

EJFAT

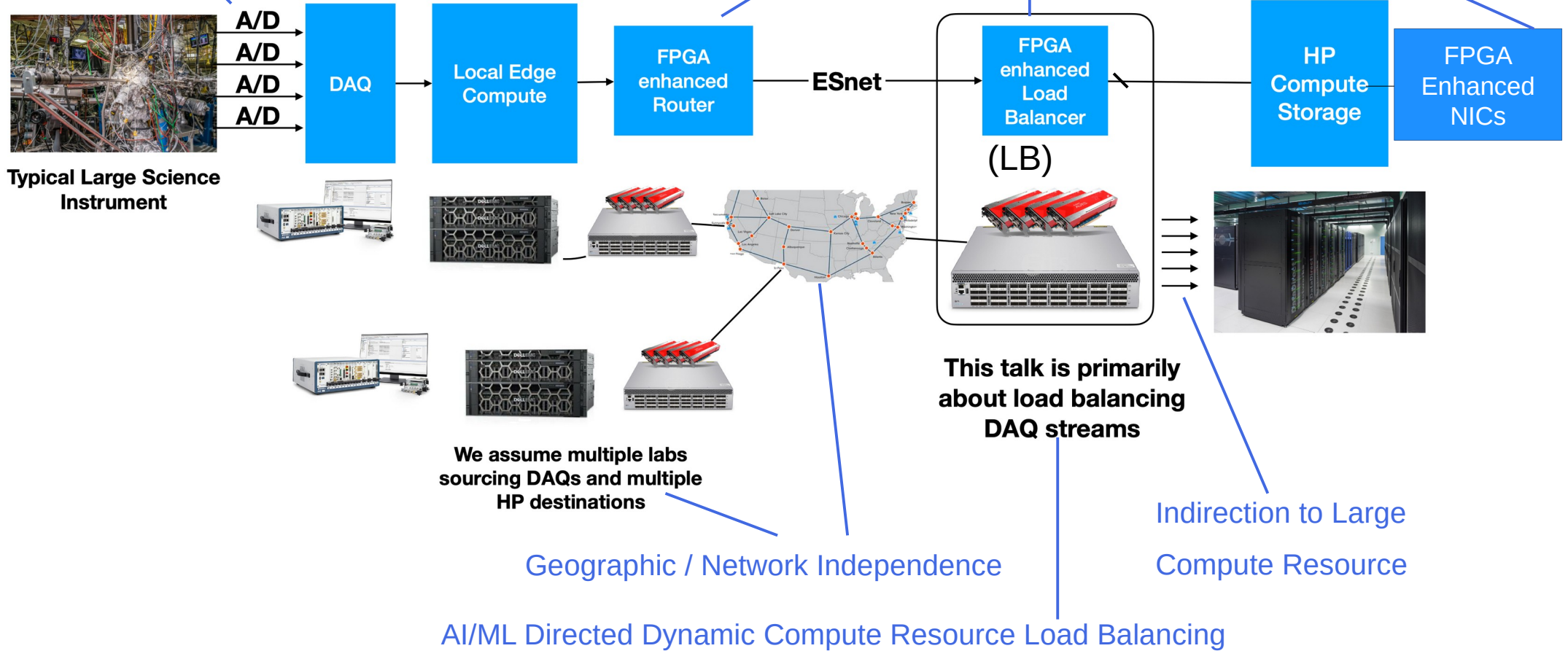
ESnet-Jefferson Lab FPGA Accelerated Transport

Michael Goodrich , Carl Timmer, Vardan Gyurjyan,
David Lawrence , Graham Hayes (JLAB)
Yatish Kumar , Stacey Sheldon (ESnet)

EJFAT = Edge to Core System Architecture to Support High Throughput Experiment Workflows Steered by AI/ML

