OAK RIDGE National Laboratory

Hot Interconnects 2026 | VIRTUAL

Diagnosing Hidden Interconnect Bottlenecks with Omnistat and Cassini Telemetry

Aaron Welch[1], Shubhendra Pal Singhal[2], Akihiro Hayashi[2], Jun Shirako[2], Oscar Hernandez[1], Vivek Sarkar[2], Steve Poole[3]

[1]Oak Ridge National Laboratory
[2]Georgia Institute of Technology
[3]Los Alamos National Laboratory

welchda@ornl.gov, ssinghal74@gatech.edu, ahayashi@gatech.edu, shirako@gatech.edu, oscar@ornl.gov, vsarkar@gatech.edu, swpoole@lanl.gov

U.S. DEPARTMENT OF ENERGY

ORNL IS MANAGED BY UT-BATELLE LLC FOR THE DEPARTMENT OF ENERGY

Introducing Yet Another Telemetry System (YATS™)

- Who cares?
- Most performance counter analysis focuses on metrics within the node
  - What about the network?
- Raw counters provide some insight on what's changing, but what about why?
- We demonstrate a telemetry-guided diagnostic capability by integrating Cassini telemetry into Omnistat

Sounds Expensive and Boring…

- Low overhead since we simply query counter data that the hardware devices already collect under the hood
- Omnistat is designed to collect system scale-out metrics in a sampling-based fashion
  - Particular emphasis on GPU performance
- We focus on adding derived network metrics that map hardware events to system questions
- When one strategy significantly affects bandwidth or related metrics, we want users to be able to answer questions like:
  - How did the unit of work change?
  - Where did the pressure move?
  - Why might NICs drain differently?

Our approach

- During the run, a separate CXI collector samples raw counters from the NIC
- At the end of the run, Omnistat combines the collected data at the job level, processes the derived metrics, and exports them as time-series data
- The new network-based metrics can be compared to other existing job info for CPU/GPU performance, power, energy, etc.
- Future integration with ROCProfiler, PCIe counters, and other sources can seamlessly extend capabilities

Omnistat CXI Derived Metrics

Category	Representative derived metrics
Throughput and traffic mix	TX/RX bandwidth, bidirectional bandwidth, TX/RX balance, protocol shares
Message rate and packet-size efficiency	Packet rates, average bytes per TX packet, small- and large-packet fractions
Link utilisation, stalls, and pipeline blockage	Link busy fraction, stall per flit, CQ blocked cycles, NIC no-work rate
Congestion control, ECN, and flow control	Pause sent/received, XOFF sent, ECN ratios, congestion discard rate
Credit pressure, resource occupancy, and queue health	Command/receive FIFO credits, PI posted credits for posted PCIe-path work, endpoint-table pressure
Dynamic routing and ordering proxies	Ordered/unordered ratio, unordered fraction, ordered request fraction
Latency, tail behaviour, and timeouts	Mean latency-bin proxy, tail-latency fraction, timeout rate
Correctness and health: errors and corruption	Bad TX/RX octets, ECC corrected and uncorrected error rates
Compute/communication overlap	Overlap score, requiring external phase labels

