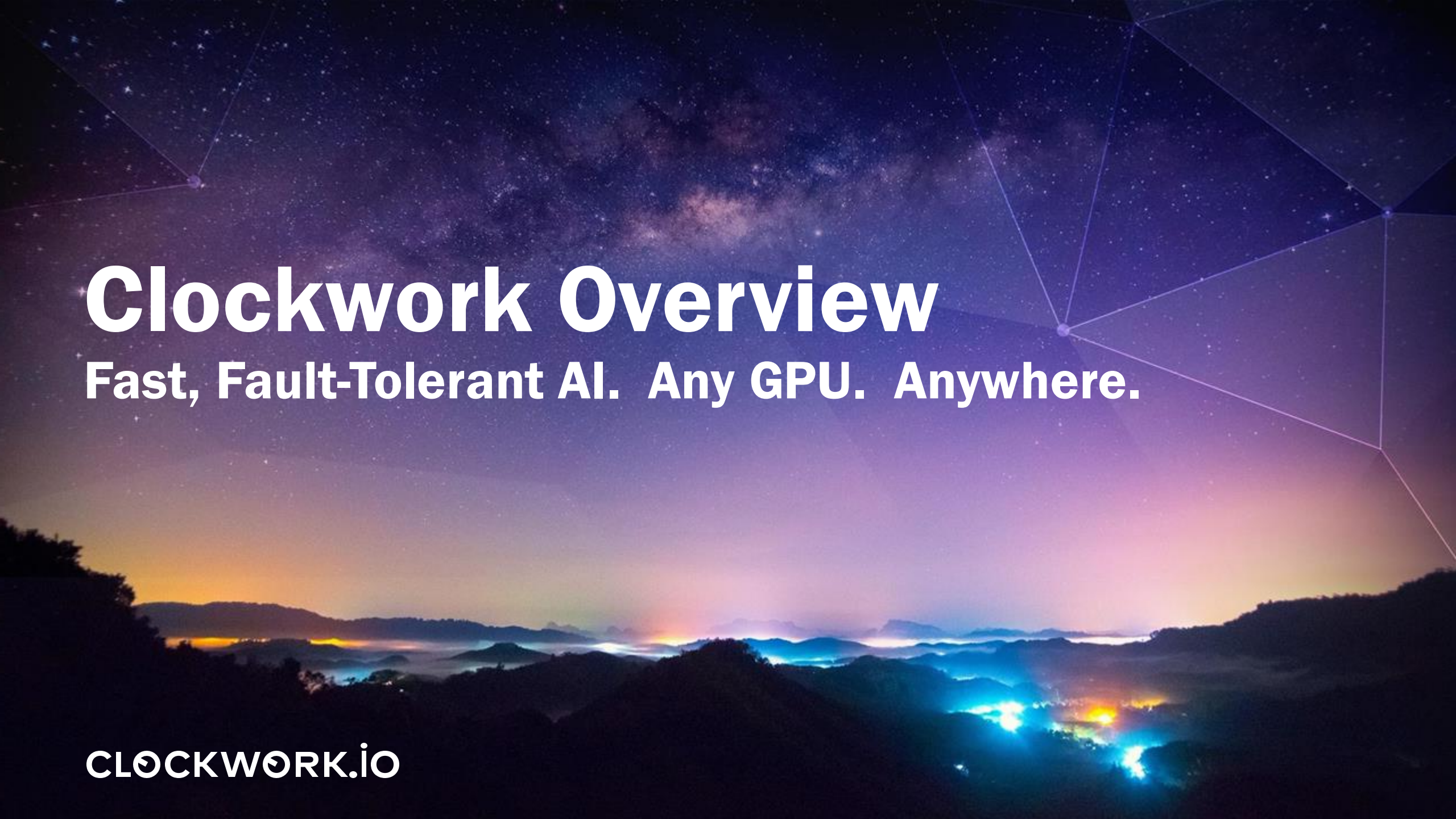
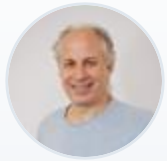

Clockwork Overview
Fast, Fault-Tolerant AI. Any GPU. Anywhere.
CLOCKWORK.IO

Combining Expertise in Distributed Systems and Category Creation

Suresh Vasudevan
CEO of Clockwork
ex-CEO: Sysdig, Nimble Storage
ex-CPO of NetApp

Balaji Prabhakar
Co-Founder, Clockwork
Professor of CS;
DCQCN Co-Inventor

Yilong Geng
Co-Founder, Clockwork
Huygens Clocksync Creator



Mendel Rosenblum
Chief Scientist, Clockwork
Professor of CS
Co-Founder of VMware

Dan Zheng
VP of Products & Solutions,
Clockwork, ex-Google

Financed by Top VCs and Angel Investors

John Hennessy
Ex-President of Stanford,
Chairman of the Board
of Alphabet

John Chambers
former CEO and
Chairman of Cisco

Lip-Bu Tan
CEO of Intel

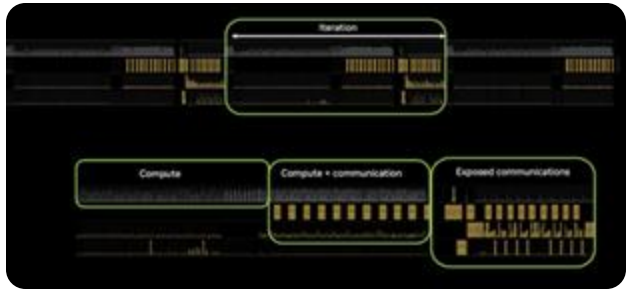
Jerry Yang
AME Cloud Ventures,
Co-Founder of Yahoo!

