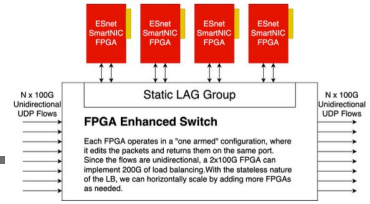
ESnet
ENERGY SCIENCES NETWORK

# EJ-FAT Real Time Streaming

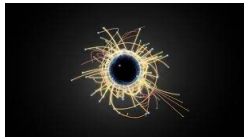
**Yatish Kumar**
ESnet


# Large Scientific Instruments

wan




wan

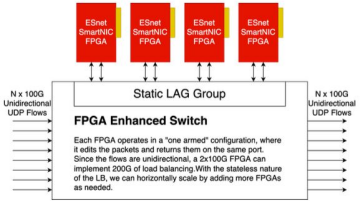
Typical DOE Accelerator

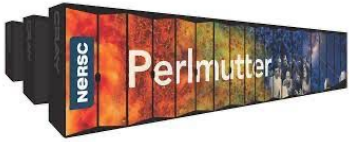
ESnet hosted EF-FAT load balancer

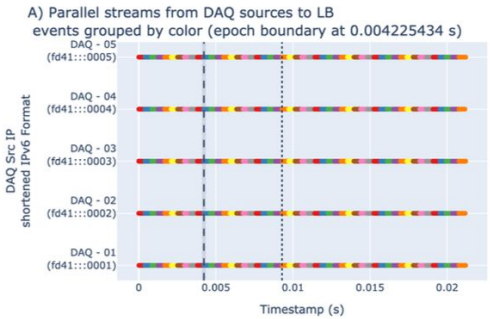
Meaning of Life

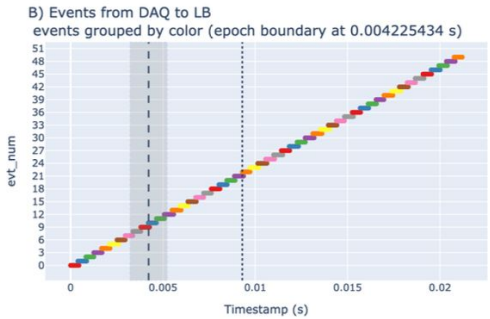
1. Unidirectional streaming @400Gbps ( No TCP / retransmissions )
2. The source does not know the dest IP address
   N sources need to be mapped to 1000 or more compute nodes
3. Nodes can appear and disappear while the accelerator is running.
   We need to juggle the balls so that not a single packet gets dropped.

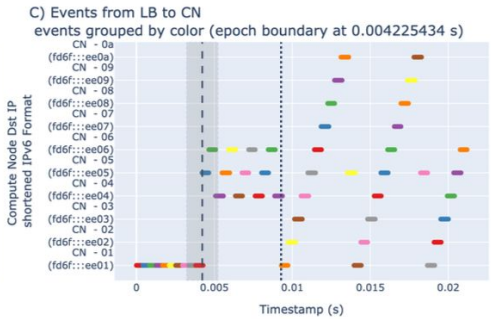
ESnet

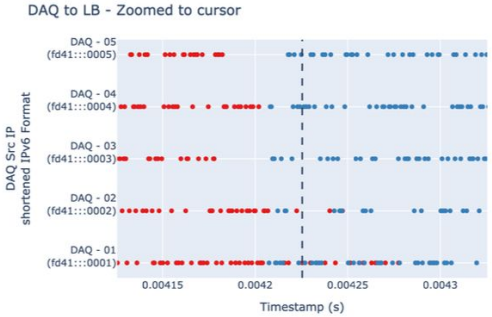

ESnet SmartNIC FPGA
ESnet SmartNIC FPGA
ESnet SmartNIC FPGA
ESnet SmartNIC FPGA
Static LAG Group
N x 100G Unidirectional UDP Flows
N x 100G Unidirectional UDP Flows
FPGA Enhanced Switch
Each FPGA operates in a "one armed" configuration, where it edits the packets and returns them on the same port. Since the flows are unidirectional, a 2x100G FPGA can implement 200G of load balancing.With the stateless nature of the LB, we can horizontally scale by adding more FPGAs as needed.