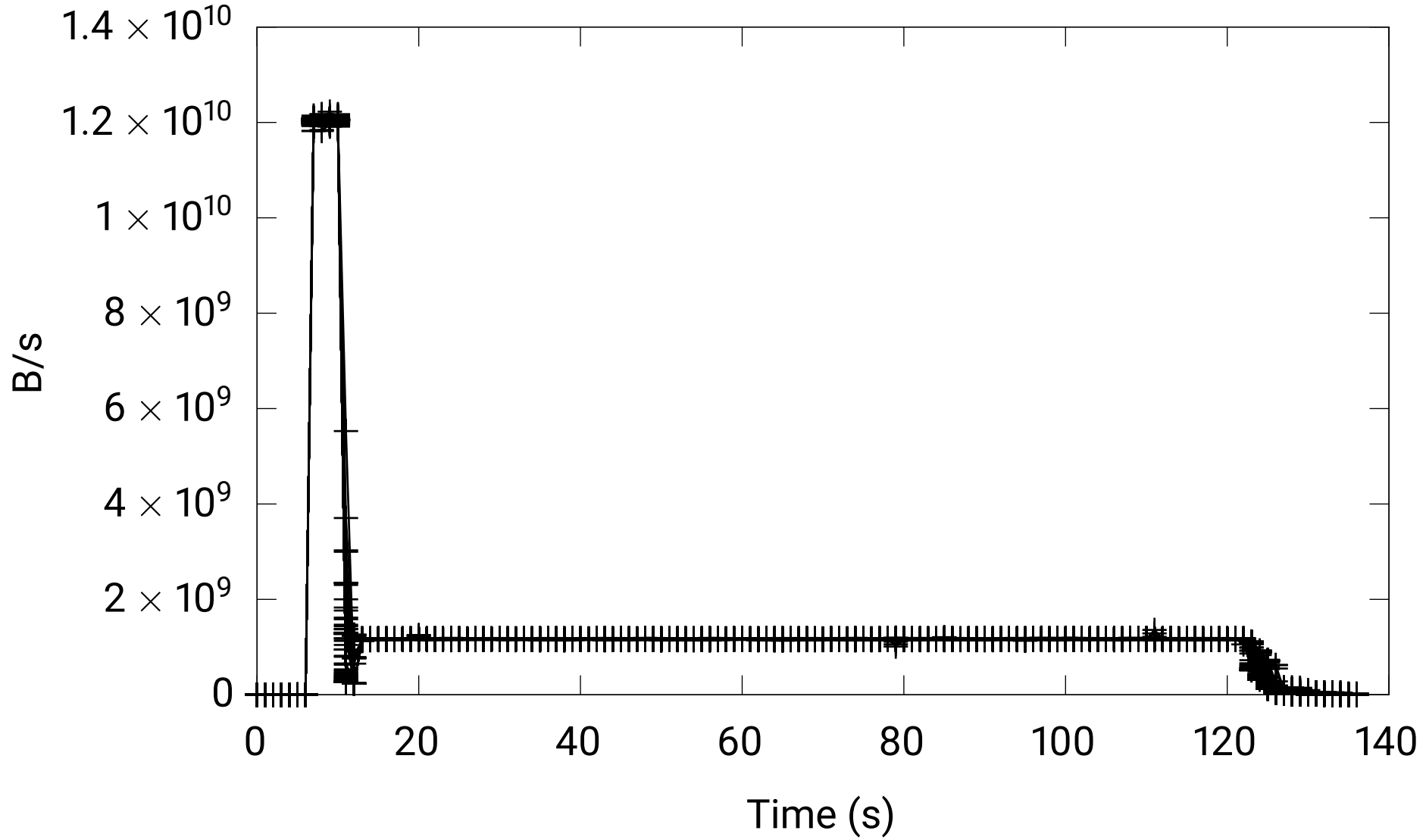
Case Study: Sparse Matrix Transpose

- Computes the transpose of a sparse matrix in compressed sparse row (CSR) format
- Difficult to aggregate due to synchronisation as processing elements (PEs) need when determing locations for writing data
- First phase uses a histogram pattern to determine the number of nonzeroes in each column
- Second phase acquires the new location for each value, then writes it to the destination PE
- We will use this as a case study for what can be achieved with this telemetry framework
  - Run using 32 nodes on Frontier with 56 processes per node (PPN)
  - Each node contains four HPE Slingshot 200 Gbit/s NICs
  - Two versions: one using small messages (AGP) directly, one aggregating them via bale conveyors

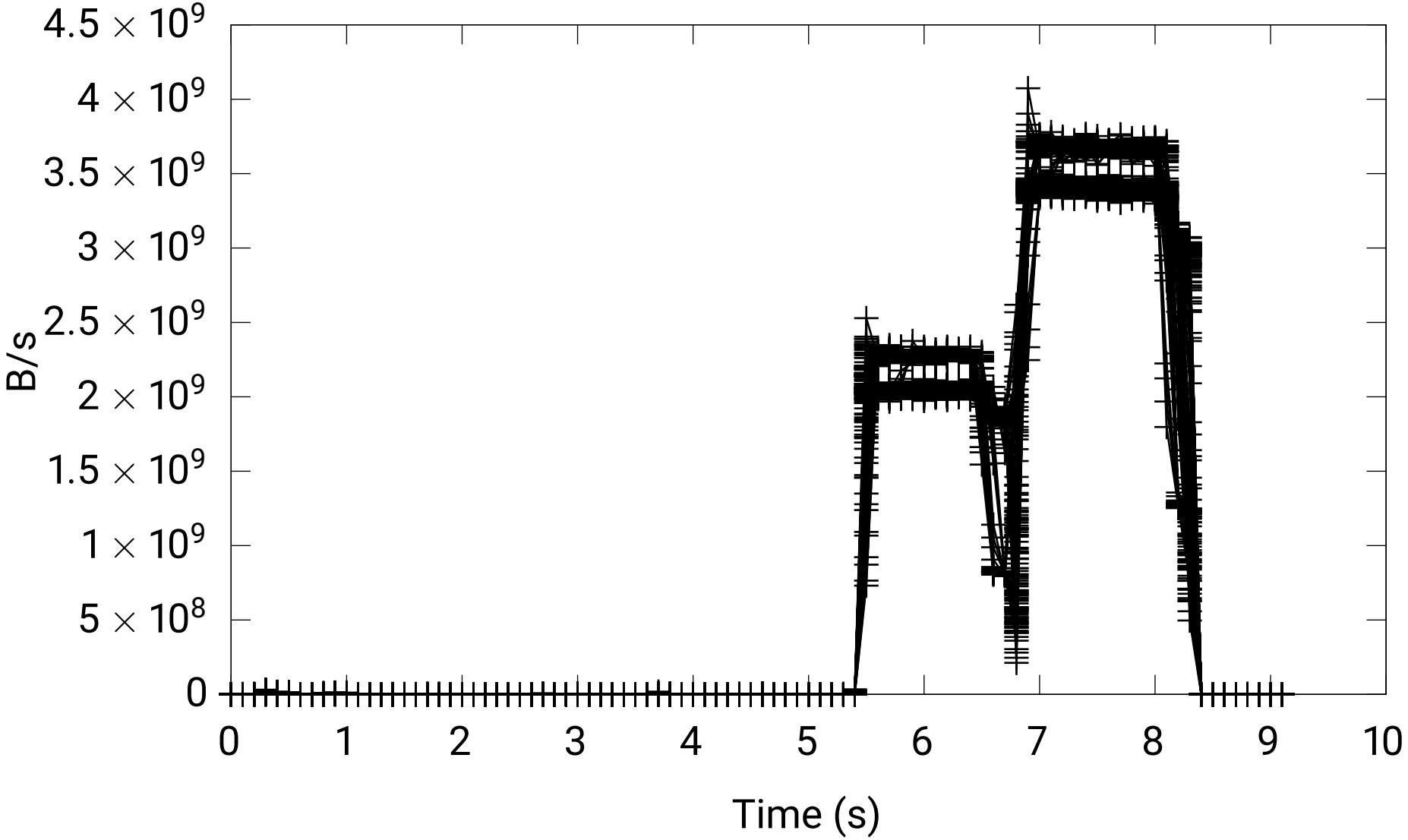
Select Metrics for Sparse Matrix Transpose

