Lack of fine-grained visibility at queue pair level


Pre-race & race time!

⛔ Visibility


NIC/link flapping

⛔ Reliability



AI-generated images

# AI/ML Networking Challenges


Lerna Ekmekcioglu
Sr. Solutions Engineer
Clockwork Systems


# Agenda

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus

AI-generated image

CLOCKWORK.IO

AI-generated image

Large scale jobs

1 DeepSpaceExplorer LLM Model

2 Deepspeed, PyTorch, etc.

3 NCCL (Nvidia Collective Communications Library)

Inter-GPU communication

4 Network Device (Infiniband or RoCE)

RDMA capable!

AI-generated image

RDMA

Fast data transfer

Memory

🐢 OS

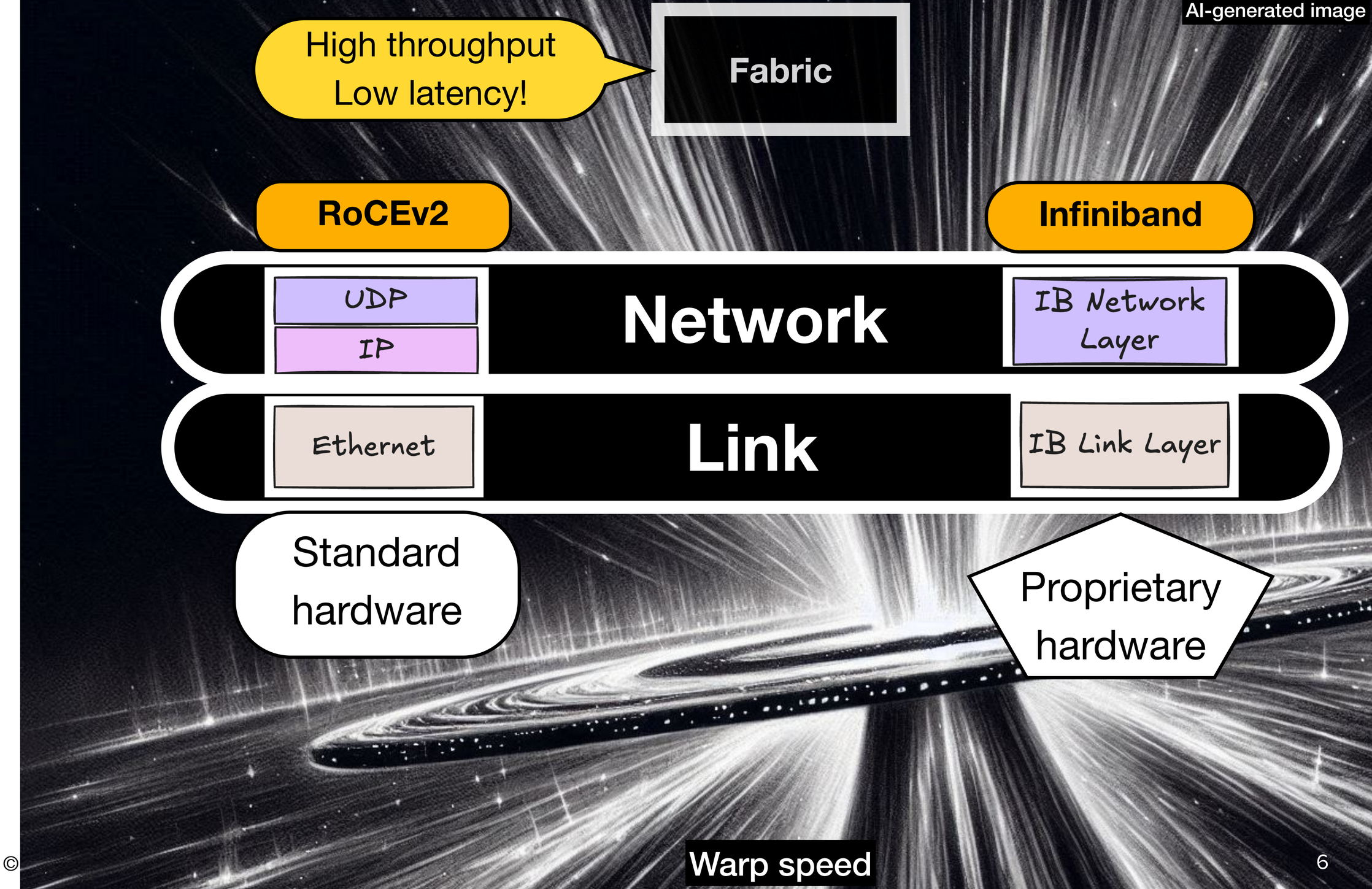
Kernel Bypass

Memory

🐢 OS

Infiniband or RoCE

Transporter


AI-generated image
High throughput
Low latency!
Fabric
RoCEv2
Infiniband
UDP
IP
Network
IB Network Layer
Ethernet
Link
IB Link Layer
Standard hardware
Proprietary hardware
Warp speed

