reliable/Acceptable Performance

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.

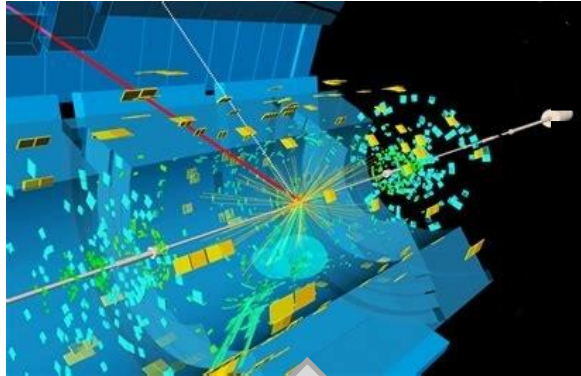


EJFAT: Network-Cluster Acceleration Fabric



Full Streaming Readout \rightarrow Greater Stability, Cluster Resilience/Scaling

Trigger: Legacy \leftrightarrow None

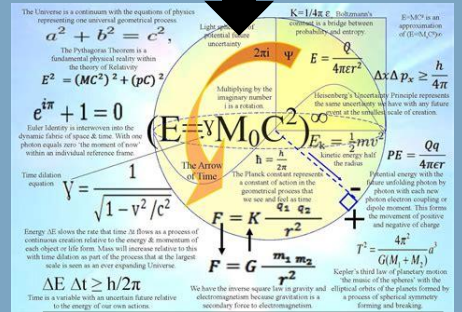


- UnAggregated
- Streaming
- COTS Edge

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

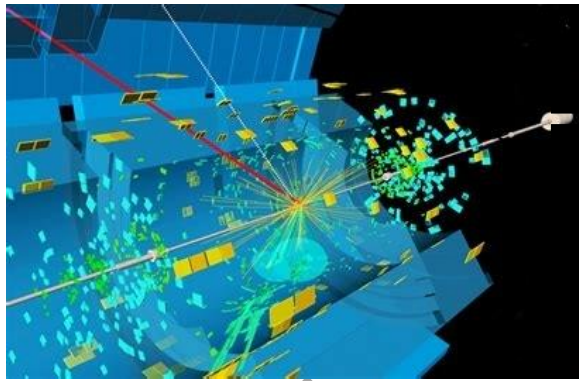
Control System
OnLine: Flexible S/W Triggers
Replace fixed H/W, F/W (?)

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

Control

Near Real-Time Feedback
Particle ID, Vertex, etc

Better Detector Feedback \rightarrow Greater Stability, Better Data



If No Trigger: Full Beam Current

If No Trigger: No Reconfig of Front End DAQ

Network / Cluster Acceleration Fabric:

