



CONFERENCE

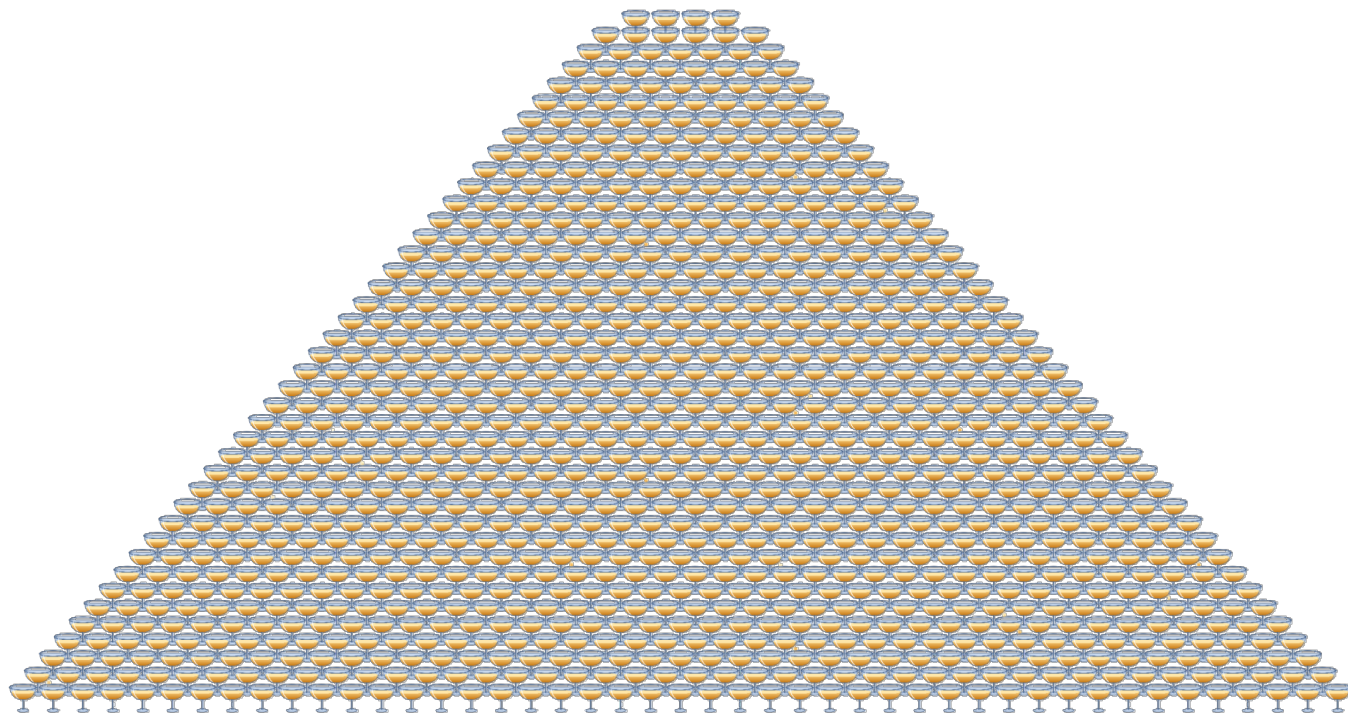
—NORTH AMERICA 2026—

Never Hit Zero

Surviving link and node failures in PyTorch training, on a real cluster
Lin She, Clockwork.io

