AI-generated images

⛔ Visibility

AI-generated images

Lack of fine-grained visibility at queue pair level

Pre-race & race time!

⛔ Visibility

AI-generated images

Lack of fine-grained visibility at queue pair level

Pre-race & race time!

⛔ Visibility

⛔ Reliability

AI-generated images

Lack of fine-grained visibility at queue pair level

Pre-race & race time!

⛔ Visibility

NIC/link flapping

⛔ Reliability

AI-generated images

Turbocharging AI/ML Workloads

Revving up Speed 💨 and Resilience 💪!

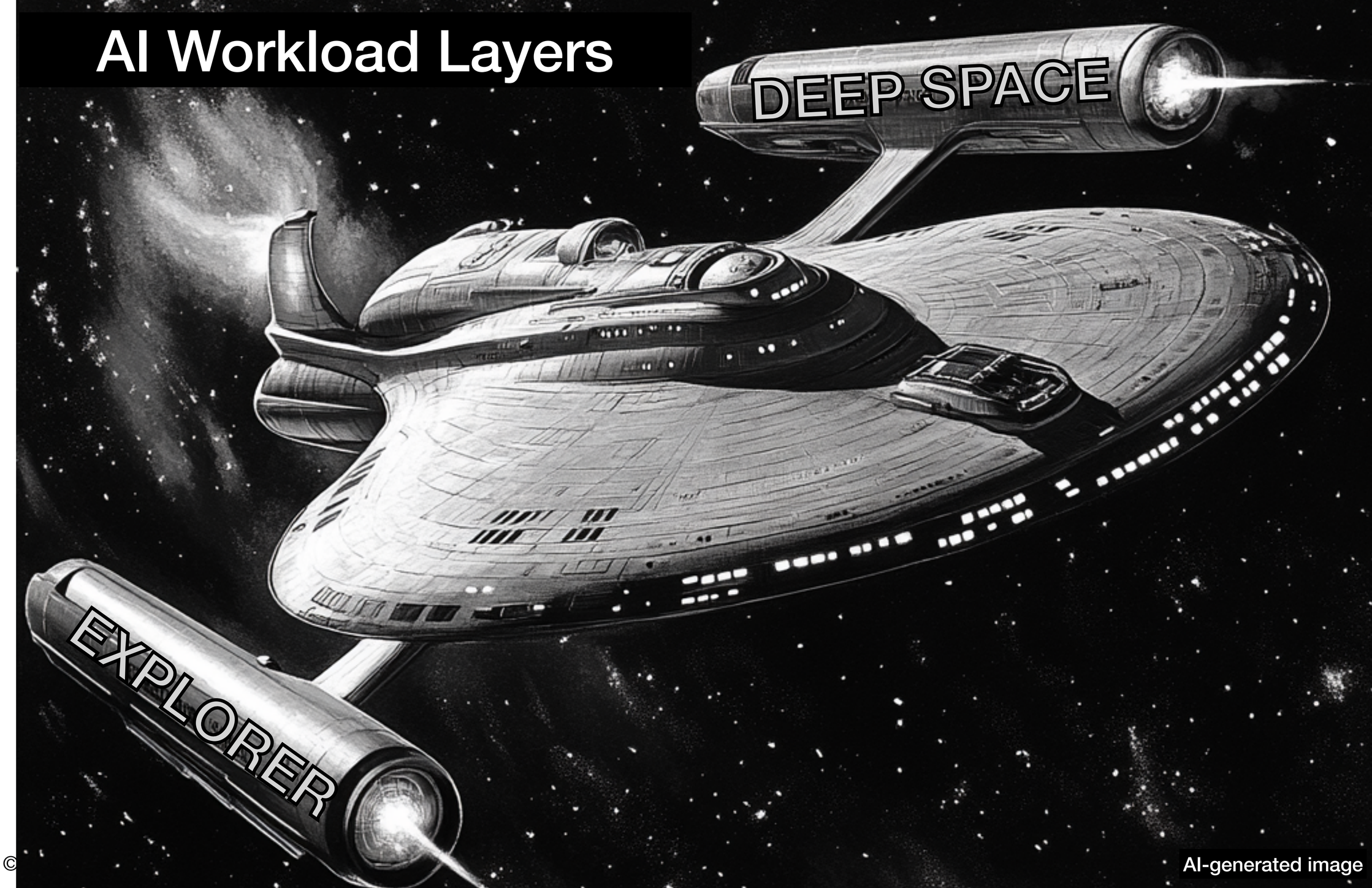
Lerna Ekmekcioglu
Sr. Solutions Engineer, Clockwork Systems

Turbocharging AI/ML Workloads

Revving up Speed 💨 and Resilience 💪!

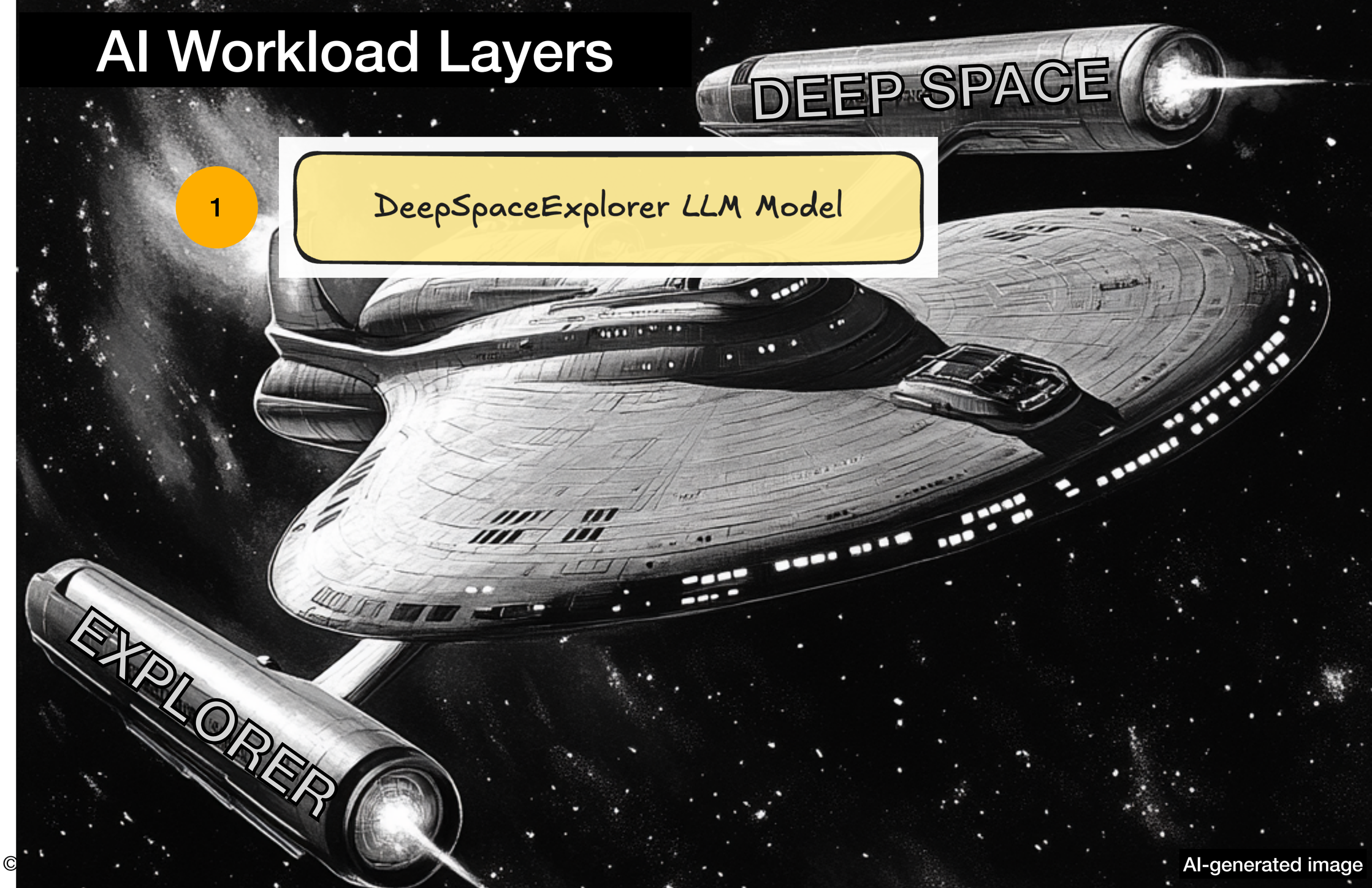
Lerna Ekmekcioglu
Sr. Solutions Engineer, Clockwork Systems

AI Workload Layers



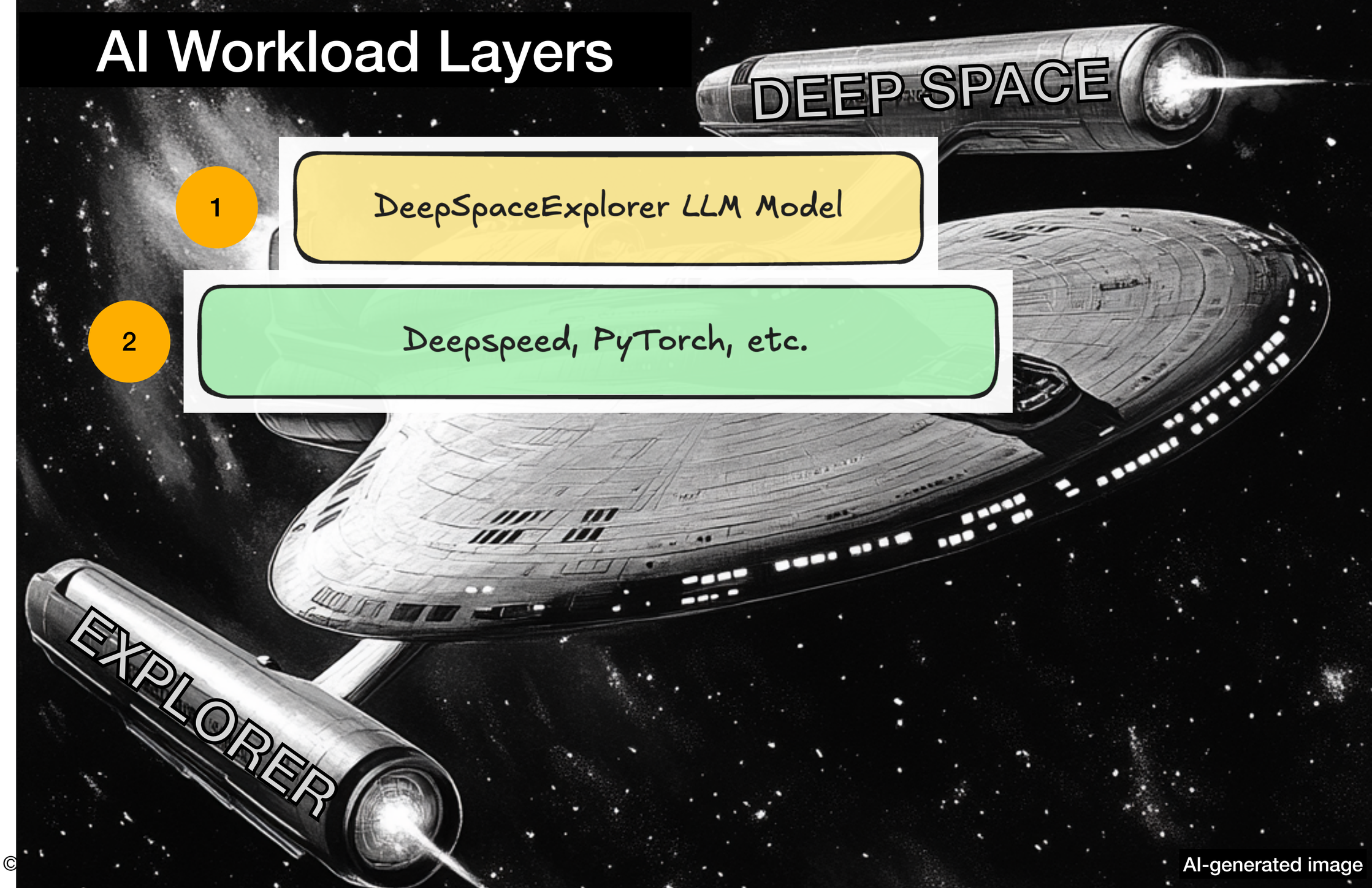
AI-generated image

AI Workload Layers



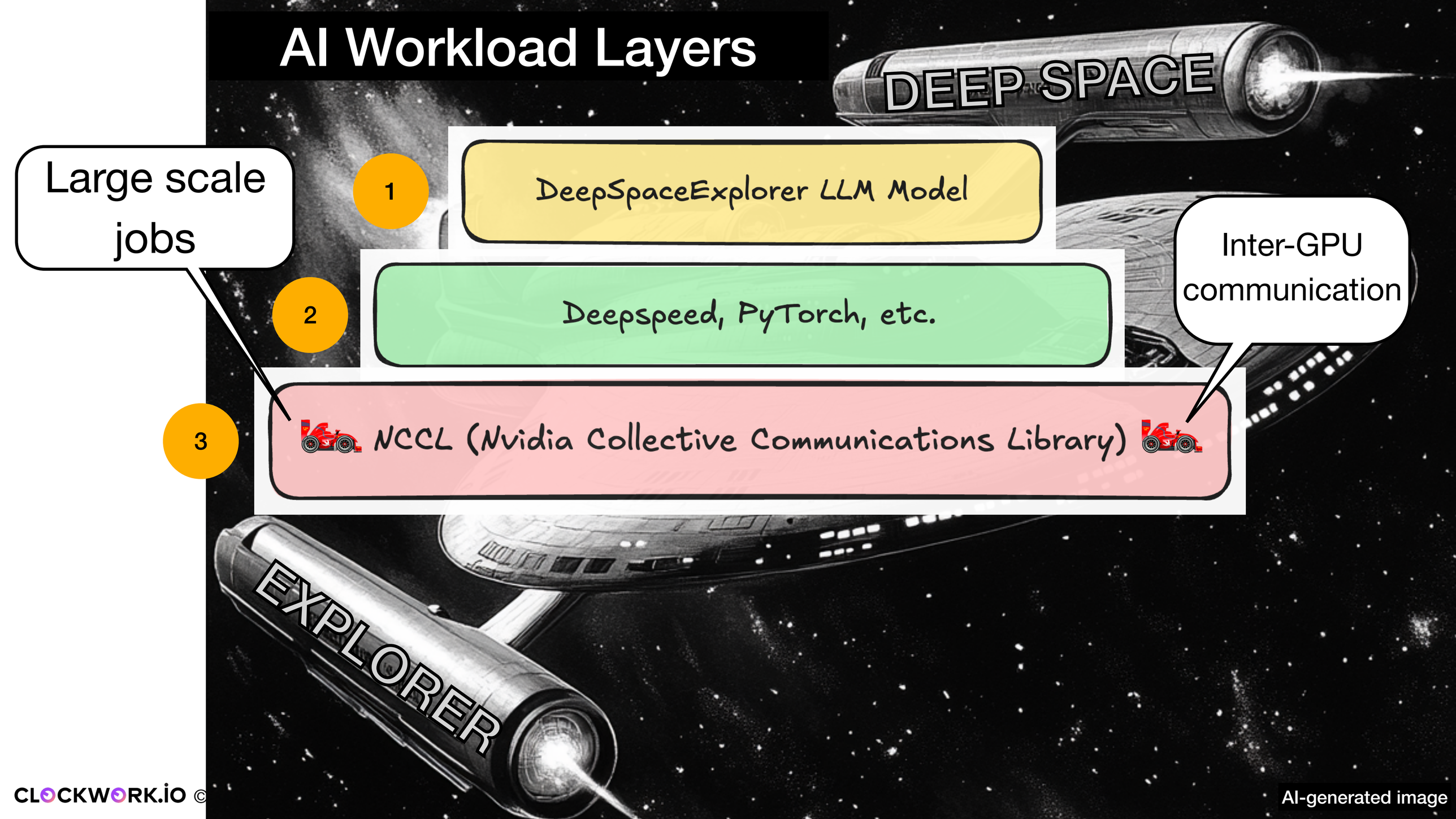
AI-generated image

AI Workload Layers



AI-generated image

AI Workload Layers



AI-generated image

AI Workload Layers

1. DeepSpaceExplorer LLM Model
2. Deepspeed, PyTorch, etc.
3. NCCL (Nvidia Collective Communications Library)
4. Network Device (Infiniband or RoCE)