Loss-less UDP Streaming Triggered

Opportunistic FPGA Based Network Acceleration



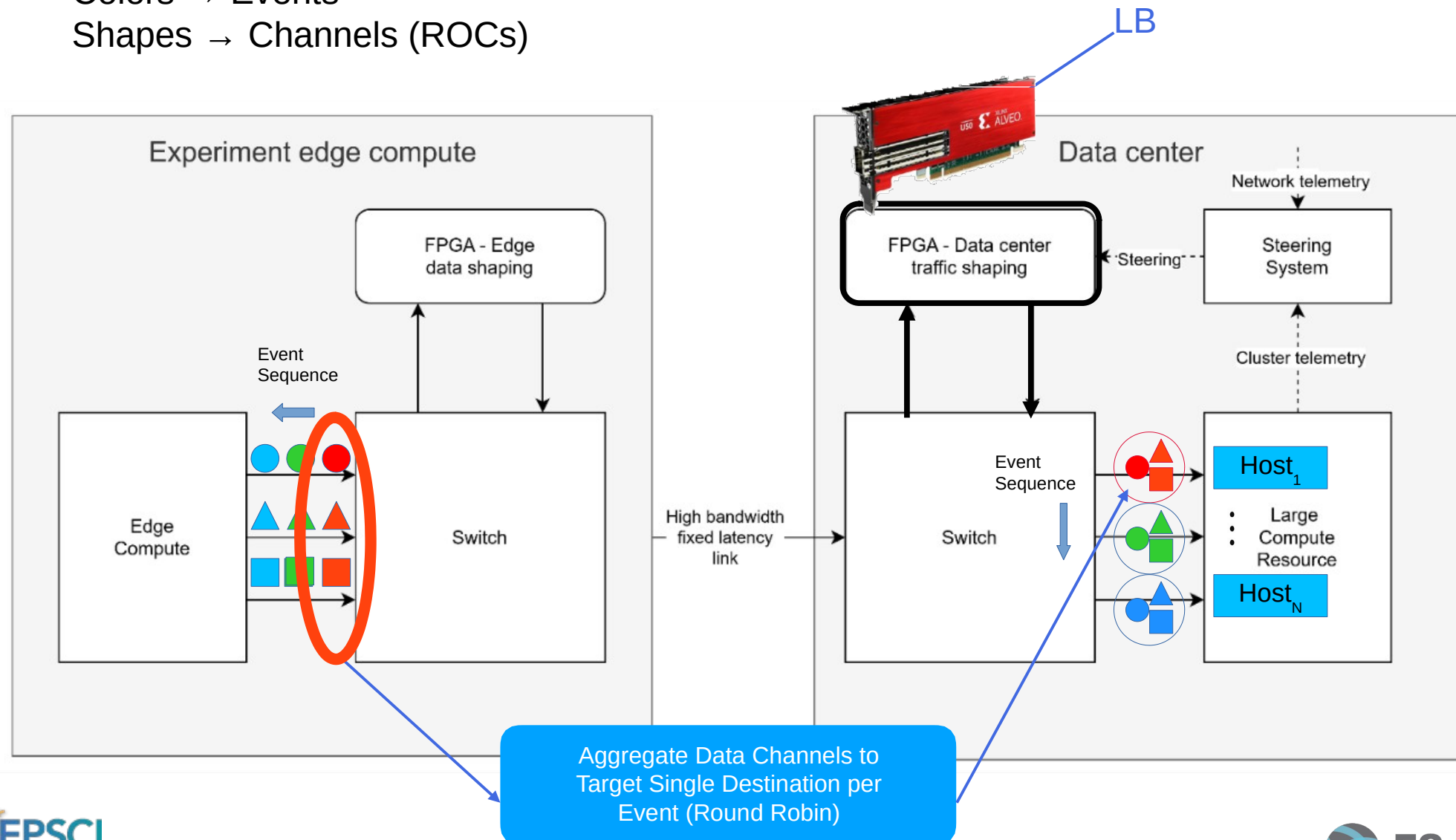
LB: Xilinx U280 FPGA (PCIe) + Host

- **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 Balancing = Data Channel Aggregation + Tiered Horizontal Scaling

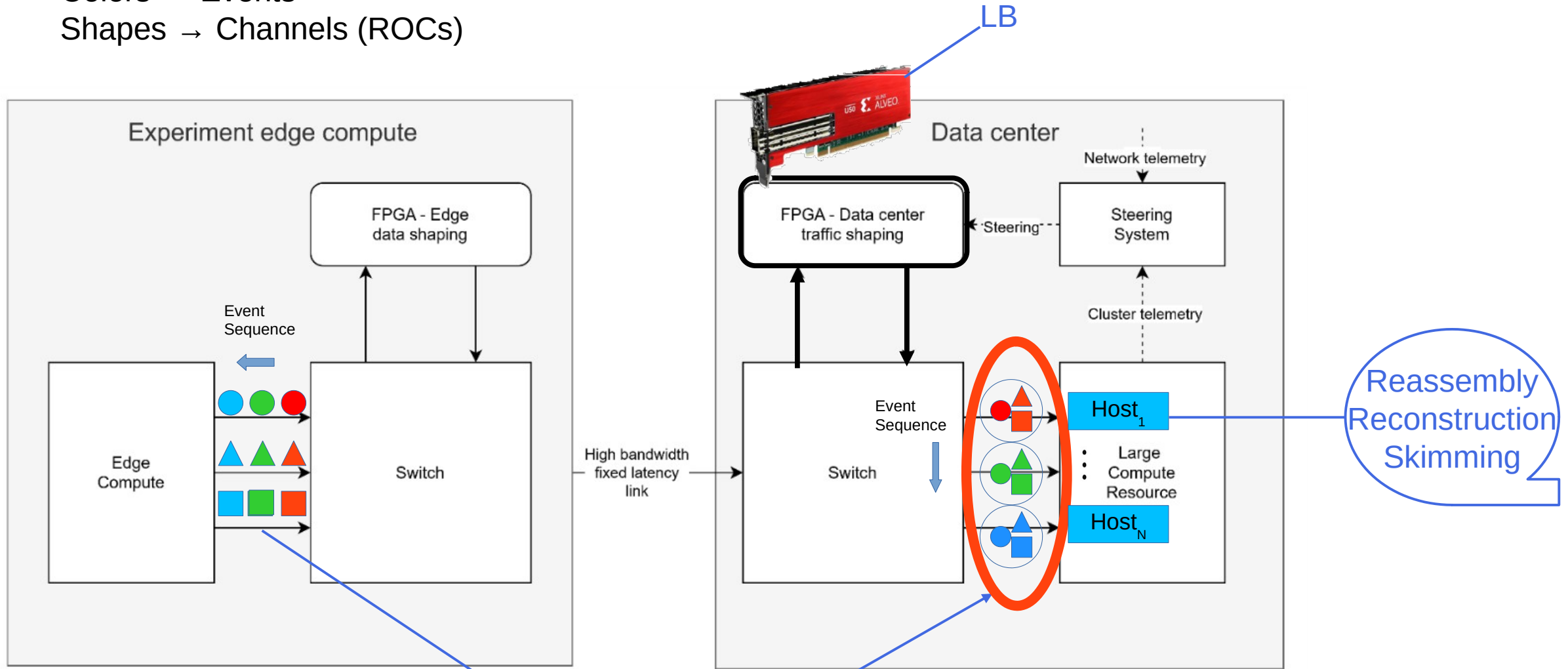
Colors → Events
Shapes → Channels (ROCs)



