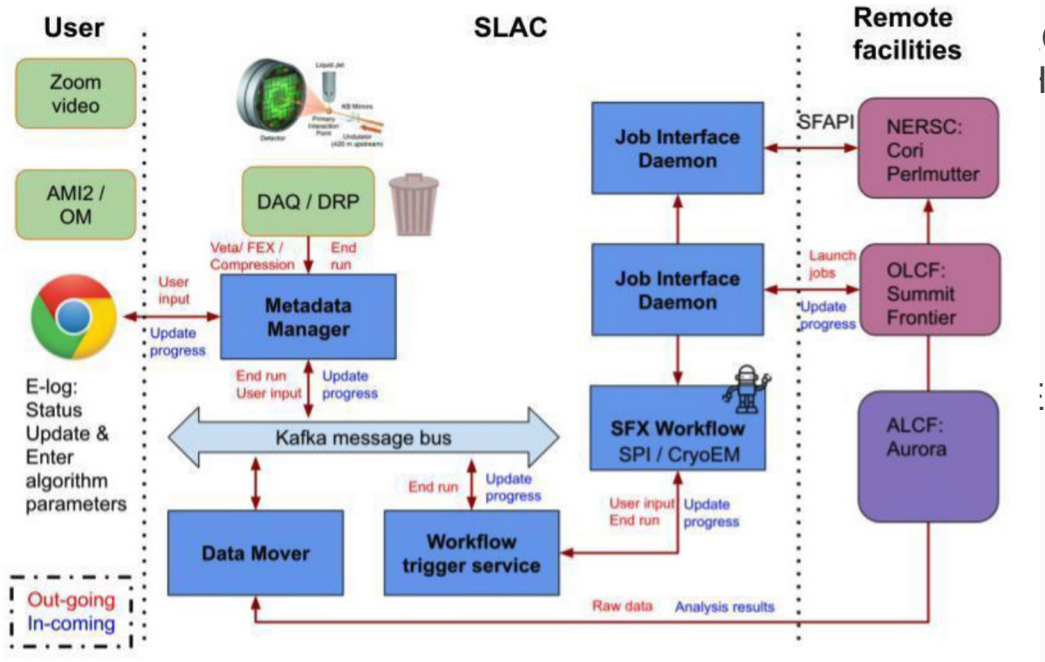
A Study of a Deterministic Networking Framework for Latency Critical Large Scientific Data Transfers

Vijeth Kumbarahally Lakshminarayana [1]
Carolina Minami Oguchi [1]
Alex Sim [2]
Kesheng Wu [2]
Dipak Ghosal [1,2]

[1]Department of Computer Science, University of California, Davis, CA
[2]Energy Sciences Network (ESnet), Lawrence Berkeley National Laboratory, Berkeley, CA

Time-Sensitive Workload from BES: LCLS-II accessing remote HPC by streaming data via ESnet



CLS II workflow is indicative of instrument to IPC flows more generally

- LCLS/NERSC workflow: ExaFEL ECP project
- Reservations made at the remote facility and in the network in advance of the experiment.

:Snet:

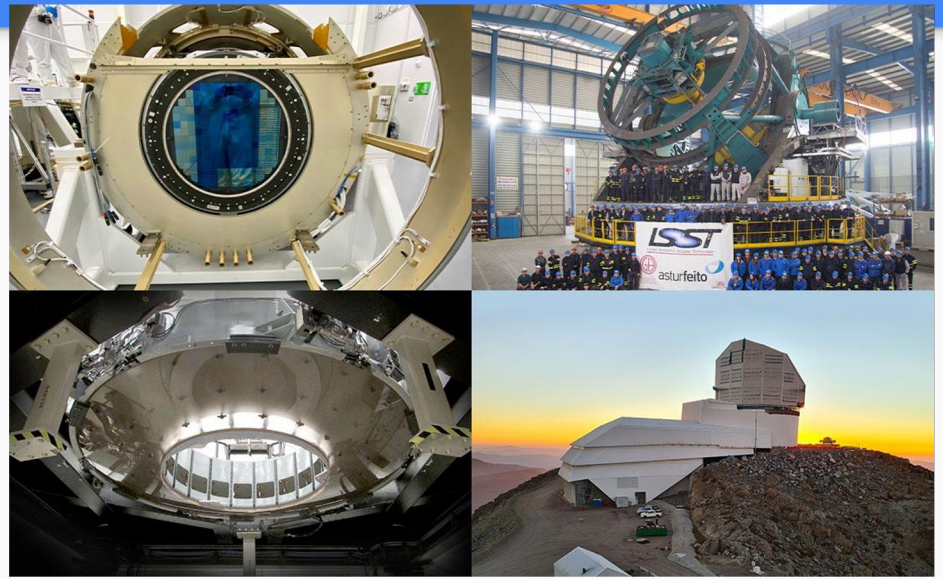
- Data transfer needs high-performance and resiliency.
- Network services should enable end-to-end system performance as raw network bandwidth is not enough.

Source: 2022 ESnet/BES Network Requirements Review, fig 5.4.10

Time-Sensitive Workload from HEP: Rubin Observatory ships data from Chile to HPC centers and completes analyses in two minutes

The Vera C. Rubin Observatory and its LSST collaboration brought DOE, NSF, SLAC, NCSA and Amlight together to form a new distributed virtual data processing center

- Timely data acquisition (traffic from Chile to USA)
- High-speed connectivity to Europe for data exchange with IN2P3, others
- Efficient cloud connectivity to support user-driven data analysis by the broader astronomy community



Rubin Observatory's mission is "to build a well-understood system that will produce an unprecedented astronomical data set for studies of the deep and dynamic universe, make the data widely accessible to a diverse community of scientists, and engage the public to explore the Universe with us."

Commonly used data transfer methods have clear limitations

Packet Switching (Best effort)

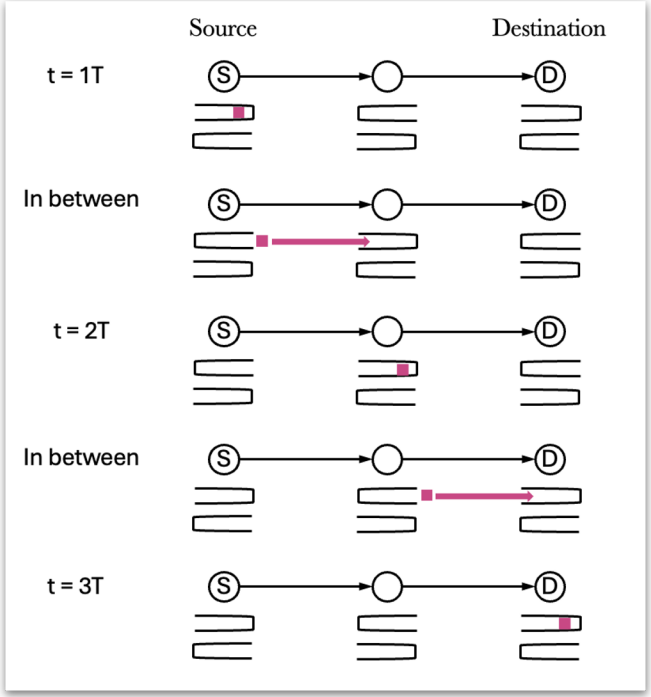
- Send packets from source with destination label.
- Routers in the network takes care of routing
- Couldn’t guarantee completion time

Circuit Switching

- Reserve path(s) for source-destination pair(s).
- No other source-destination pairs can use the reserved portion

Alternative to Consider -- Time Sensitive Networks (TSN)

- IEEE 802.1 Working Group
- Link layer (Layer 2)
- Strict latency and reliability requirements
  - Low latency with guaranteed upper bound
  - Small variety in delivery time (small jitter)
  - Reliable delivery
- All nodes are time synchronized
- Routing and Scheduling
- Path for delivery is determined beforehand
  - Deterministic network
- Guaranteed completion time