AI-generated image

Traditional networks

AI networks

AI-generated images

Traditional networks

AI networks

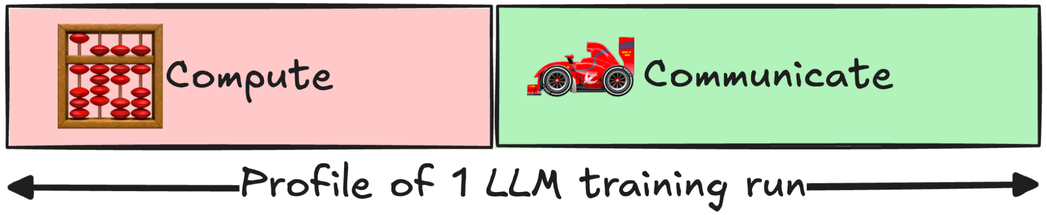
AI-generated images

Traditional networks

AI networks

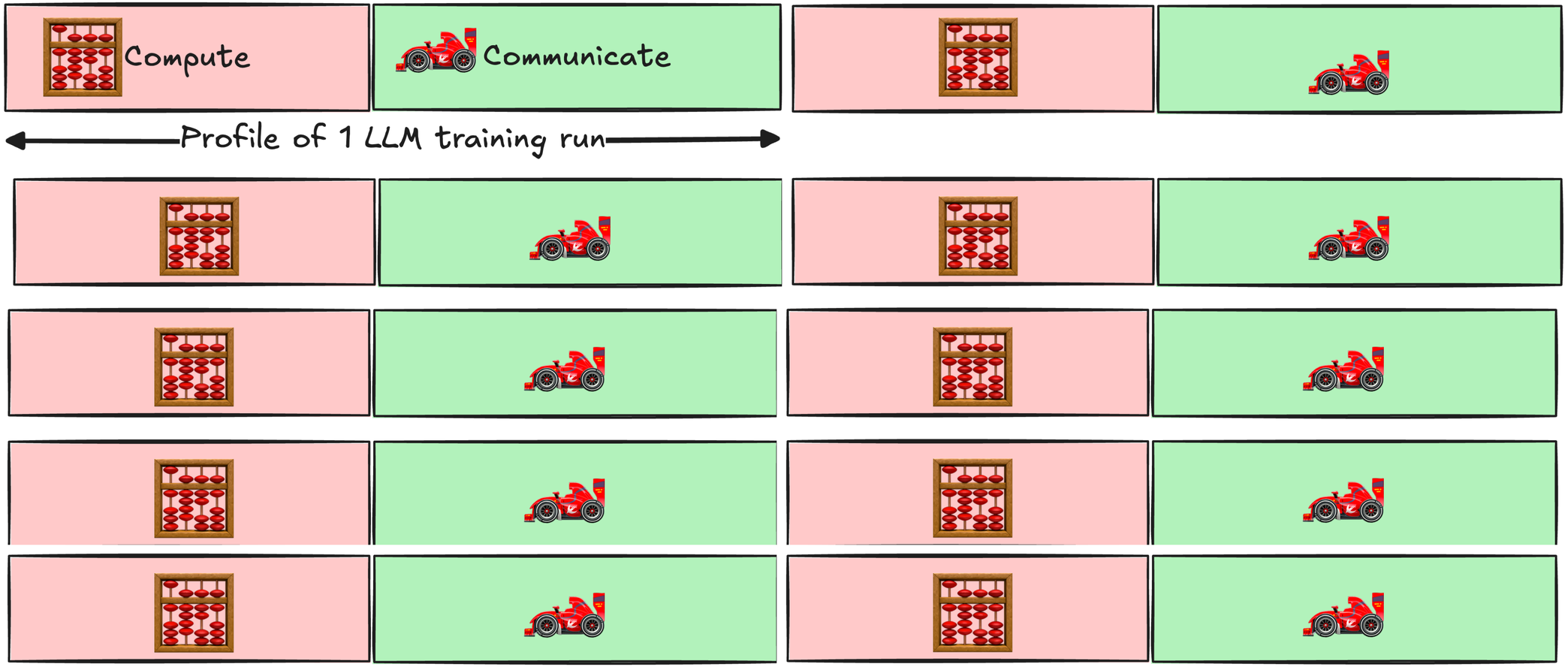
AI-generated images

Profile of an LLM training job

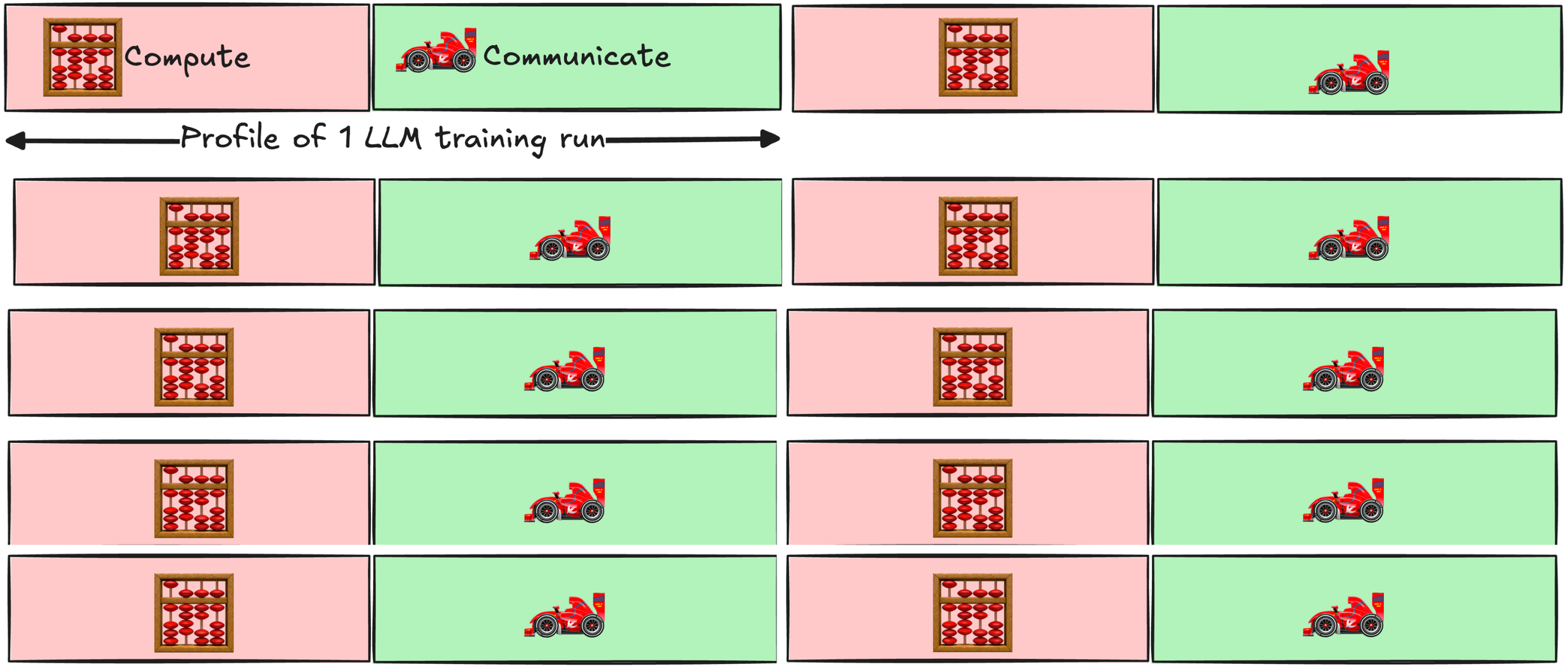


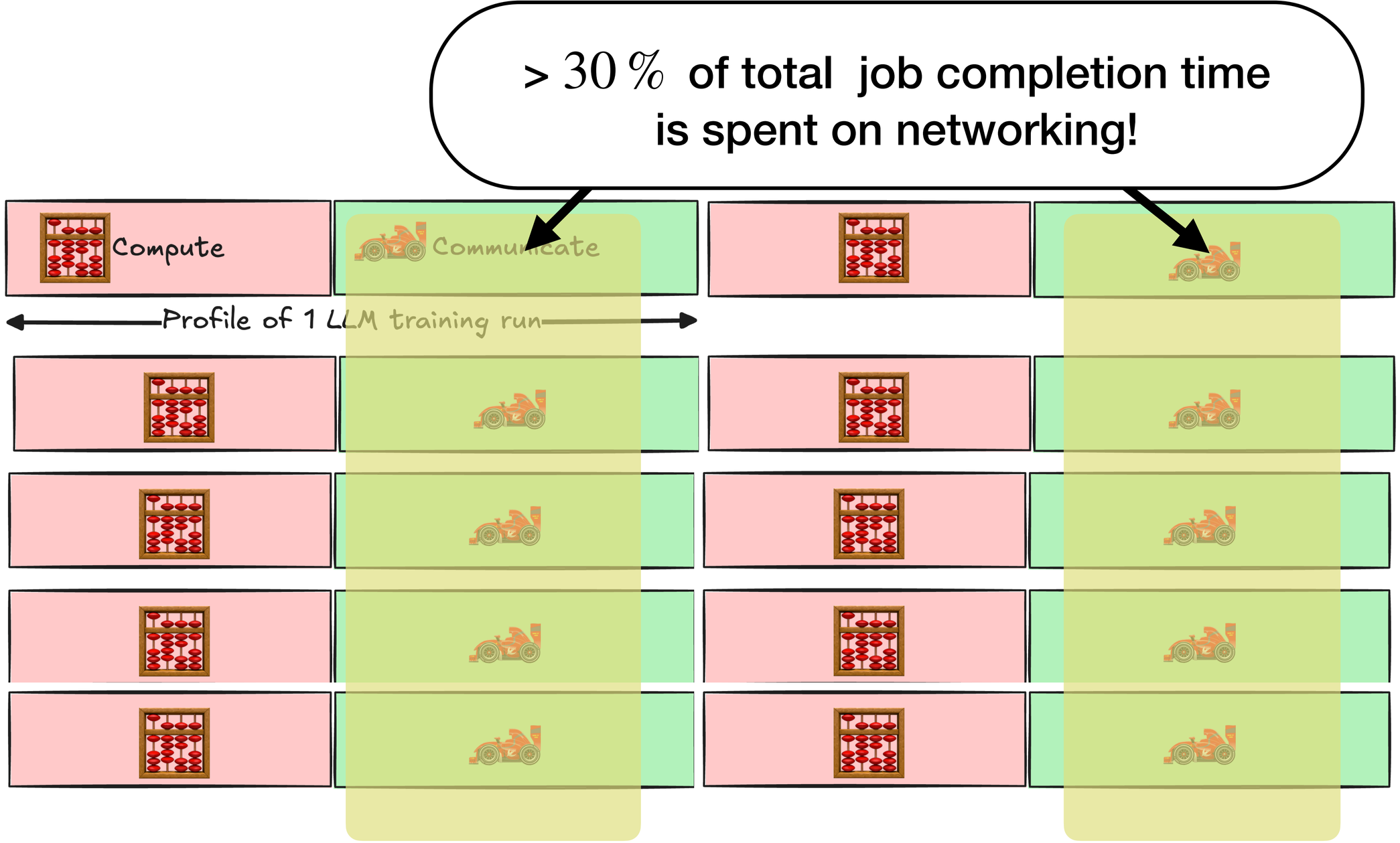
Source: Nvidia

Profile of an LLM training job



Source: Nvidia

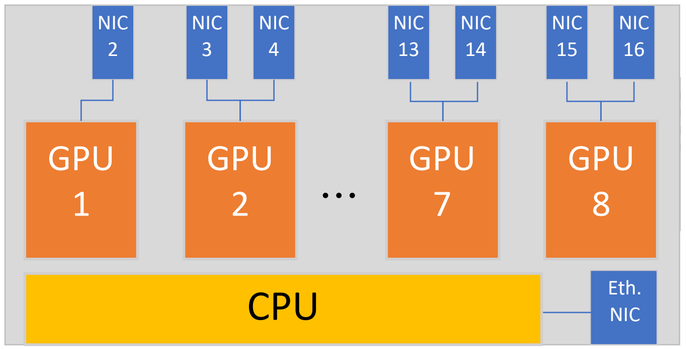


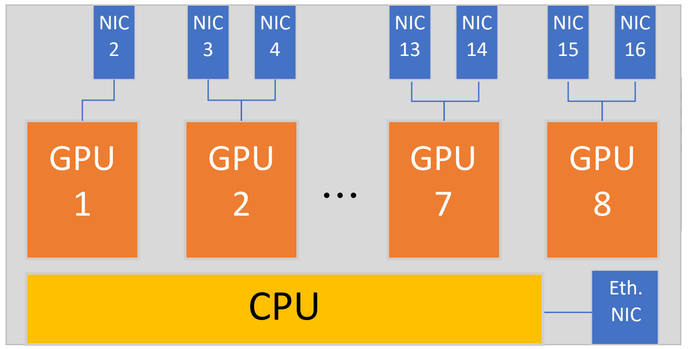


Source: AMD

Under the hood: NCCL or other communication library

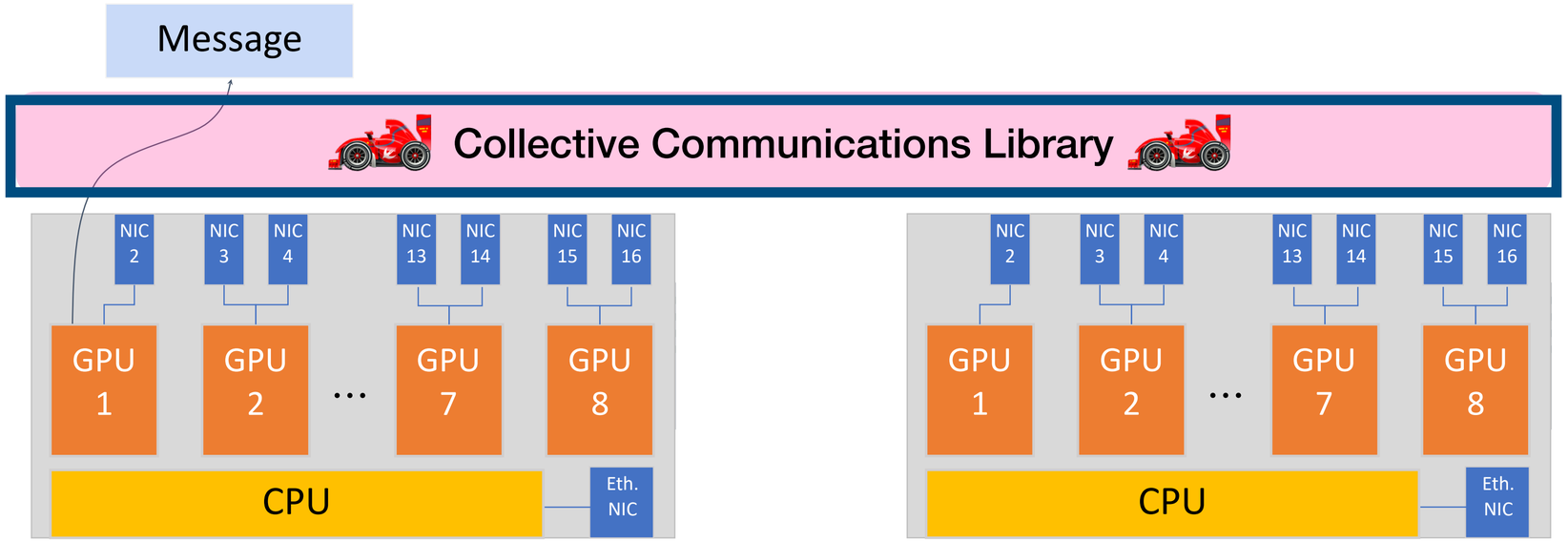
Inter-GPU communication





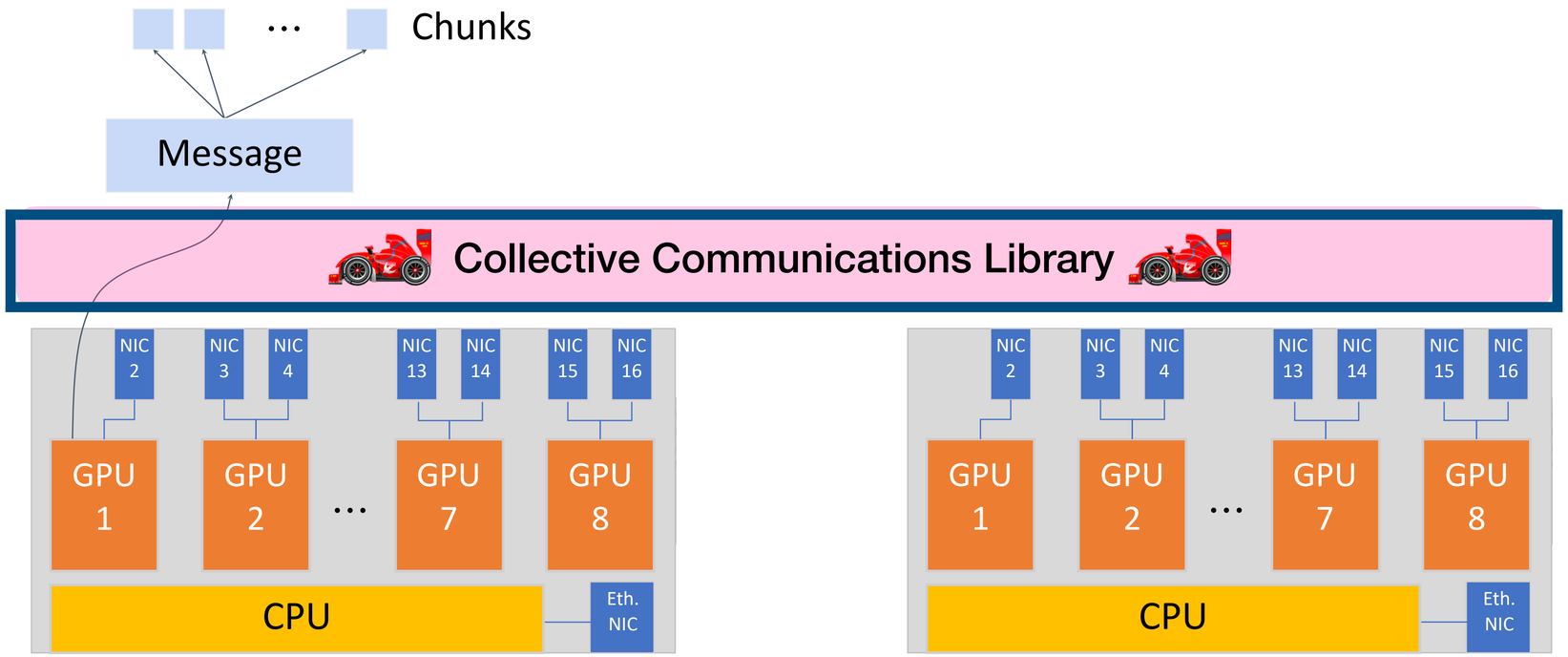
Under the hood: NCCL or other communication library

Inter-GPU communication



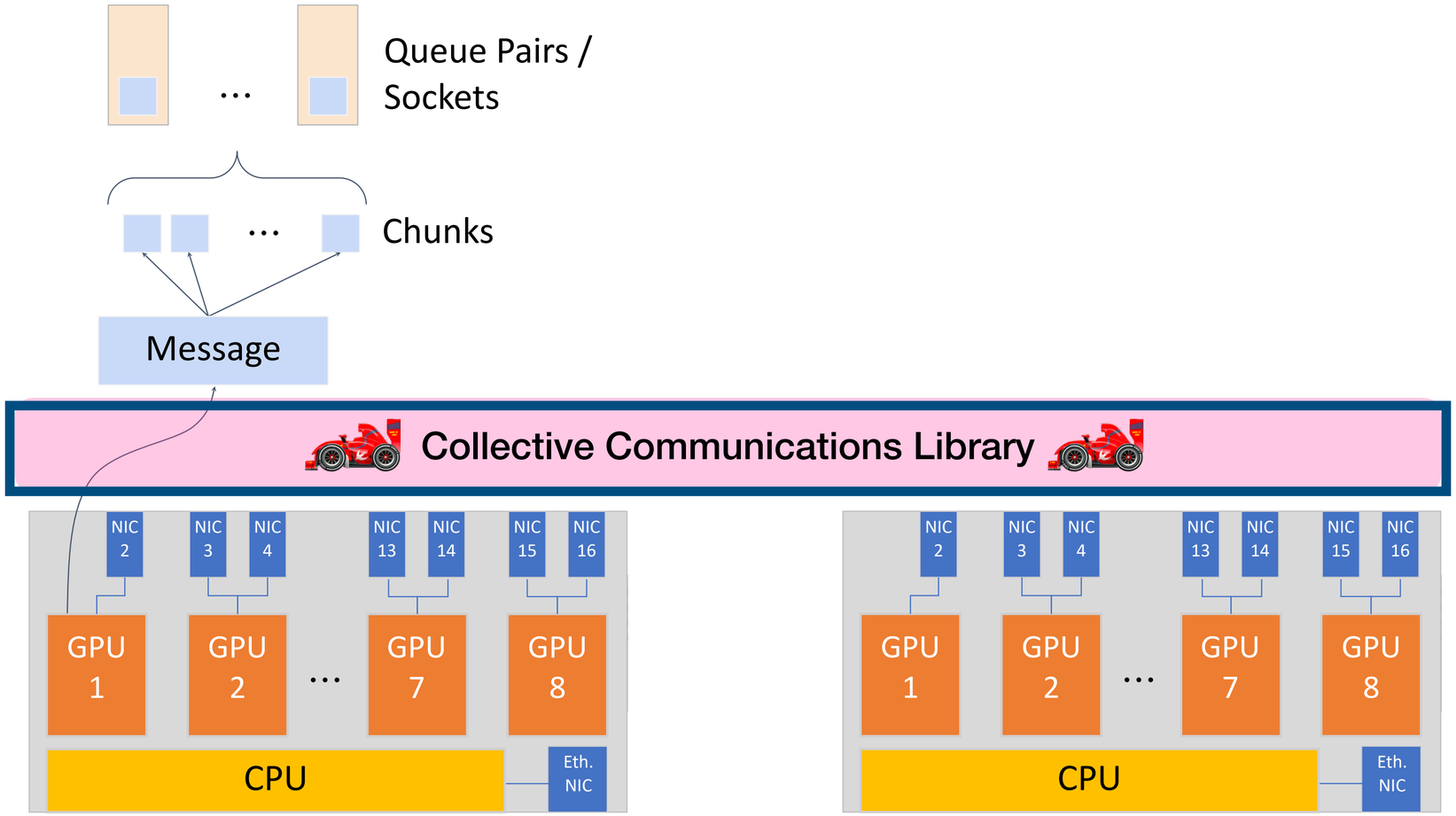
Under the hood: NCCL or other communication library