Tiered Horizontal Scaling: Tier-1 – Across Hosts

Colors → Events

Shapes → Channels (ROCs)

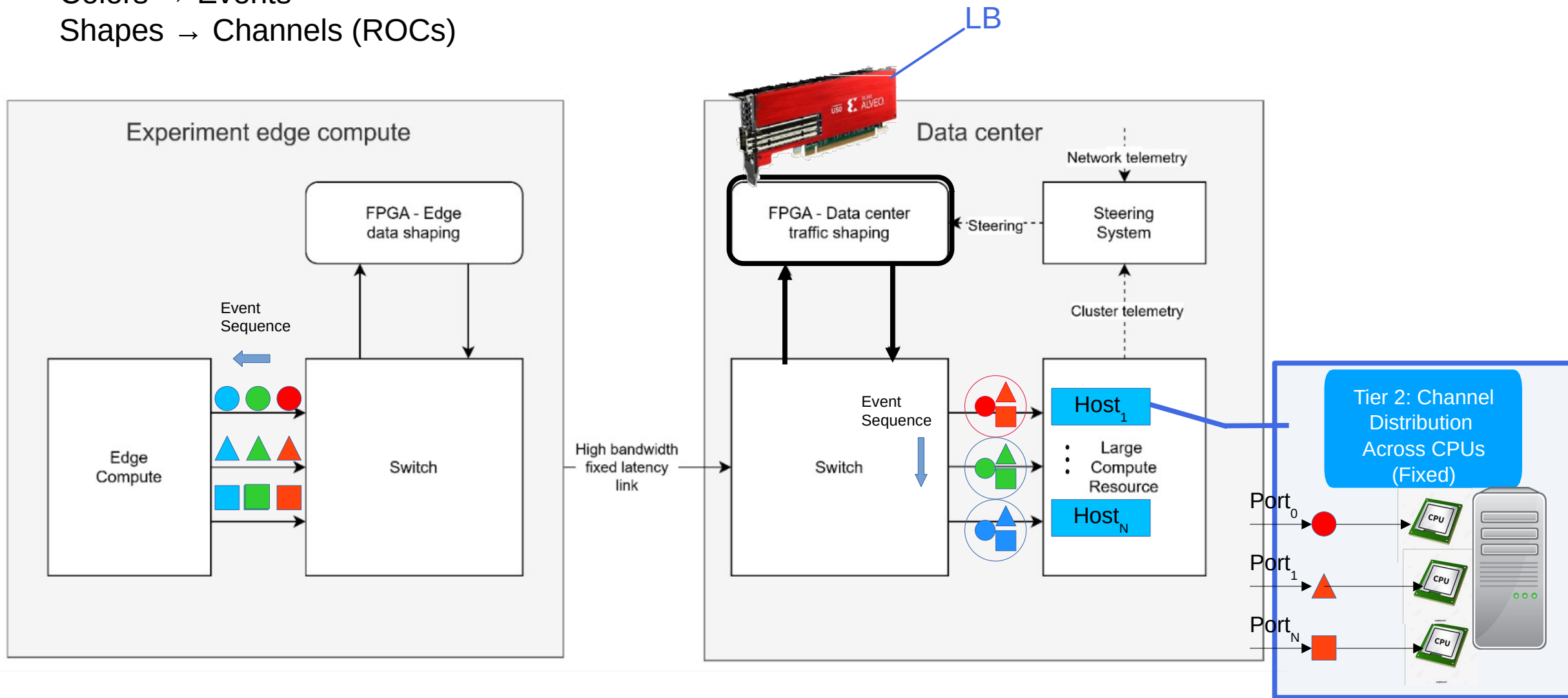


Tier 1: Distribution Across Hosts (Round Robin)



