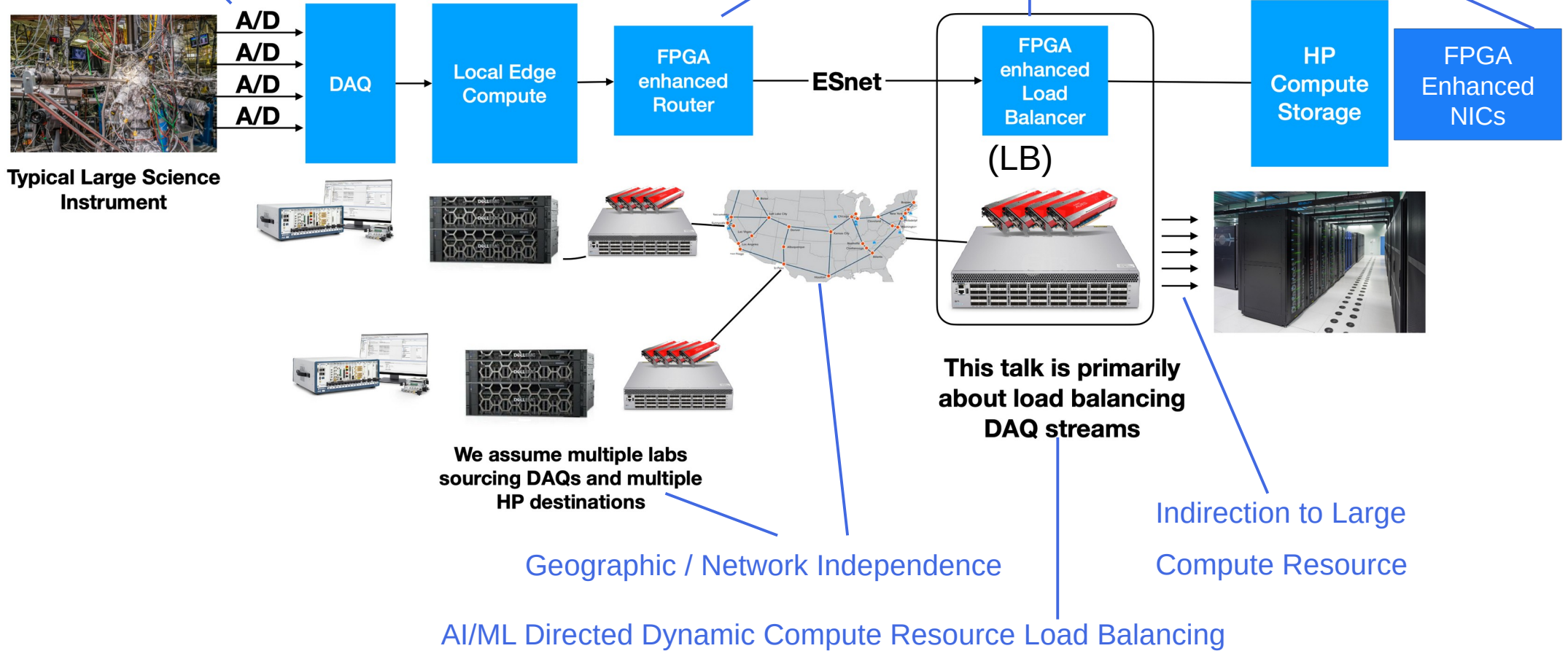**EJFAT**

ESnet-Jefferson Lab FPGA Accelerated Transport

Michael Goodrich , Carl Timmer, Vardan Gyurjyan,

David Lawrence , Graham Hayes (JLAB)

Yatish Kumar , Stacey Sheldon (ESnet)