Metric	AGP	conveyors
Avg. bidirectional BW	1.3 GB/s	2.7 GB/s
Avg. bytes per TX packet	68	141
Large-packet fraction	0.0005	0.0476
Packet-rate behaviour	burst plus long tail	compact late burst
Credit/endpoint pressure	persistent drain tail	two shorter pressure episodes
Health/loss signals	zero discard, timeout, bad octets	zero discard, timeout, bad octets

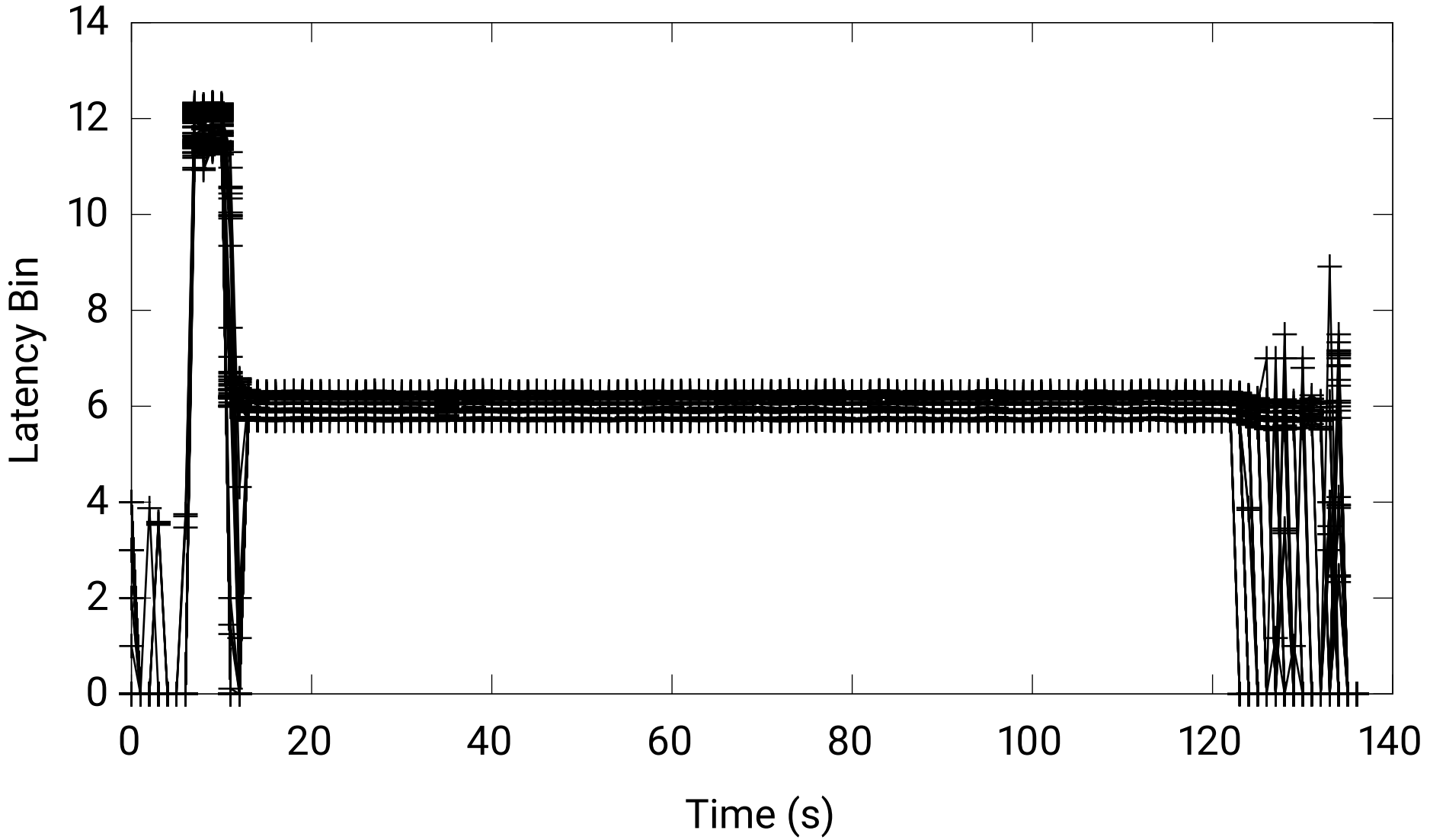
Bidirectional Bandwidth (AGP)