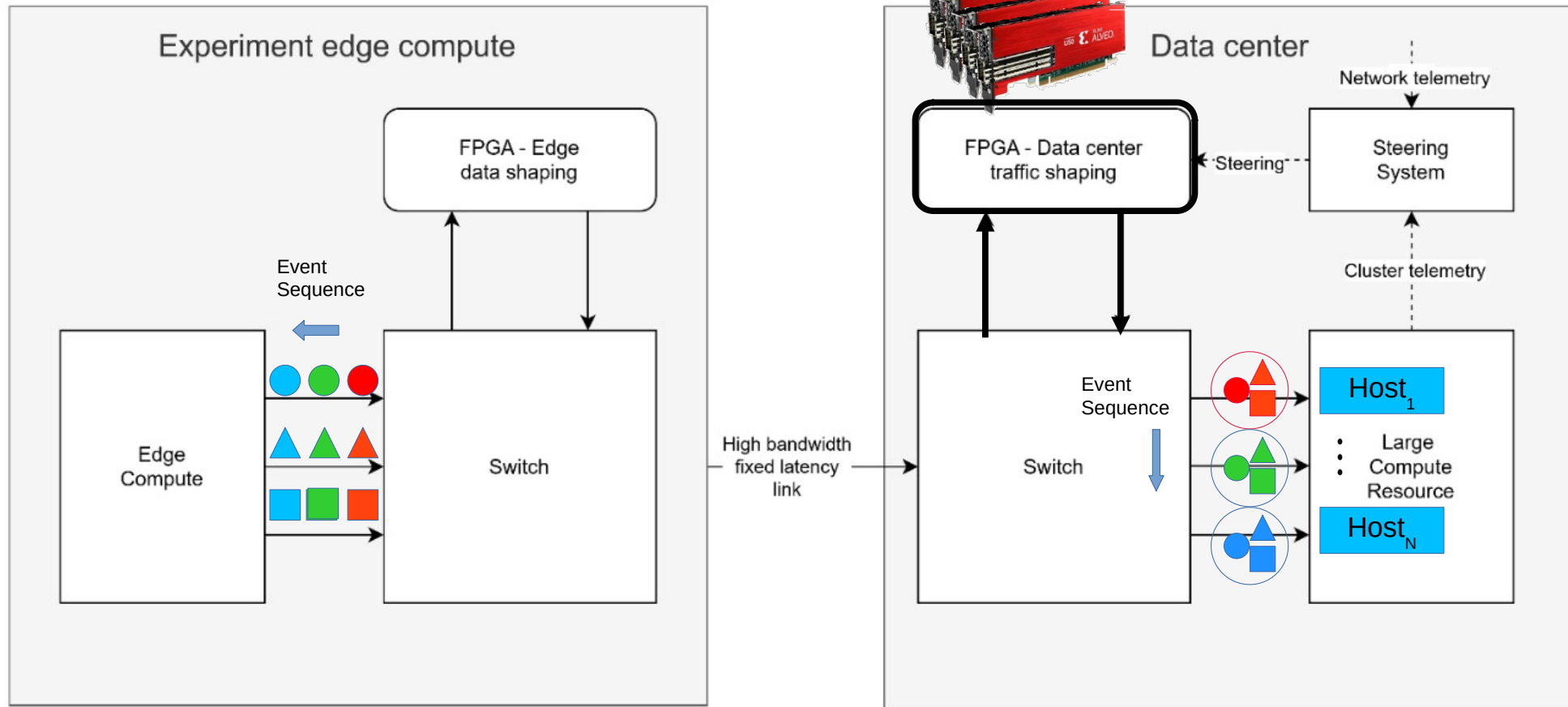
Tiered Horizontal Scaling: Tier-2 – Across Host CPUs via Ports

Colors → Events
Shapes → Channels (ROCs)



Tiered Horizontal Scaling: Tier-3 – Across LBs

Colors → Events
Shapes → Channels (ROCs)



Load Balancer: Data Plane

LB Control Plane

Epoch: Contiguous Event ID Subspace

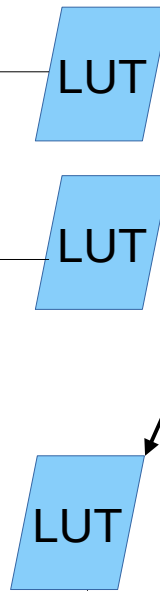
ID	Epoch
0	0
...	0
1000	0
1001	1
...	1
5000	1
5001	2
...	2
10000	2

Primary Function: Perform LU Functions to Re-write UDP Headers in Real-Time.

(Primary Keys in **bold**)

Epoch	ID (LSB*)	Core Slot
...

*Core Slot Round Robins over LSBs of ID for Each Epoch



Network / Compute Telemetry



Primary Concern: Maintain / Use Epoch Advance to Dynamically and Predictively Progress Event ID to Core Mappings for Changing Conditions

Core Host

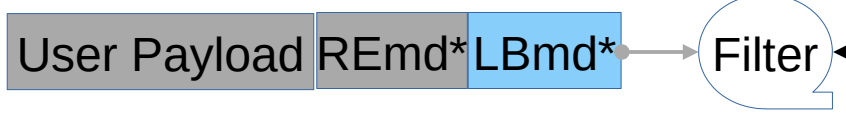
Provide Telemetry on Availability / Capacity to LB Control Plane

IP Type	Core Slot	Host IP	Host MAC	Base Port	• Port Entropy Bits
...

Load Balancer: Data Plane

Control Plane Host

Incoming UDP Packet Payload



LB IP, MAC,...