High-level Analysis Driven Feedback Loops Require Aggregate Data



Farm Allows Increasingly Complex Trigger With Increased Flexibility

'Hit By Hit' Drifts $O(\mu s)$

Suppress Noise And Dramatically Increase The SNR

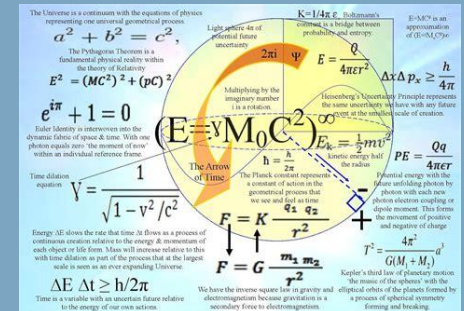
Baseline Drifts $O(s)$

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

Statistics For Fit To Apply Corrections

More Robust Algorithms, Easier Algorithm Monitoring

Control System

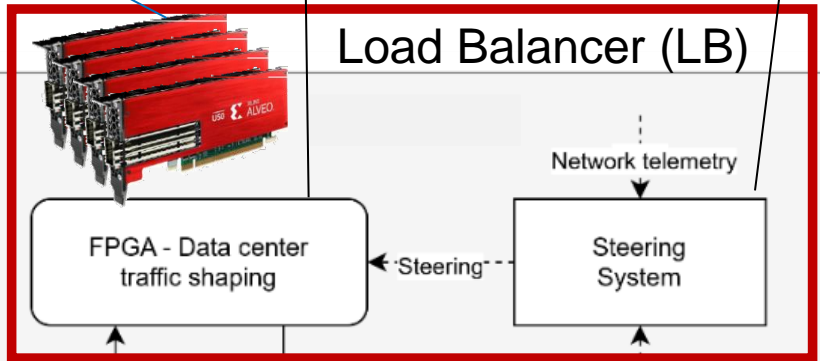


EJFAT LB: Horizontal Scaling

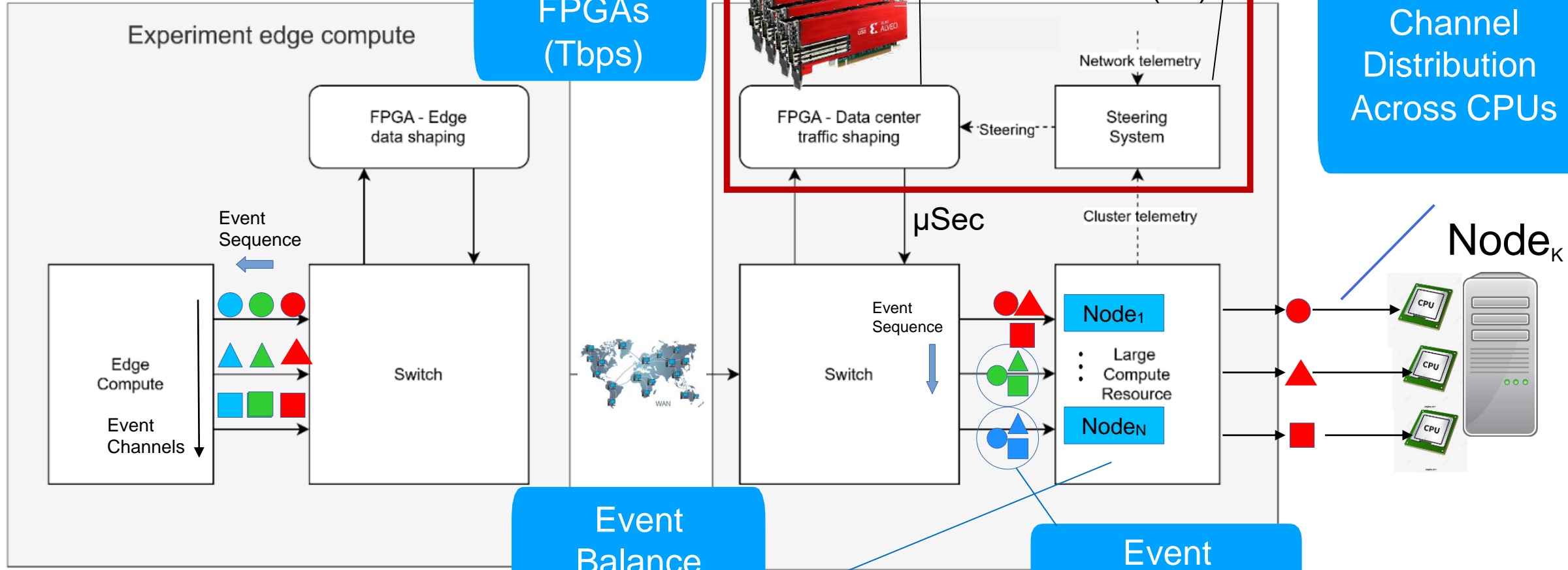
Colors → Events
Shapes → Channels (ROCs)

Multiple Load Balancer FPGAs (Tbps)

Data Plane (DP) Control Plane (CP)