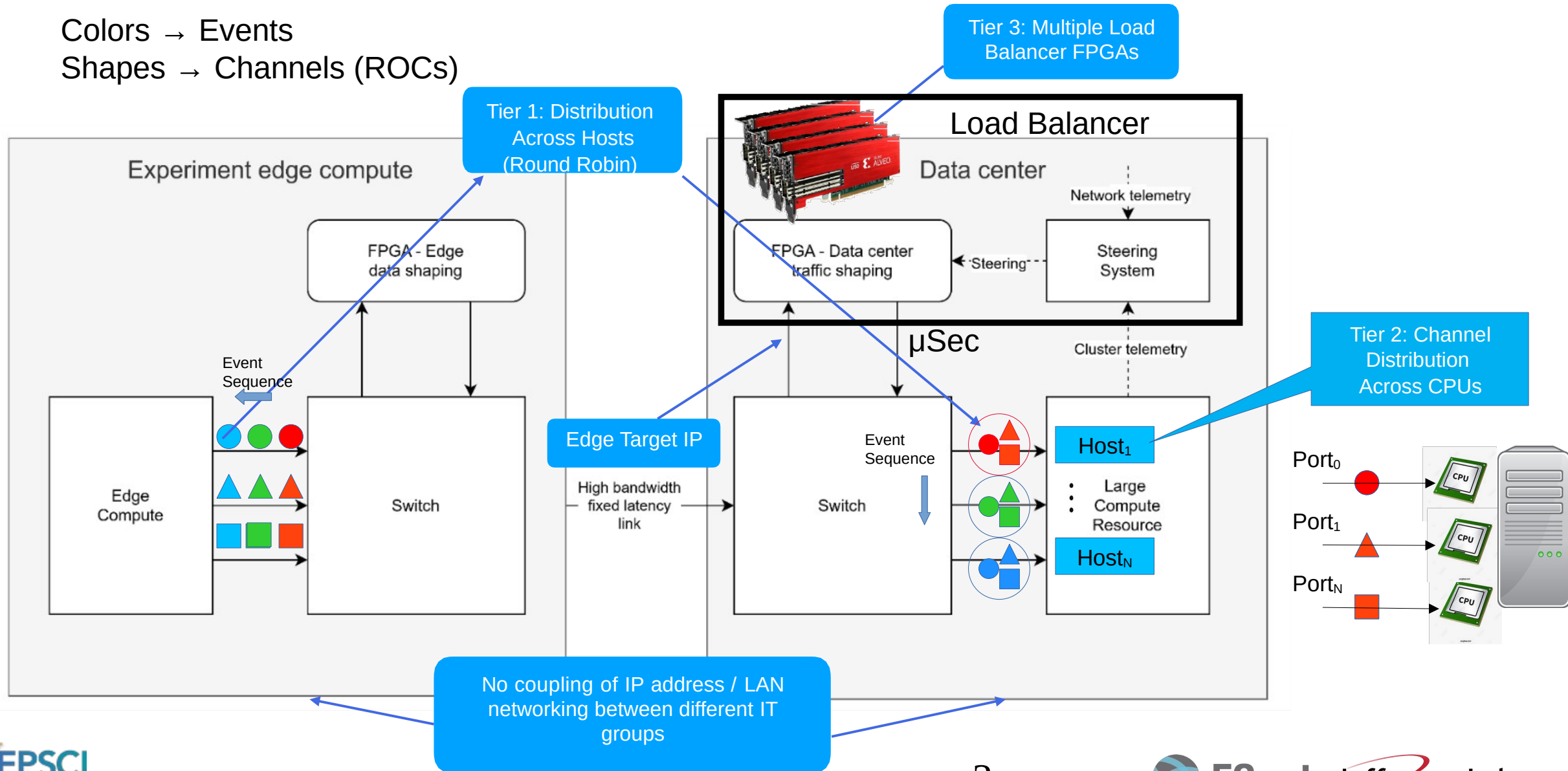# EJFAT = Edge to Core System Architecture: Workflows Steered by AI/ML

Loss-less UDP Streaming ~~Triggered~~

Opportunistic FPGA Based Network Acceleration

Typical Large Science Instrument

A/D
A/D
A/D
A/D

DAQ → Local Edge Compute → FPGA enhanced Router → ESnet → FPGA enhanced Load Balancer (LB) → HP Compute Storage → FPGA Enhanced NICs

This talk is primarily about load balancing DAQ streams

We assume multiple labs sourcing DAQs and multiple HP destinations

Indirection to Large Compute Resource

Geographic / Network Independence

AI/ML Directed Dynamic Compute Resource Load Balancing

EPSCI   ESnet Jefferson Lab

Channel Aggregation + Three Tier Horizontal Scaling

Colors → Events
Shapes → Channels (ROCs)

Tier 3: Multiple Load Balancer FPGAs

Tier 1: Distribution Across Hosts (Round Robin)

Load Balancer

Data center

Experiment edge compute

FPGA - Edge data shaping

Network telemetry

FPGA - Data center traffic shaping

Steering

Steering System

Event Sequence

µSec

Cluster telemetry

Tier 2: Channel Distribution Across CPUs

Edge Compute

Switch

High bandwidth - fixed latency link

Event Sequence

Switch

Host$_1$

Large Compute Resource

Host$_N$

Edge Target IP

Port$_0$

CPU

Port$_1$

CPU

Port$_N$

CPU

No coupling of IP address / LAN networking between different IT groups

EPSCI

ESnet

Jefferson Lab

LB: Xilinx U280 FPGA (PCIe) + Host

- **Data Plane** (DP): FPGA FW = RTL + P4
  - Packet Filtering, ARP, Ping
  - P4: Data Base for UDP Hdr Rewrites

- **Control Plane** (CP): Host
  - DP DB Maintenance
  - Monitor Network / Core Telemetry
  - AI/ML Steerage / Feedback
    - Upstream: Experiment / DAQ
    - Downstream: Core Computing
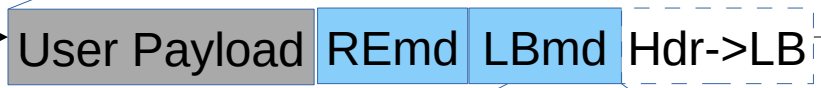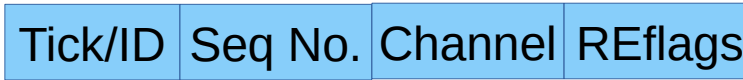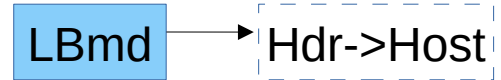   - Core Resource Provisioning

Load Balancer: User Interface
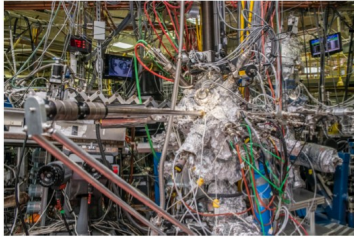
Data Source UDP Packet Payload

| User Payload | REmd | LBmd | Hdr->LB |

Typical Large Science Instrument

| Tick/ID | Channel | LBflags |

Load Balancer meta-data

| LBmd | → | Hdr->Host |

| User Payload | REmd | Hdr->Host |

| Tick/ID | Seq No. | Channel | REflags |

Reassembly meta-data

Compute Host

5

LB Data Plane (FPGA):   Network / Calendar Setup

Network Coordinates:

IPv4 IP	IPv6 IP	MAC	...