Inter-GPU communication

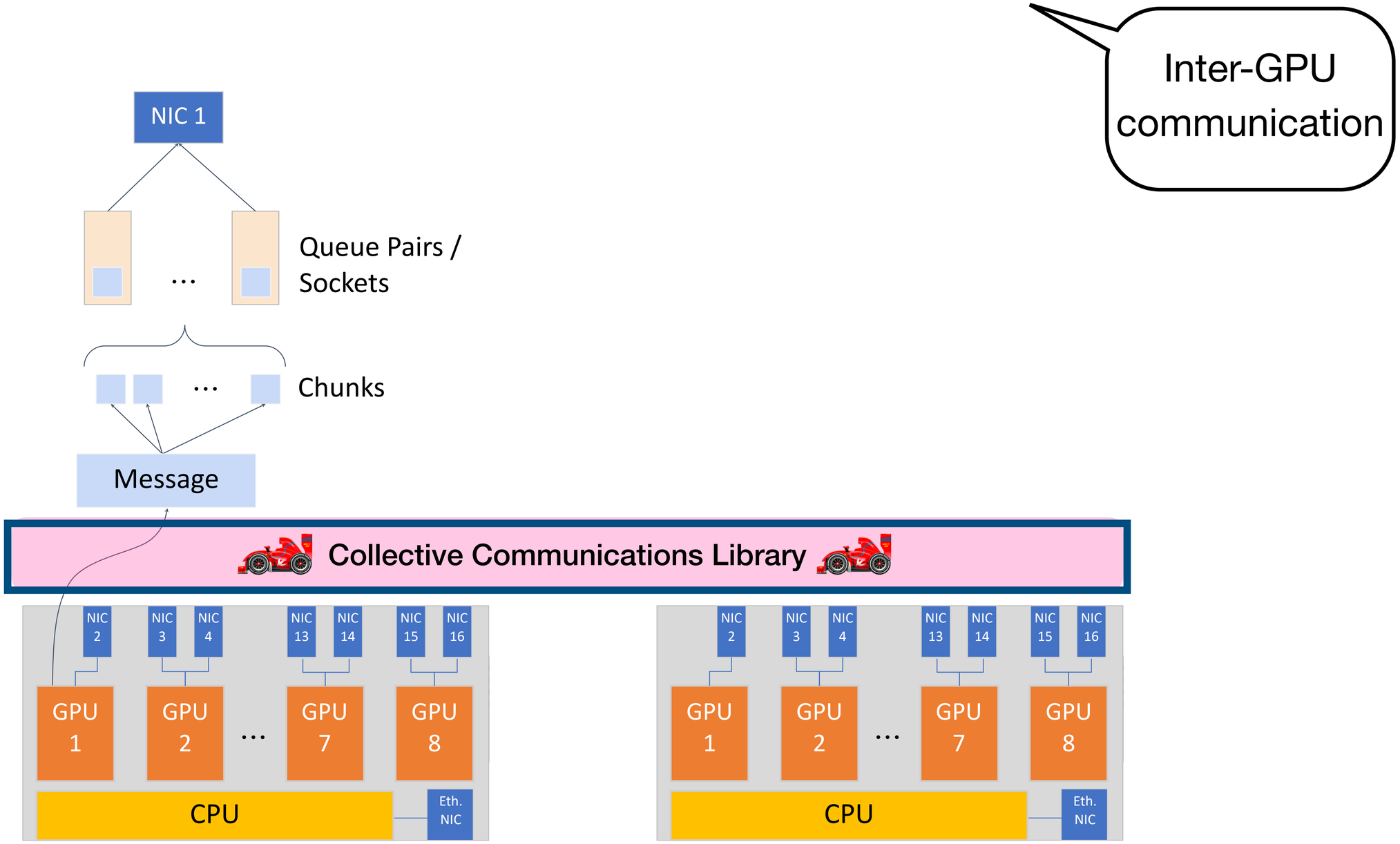


Under the hood: NCCL or other communication library

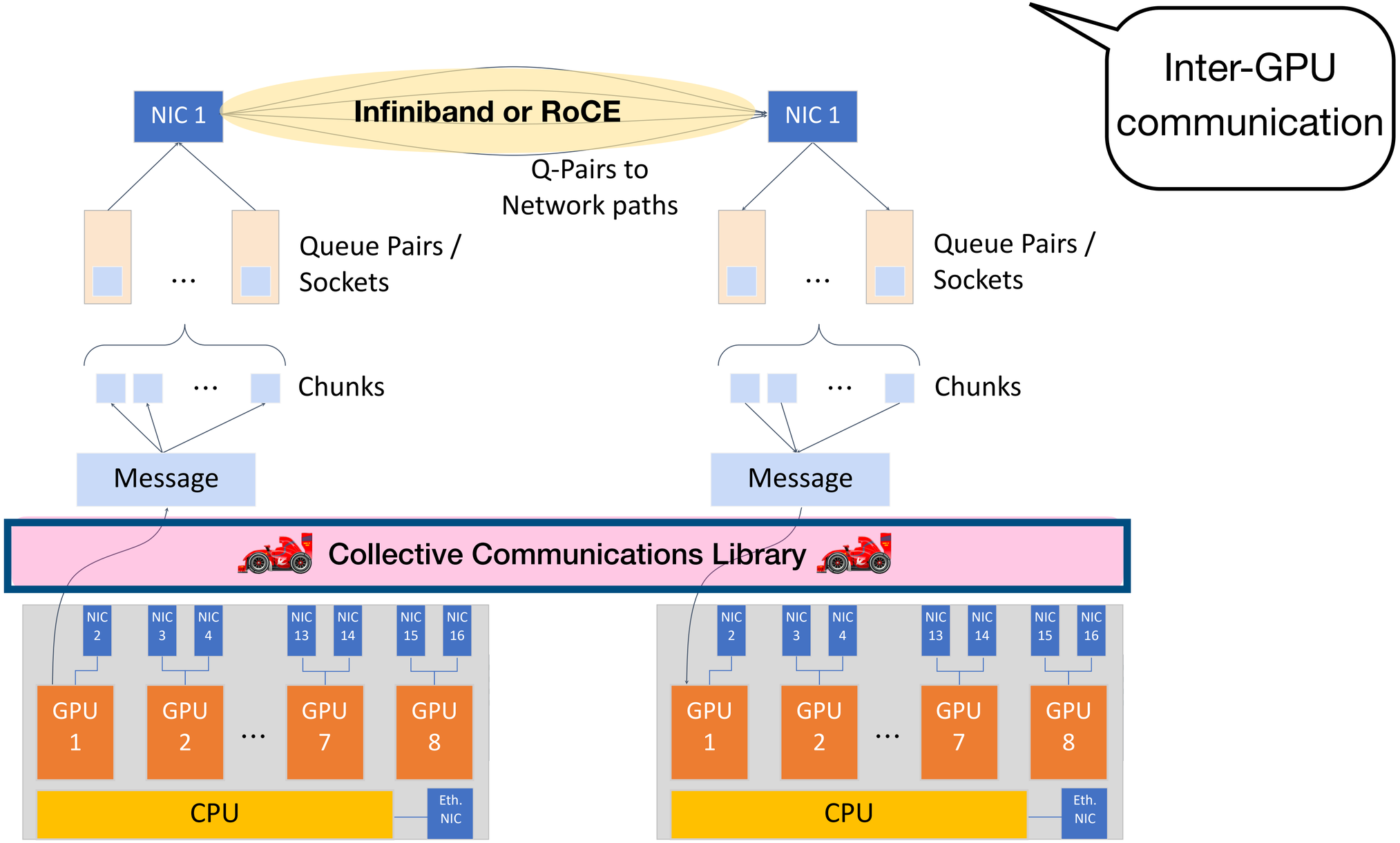
Inter-GPU communication

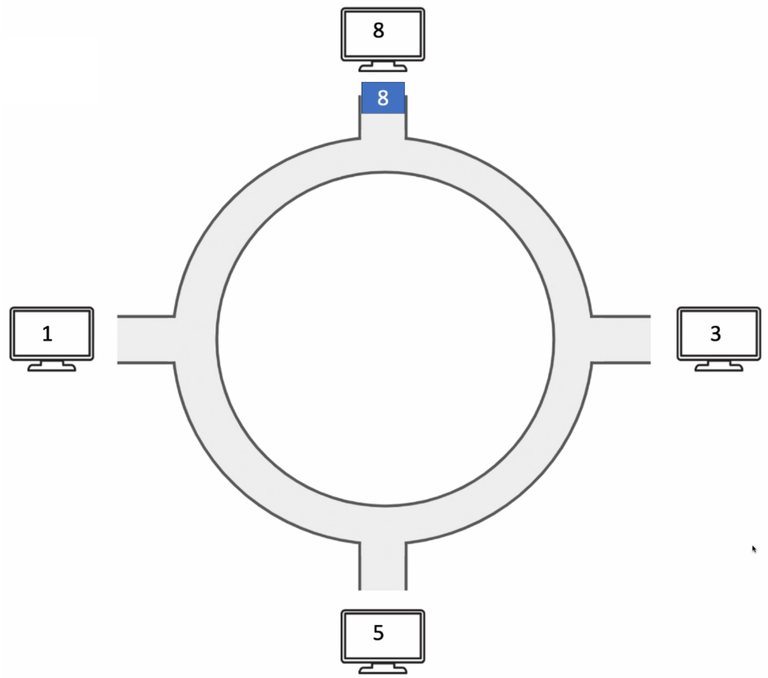


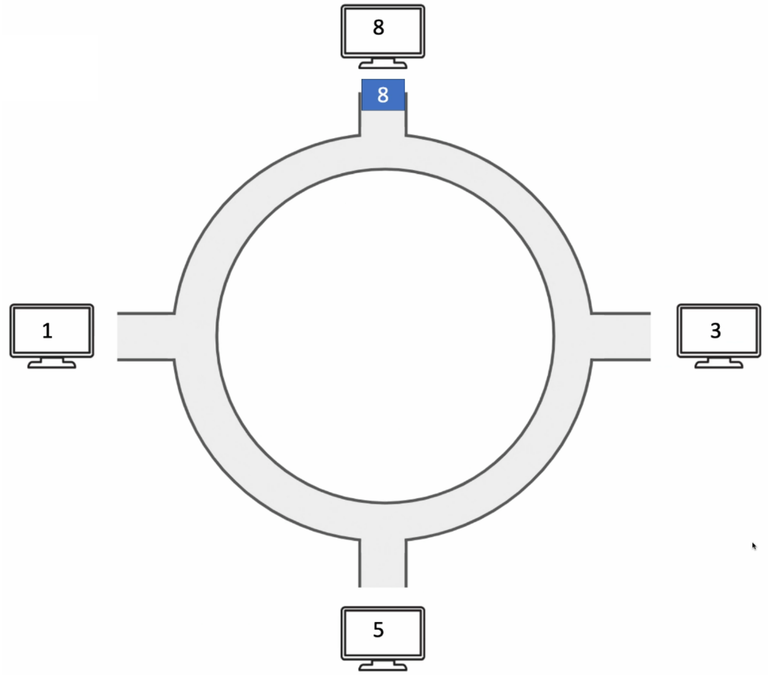
Under the hood: NCCL or other communication library



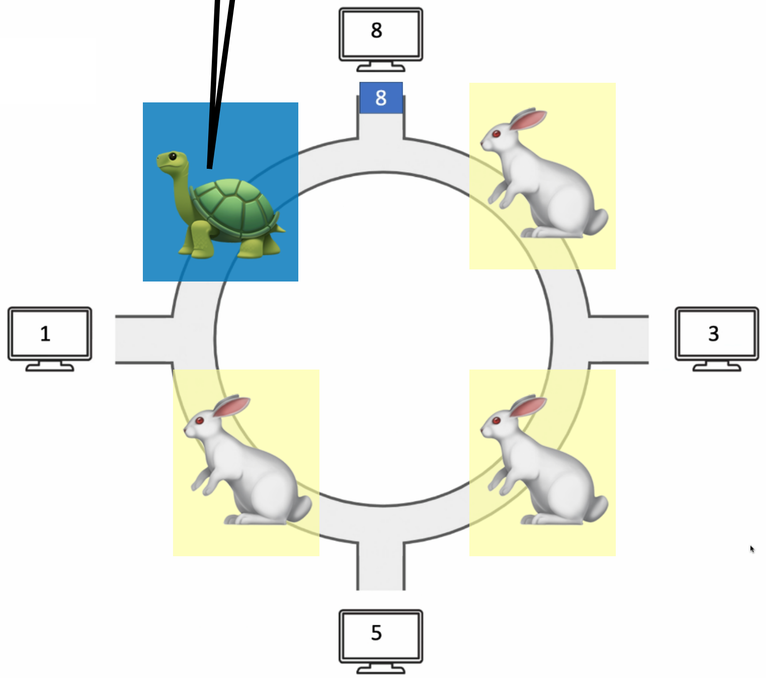
Under the hood: NCCL or other communication library







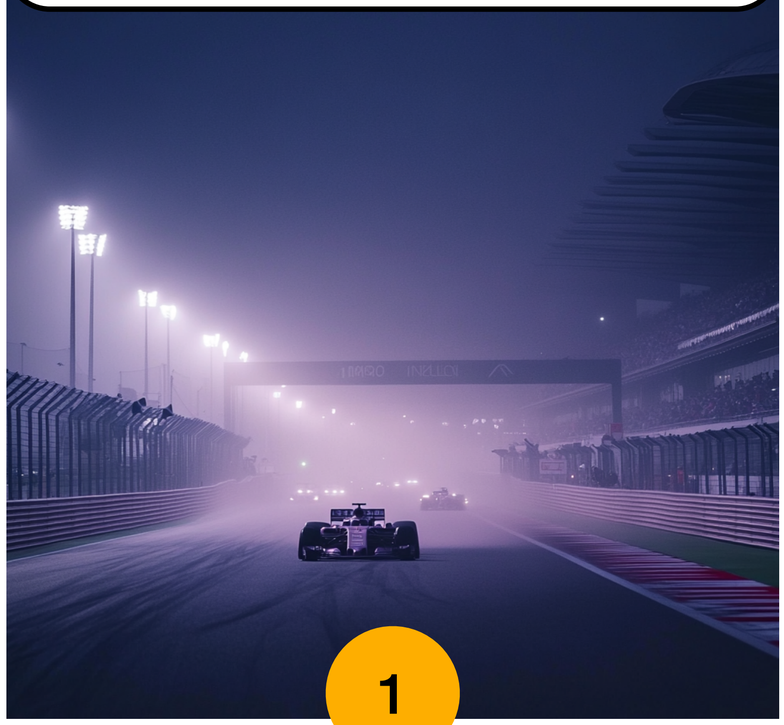
Slowest flow (or node) determines throughput.



Slowest team member determines speed.

Networking challenges of AI/ML workloads