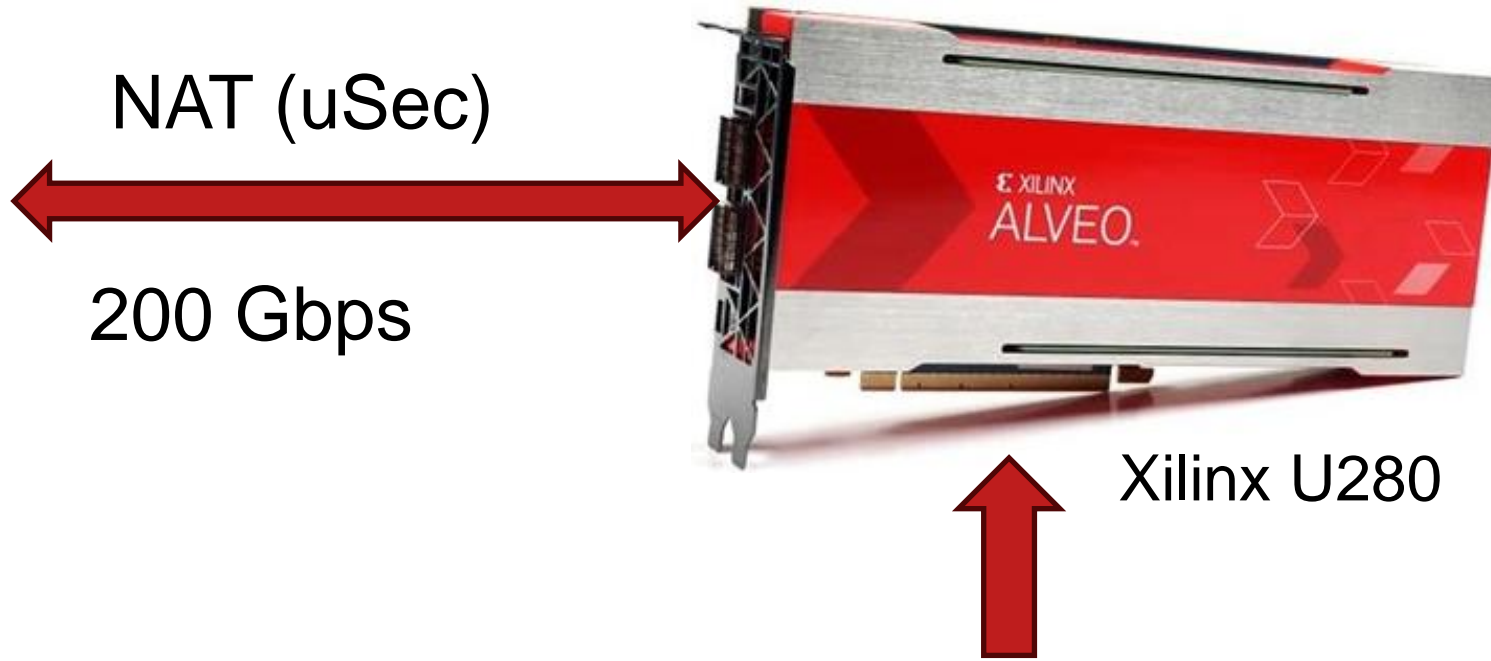
Event Channel Distribution Across CPUs



Event Balance Across Nodes (NAT)

Event Aggregation

EJFAT LB FPGA Data Plane (DP)