Control Plane
(FPGA Host)

Configure

Epoch: Contiguous Event ID Subspace

Epochs:

Epoch ID$_1$	Start ID	End ID
Epoch ID$_N$	Start ID	End ID

Configure

EPSCI

6

ESnet *Jefferson* Lab
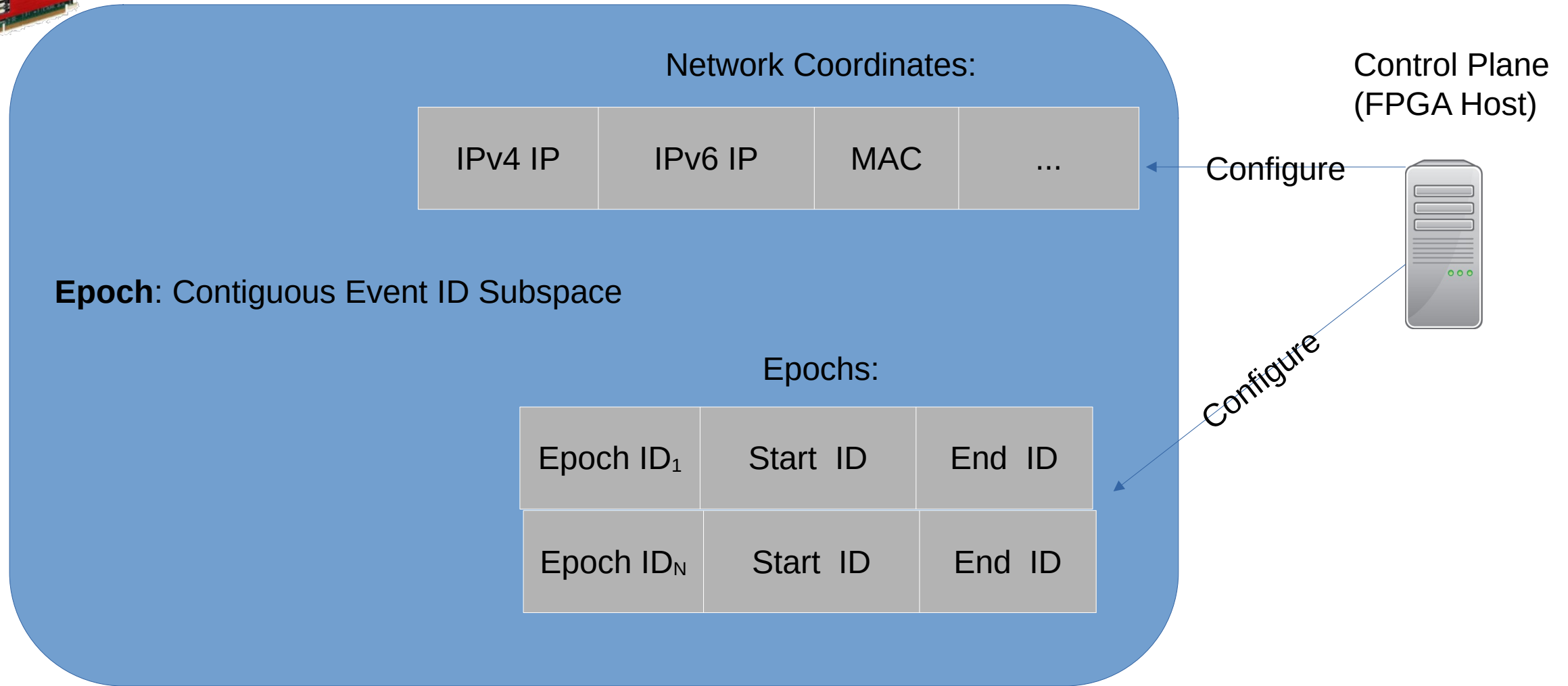
LB Data Plane (FPGA):   Destination Setup for Each Epoch

Event → Destination Index

Epoch ID	LSB_M (Event ID)	Destination Slot Index

Configure

Control Plane (FPGA Host)