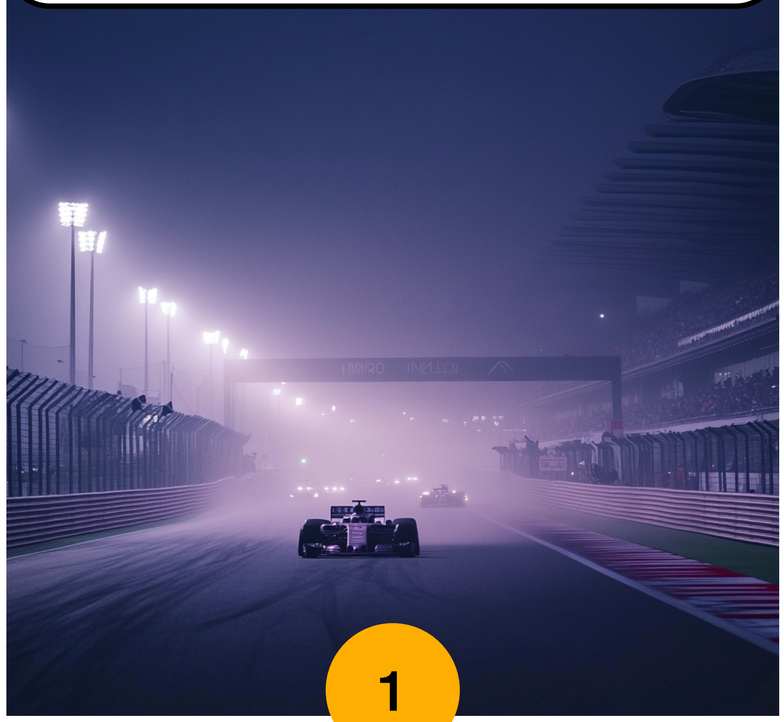
⛔ Visibility



AI-generated images

Networking challenges of AI/ML workloads

⛔ Visibility



1

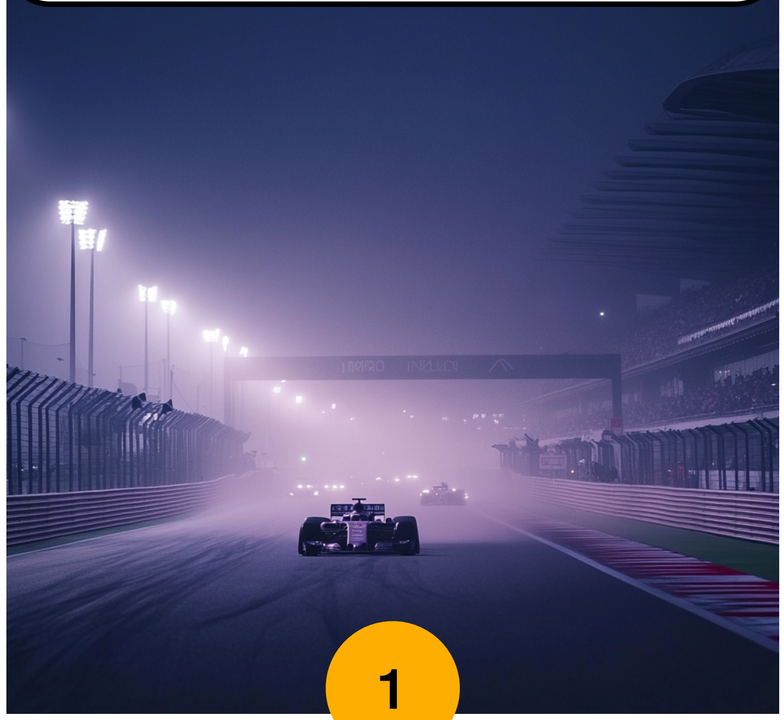
⛔ Reliability

2

AI-generated images

Networking challenges of AI/ML workloads

⛔ Visibility

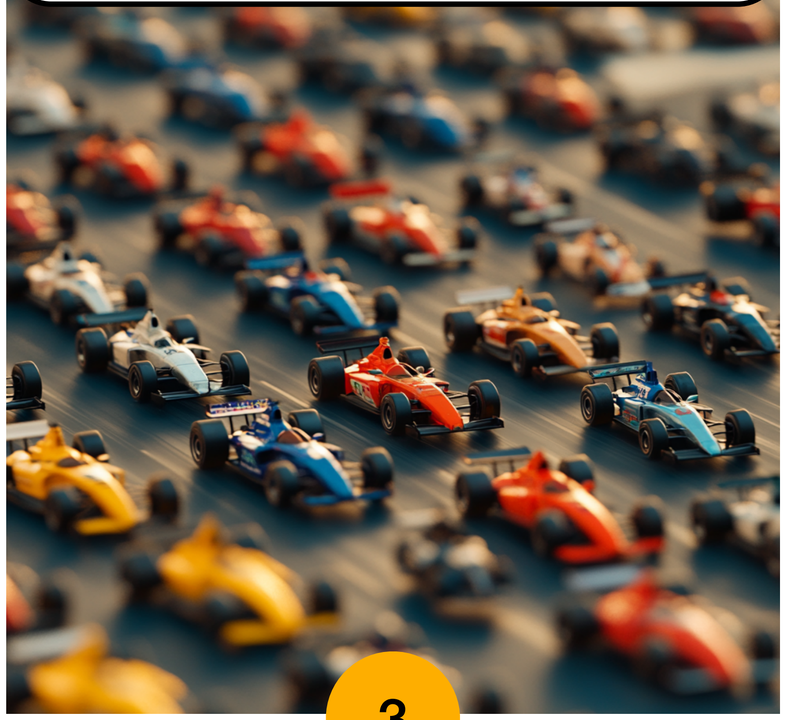


1

⛔ Reliability

2

⛔ Performance



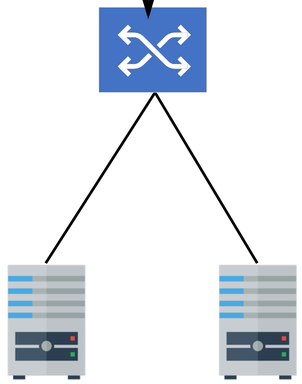
3

AI-generated images

1 Lack of visibility

🚨Alert

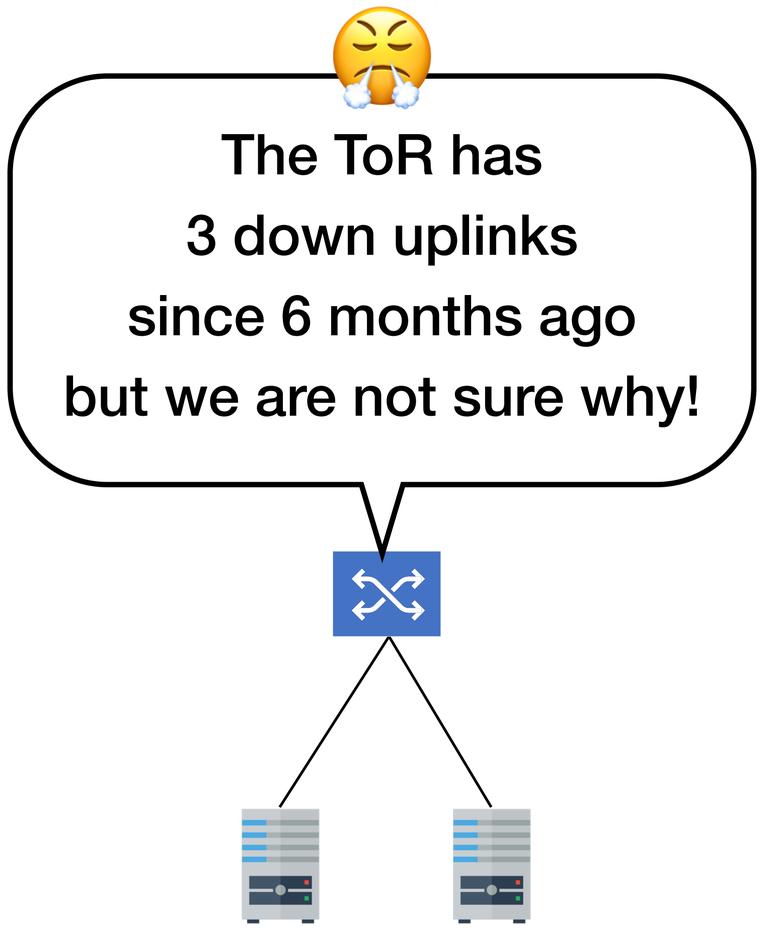
The ToR has
3 down uplinks
since 6 months ago
but we are not sure why!