Greg Papadopoulos
Lead Series A investor

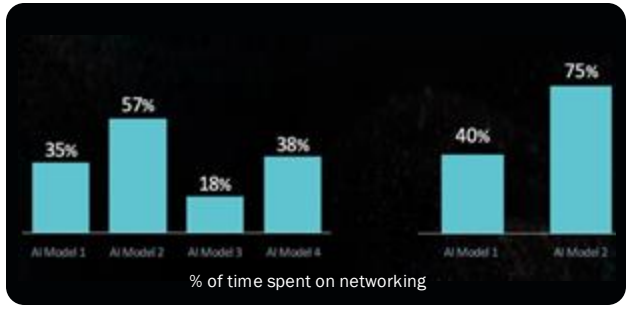
Forest Baskett
Lead Series A investor

AI Teams Have To Grapple With Dysfunctional Infrastructure

GTC 2024: Bursty Communication Causes GPU Wait Times



AMD Advancing AI 2024: Networking Drives GPU Idle Time



AI Infra Are Different

- Separate Back-end and Front-end Networks
- Highly Demanding:
  - Lossless
  - Very high-bandwidth
  - Low latency and jitter
  - In-order delivery
- Frequent GPU and network failures, memory errors, and data corruptions, resulting in job interruptions.

Dysfunctional Infrastructure:

Visibility Gap

Lack real-time detection, e2e correlation, and failure attribution

Resiliency Gap

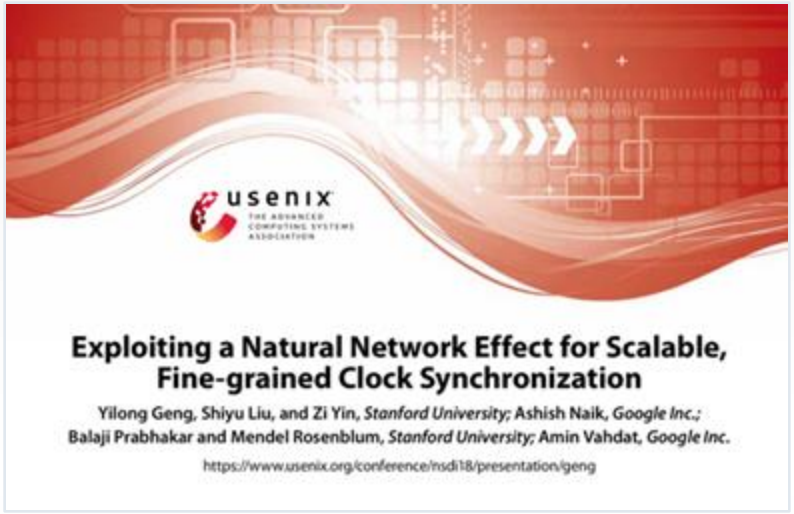
Link failures/flaps cause job restarts
Node failures cause job interruptions

Performance Gap

Contention and congestion undermines GPU utilization

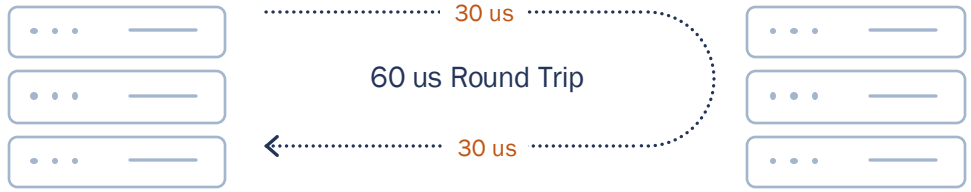
➔ **Lower GPU utilization, longer JCT, lower ROI**