CLOCKWORK.IO

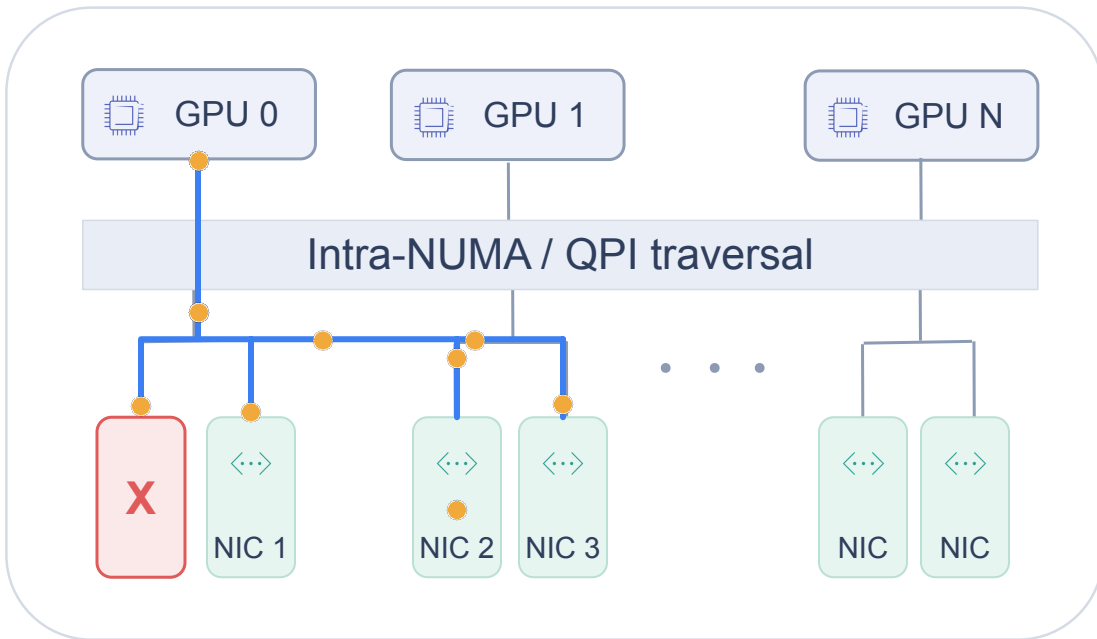
A Reminder of the Problem!



LinkPass: Clockwork's Link Failover Plugin

Prevent link flaps and failures from interrupting training, RL and multi-node inference workloads

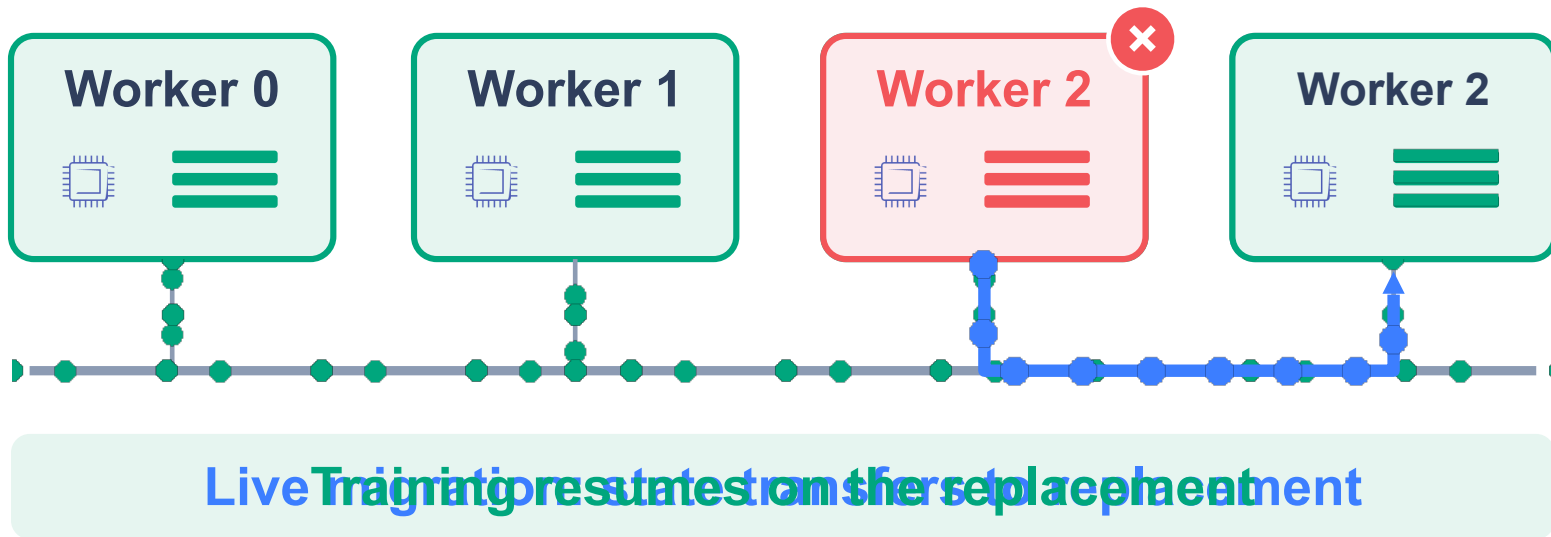
- 1 GPU sends traffic through NIC 0
- 2 NIC 0 fails
- 3 Traffic rerouted through NICs 1, 2 & 3
 - Workload continues as if nothing happened
- 4 Traffic returns to original path if link becomes healthy again