AI-generated images

1 Lack of visibility

🚨Alert



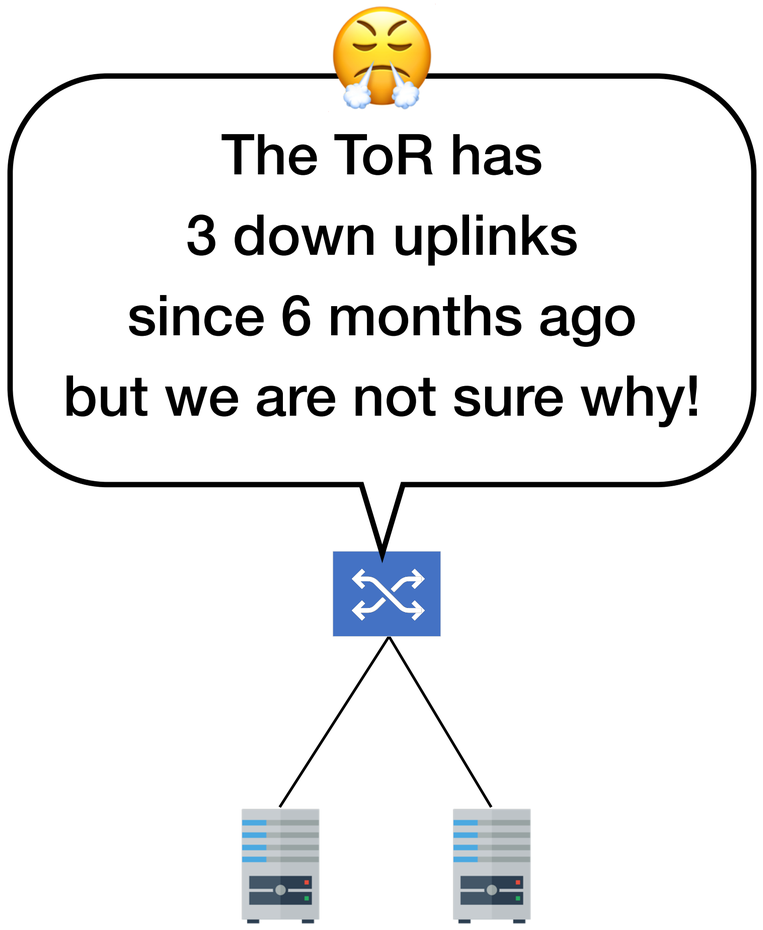
Why?

- Aggregate level metrics
- No fine grained instrumentation
- No visibility down to the queue pair level

AI-generated images

1 Lack of visibility

🚨Alert



Why?

- Aggregate level metrics
- No fine grained instrumentation
- No visibility down to the queue pair level

So what?

- Unable to identify root cause to resolve issues quickly.
- Not only at fleet level but at workload level

AI-generated images

Visibility

Huygens clock synchronization

- Software based
- High precision at scale 🎯

https://www.usenix.org/conference/nsdi18/presentation/geng

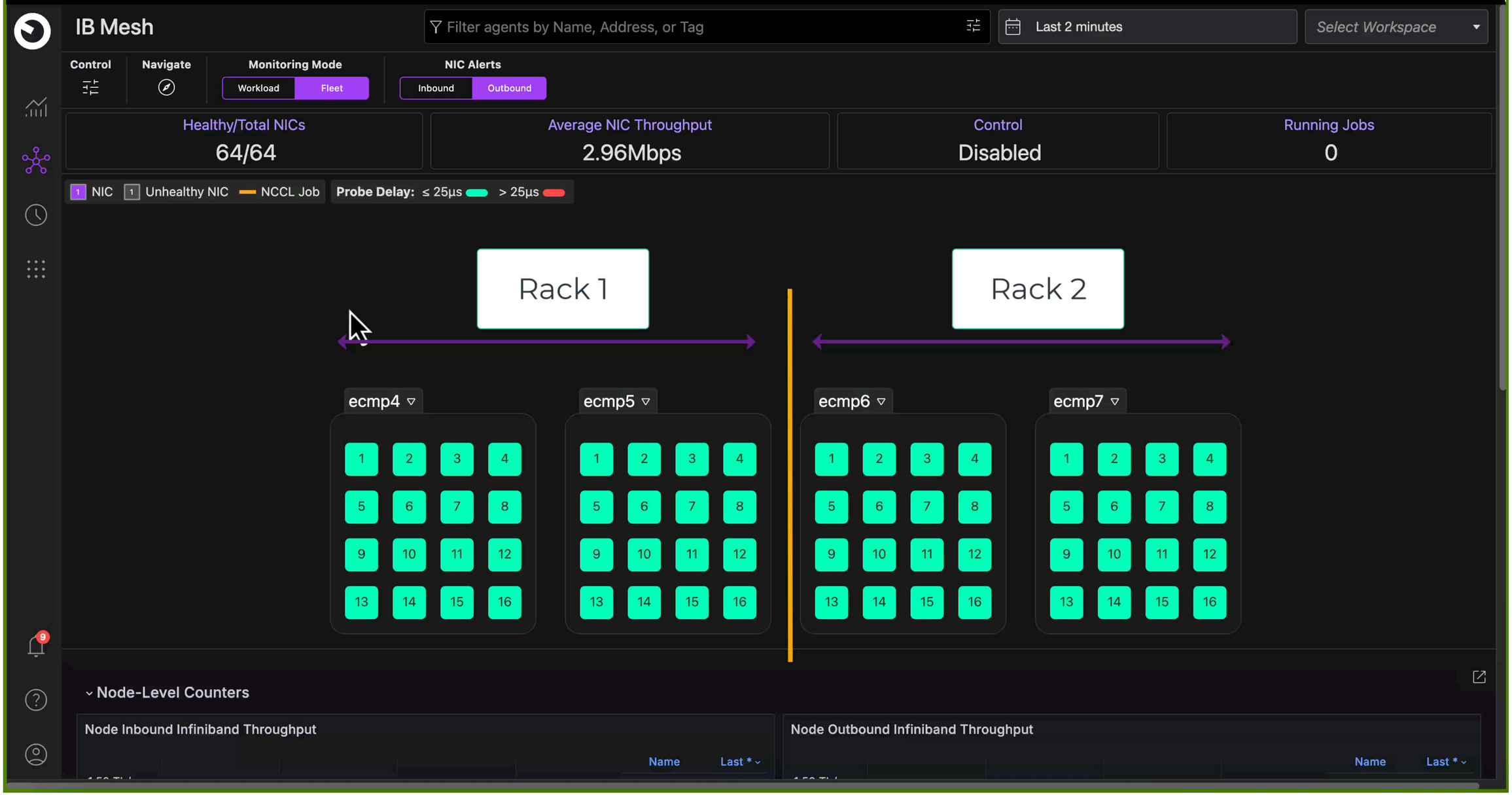
Huygens clock synchronization

- Software based
- High precision at scale 🎯

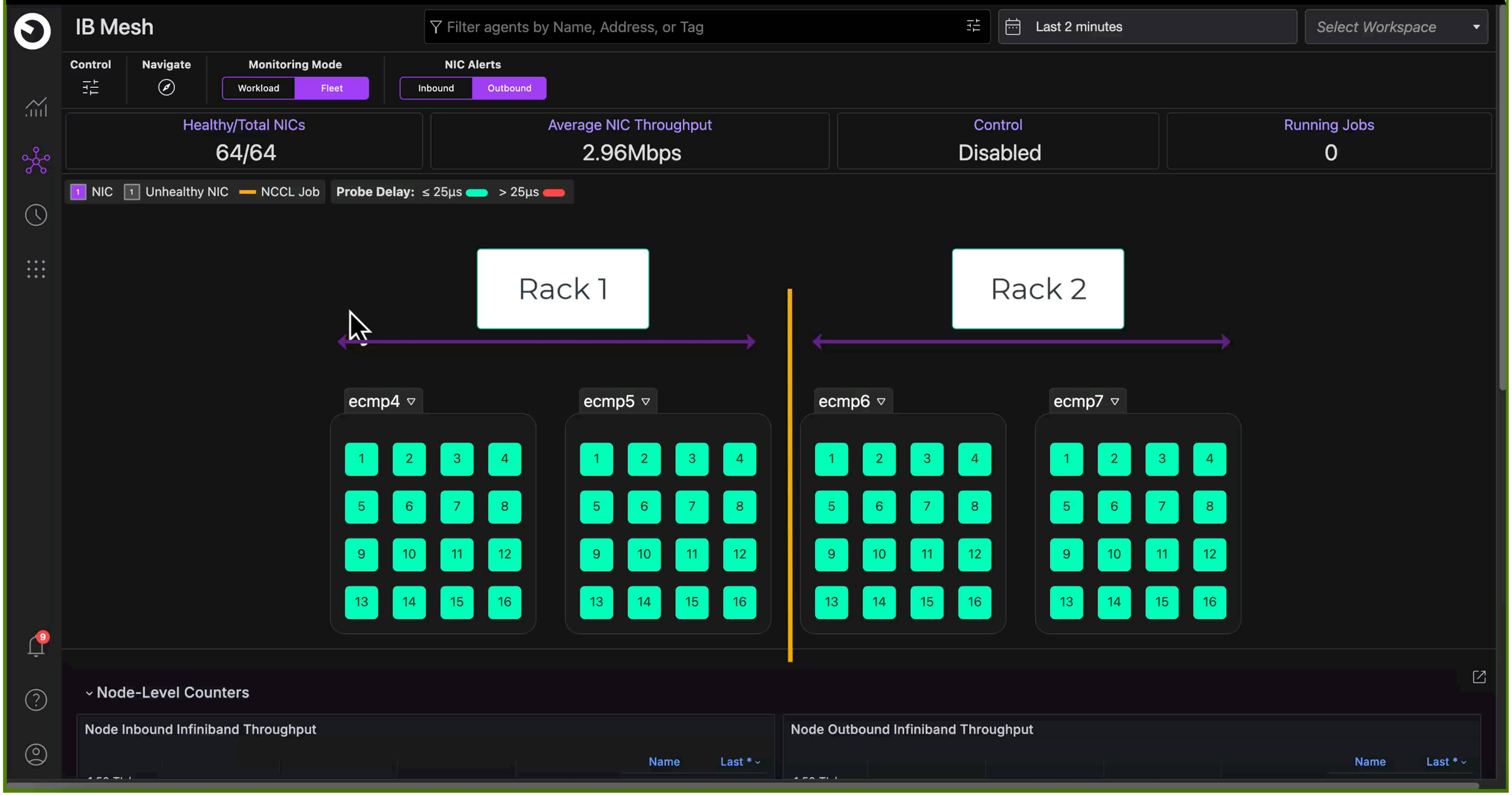
Measures
one way delays

https://www.usenix.org/conference/nsdi18/presentation/geng

Visibility demo



Visibility demo

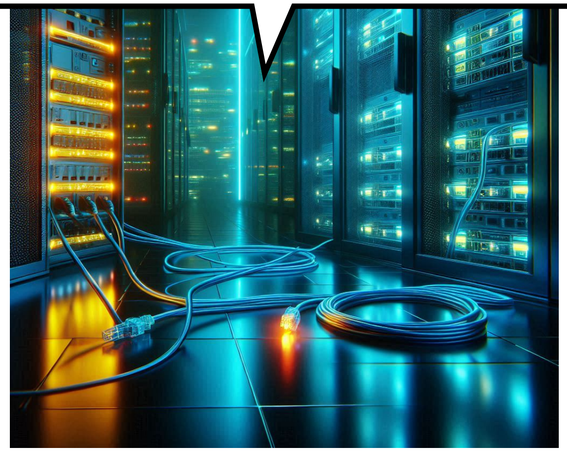


2 Lack of reliability

🚨 Alert

😤

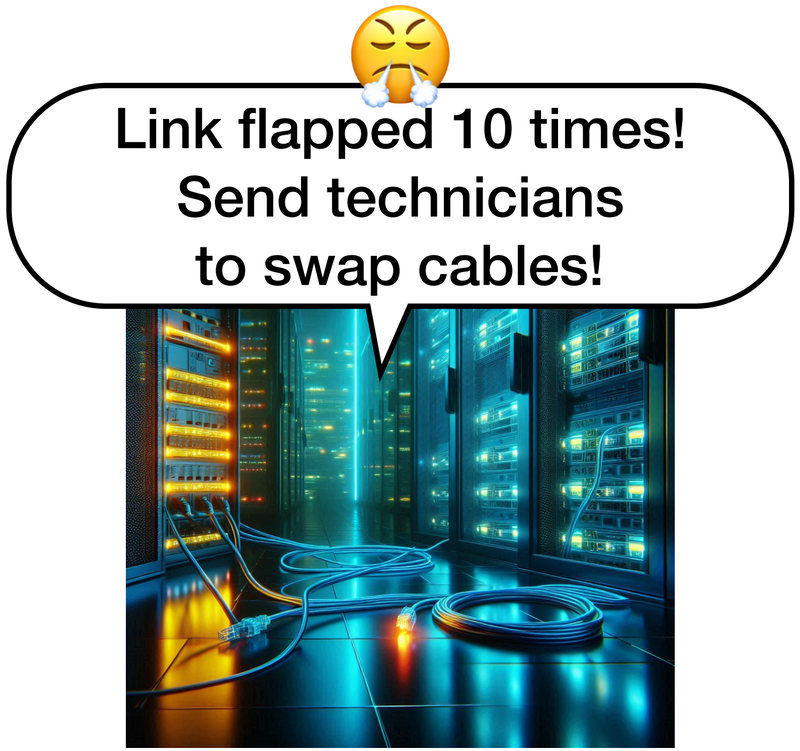
Link flapped 10 times!
Send technicians
to swap cables!