Set LB Network Coordinates

*Meta Data

- Prepared by LB Application
- Pre-pended to User Payload
- LBmd = Load Balancer - use case independent
 - Event ID (Timestamp)
 - Channel ID
- REmd = Reassembly – Use Case Defined
 - Event ID (Timestamp)
 - Channel ID
 - Sequence Number

Uses LBmd for Event to Core/Port Mapping

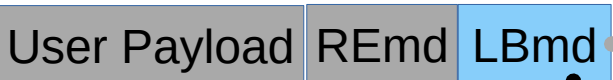
Core Host

Use REmd for Reassembly

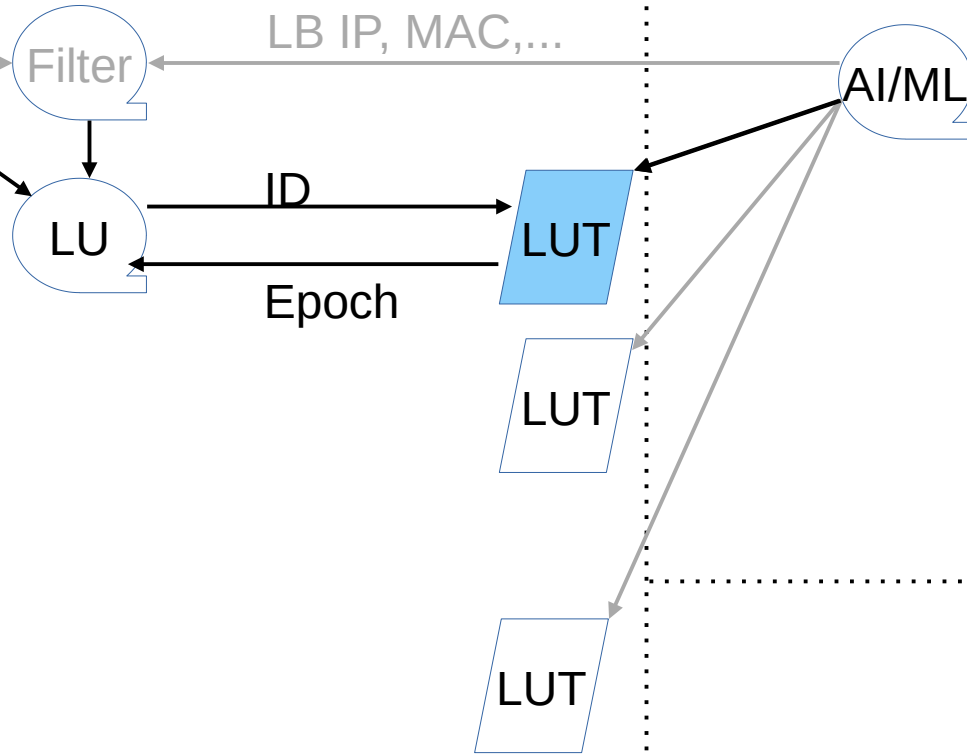
Load Balancer: Data Plane

Control Plane Host