Destination

IP	Port	MAC	Port Count

Configure

Epoch: Contiguous Event ID Subspace

EPSCI   ESnet   Jefferson Lab
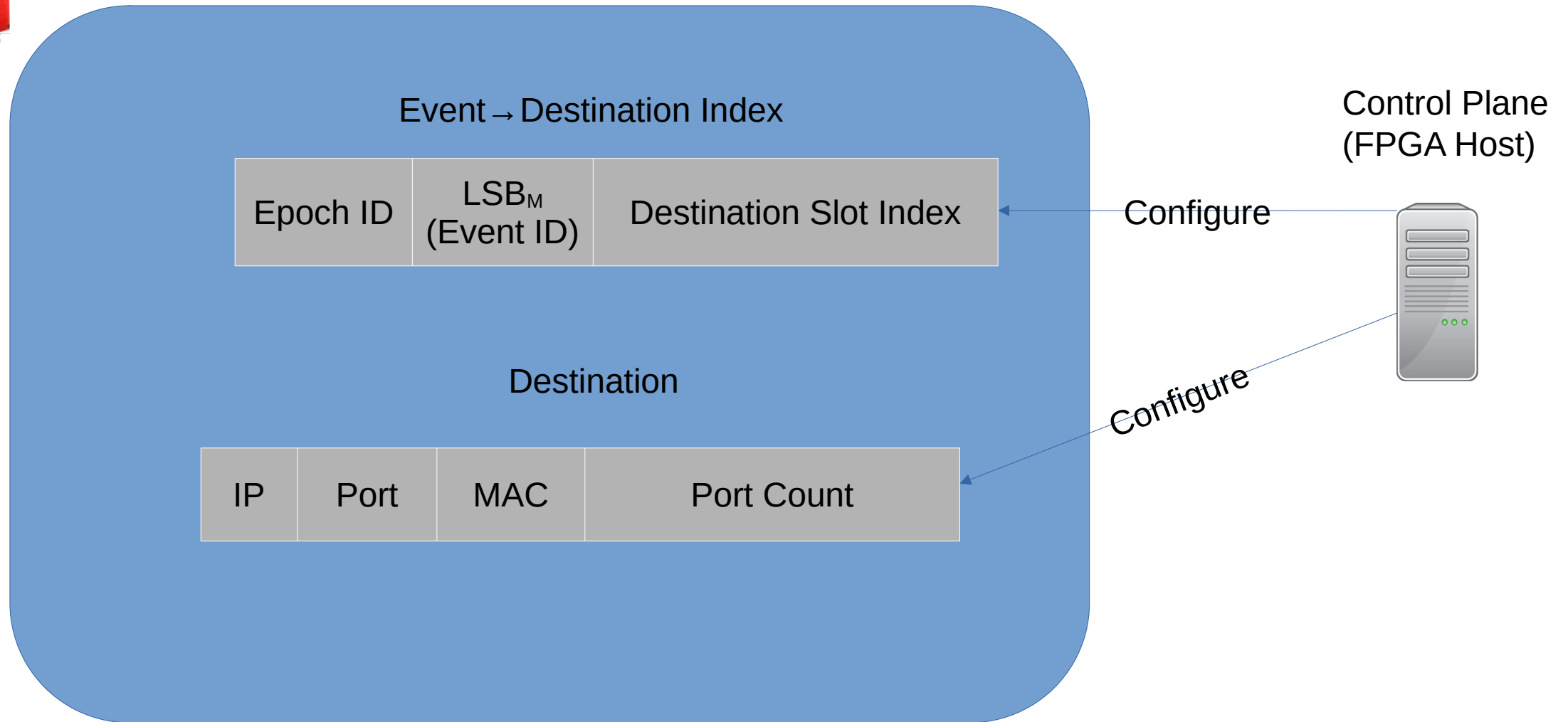
LB Data Plane (FPGA): Packet Processing
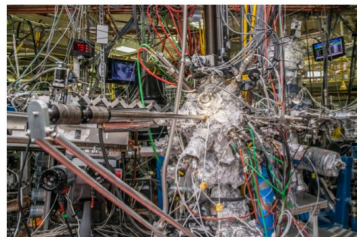
Typical Large Science Instrument

*Event ID = Tick

Event ID*

Epoch ID$_X$	Start ID	End ID

Epoch

Epoch ID	LSB$_M$ (Event ID)	Destination Slot Index

Destination Slot Index

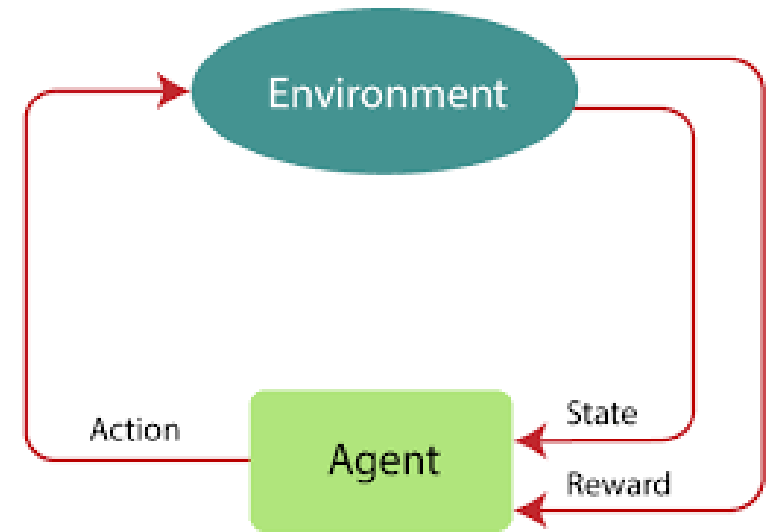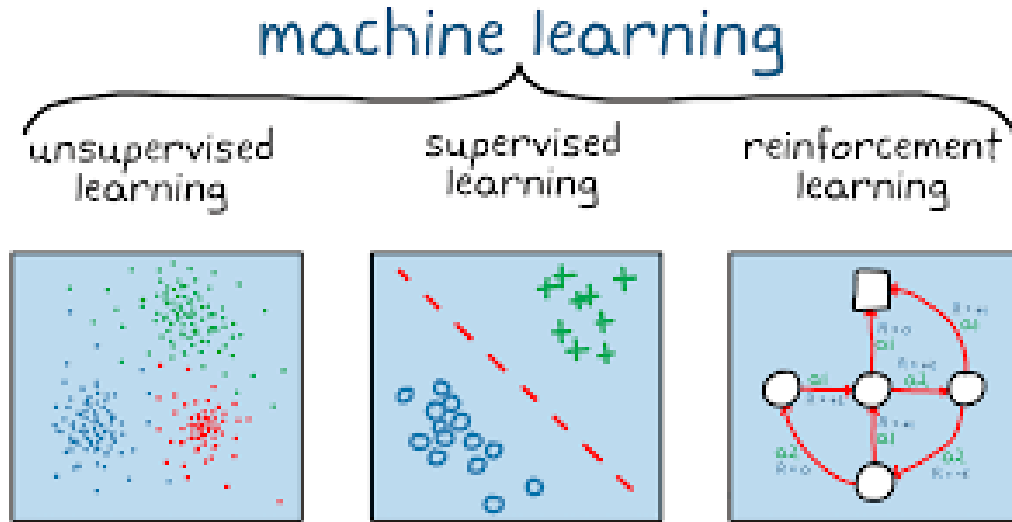| IP	Port	MAC	Port Count