Founding Inspiration: Software Based Nanoseconds-Accurate ClockSync

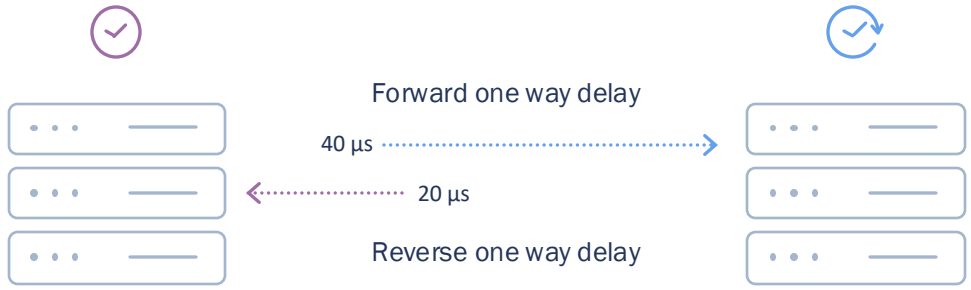


ebay WELLS FARGO coinbase

Traditional Approach: Round Trip Times as an Estimate of True One-Way Delays

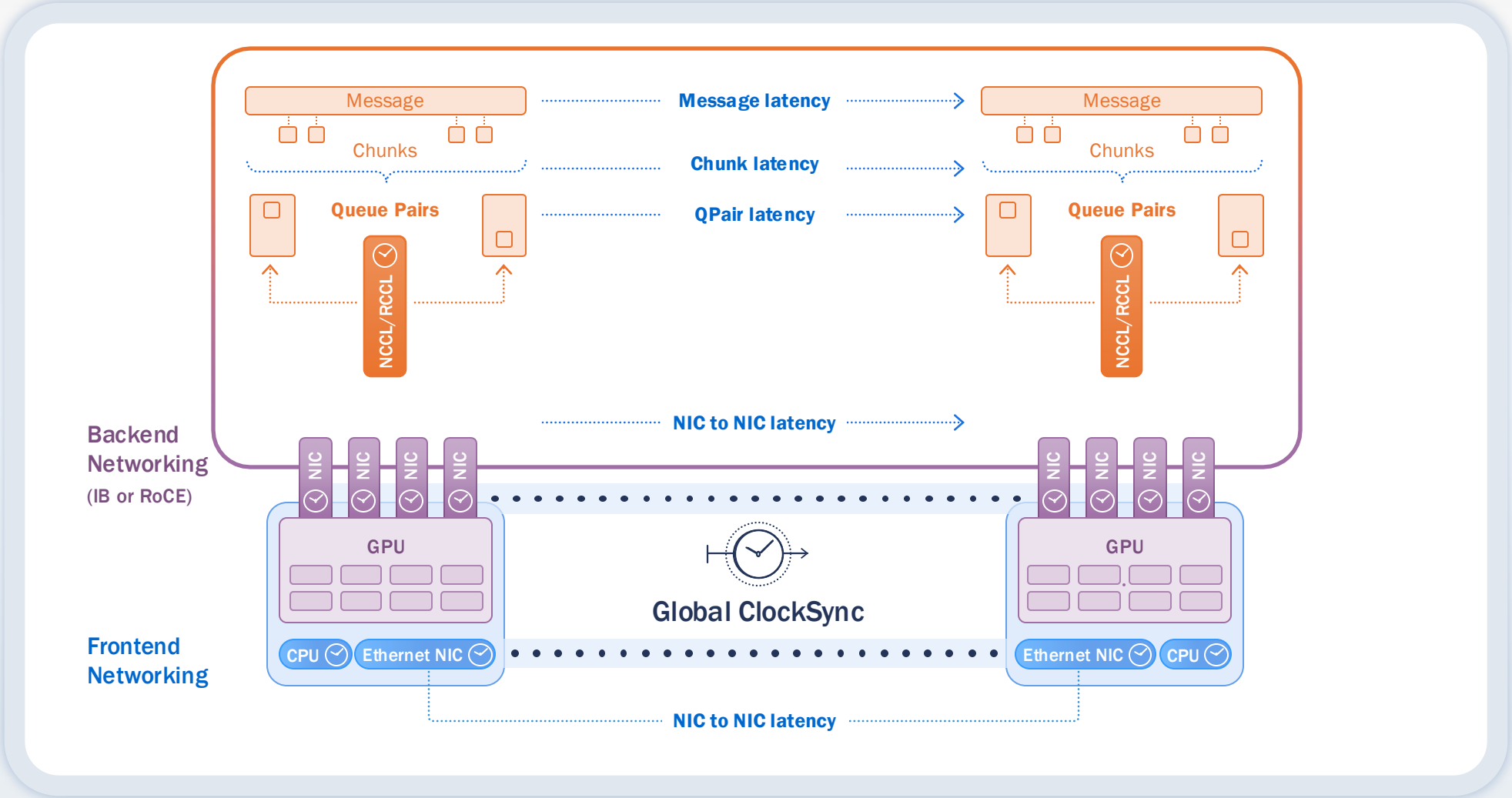


Data Centers with Clockwork
Measures of True One-Way Delays



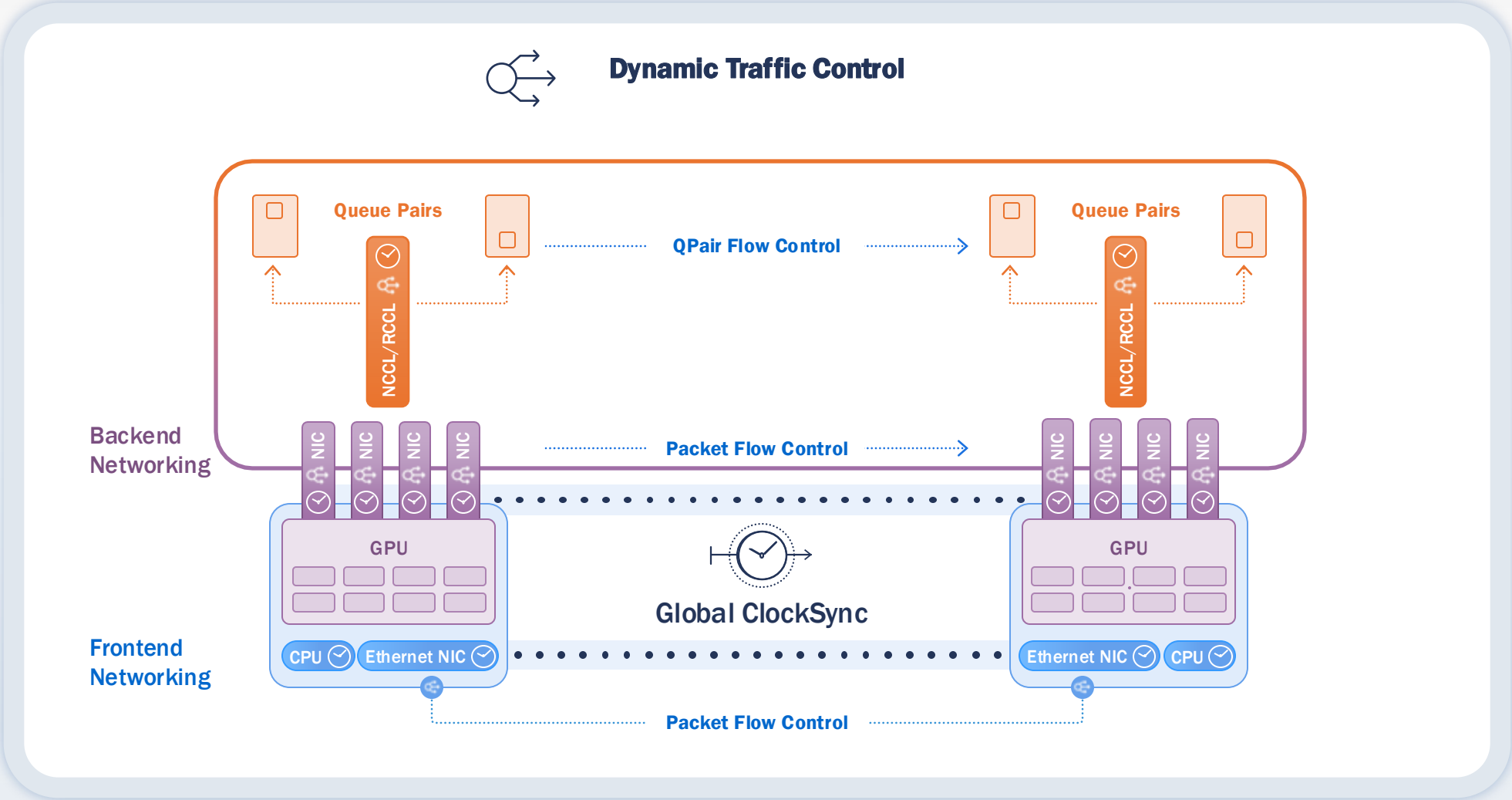
Clockwork FleetIQ Platform Foundation: Global ClockSync