Incoming UDP Packet Payload



Event ID: Timestamp
or Contiguous One-Up
Counter: Monotonic +
Non-Repeating



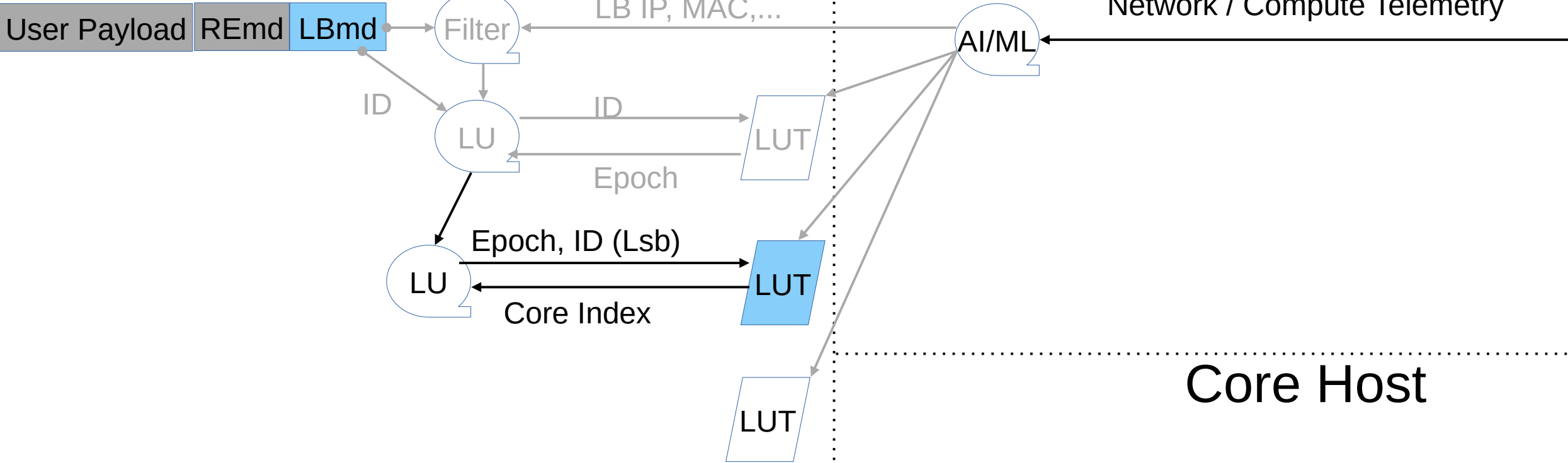
Network / Compute Telemetry

Core Host

Load Balancer: Data Plane

Control Plane Host

Incoming UDP Packet Payload

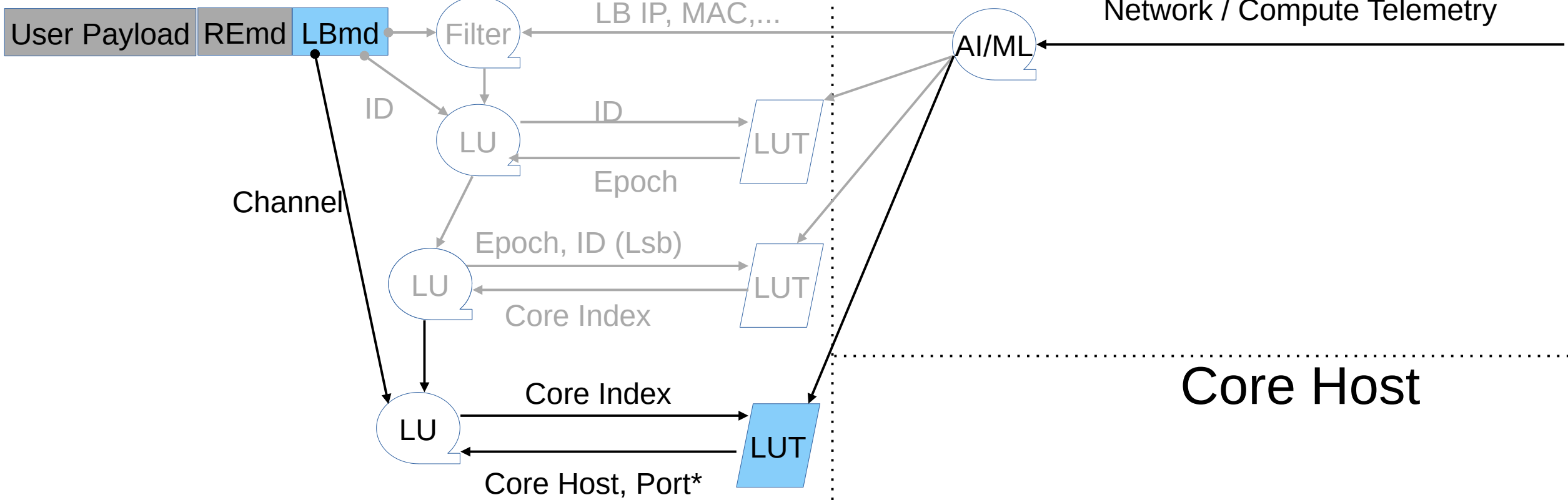


Core Host

Load Balancer: Data Plane

Control Plane Host

Incoming UDP Packet Payload

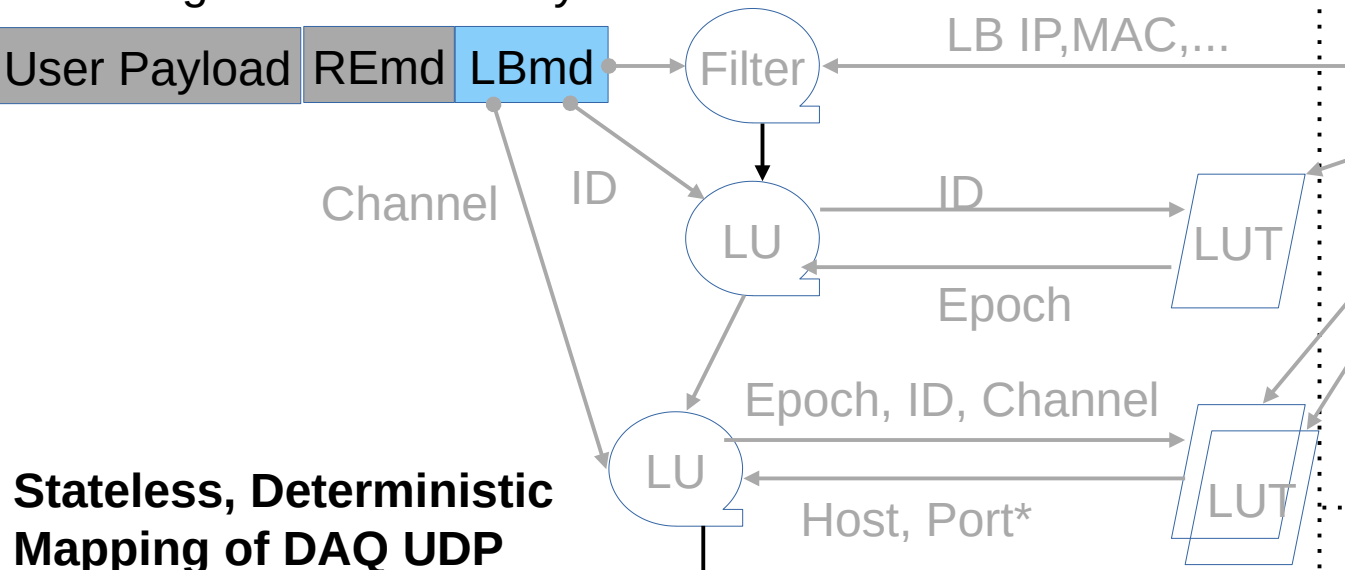
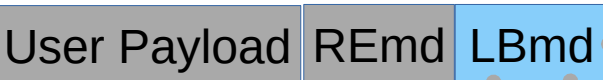