New UDP Header

EPSCI

ESnet Jefferson Lab
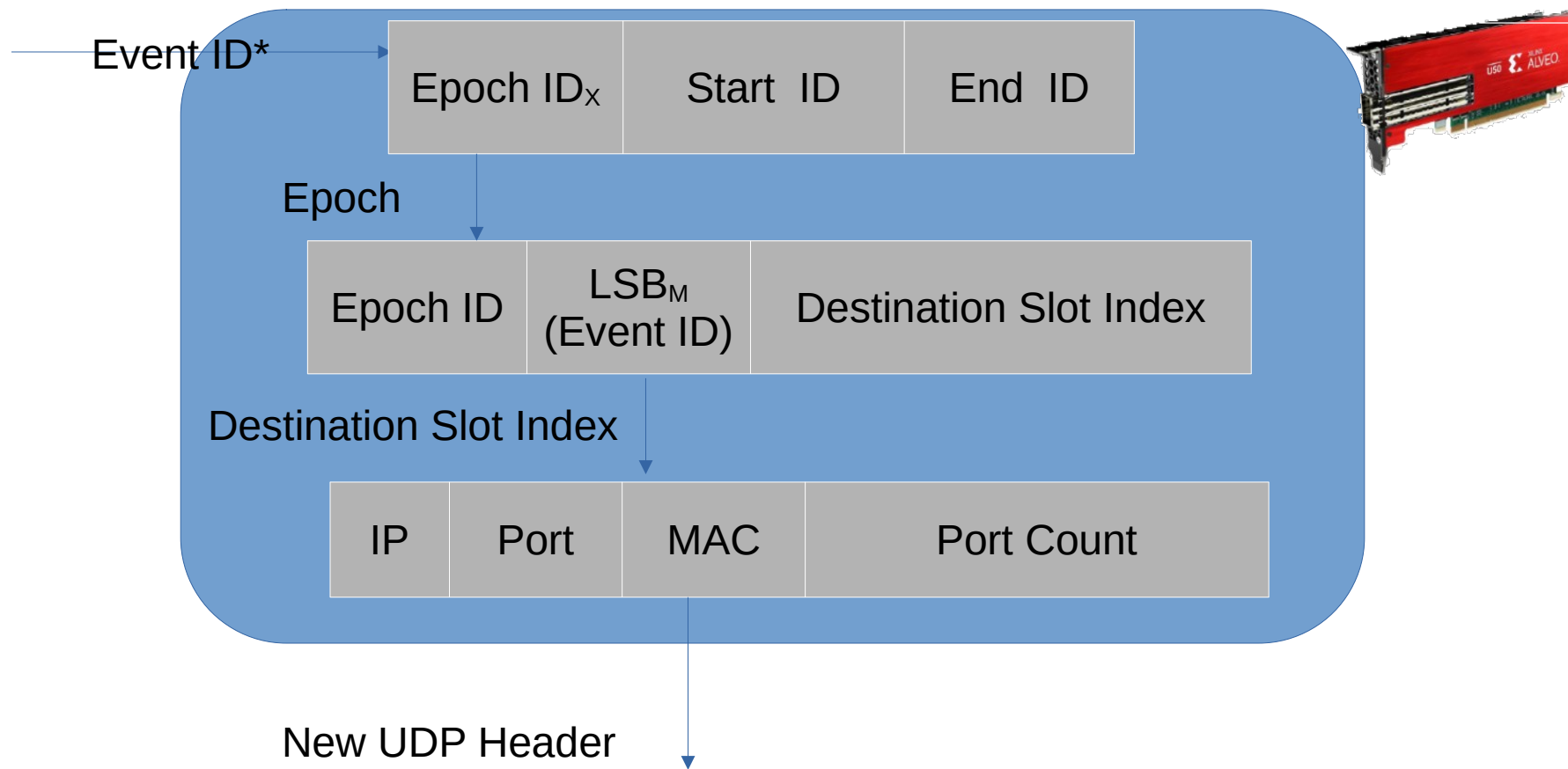
LB Control Plane (FPGA Host):
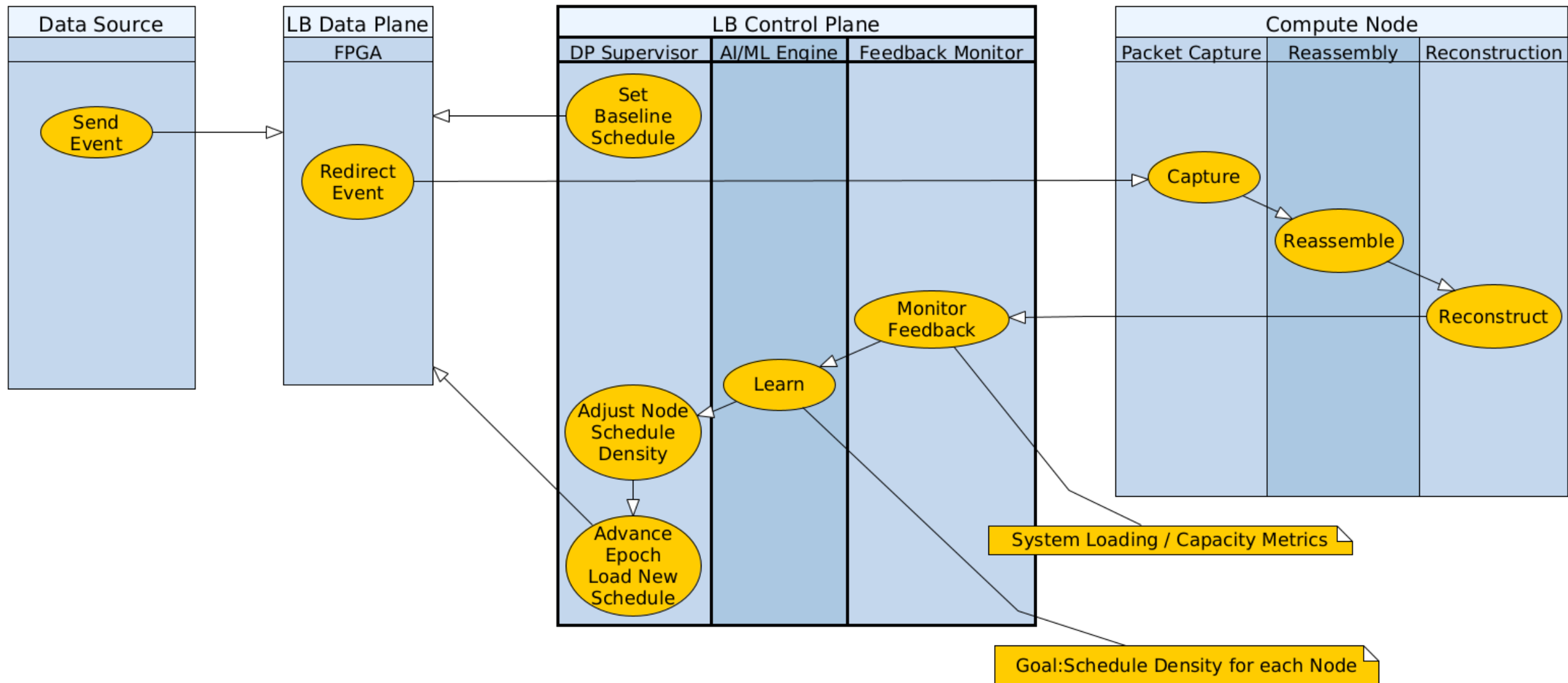
Reinforcement Learning