TSN Algorithm:
Cyclic Queuing and Forwarding (CQF)

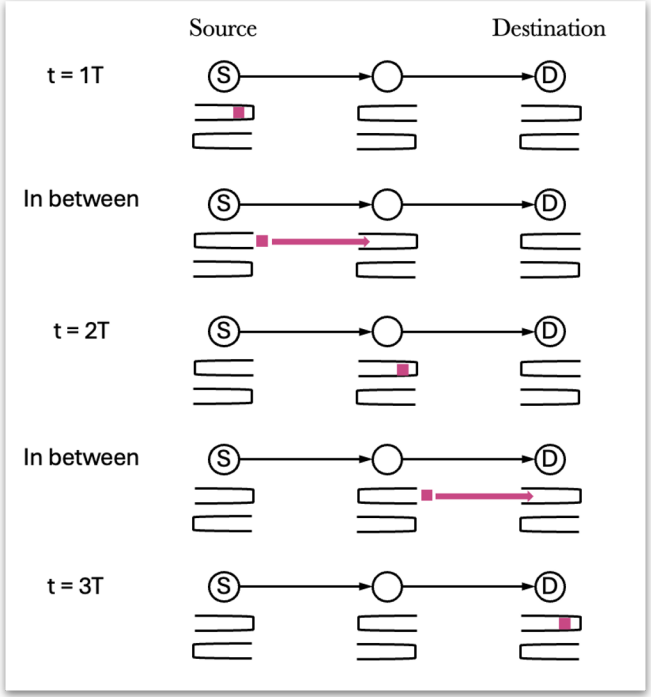
- One of TSN algorithms
- Cycle time T

$$T = \max_{(i,j) \in S} \{T_{i,j}\}$$

  - n: number of packets to transmit
  - l : length of a packet [bits]
  - T_{i,j} : minimum time required by node i to transmit n packets to node j
  - δ_{i,j}: propagation delay
  - r_{i,j}: data rate

$$T_{i,j} = \frac{n \times l}{r_{i,j}} + \delta_{i,j}$$

Provides completion time guarantee!



Ultimate objective: Integrate CQF into ESnet data transmission management

Current ESnet

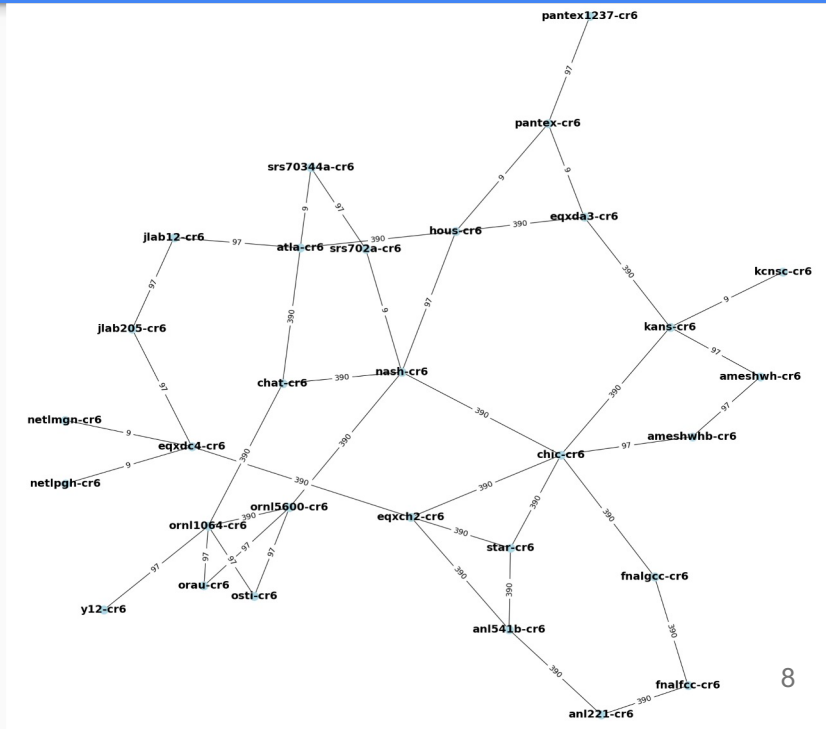
- Online Services for Circuit Provisioning and Reservation (OSCARS)
  - Create, manage, and monitor dedicated network pathways (circuit switching)
- Reserve full path or partial path throughout the transmission