*Port = $function(\text{Base Port, Port Entropy Bits, Channel})$

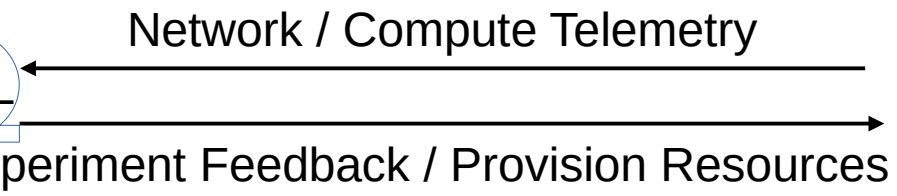
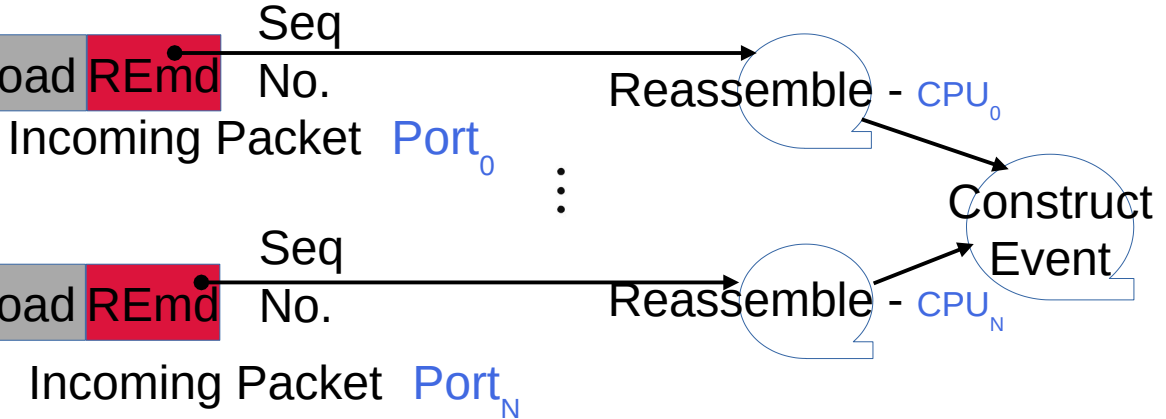
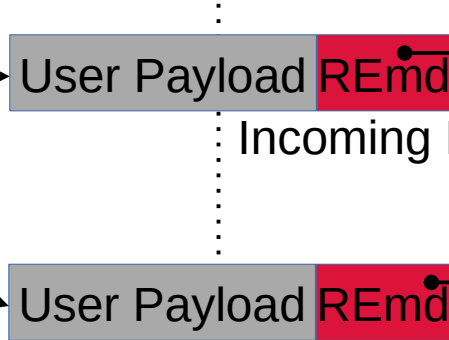
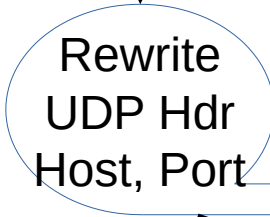
Load Balancer: Data Plane

Control Plane Host

Incoming UDP Packet Payload



Stateless, Deterministic Mapping of DAQ UDP Packets to Core Host Ports (CPUs) with uSec Latency, nSec Interval