Q Learning:
- Many Variants
- Exploration / Exploitation
- Exploration → Learning
- Exploitation → Control
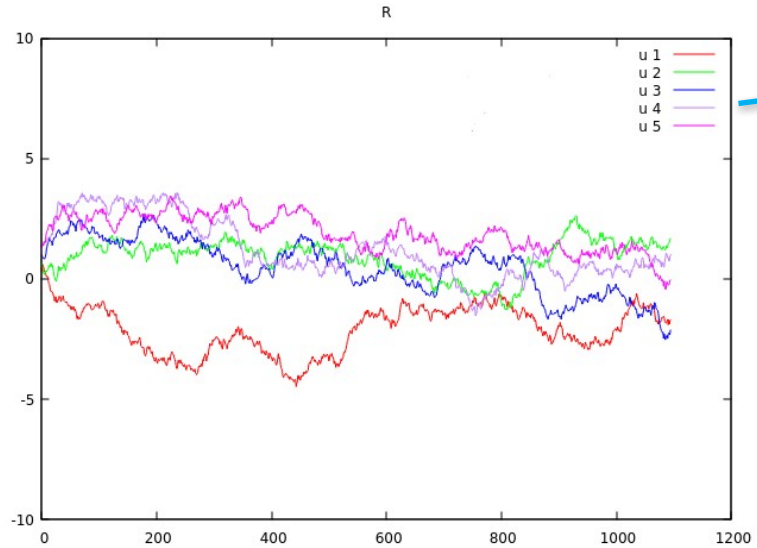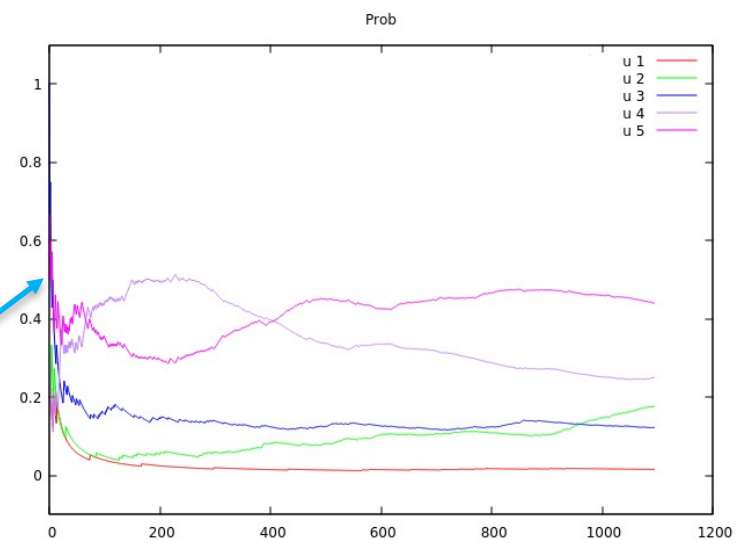
Reinforcement Learning -> Farm Scheduling
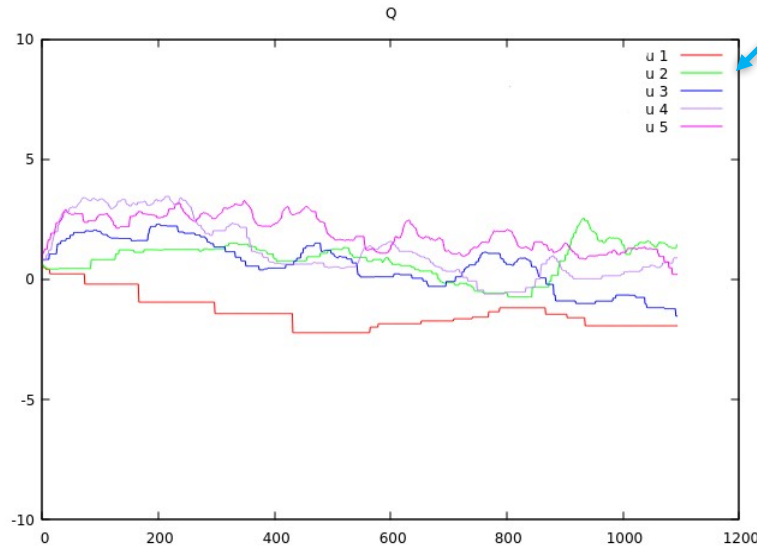