AI-generated images

2 Lack of reliability

🚨 Alert



Why?

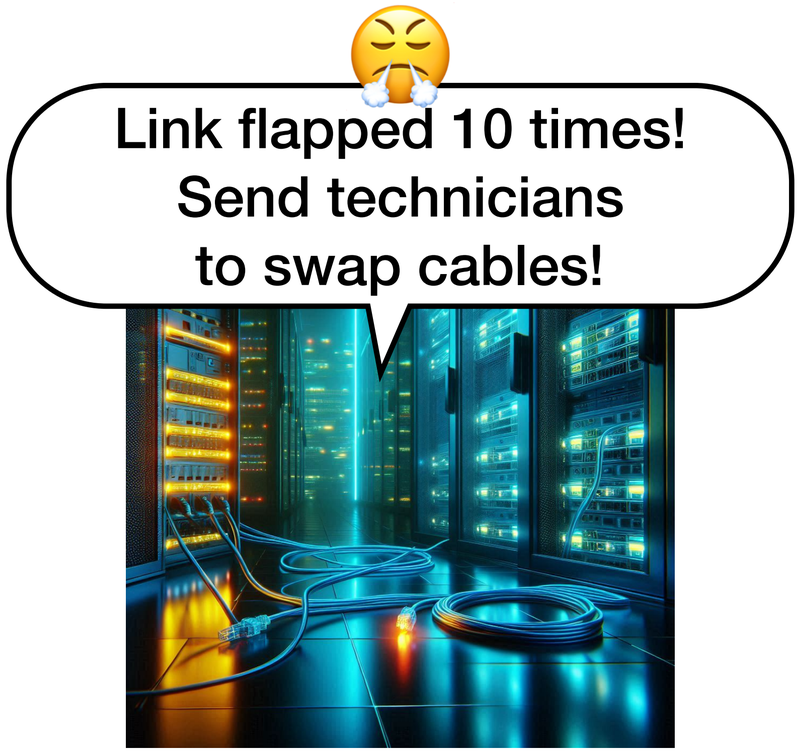
🔥 Common in large clusters

Source: semianalysis

AI-generated images

2 Lack of reliability

🚨 Alert



Why?

🔥 Common in large clusters

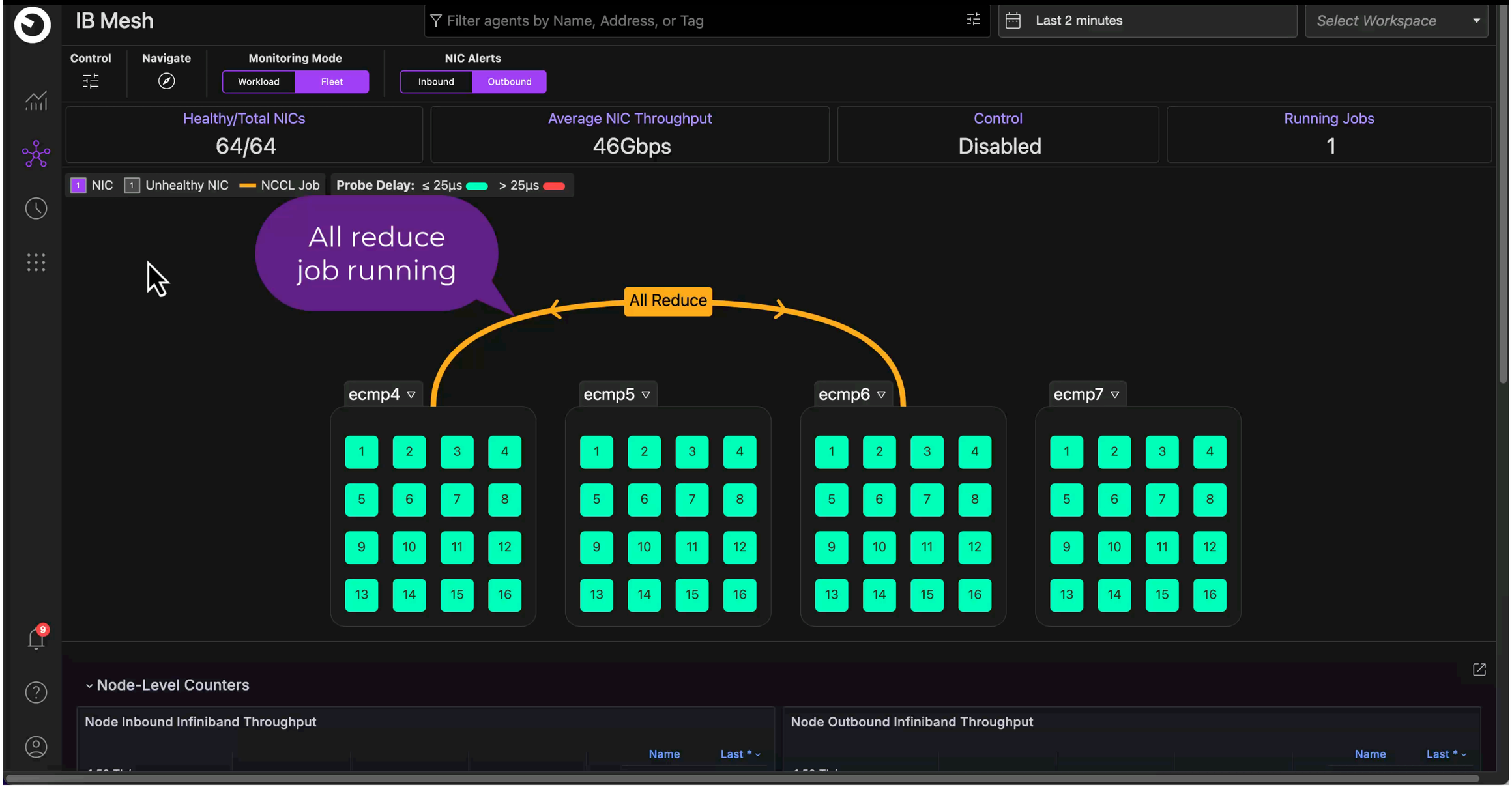
So what?

- Causes job to crash.
- Forces to restart from previous checkpoint.
- Lost ⏳💰

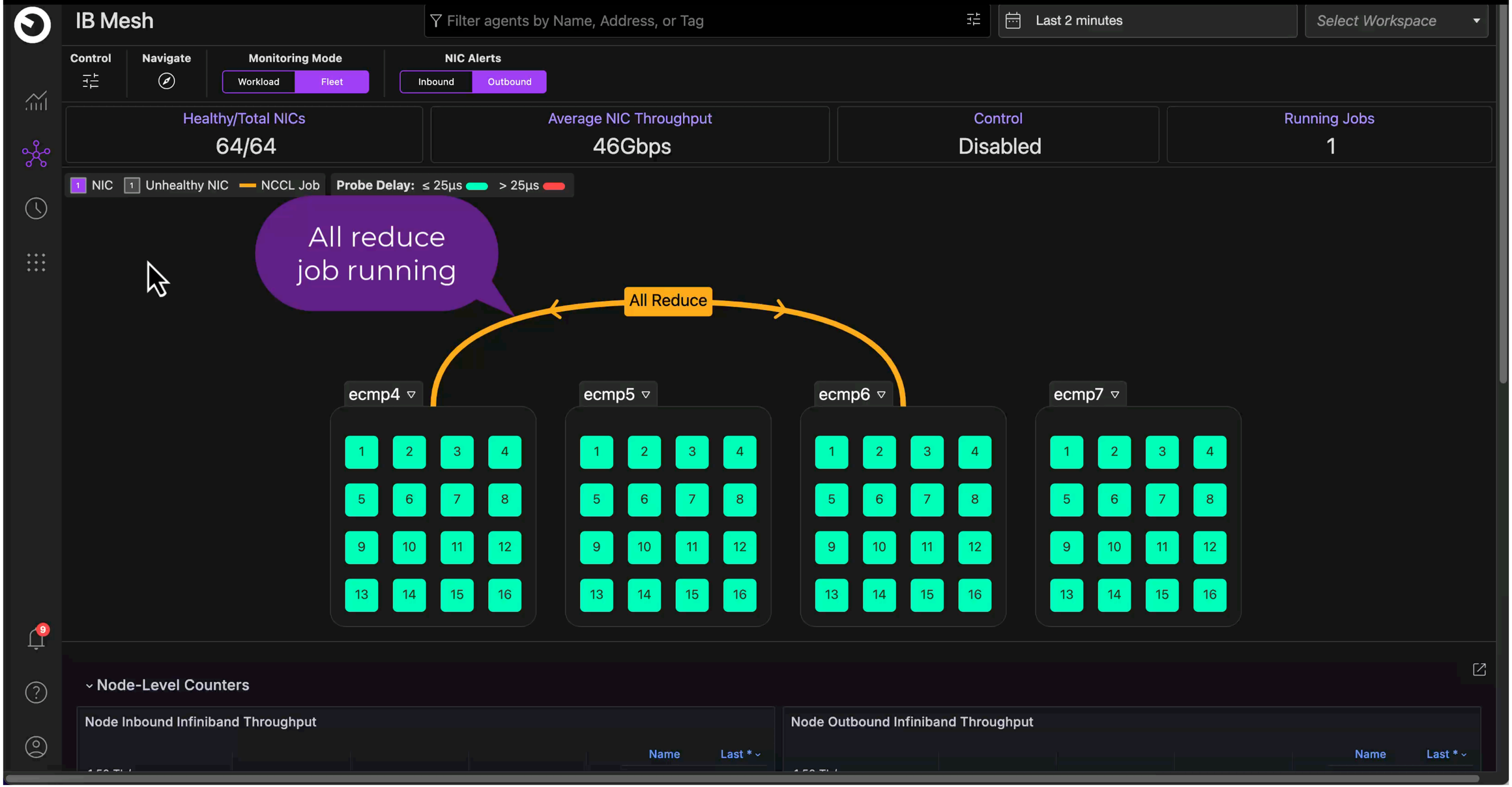
Source: semianalysis

AI-generated images

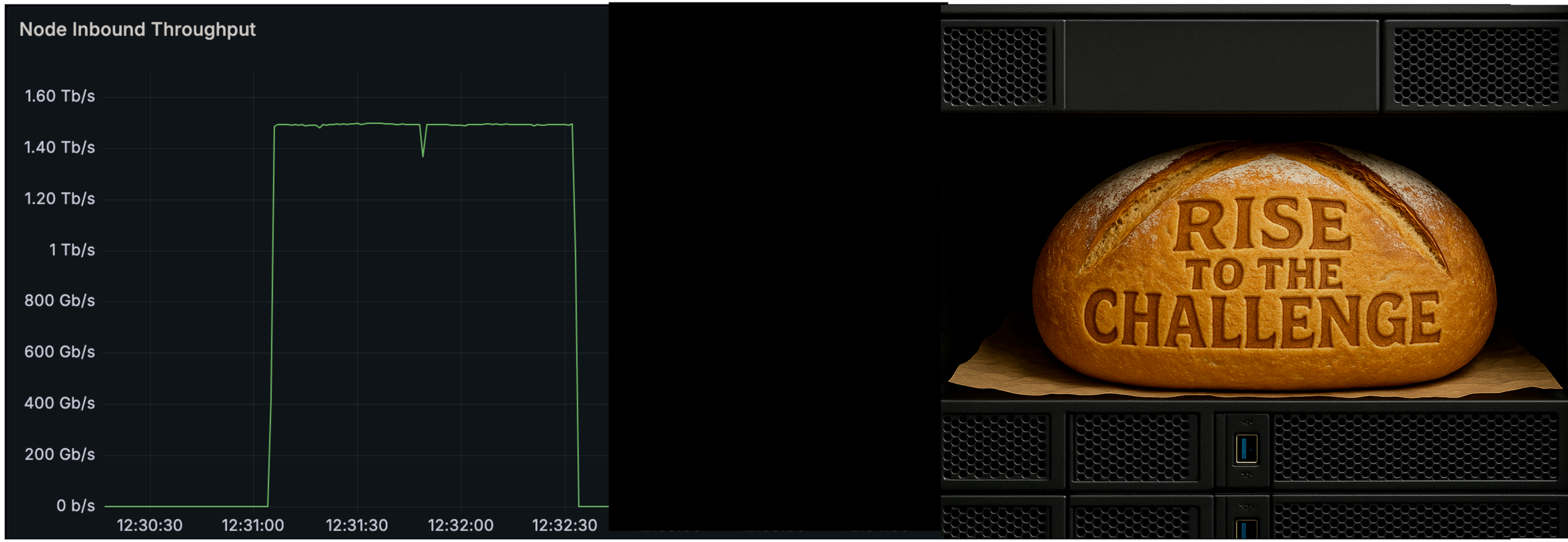
NIC flapping demo



NIC flapping demo



Resilience to NIC Flaps in Oracle Cloud

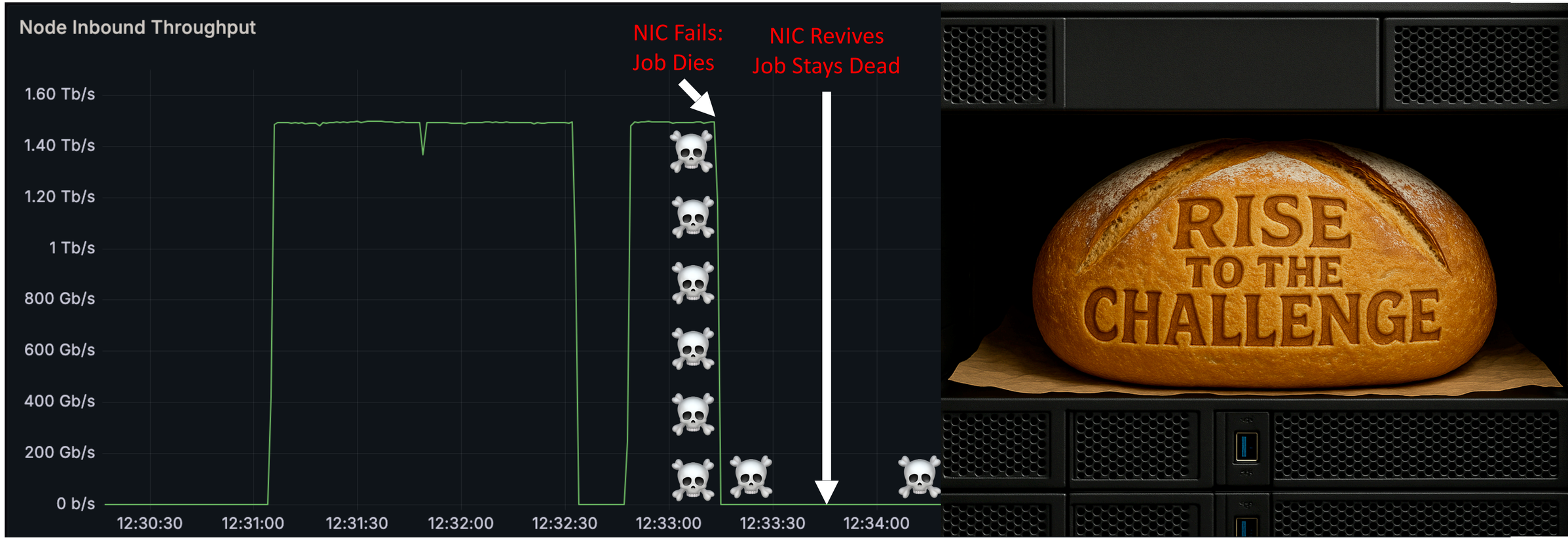


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: Clockwork solution
NIC Failure + Auto recovery