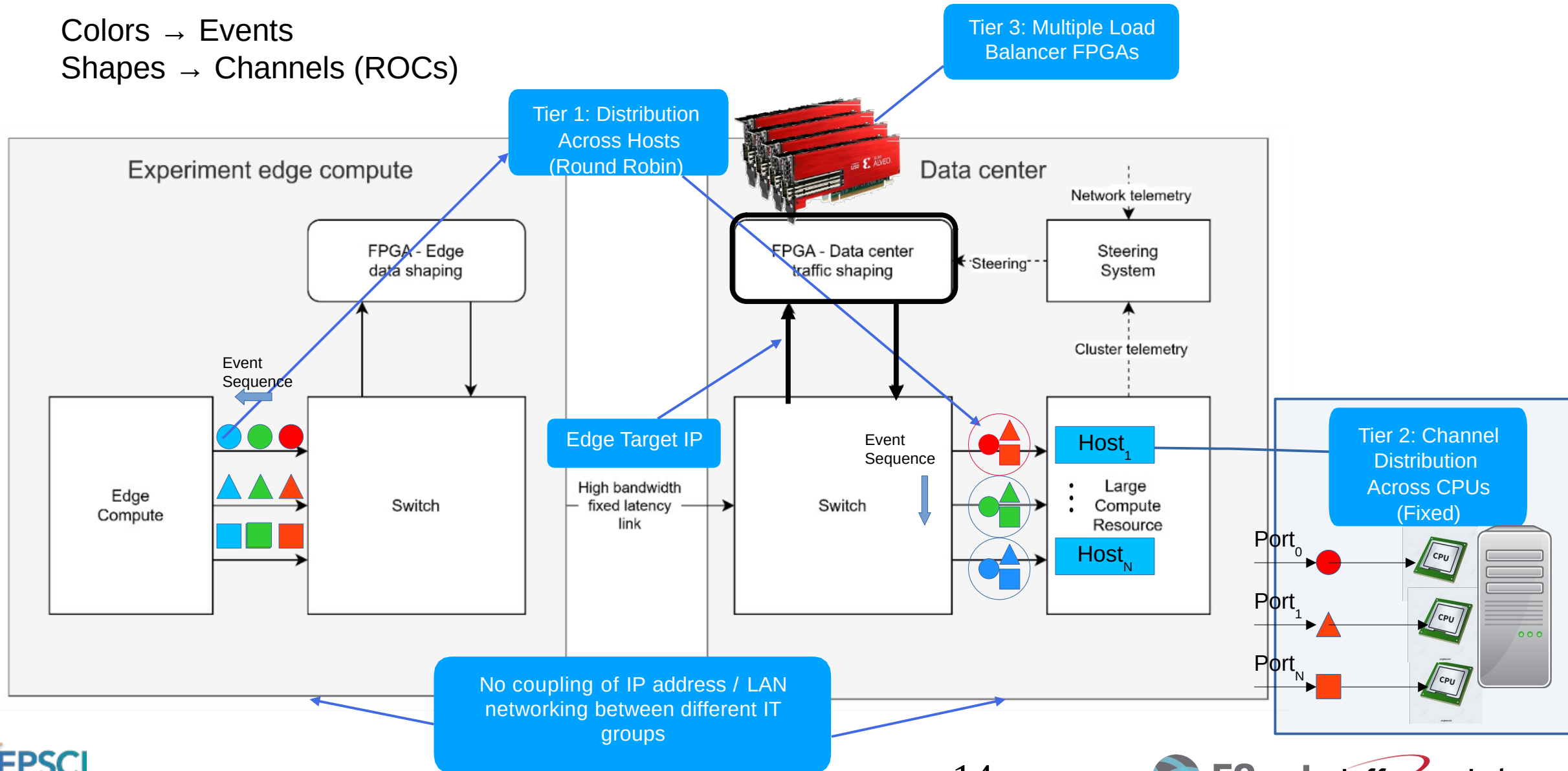
Dynamically Maintain Packet Rewrite Coherence in LB LUTs

Core Host - Parallel Reassembly



Summary : Channel Aggregation + Three Tier Horizontal Scaling

Colors → Events
Shapes → Channels (ROCs)



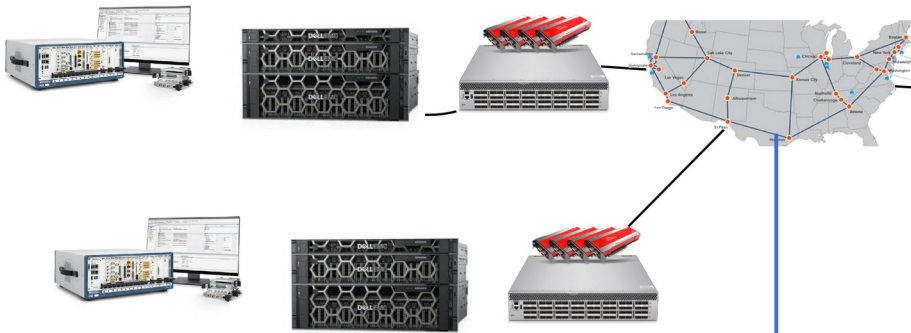
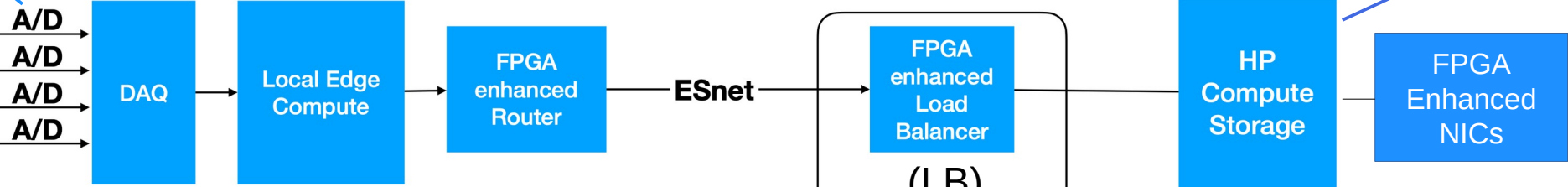
Benefits

Simpler Front End Electronics

Near R/T Experiment Data Processing
Reduce Archived Data Volume



Typical Large Science Instrument



We assume multiple labs sourcing DAQs and multiple HP destinations

This talk is primarily about load balancing DAQ streams

Real-Time Feedback for Experiment Steering

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

Questions ?