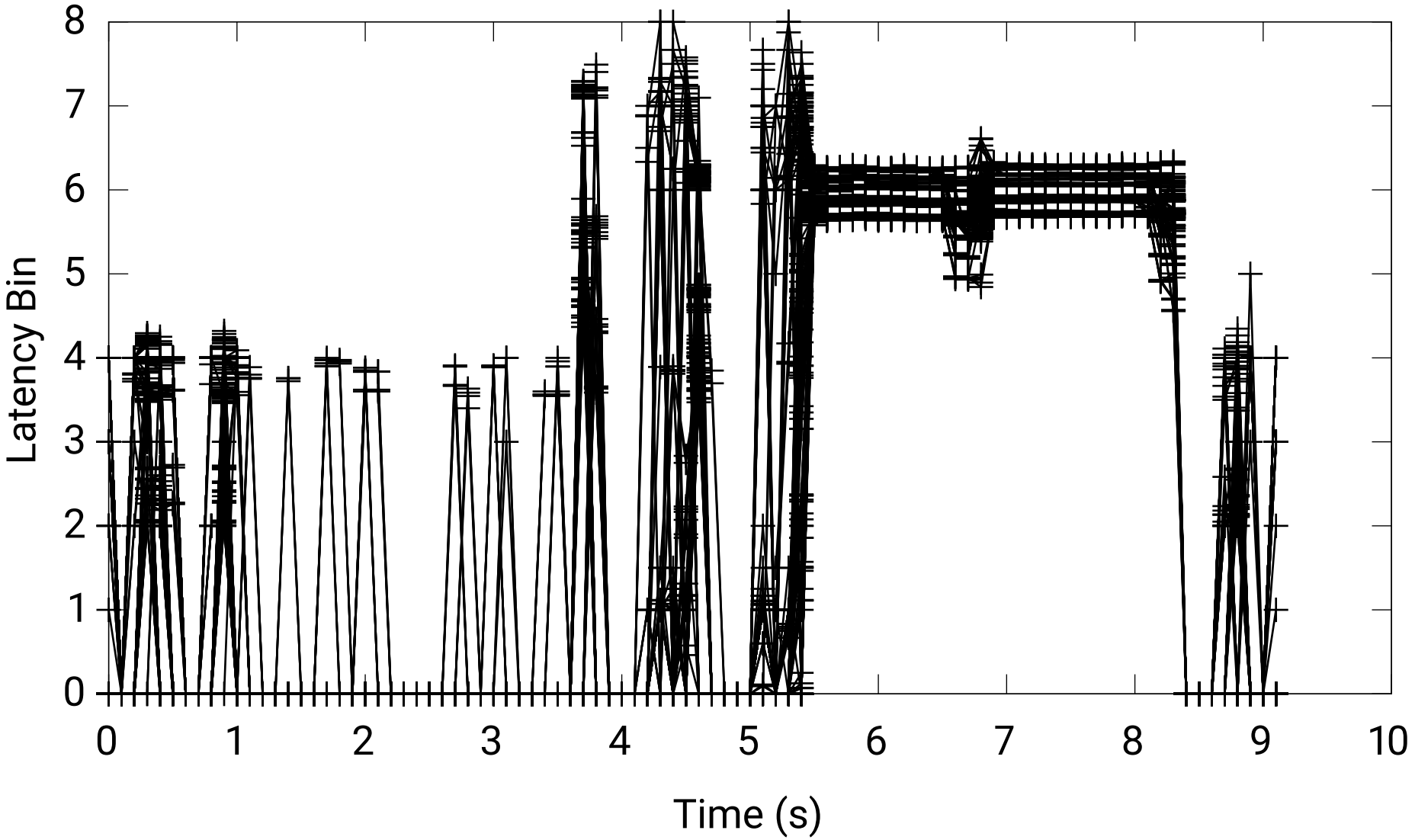
Bidirectional Bandwidth (Conveyors)