TorchPass - Live GPU Migration

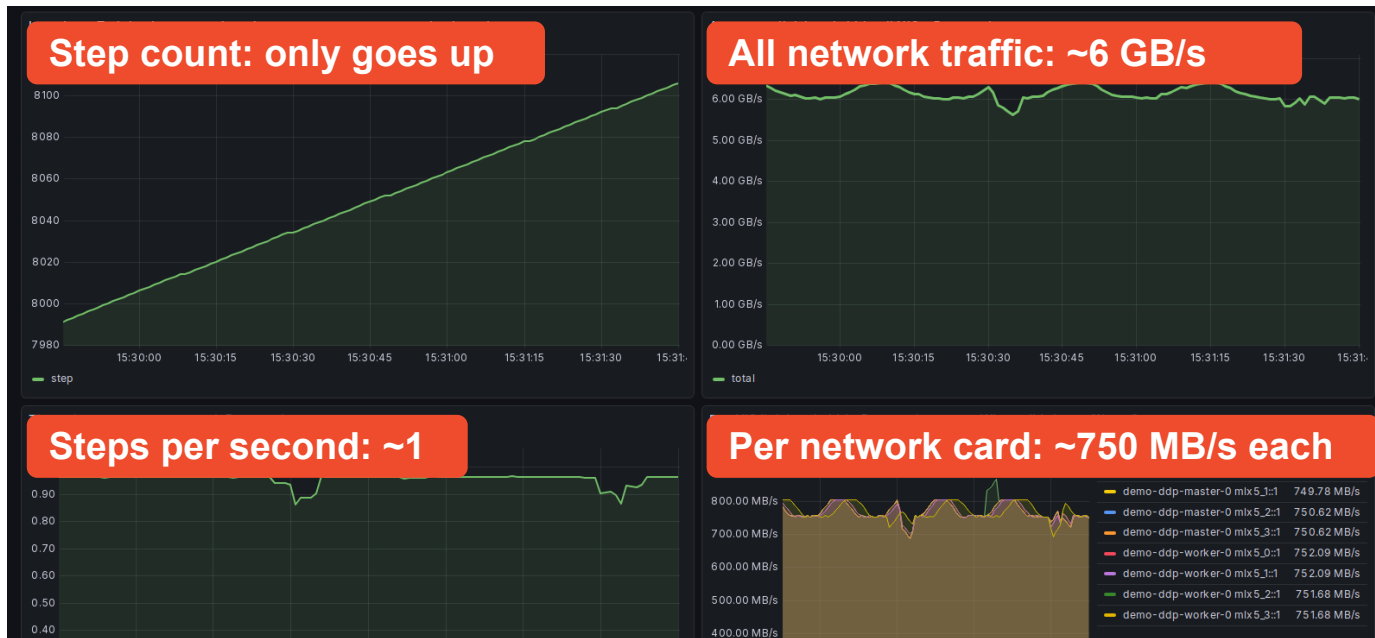
Prevent GPU, node and software failures from crashing training jobs

Training



One job. Two failures. No checkpoint restart.