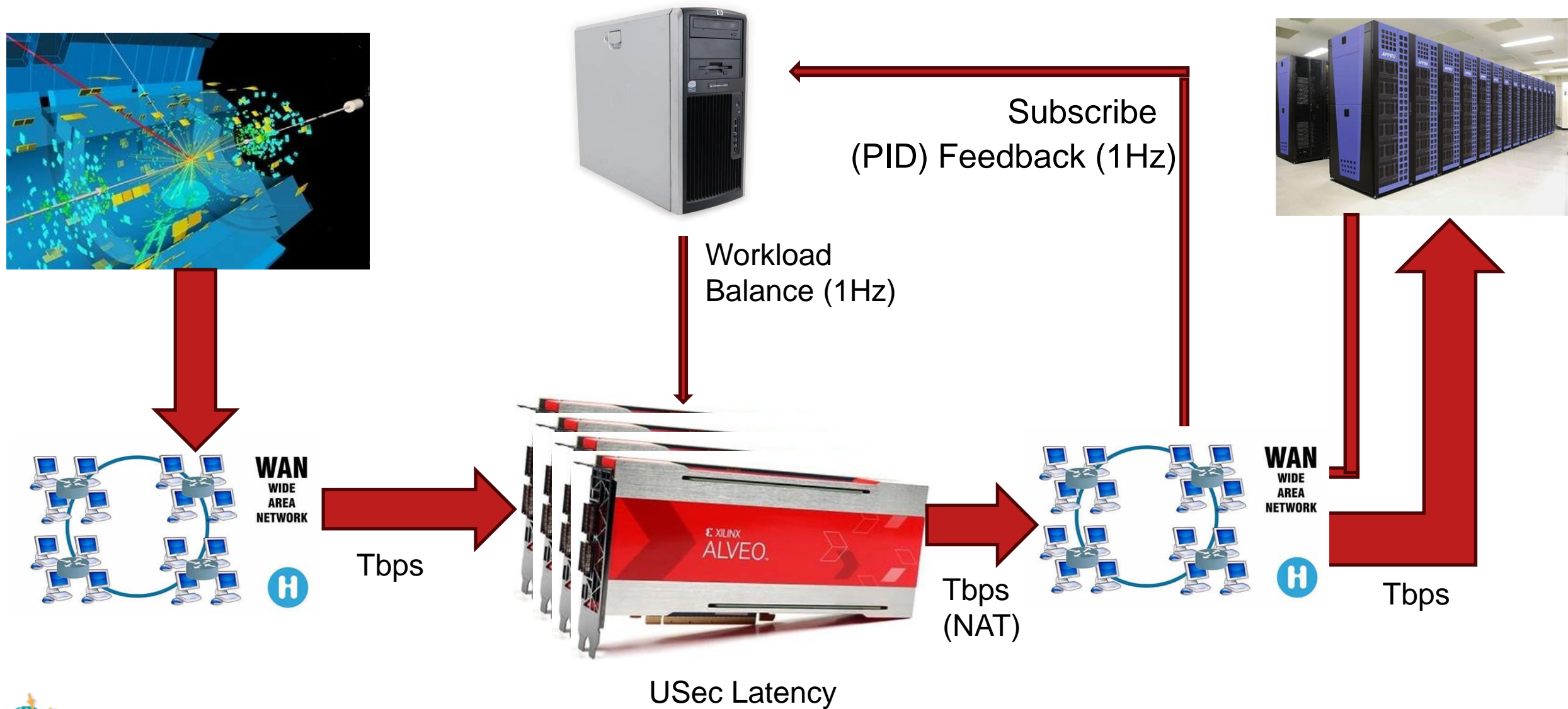


Network Device:

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

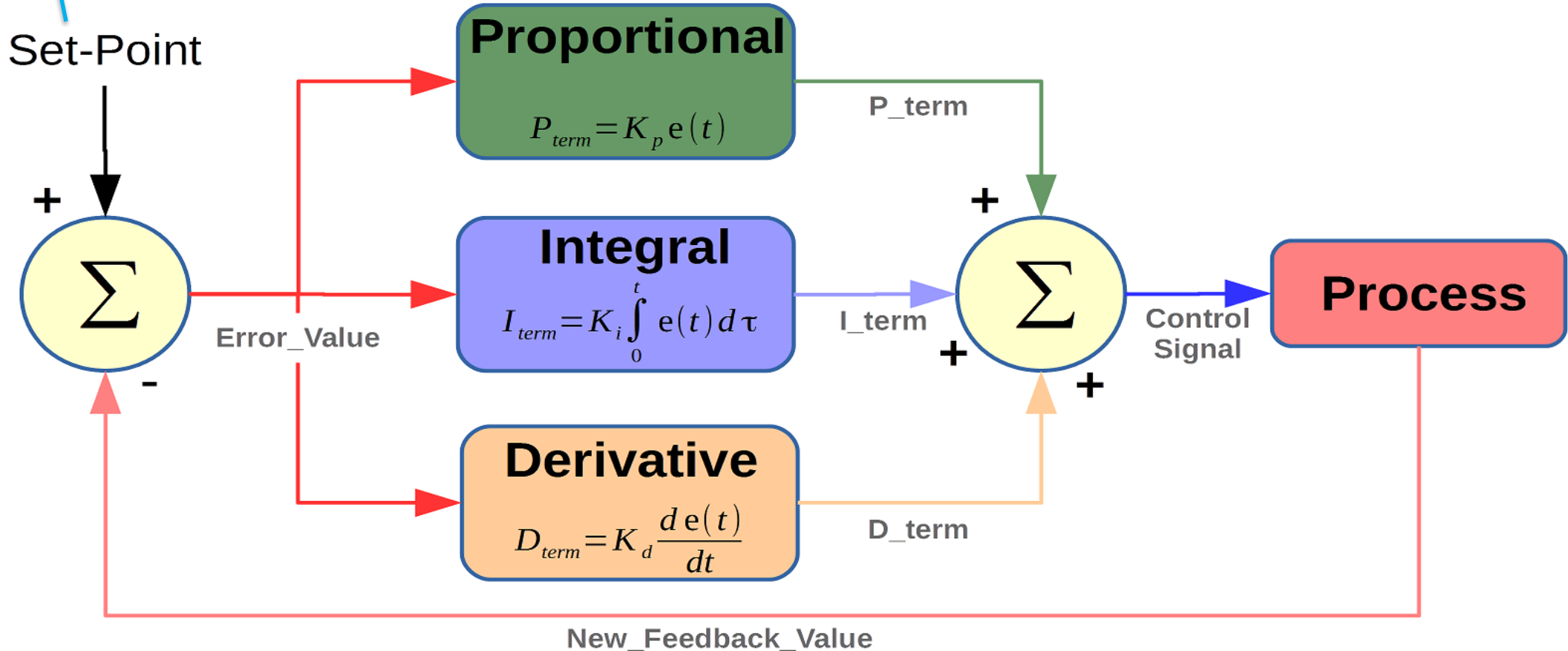
- Supports Four Virtual DP Pipelines / Separate Experiments
- NAT Look Up Tables Configured by Control Plane
 - Node Network Coordinates
 - Event to Node Dynamic Balancing (1Hz)
 - Destination Ports for Channels

EJFAT LB Control Plane (CP)