Delivers Insane Visibility

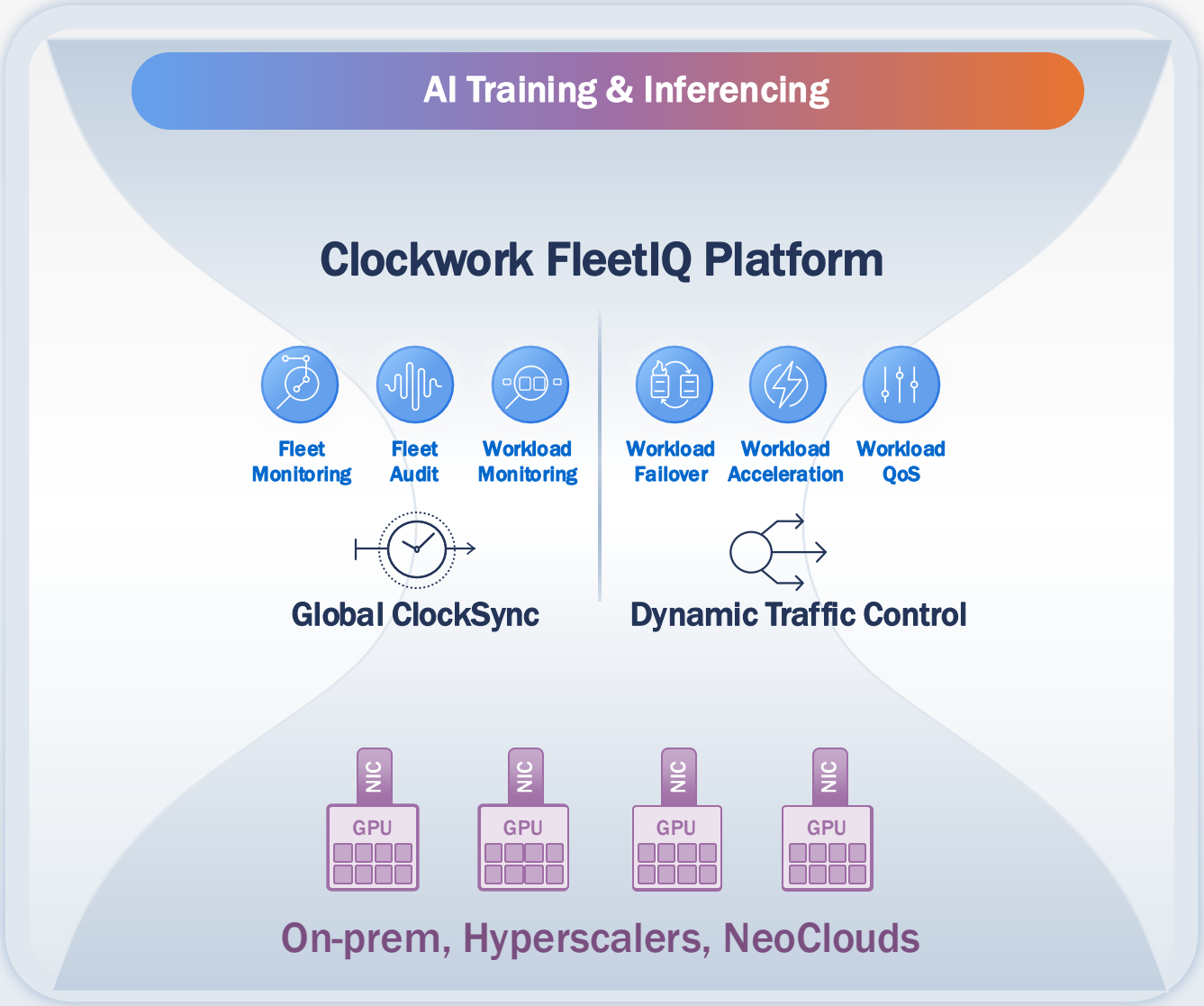


Clockwork FleetIQ Platform Foundation: Dynamic Traffic Control

Delivers Network Failover, Congestion Control and Load Balancing



Clockwork FleetIQ: Accelerate AI around the Clock



Deep Visibility

Quickly identify **WHY** your jobs are slow, inefficient or failing with cluster-wide health checks and workload monitoring.

Fault Tolerance

Auto failover and recovery for link / NIC flapping. No job crashes and restarts.

Performance Acceleration

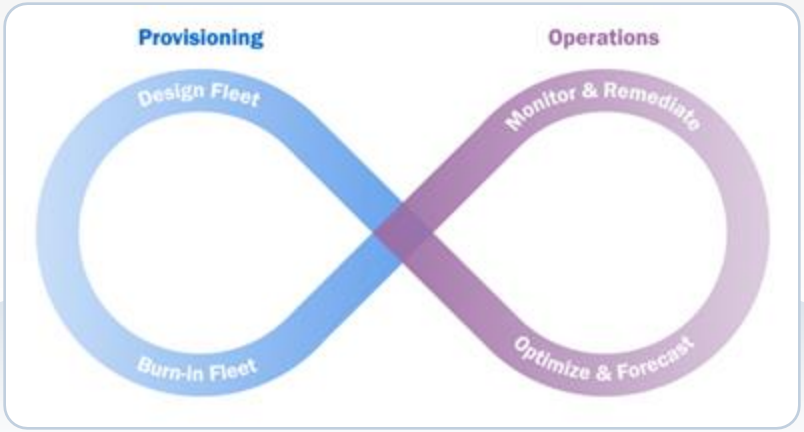
Auto eliminate contention & congestion. Application-level Quality-of-Service. No job slowdown.

Working with GPU Operators & Enterprises:

- A large cloud provider
- A large EV company
- A large video communication company
- A large social networking platform
- ....