Resilience to NIC Flaps in Oracle Cloud

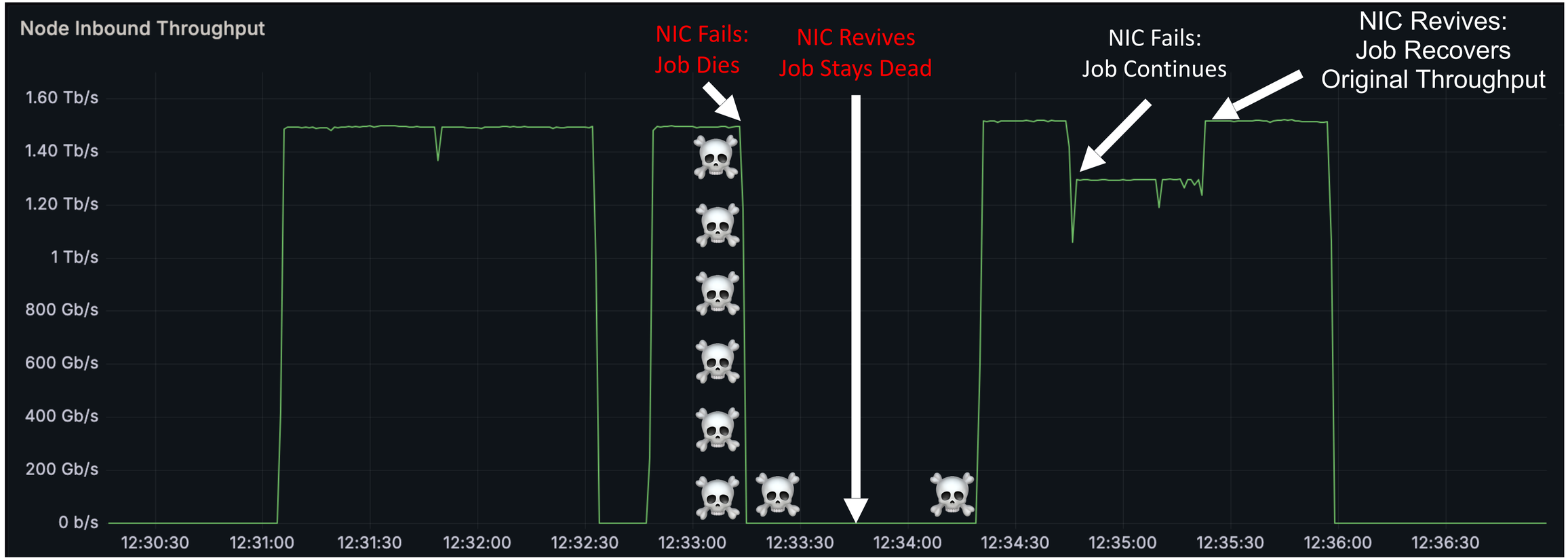


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: Clockwork solution
NIC Failure + Auto recovery

Resilience to NIC Flaps in Oracle Cloud

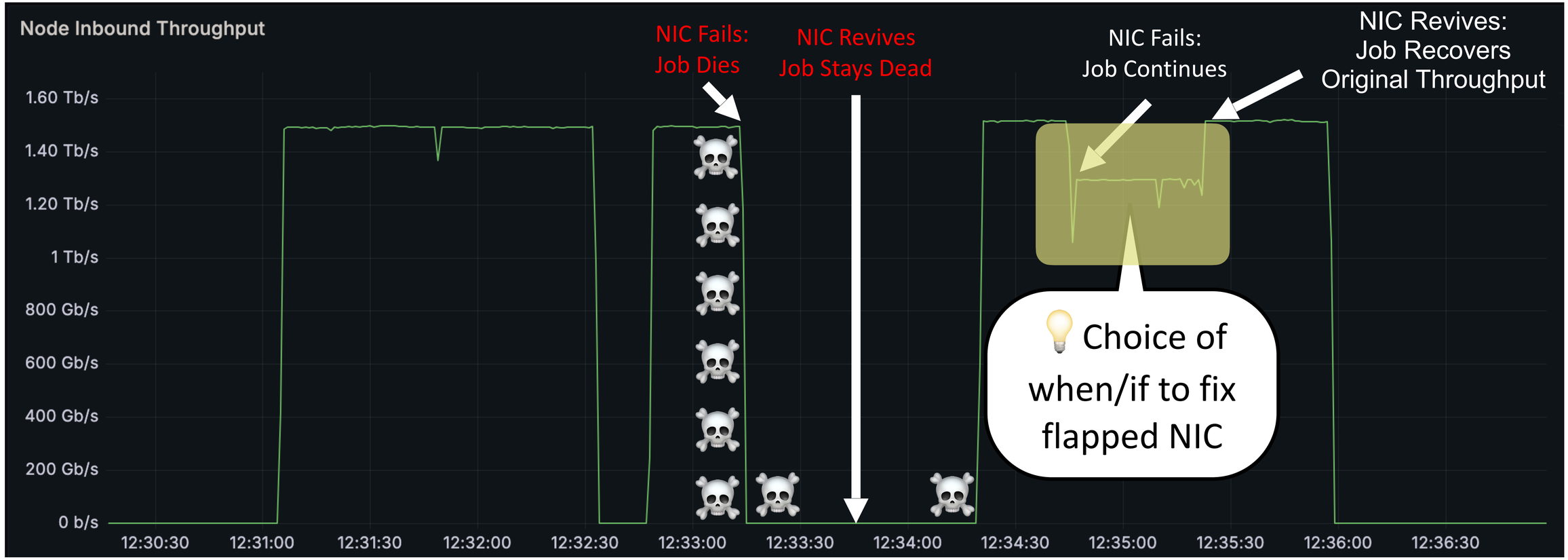


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: Clockwork solution
NIC Failure + Auto recovery

Resilience to NIC Flaps in Oracle Cloud

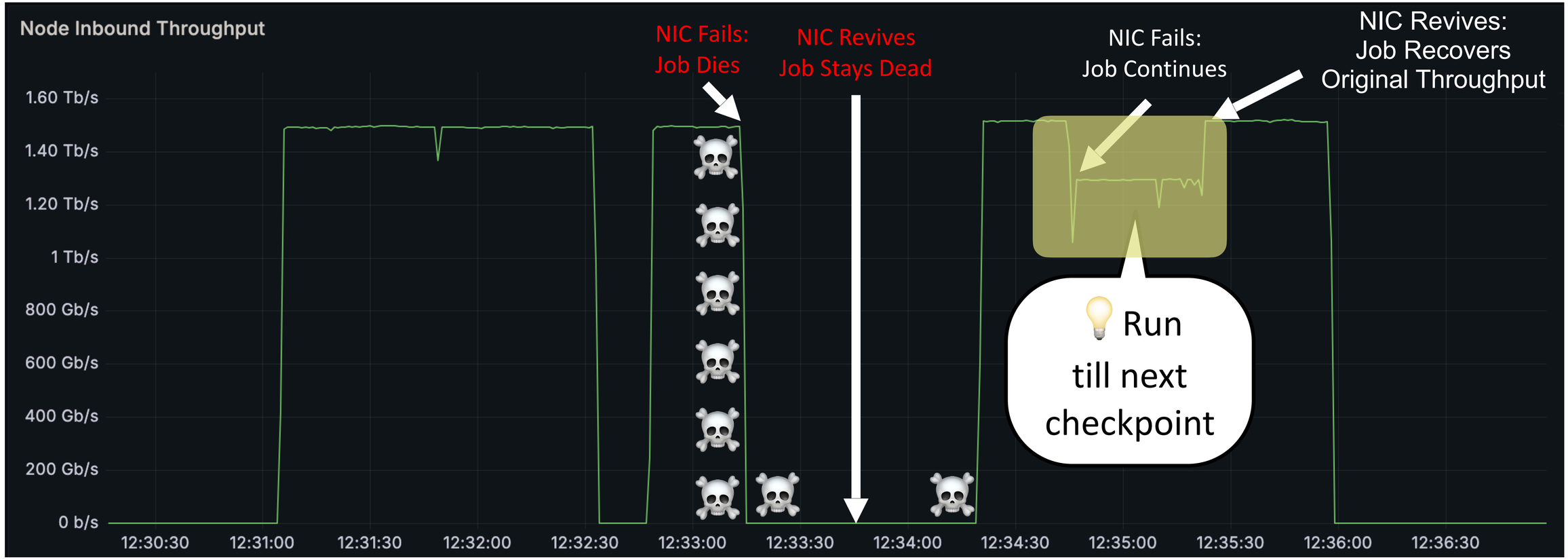


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: Clockwork solution
NIC Failure + Auto recovery

Resilience to NIC Flaps in Oracle Cloud



Run 1: No Failure

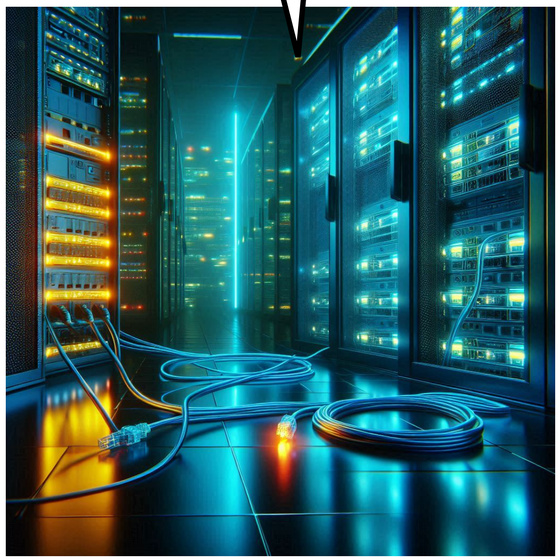
Run 2: NIC Failure
(What happens today)

Run 3: Clockwork solution
NIC Failure + Auto recovery

3 Lack of predictable performance

🚨 **Alert**

😤 The app is slow!



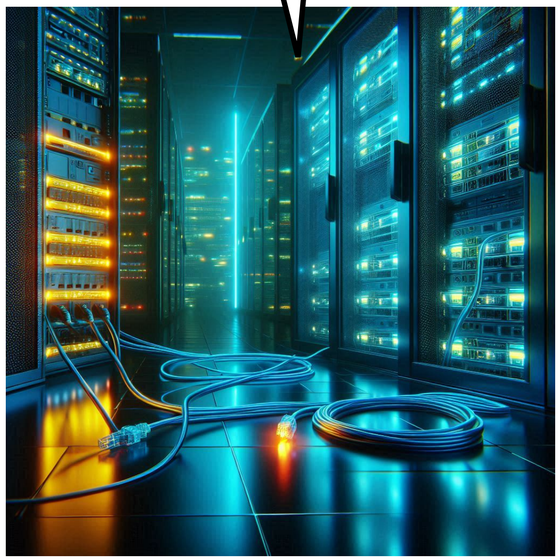
AI-generated images

3 Lack of predictable performance

🚨 **Alert**

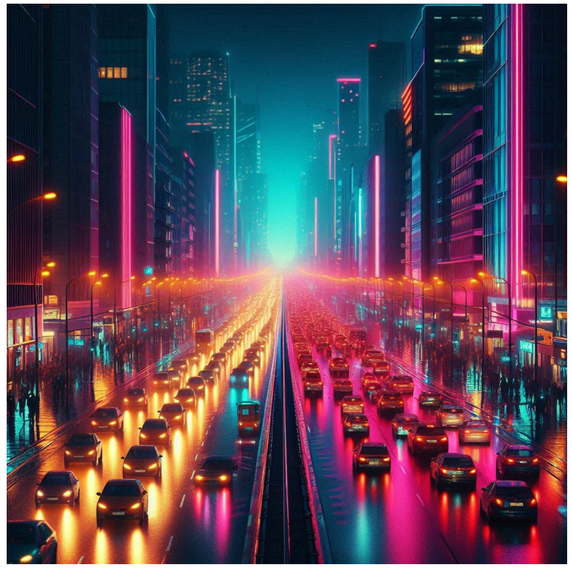
😤

The app is slow!



Why?

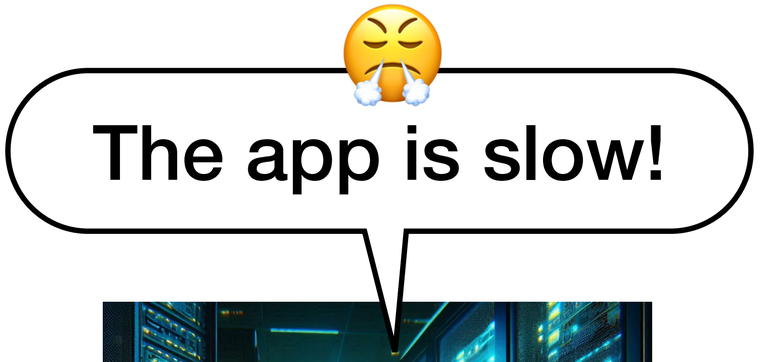
Flows contending for bandwidth

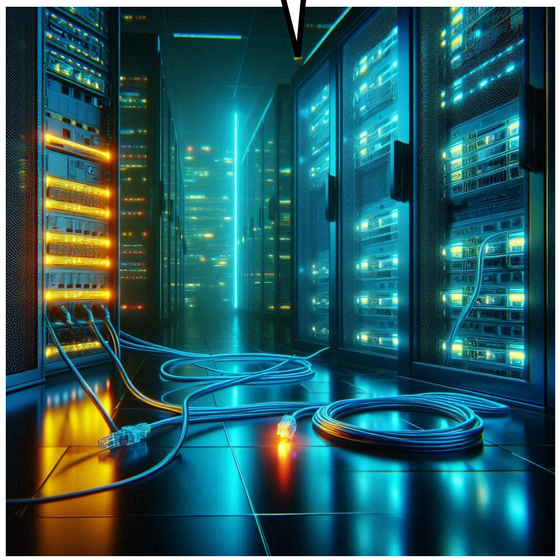


AI-generated images

3 Lack of predictable performance

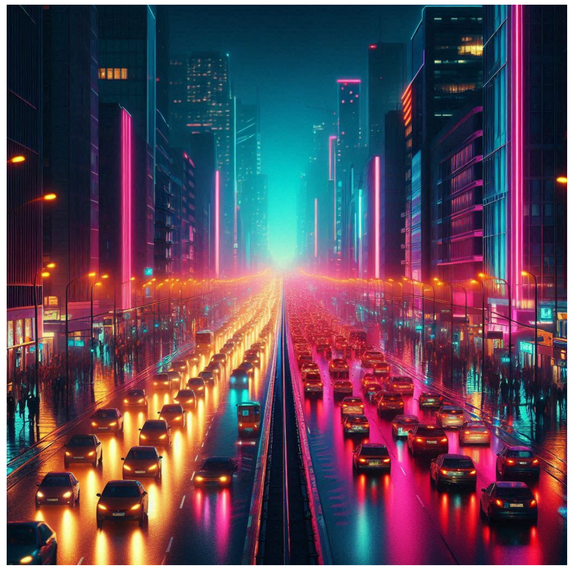
🚨 **Alert**





Why?