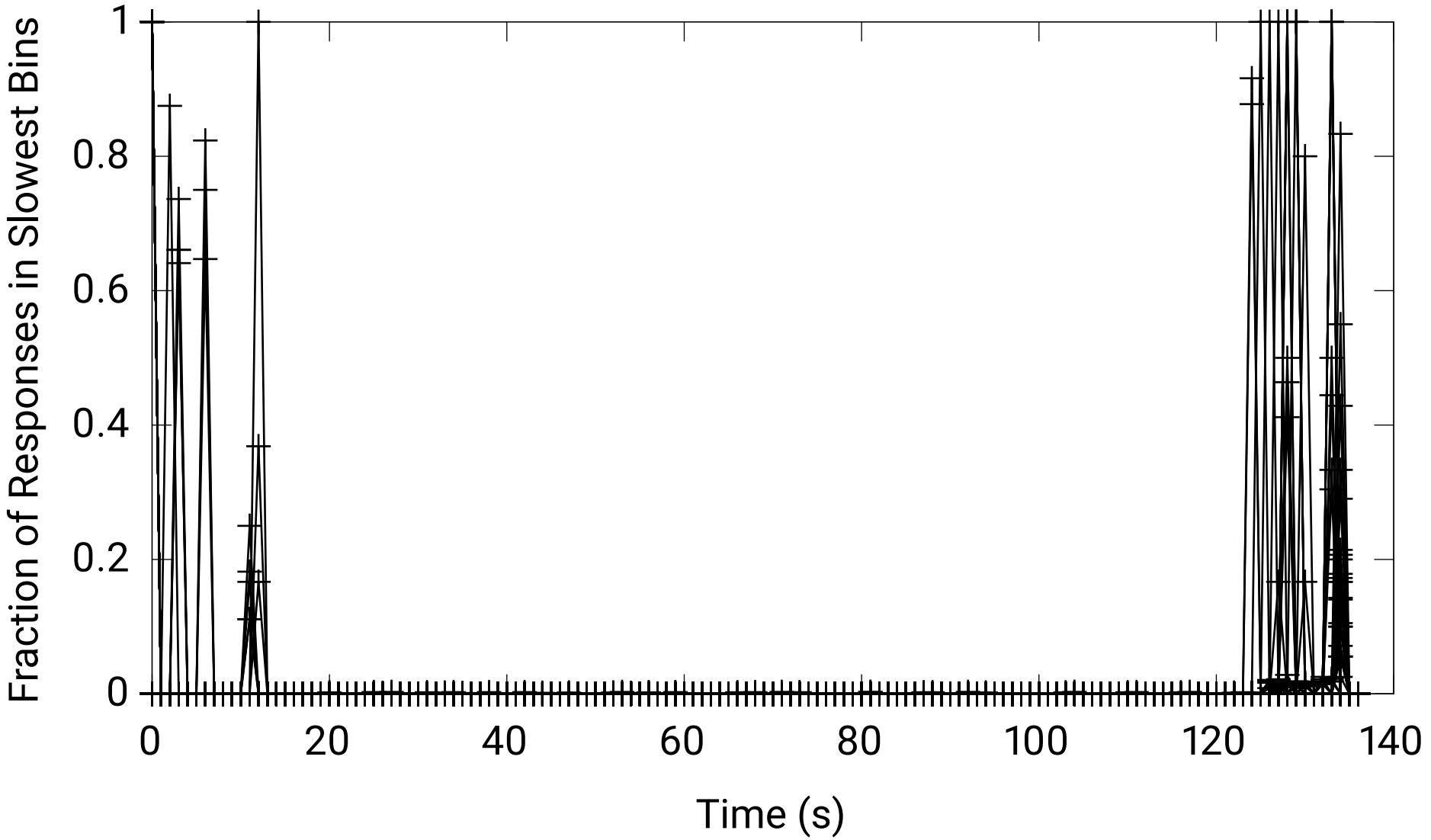
Sparse Matrix Transpose Mean Response Latency (AGP)



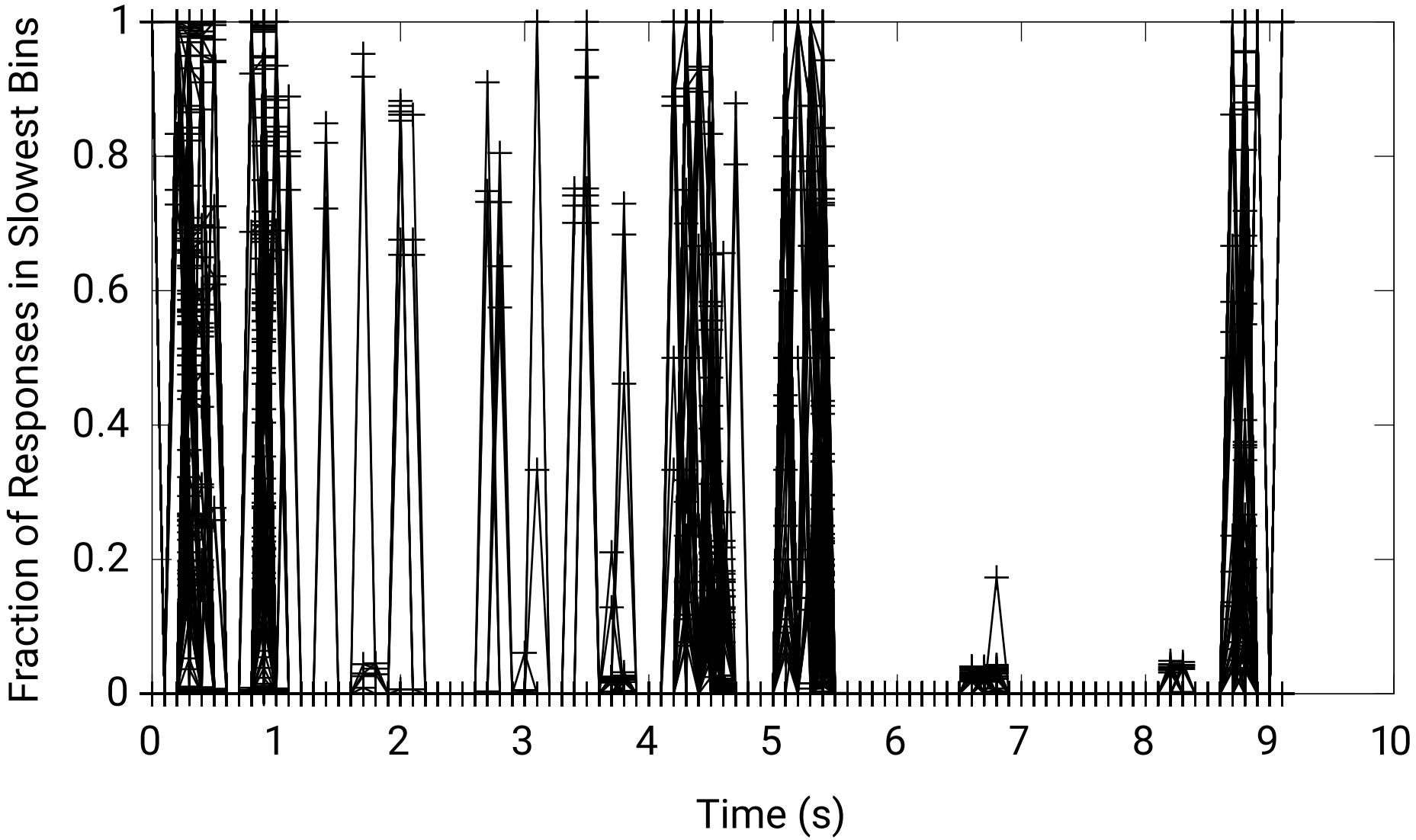
Sparse Matrix Transpose Mean Response Latency (Conveyors)