Addressing the Visibility Gap:

Clockwork Fleet Audit, Fleet Monitoring, Workload Monitoring



Fleet Audit
(active health checks)

- Software checks
- Node checks
- Front-end network validation
- Back-end GPU network validation

Fleet Monitoring
(infra telemetry)

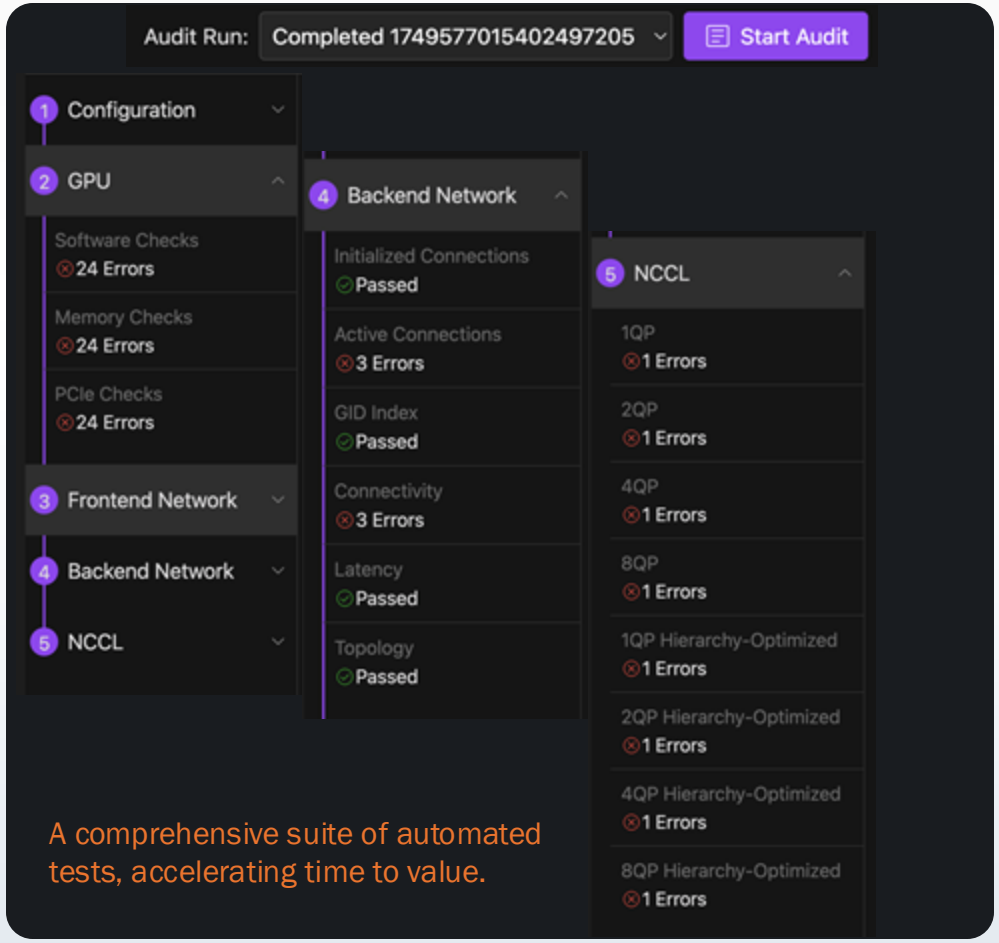
- Runtime link failures/flaps
- Runtime fabric topology
- Runtime fabric performance
- Congestion/contention monitoring

Workload Monitoring
(in-band telemetry)

- Deep visibility into communication flows associated with AI jobs
- Correlation of job, data path and network metrics to detect slow downs and diagnose root cause

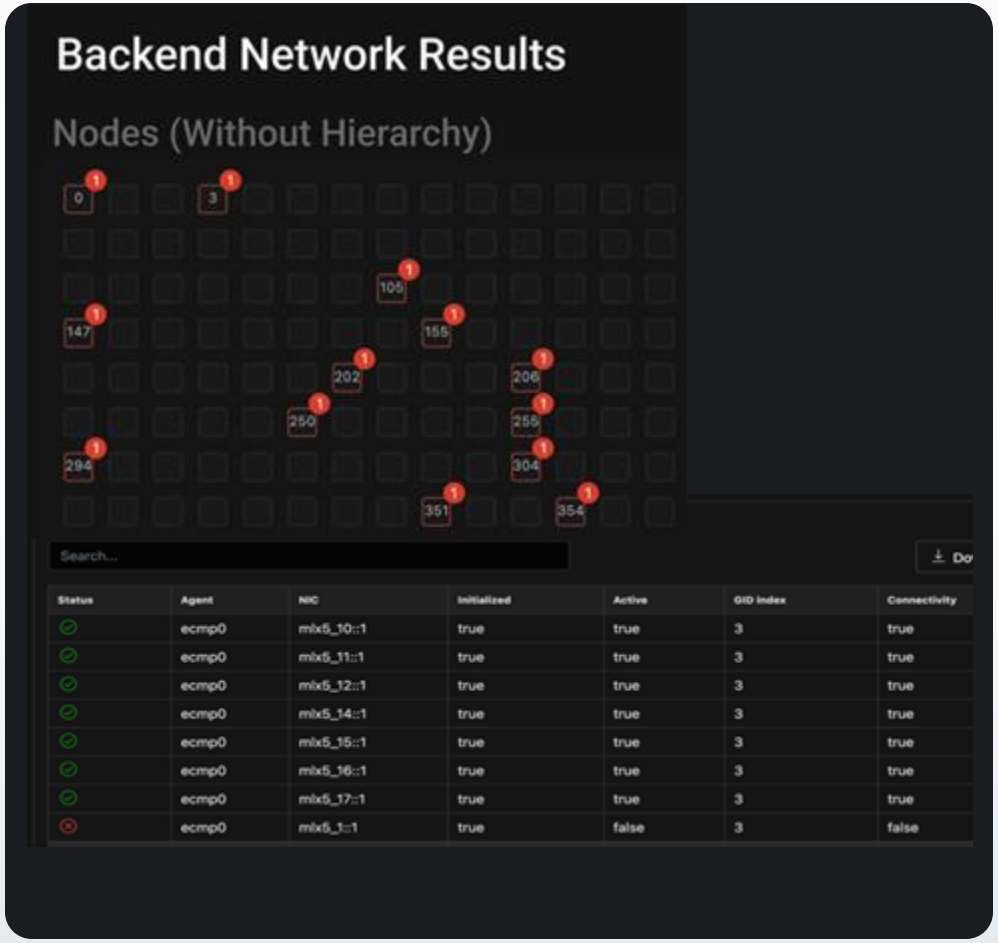
Clockwork Fleet Audit: Illustrative Customer Value

"I want to make sure my cluster is configured correctly before I run a week-long training job."