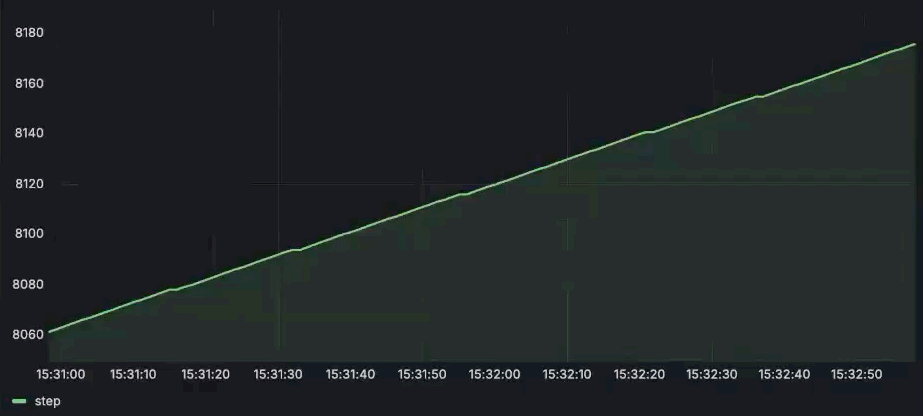
Recorded live on Oct 1 · 2 servers × 4 GPUs, 4 × 100G links each · 1 spare standing by





<< Last 2 minutes PDT >>

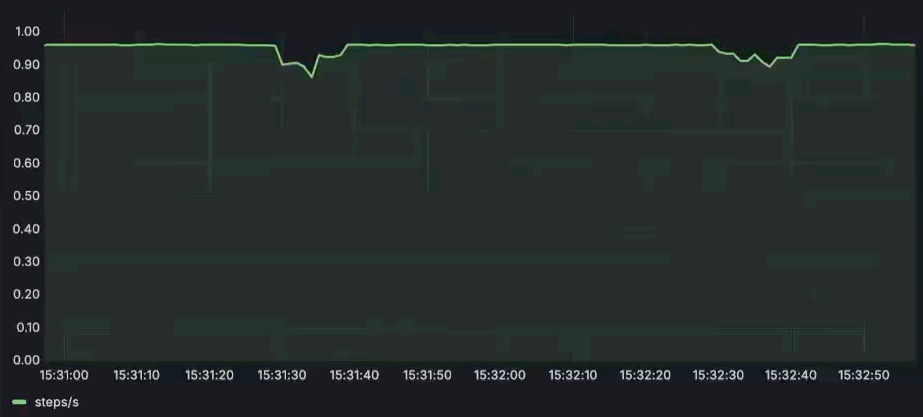
Iterations. Training keeps running: the step counter never goes backwards



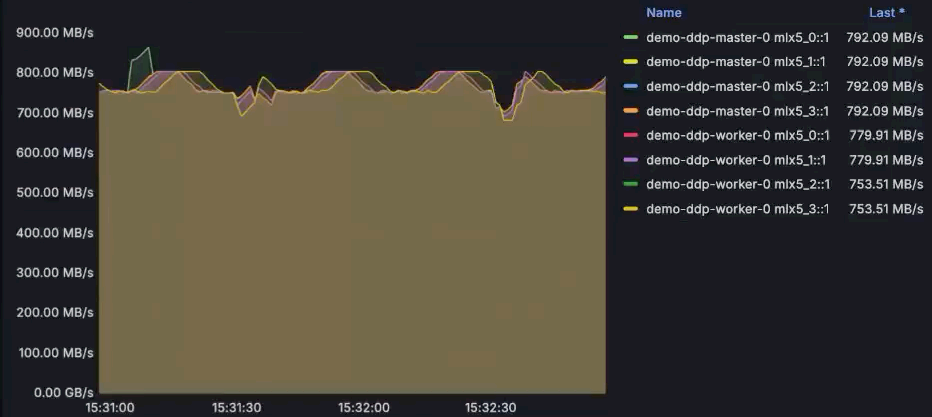
Aggregate link bandwidth, all NICs, 5-second average