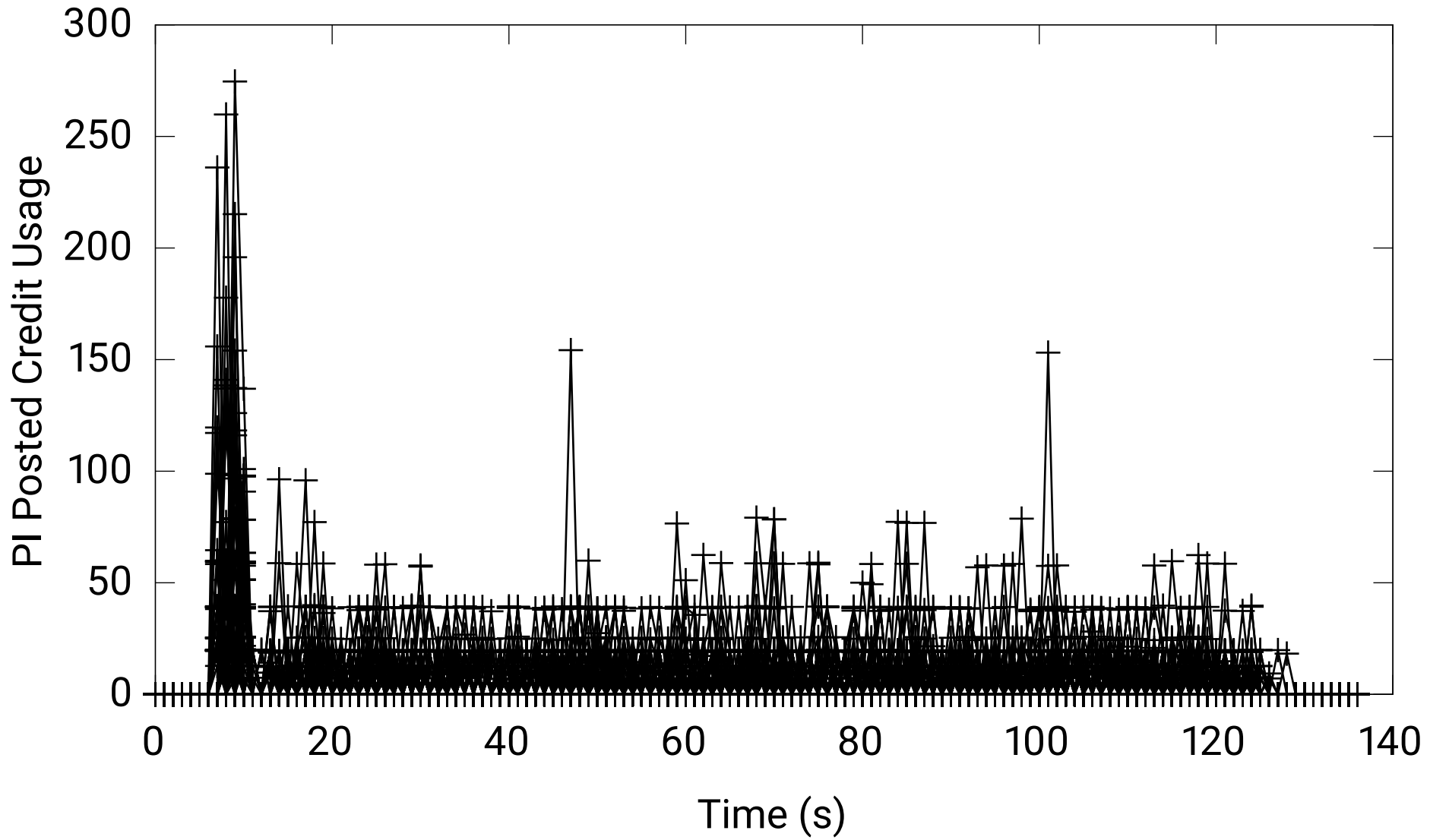
Tail Latency (AGP)