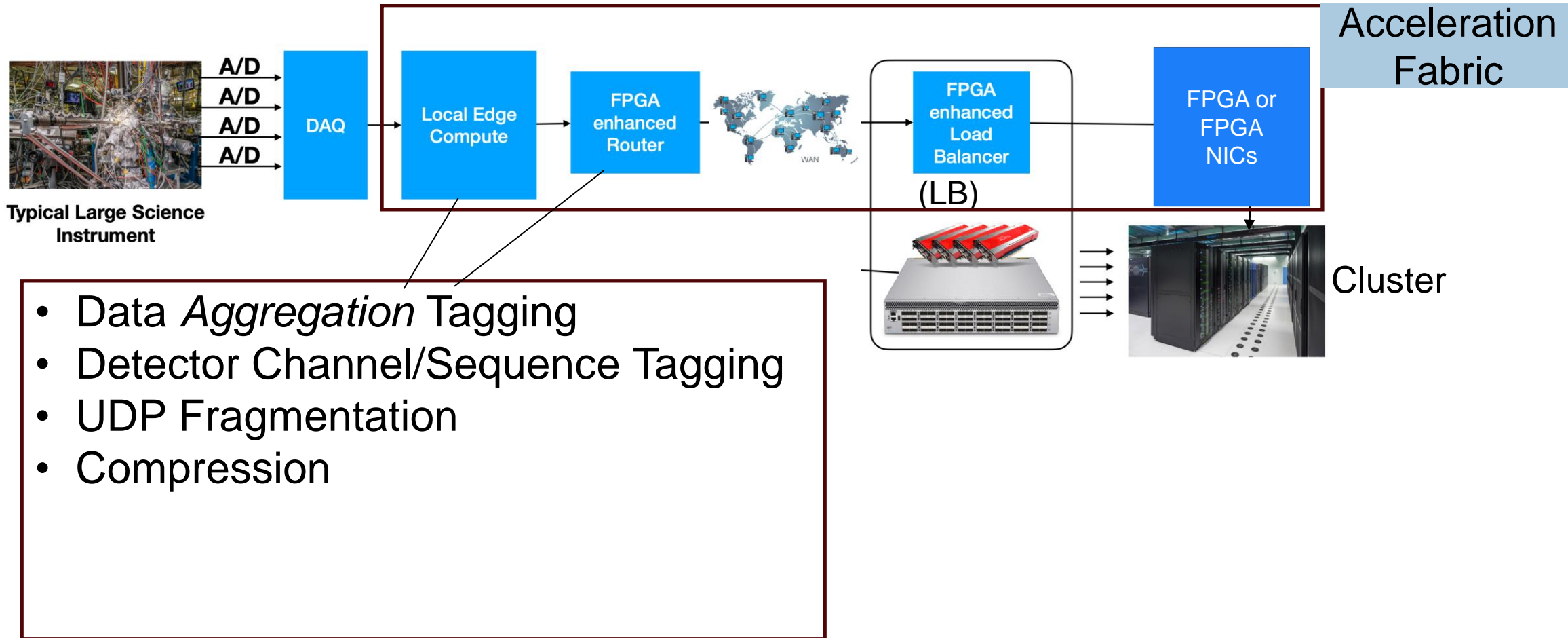
CP Load Balancing: PID Control

FIFO Full = X%

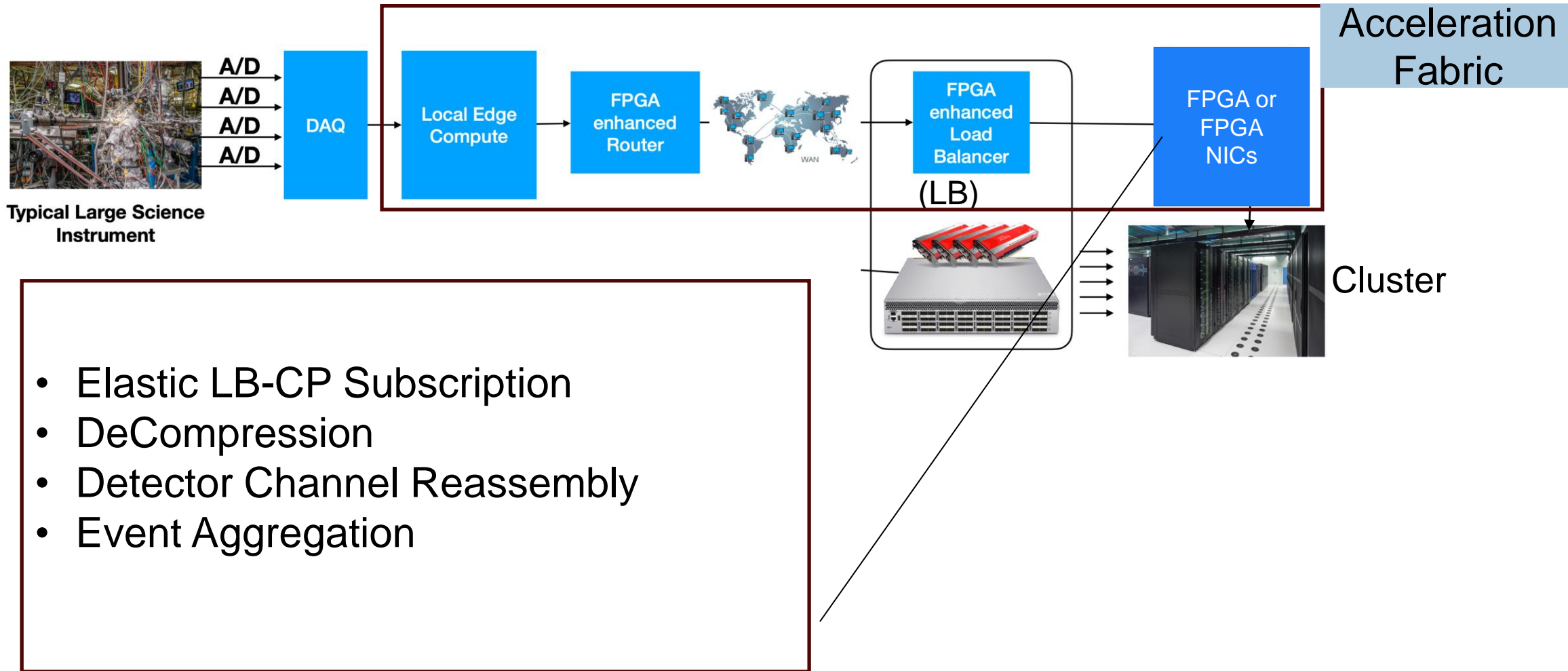


Node Event Rate = DP Schedule Occupancy