Throughput, steps per second, 5-second average



Per-NIC link bandwidth, 5-second average. Where did the traffic go?



What you just saw

Cable pull: LinkPass moved the traffic onto the other three links. Training never paused.

Power pull: half the GPUs gone. No warning, no chance to save.

TorchPass rebuilt the lost ranks on the spare from a surviving data-parallel replica and resumed at the same step*. No checkpoint loaded from disk.

**Needs a replica of every state key.*

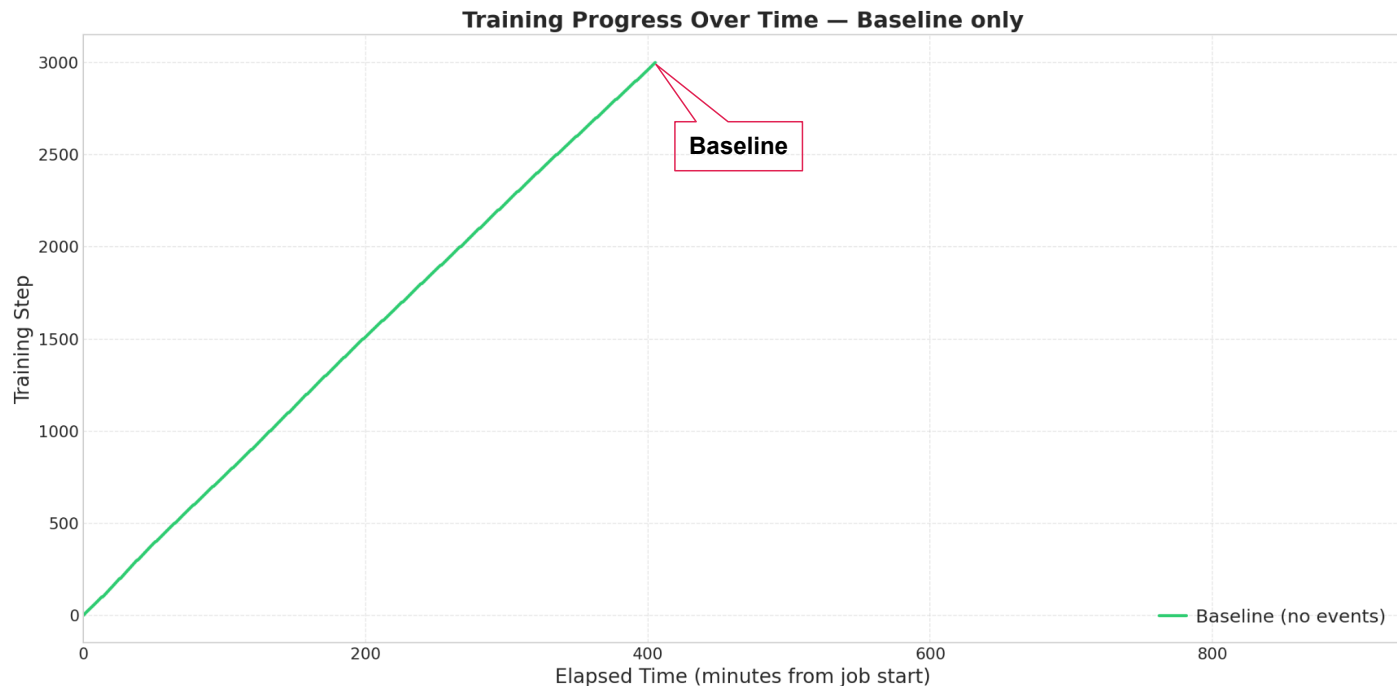
Pull to resume