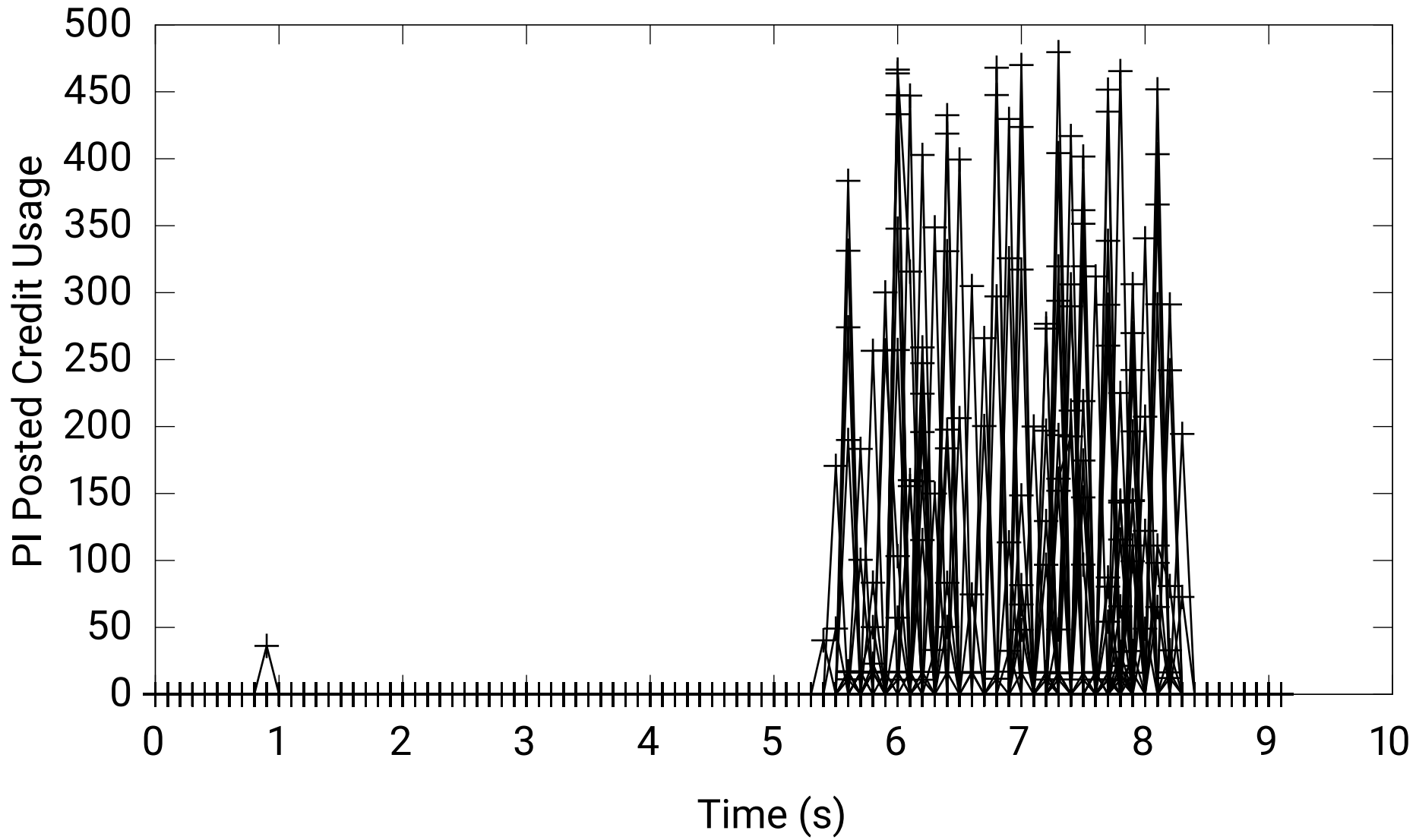
Tail Latency (Conveyors)



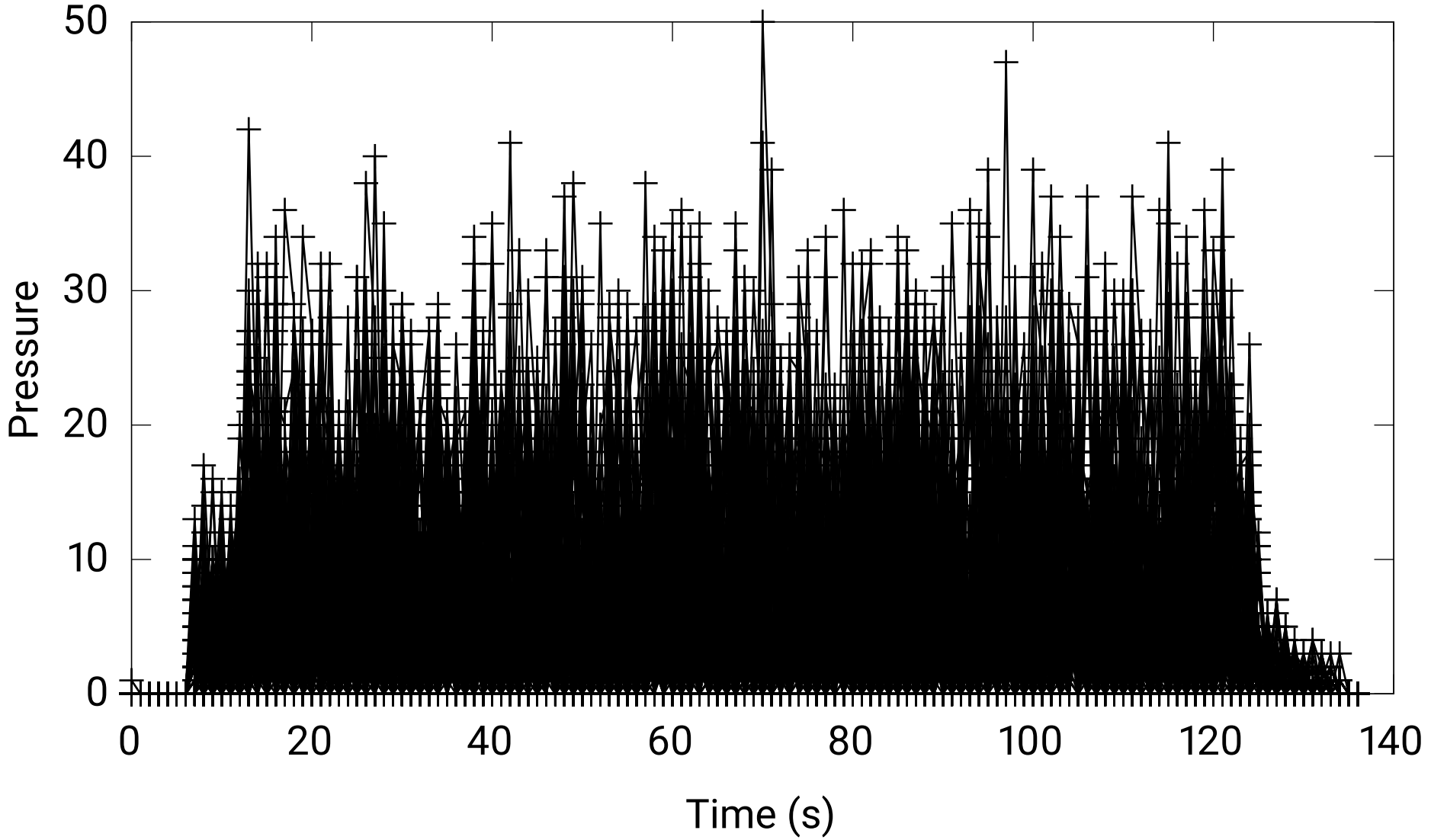
PI Posted Credit Pressure (AGP)