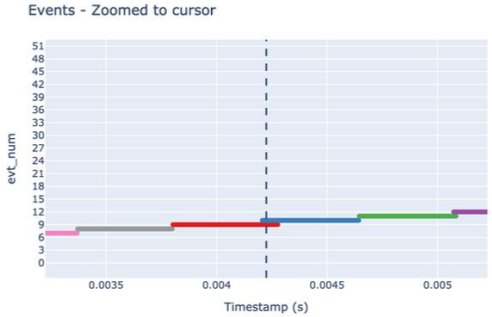

NERSC
Perlmutter

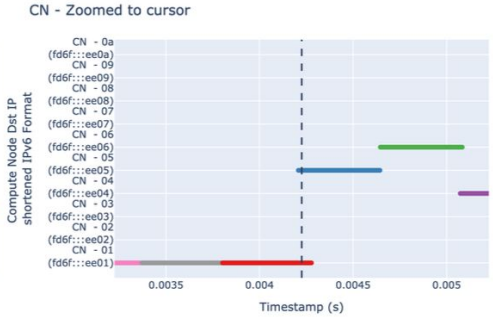

A) Parallel streams from DAQ sources to LB
events grouped by color (epoch boundary at 0.004225434 s)
DAQ Src IP shortened IPv6 Format
DAQ - 05 (fd41:::0005)
DAQ - 04 (fd41:::0004)
DAQ - 03 (fd41:::0003)
DAQ - 02 (fd41:::0002)
DAQ - 01 (fd41:::0001)
0
0.005
0.01
0.015
0.02
Timestamp (s)


B) Events from DAQ to LB
events grouped by color (epoch boundary at 0.004225434 s)
evt_num
51
48
45
42
39
36
33
30
27
24
21
18
15
12
9
6
3
0
0
0.005
0.01
0.015
0.02
Timestamp (s)


C) Events from LB to CN
events grouped by color (epoch boundary at 0.004225434 s)
Compute Node Dst IP shortened IPv6 Format
CN - 0a (fd6f:::ee0a)
CN - 09 (fd6f:::ee09)
CN - 08 (fd6f:::ee08)
CN - 07 (fd6f:::ee07)
CN - 06 (fd6f:::ee06)
CN - 05 (fd6f:::ee05)
CN - 04 (fd6f:::ee04)
CN - 03 (fd6f:::ee03)
CN - 02 (fd6f:::ee02)
CN - 01 (fd6f:::ee01)
0
0.005
0.01
0.015
0.02
Timestamp (s)


DAQ to LB - Zoomed to cursor
DAQ Src IP shortened IPv6 Format
DAQ - 05 (fd41:::0005)
DAQ - 04 (fd41:::0004)
DAQ - 03 (fd41:::0003)
DAQ - 02 (fd41:::0002)
DAQ - 01 (fd41:::0001)
0.00415
0.0042
0.00425
0.0043
Timestamp (s)


Events - Zoomed to cursor
evt_num
51
48
45
42
39
36
33
30
27
24
21
18
15
12
9
6
3
0
0.0035
0.004
0.0045
0.005
Timestamp (s)


CN - Zoomed to cursor
Compute Node Dst IP shortened IPv6 Format
CN - 0a (fd6f:::ee0a)
CN - 09 (fd6f:::ee09)
CN - 08 (fd6f:::ee08)
CN - 07 (fd6f:::ee07)
CN - 06 (fd6f:::ee06)
CN - 05 (fd6f:::ee05)
CN - 04 (fd6f:::ee04)
CN - 03 (fd6f:::ee03)
CN - 02 (fd6f:::ee02)
CN - 01 (fd6f:::ee01)
0.0035
0.004
0.0045
0.005
Timestamp (s)