# Agenda

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus

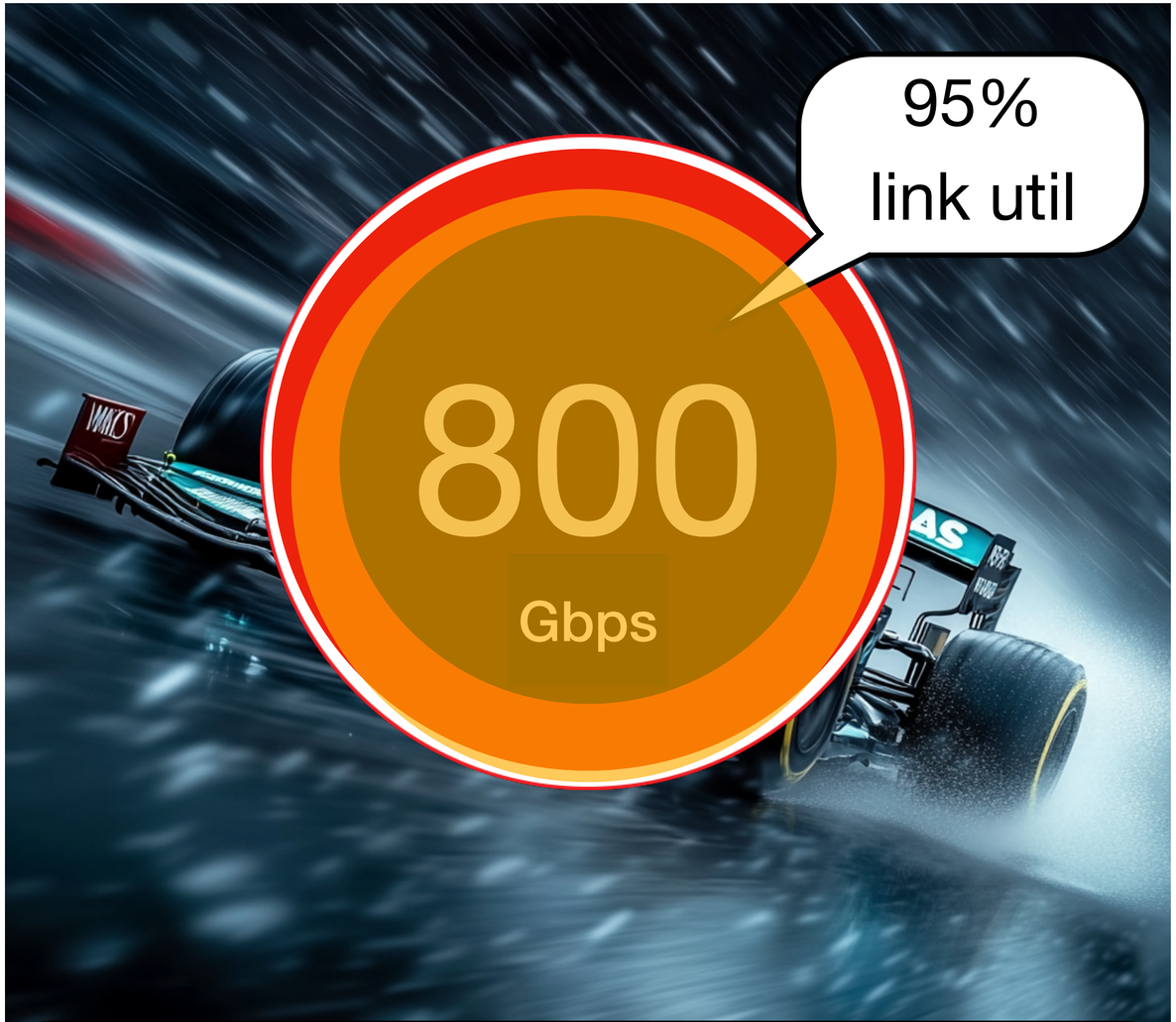

AI-generated image