Flows contending for bandwidth



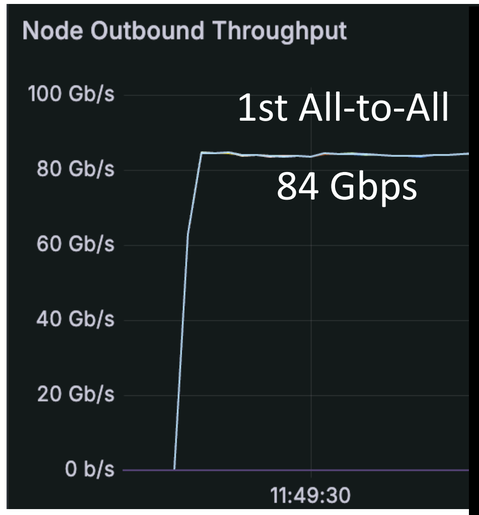
So what?

Low throughput, high latency

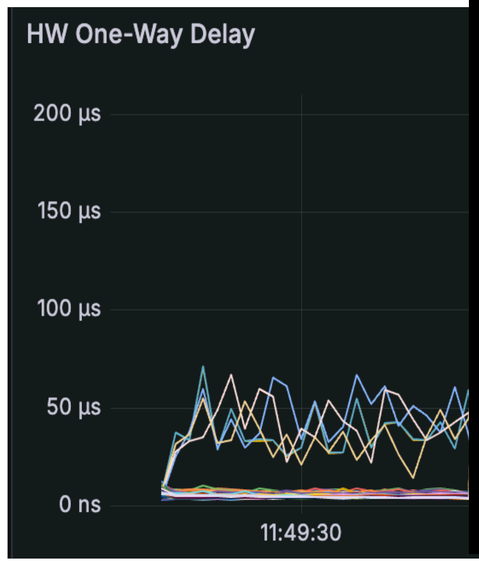
AI-generated images

Detecting and eliminating contention on ECMP Test Bed

Throughput



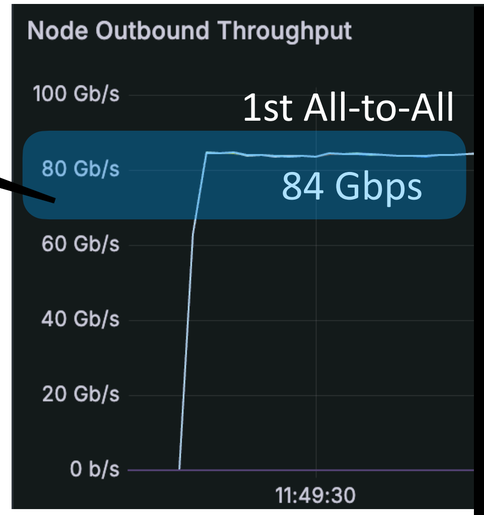
One-way Delays



Detecting and eliminating contention on ECMP Test Bed

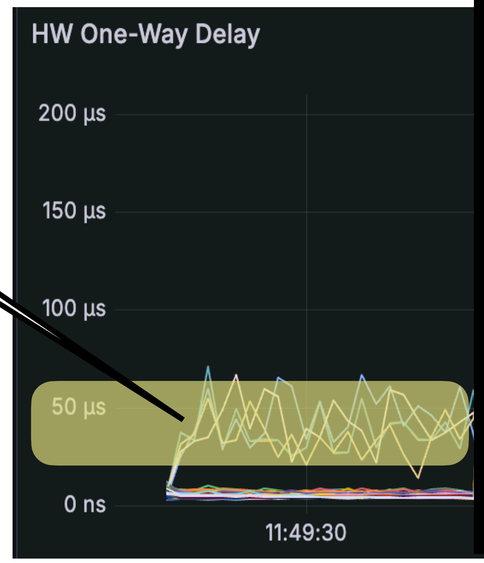
⬇️ Lower Throughput!

Throughput



🐢 Queue pairs with high one way delays!

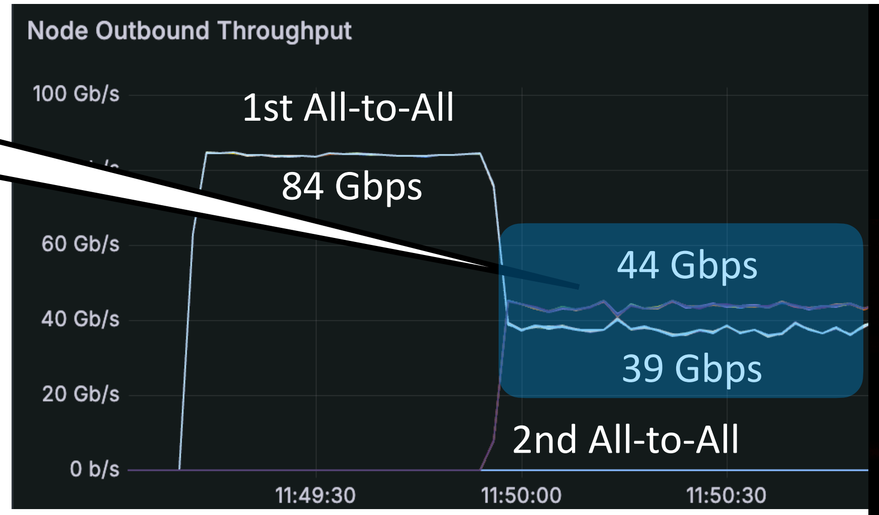
One-way Delays



Detecting and eliminating contention on ECMP Test Bed

⬇️ Lower Throughput!

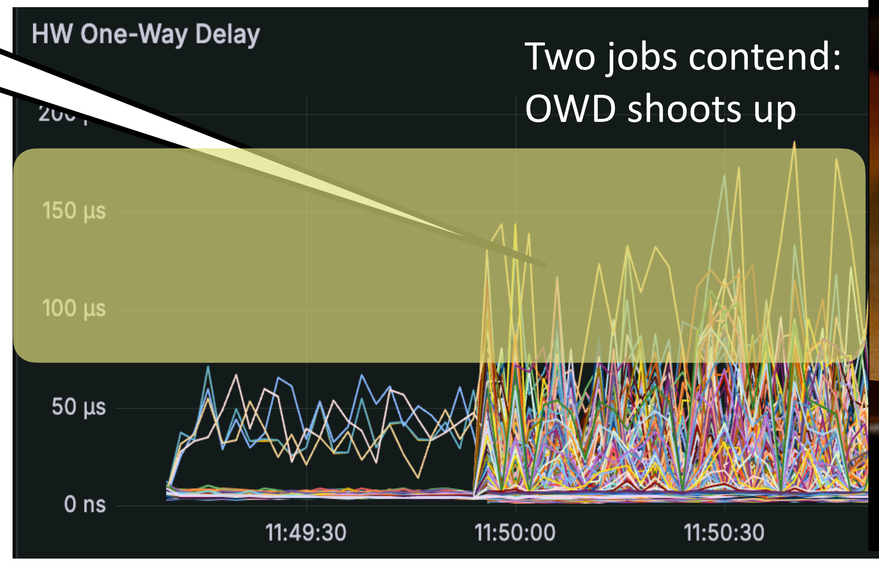
Throughput



🕵🏽 Contention!

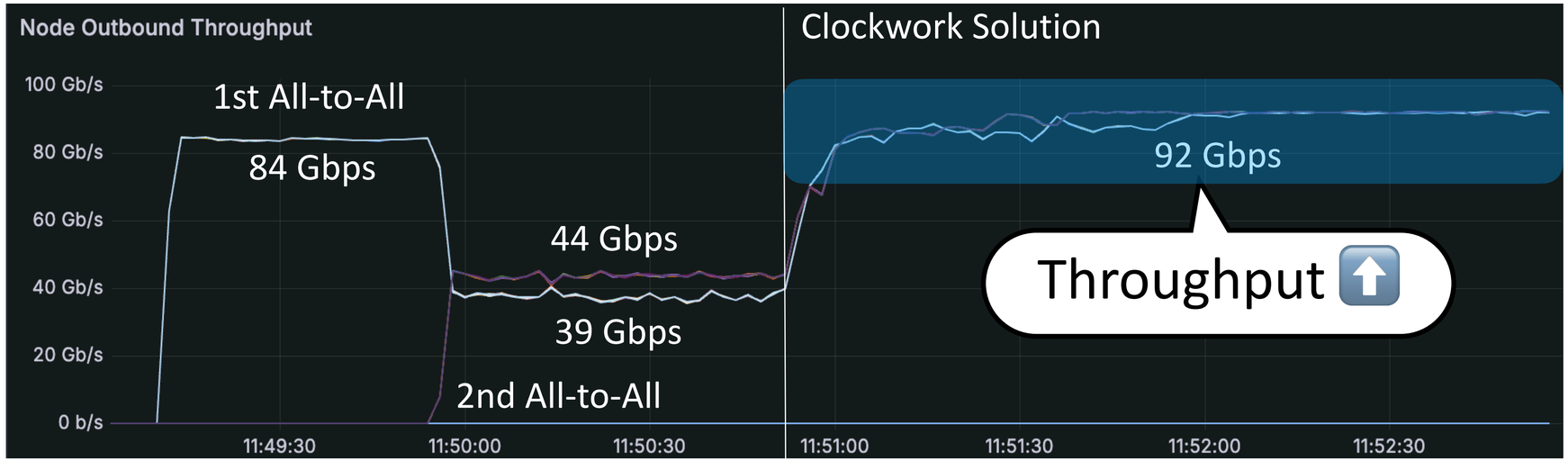
🐢 Queue pairs with high one way delays

One-way Delays

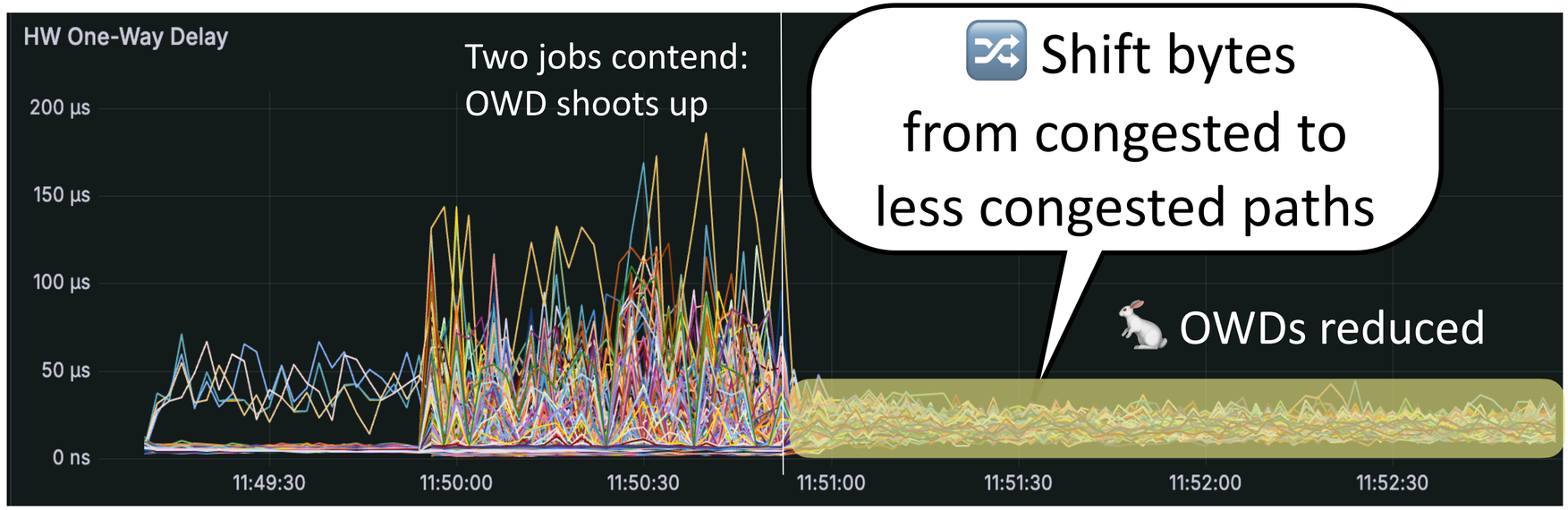


Detecting and eliminating contention on ECMP Test Bed

Throughput



One-way Delays



Key Takeaways

🔎 Fine-grained visibility into queue pairs speeds up diagnosis.

✅ Visibility

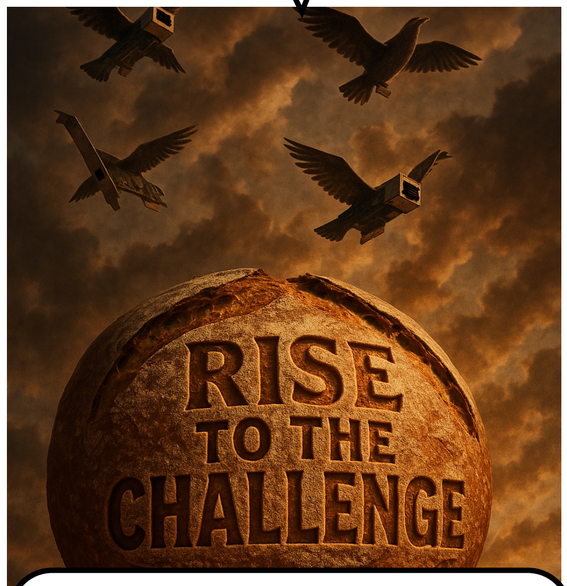
AI-generated images

Key Takeaways

🔎 Fine-grained visibility into queue pairs speeds up diagnosis.

✅ Visibility

Checkpoint to recover from crashes due to NIC/link flapping.



✅ Reliability

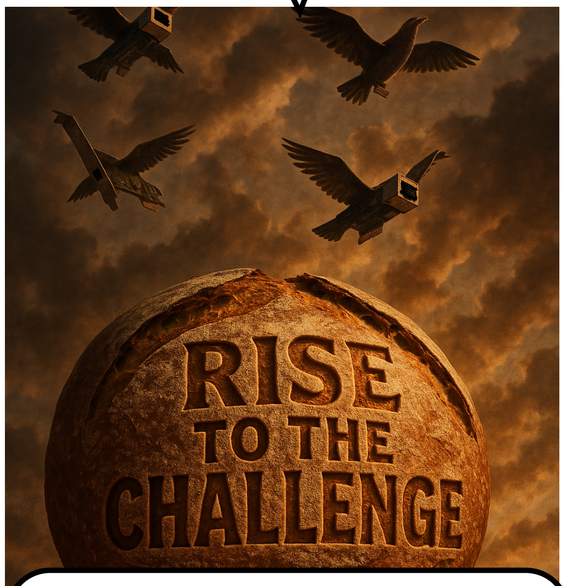
AI-generated images

Key Takeaways

🔎 Fine-grained visibility into queue pairs speeds up diagnosis.

✅ Visibility

Checkpoint to recover from crashes due to NIC/link flapping.



✅ Reliability

💨 Minimize delays to slow flows to optimize throughput.

✅ Performance

AI-generated images

Thank You

Huygens Usenix Paper
& Slides

🖖🏽lerna@clockwork.io

Thank You

CHECK OUT
MY FULL 📽️ TALK!

Huygens Usenix Paper
& Slides

🖖🏽lerna@clockwork.io