Rewards: Feedback from Nodes ->

Relative Availability for New Work

Q Learning: SARSA
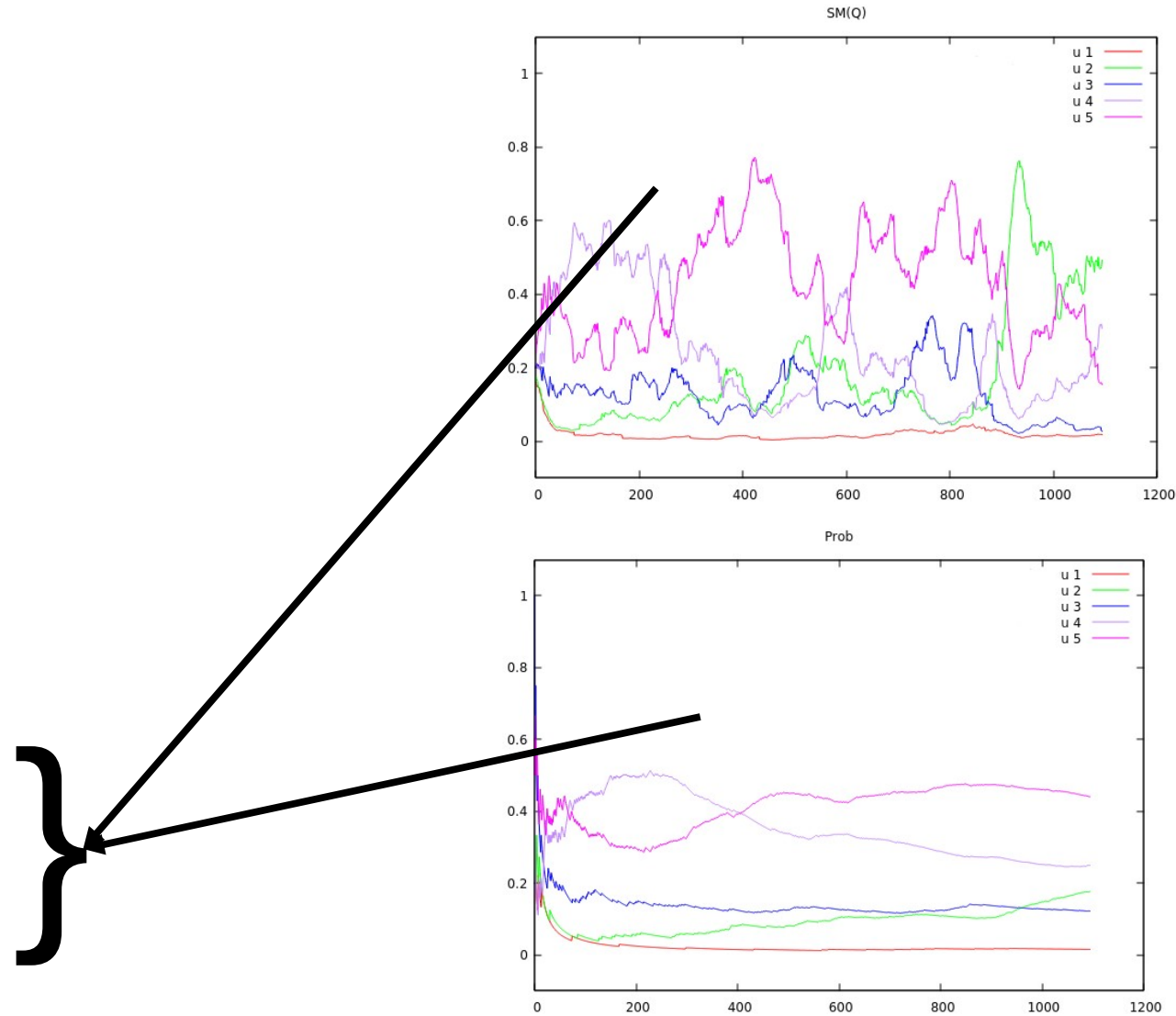
Action Selection(Q):