Under 1 min

Steps lost

0

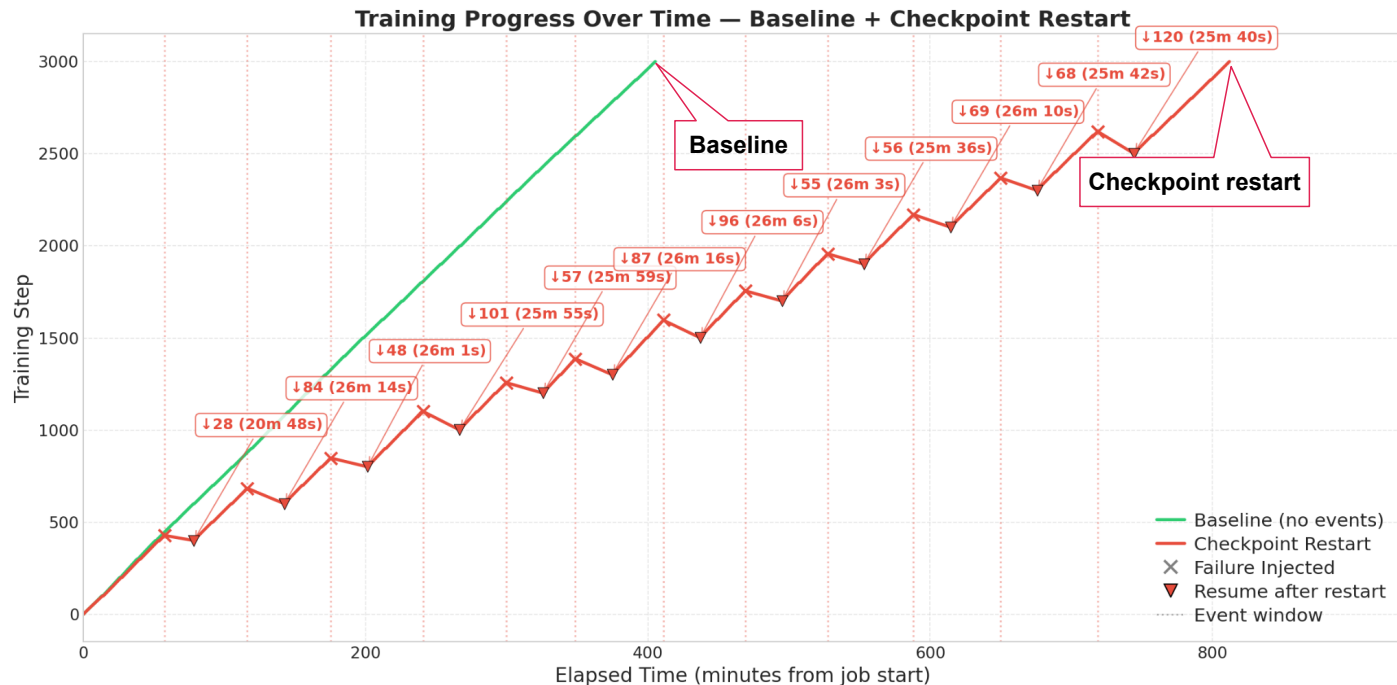
At scale: Llama 4 Scout on 64 H200 GPUs



No failures

411 min

At scale: Llama 4 Scout on 64 H200 GPUs



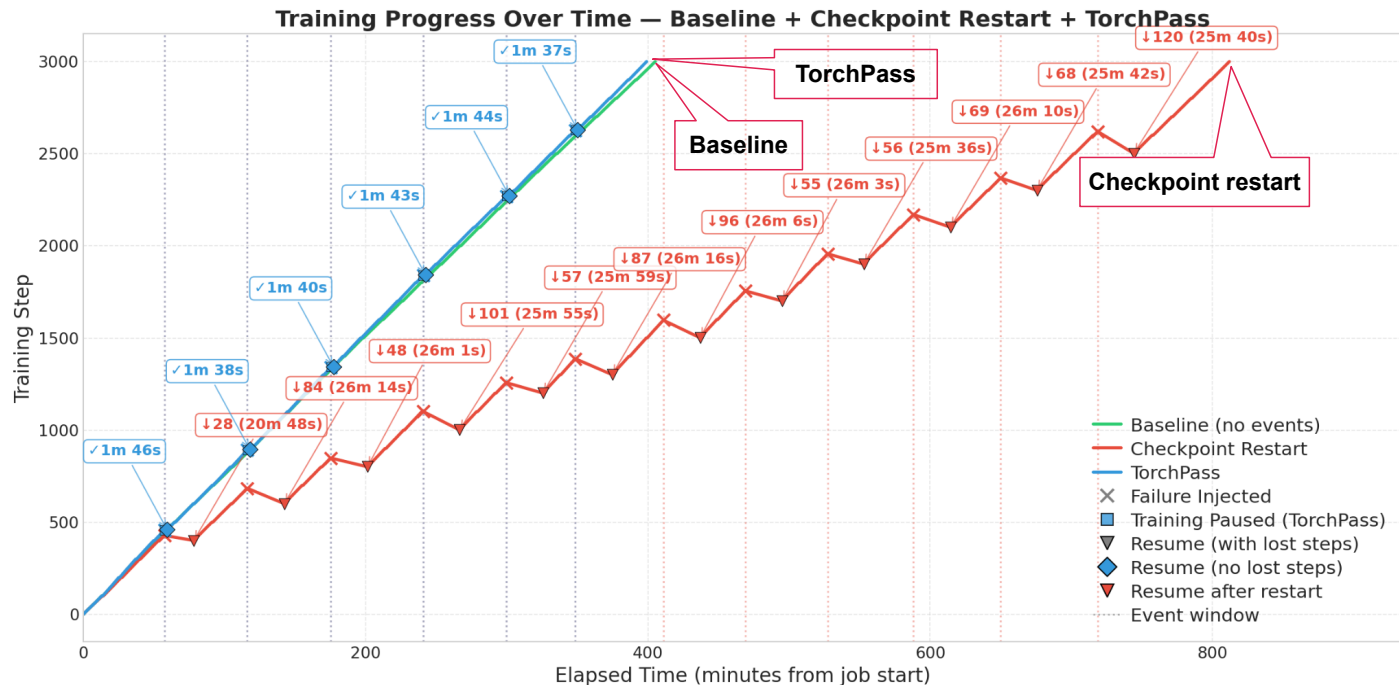
Checkpoint restart

818 min

Steps recomputed

869

At scale: Llama 4 Scout on 64 H200 GPUs



TorchPass

405 min

Steps lost

0

Integration: Deploy at the Platform Layer or Model Layer

Schedulers and Job Managers



kubernetes



slinky



VOLCANO



KAI Scheduler

1 Platform layer

Zero code changes required

Works with any of the platforms above

2 Model layer two options

Out of the box

Common training frameworks



PyTorch

torchtitan

Megatron-LM



DeepSpeed

and more to come

Custom

3 tightly scoped changes

- 1 `from torchpass import Manager`
- 2 `Define get_state / set_state`
- 3 `Define crit_region`

Never hit zero.

Come find us:

Clockwork · Booth P8

- ▶ Live demos: LinkPass, TorchPass, FleetAudit, FleetLens
- ▶ The engineers who built LinkPass and TorchPass
- ▶ Book a 1:1 or a TorchPass demo →



Book a 1:1

CLOCKWORK.io