ESnet with CQF

- Path reservation in granular time
- Reserve a part of path for only when packets are scheduled to pass through
- Time-synchronized for latency guarantee

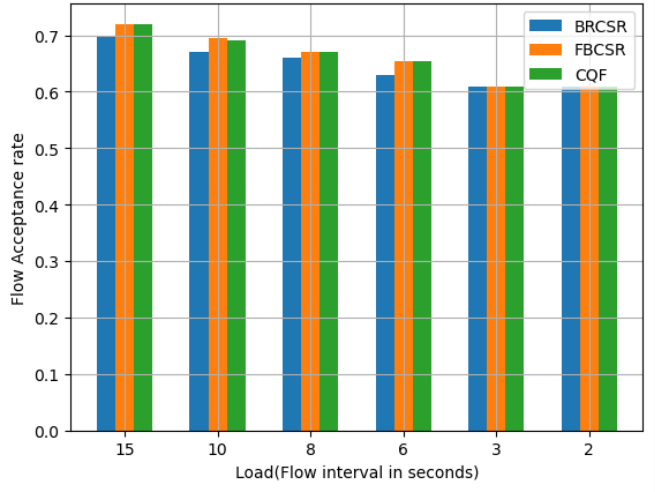
Current Exploration: Understanding the effectiveness of CQF at ESnet scale

- Settings: Large Scientific Network (ESnet)
  - Long links, geographically spread,
- Comparative Analysis of 3 methods
  - Circuit switching
    - Bandwidth-Reserved Circuit-Switched Routing(BRCSR)
    - Full-Bandwidth Circuit-Switched Routing (FBCSR)
  - Packet switching
    - Cyclic Queuing and Forward (CQF)

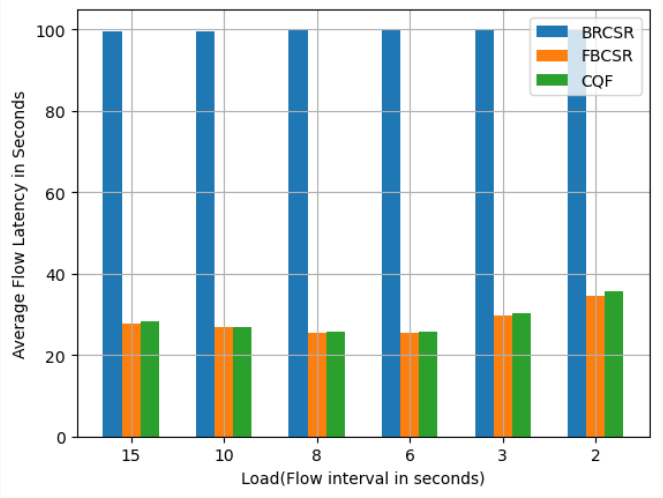


Current Observation: CQF could achieve the same level of performance as the best-known circuit-switch routine algorithm

Measured in flow acceptance rate, CQF achieves similar performance as best-known circuit-switching approaches (BRCSR and FBCSR)



Measured in task completion time, CQF uses much less time than the relatively static BRCSR, but achieves similar performance as the dynamic FBCSR



Conclusion

- Compared Bandwidth-Reserved Circuit-Switched Routing (BRCSR), Full-Bandwidth Circuit-Switched Routing (FBCSR), and Cyclic Queueing and Forwarding (CQF).
- The current simple implementation of CQF performs similarly to circuit-switched routing methods
- Future work: Use novel optimization techniques and reinforcement learning to tune several network parameters