CLOCKWORK.IO


Traditional networks

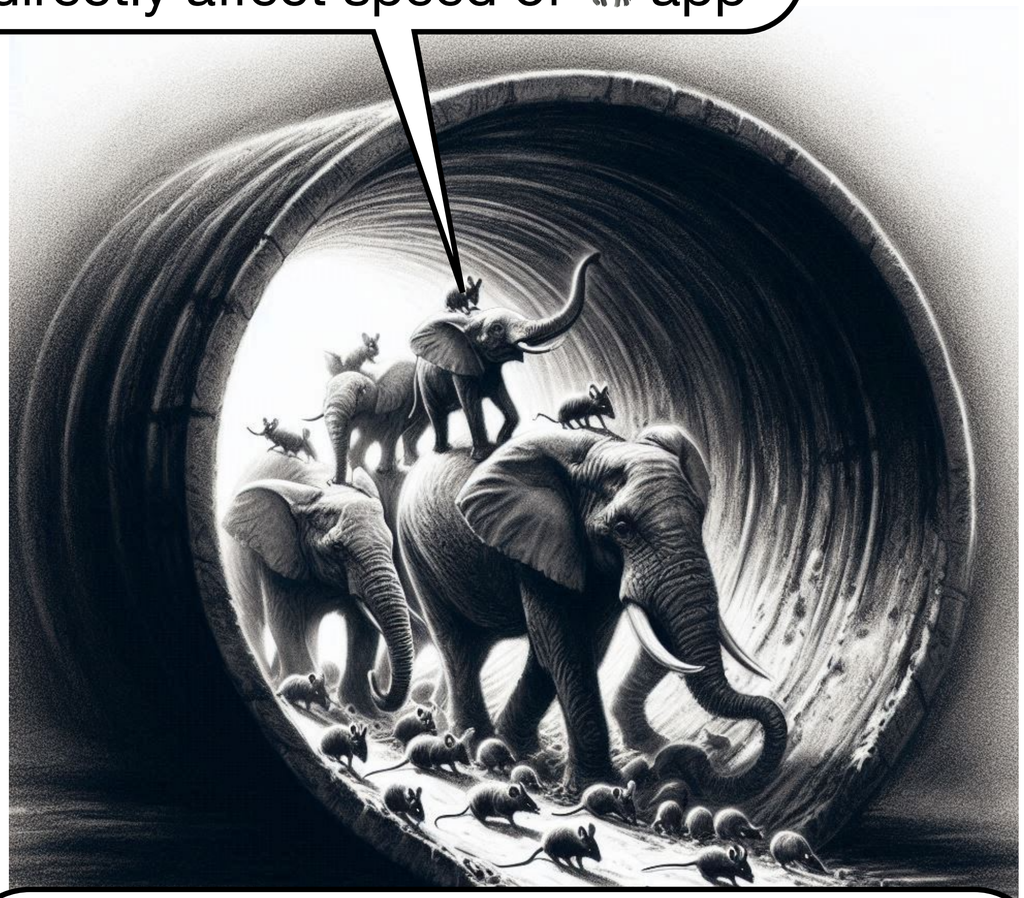

AI networks


Traditional networks