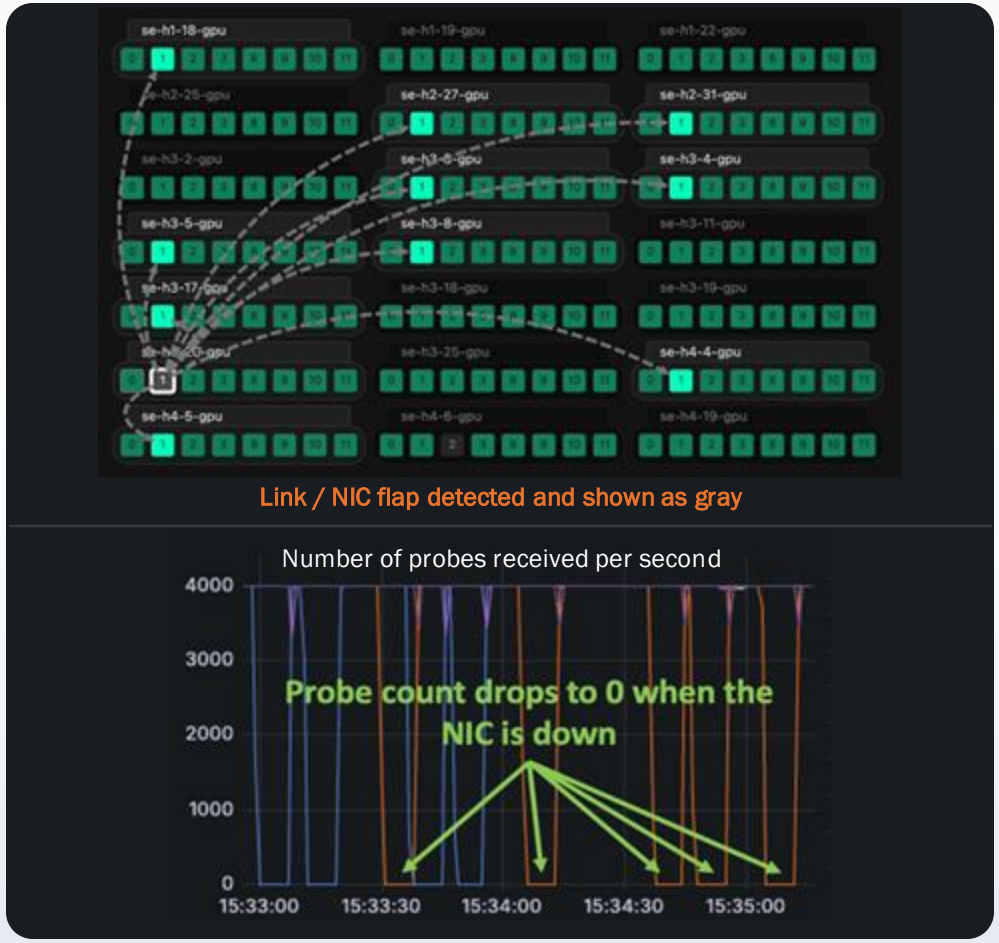
A comprehensive suite of automated tests, accelerating time to value.

"I found (i) 3 through cabling checks; (ii) 7 through cross-cluster ping6 test; BUT (iii) 3 are unique that I would not have found!"



Clockwork Fleet Monitoring: Illustrative Customer Value

"We want to detect network failures/link flap as soon as they happen, and not when our jobs stall!"

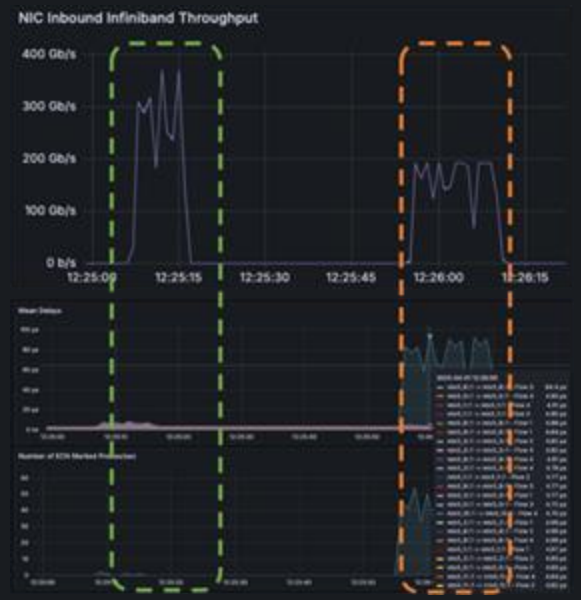


"We want to be sure that replacement GPUs in the cloud are meeting topology/latency SLAs?"

"We'd like to track latency continuously and get alerted when it goes above our set thresholds"

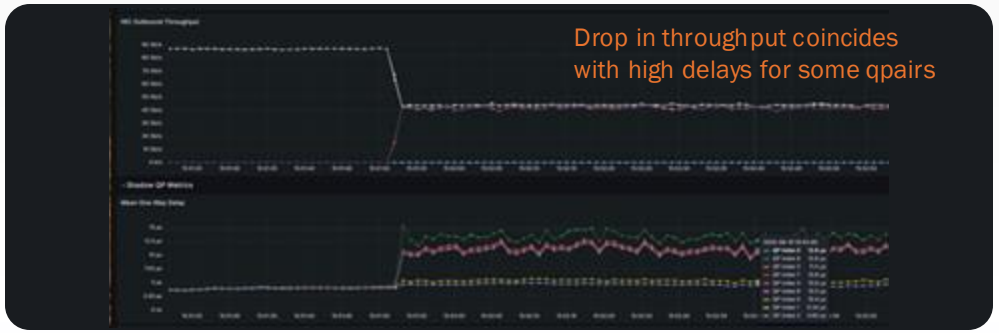
Clockwork Workload Monitoring: Illustrative Customer Value

"I ran all_reduce_perf workload twice, 1st run ~360Gbps, 2nd run only ~190Gbps. What could be the problem?"