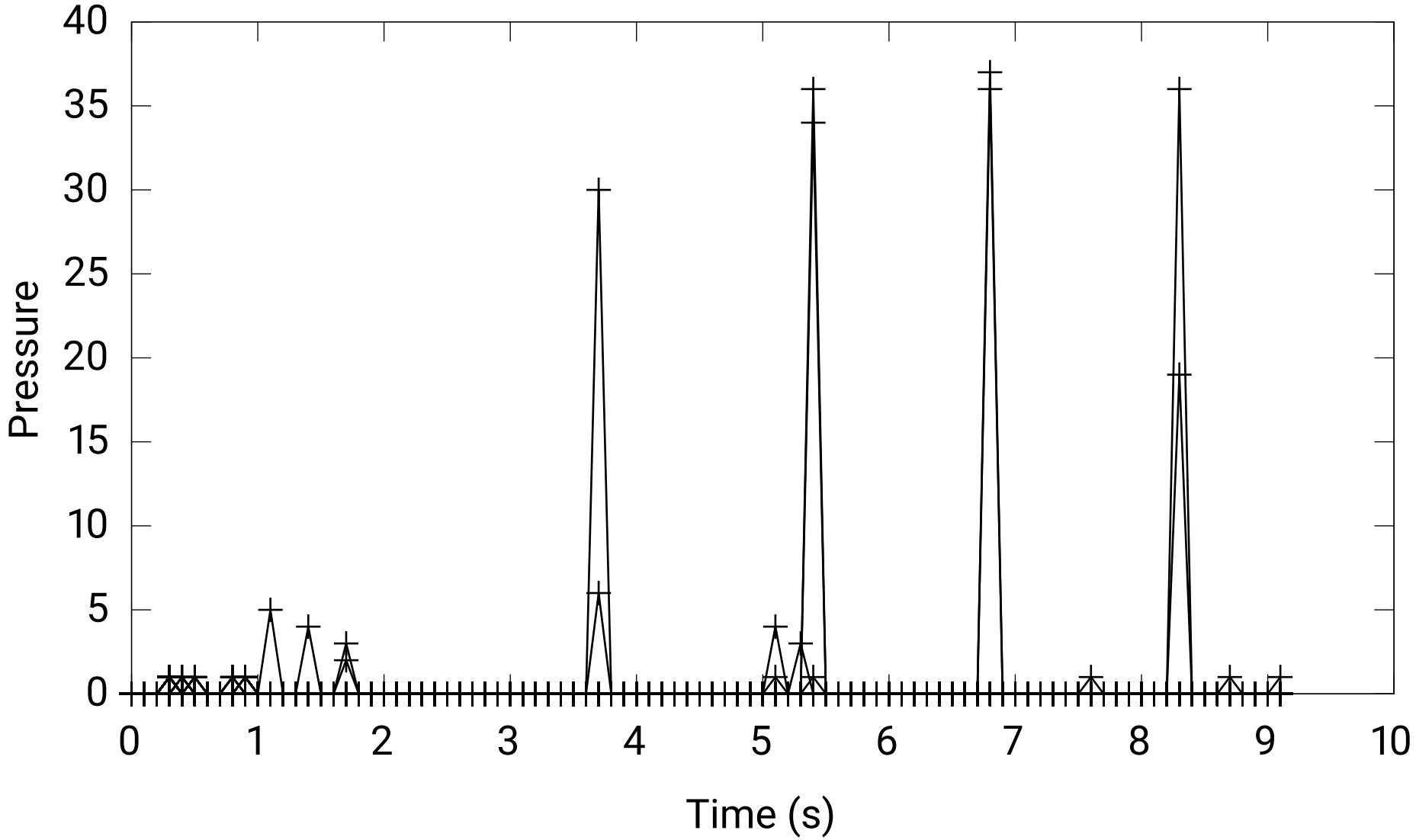
PI Posted Credit Pressure (Conveyors)



Endpoint Table Pressure (AGP)



Endpoint Table Pressure (Conveyors)



Questions?

welchda@ornl.gov