SoftMax Policy +
Weighted Sampling

LaPlace's Rule of Succession(Action) ->

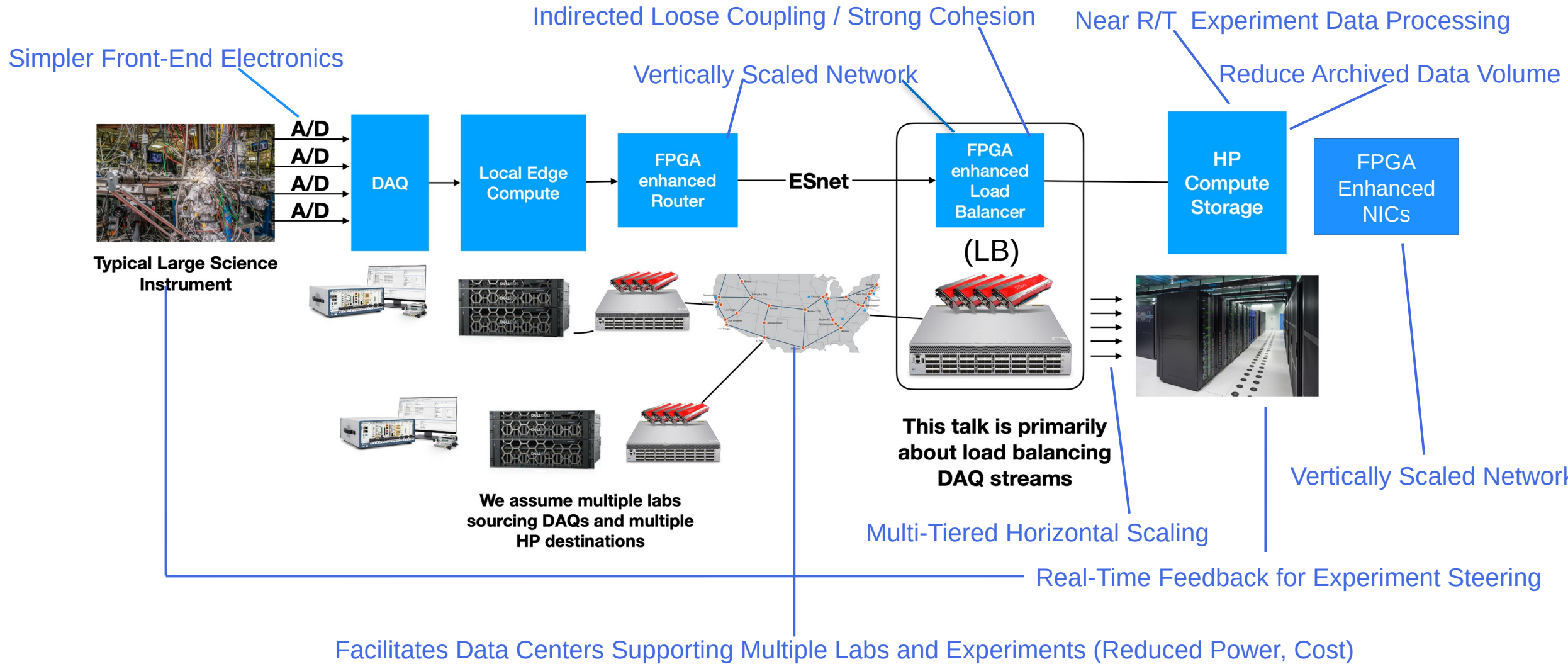Probability of Scheduling Each Node
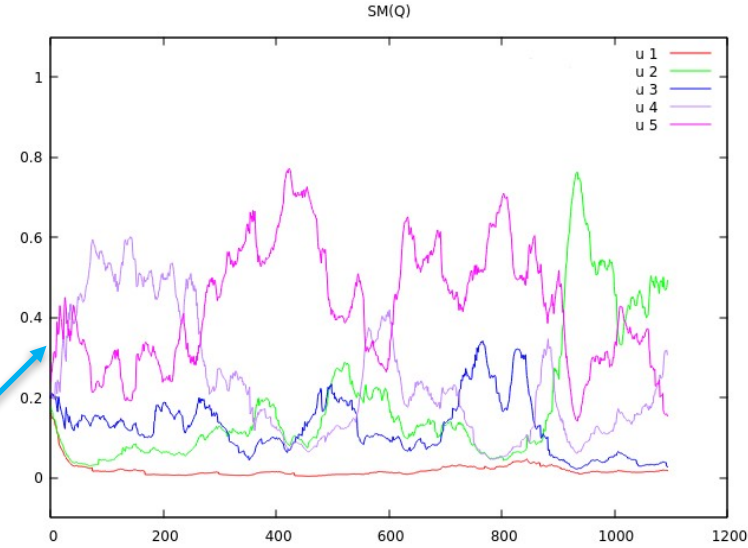
Q Learning -> LB Scheduling

Epoch ID	LSB$_M$ (Event ID)	Destination Slot Index
0	LSB$_M$ (Event ID)	Destination Slot Index
0		
0		
...		
1		
1		
1		

}

Benefits

Indirected Loose Coupling / Strong Cohesion

Near R/T Experiment Data Processing

Simpler Front-End Electronics

Reduce Archived Data Volume

Vertically Scaled Network

A/D
A/D
A/D
A/D

DAQ → Local Edge Compute → FPGA enhanced Router → ESnet → FPGA enhanced Load Balancer (LB) → HP Compute Storage → FPGA Enhanced NICs

Typical Large Science Instrument

This talk is primarily about load balancing DAQ streams

We assume multiple labs sourcing DAQs and multiple HP destinations

Vertically Scaled Network

Multi-Tiered Horizontal Scaling

Real-Time Feedback for Experiment Steering

Facilitates Data Centers Supporting Multiple Labs and Experiments (Reduced Power, Cost)

EPSCI

13

ESnet Jefferson Lab

EJFAT - Summary

- Supports Triggered + Streaming Detector, Data Center Workflows
- Data Event Aggregation
- Data Channel to Port Distribution
- Real-Time UDP Packet Re-Direction with Fixed μSec Latency
- Real-Time AI/ML Guided Destination Load Balancing
- Real-Time AI/ML Guided Cluster Resource Provisioning
- Decouples Edge and Cluster, Geographically and Network
- Running at 100Gbs – can support Up To 200Gbs

EPSCI

ESnet Jefferson Lab

Questions ?