The performance drop in the second run was caused by a significant delay on **one specific flow** (94 µs), while all other flows remained low-latency (~4 µs).

"We saw a sudden slowdown in job performance, could it be network-related?"



"Out-of-band and in-band Qpair one-way-delays are very different. The workload was mistakenly configured to use RoCEv1 instead of RoCEv2"

Disruptive Network Failures and Link Flaps Are Common and Expensive

Job Restarts Due To Disruptive Events Per Year

Number of GPUs	Job restarts/year	Mean time to failure
1,000 GPUs	100 - 250	35 - 87 hours
5,000 GPUs	500 – 1,250	7 - 18 hours
10,000 GPUs	1,000 – 2,500	3.5 - 9 hours
50,000 GPUs	5,000 – 12,500	42 - 105 minutes

"One of the most common problems encountered is Infiniband/RoCE link failure. Even if each NIC-to-leaf switch link" had a mean to failure rate of 5 years, due to the high number of transceivers, ***it would only take 26.28 minutes for the first job failure***

semianalysis

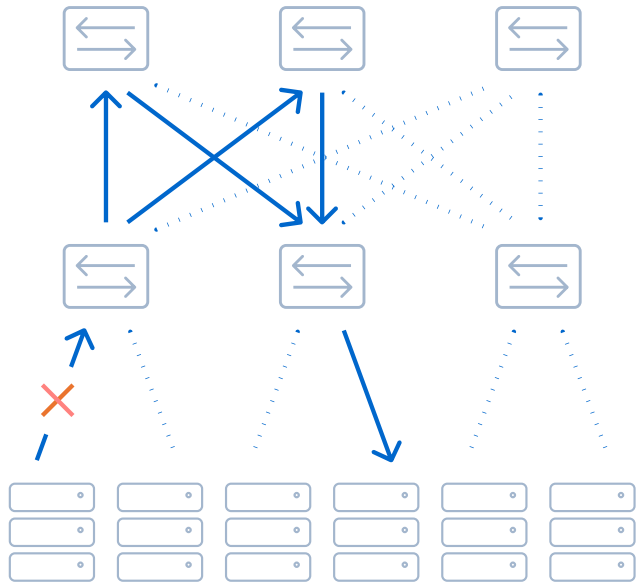
GPU Hours Lost Per Disruptive Event

	GPUs impacted	Checkpoint loss *	Recovery time	GPU hours lost
Job 1	256	2 hour	30 mins	640 hours
Job 2	512	2 hour	30 mins	1,280 hours
Job 3	1,024	2 hour	30 mins	2,560 hours

8-24 engineer hours & many 1,000s of dollars lost per incident

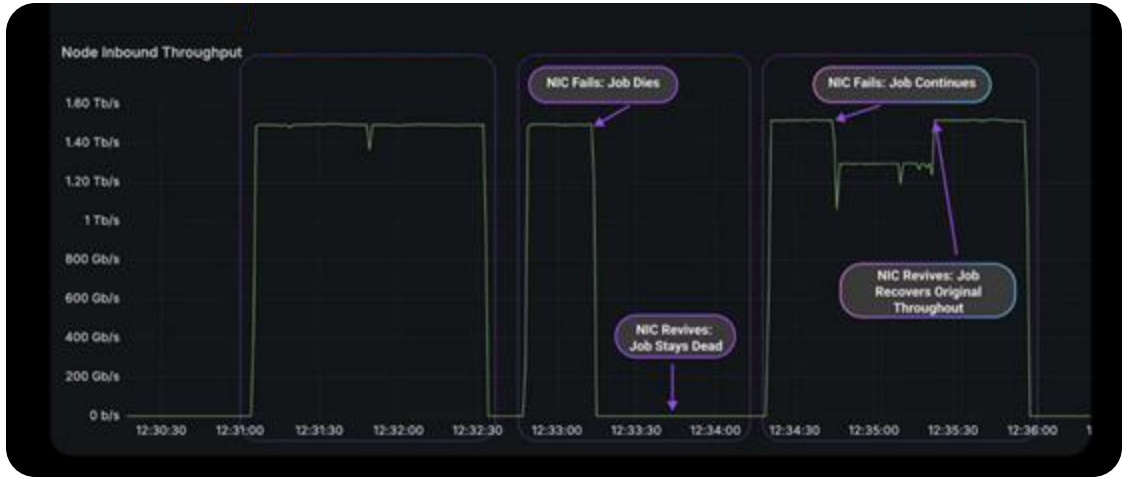
Source: Falcon: Pinpointing and Mitigating Stragglers for Large-Scale Hybrid-Parallel Training, 2024 · The Llama 3 Herd of Models, 2024 · "Alibaba HPN: A Data Center Network for Large Language Model Training", ACM SIGCOMM '24 · Gemini: Fast Failure Recovery in Distributed Training with In-Memory Checkpoints, 2023