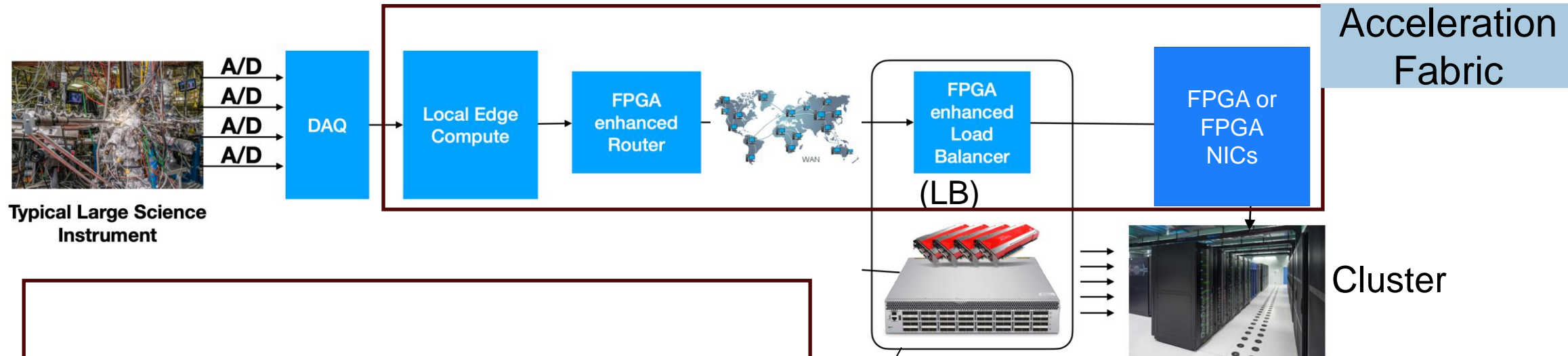
EJFAT: Data *Producer* Acceleration



EJFAT: Data Consumer Acceleration



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?