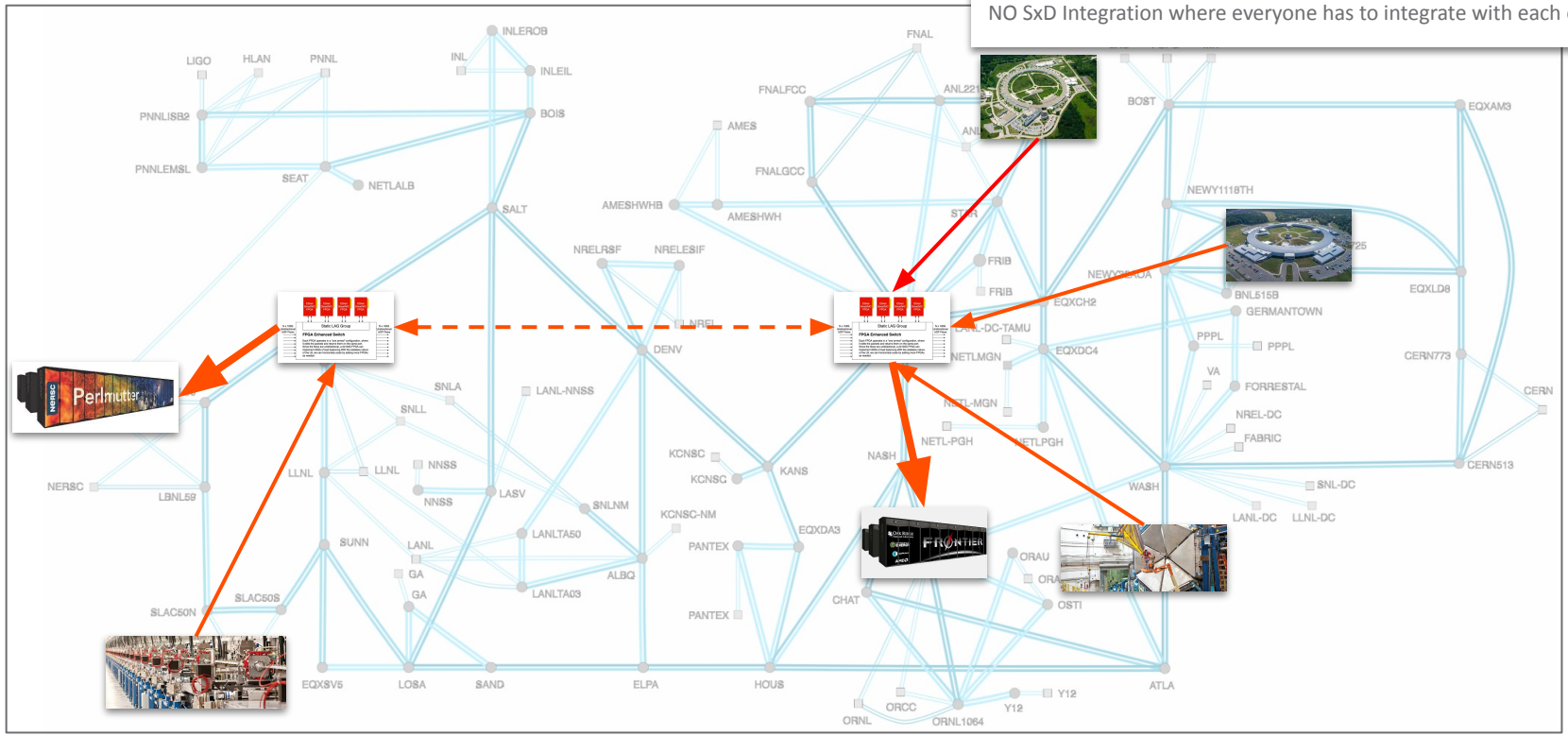
ESnet

# Benefit of ESnet host the LB ?

There are lots of sites. Every national lab. NO ONE needs to buy and operate any FPGA based funky thing.

SOURCES: "just send all your packets to 1 or 2 well known ESnet LB addresses"
SINKS: "just register with ESnet when a CN is ready, and we can send work from ANY source"

NO SxD Integration where everyone has to integrate with each other.



# Summary - Real Time Streaming as a network service

1. Provide Location and ID separation. By dynamically mapping each packet to a final destination based on feedback from 1000's of individual compute nodes.
2. Operate on a unidirectional UDP protocol. It isn't elephant flows, it isn't mouse flows, it's a swarm of locusts.
   - does not care about packet ordering.
     We can scale to Nx400G parallel links with no impact on reassembly or forwarding errors.
3. Eliminate the Source Site to Dest Site. [SxD] matrix of integration problems.
   - sources just need to integrate with ESnet to access any compute resource
   - destinations just need to integrate with ESnet to serve any source
4. ESnet can upgrade the capacity using L1 optical / L2 VPN / L3 whatever makes sense as demand increases

ESnet