Clockwork's Workload Failover Provides Resilience To Link Flaps

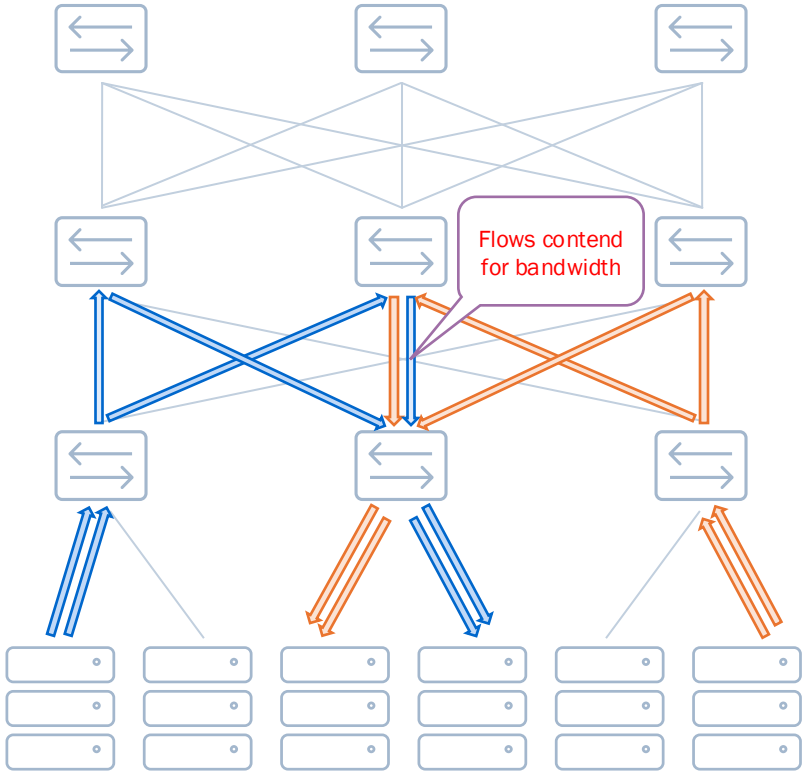


Link/NIC flapping

- Quickly detect link/NIC failure
- Use an alternate path
- Monitor failed paths and reuse them on recovery

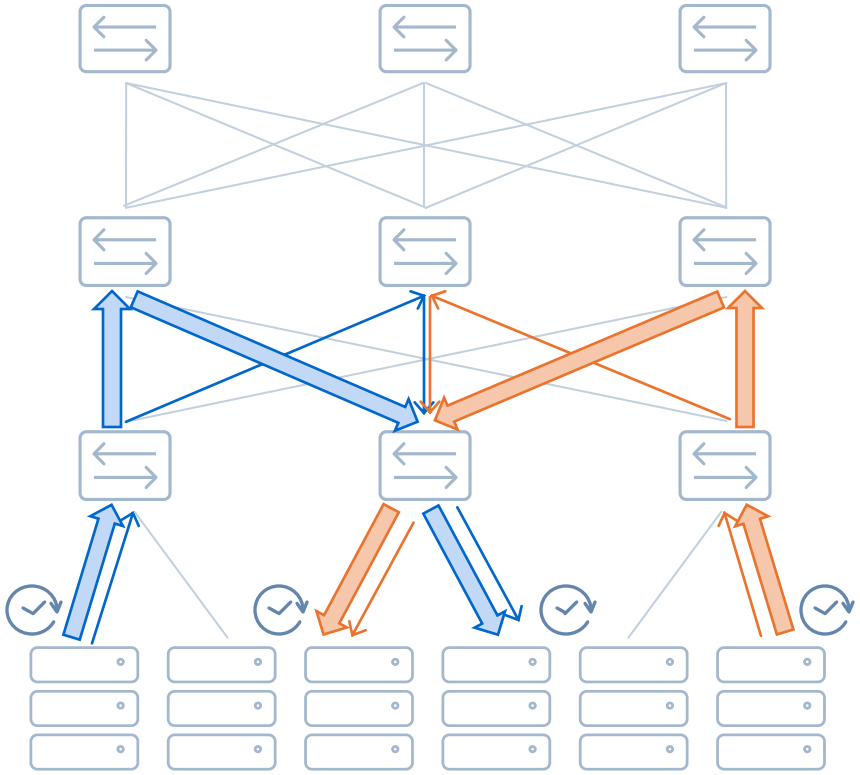


Detecting & Eliminating Contention



Contention:

- QPairs collide on links and contend for network bandwidth

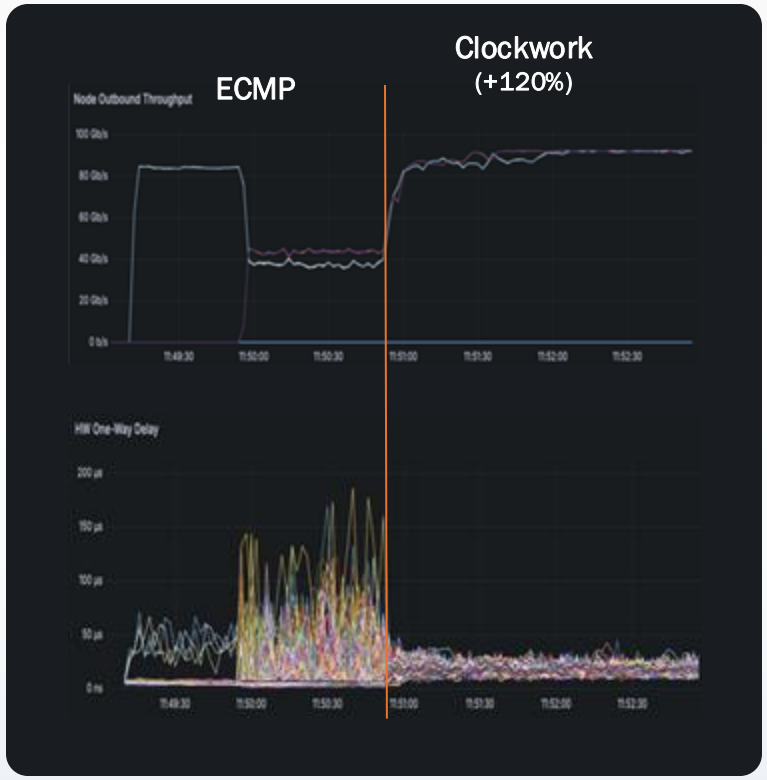


Clockwork's Workload Acceleration:

- QPairs with contentions have high one-way delays
- Shift traffic from congested paths to uncongested paths

Clockwork's Workload Acceleration: Example Use Cases

OCI: 2 all-to-all jobs (vs ECMP)



OCI: 2 all-to-all jobs (vs DLB)

Meta: 2 all-reduce jobs

DEMO

Questions?

Contact: hello@clockwork.io