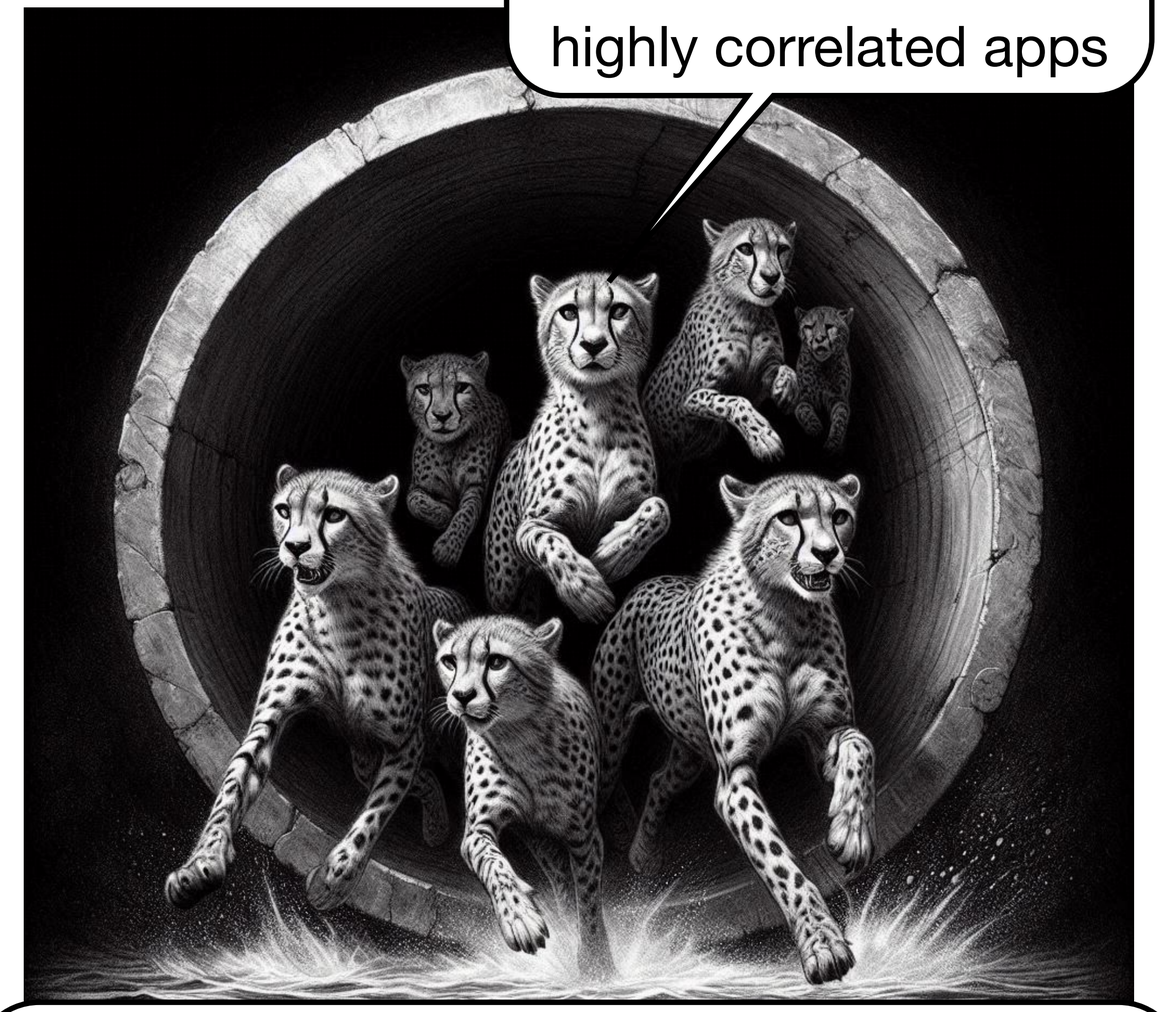

AI networks

Speed of 🐭 app doesn’t directly affect speed of 🦣 app



Traditional networks

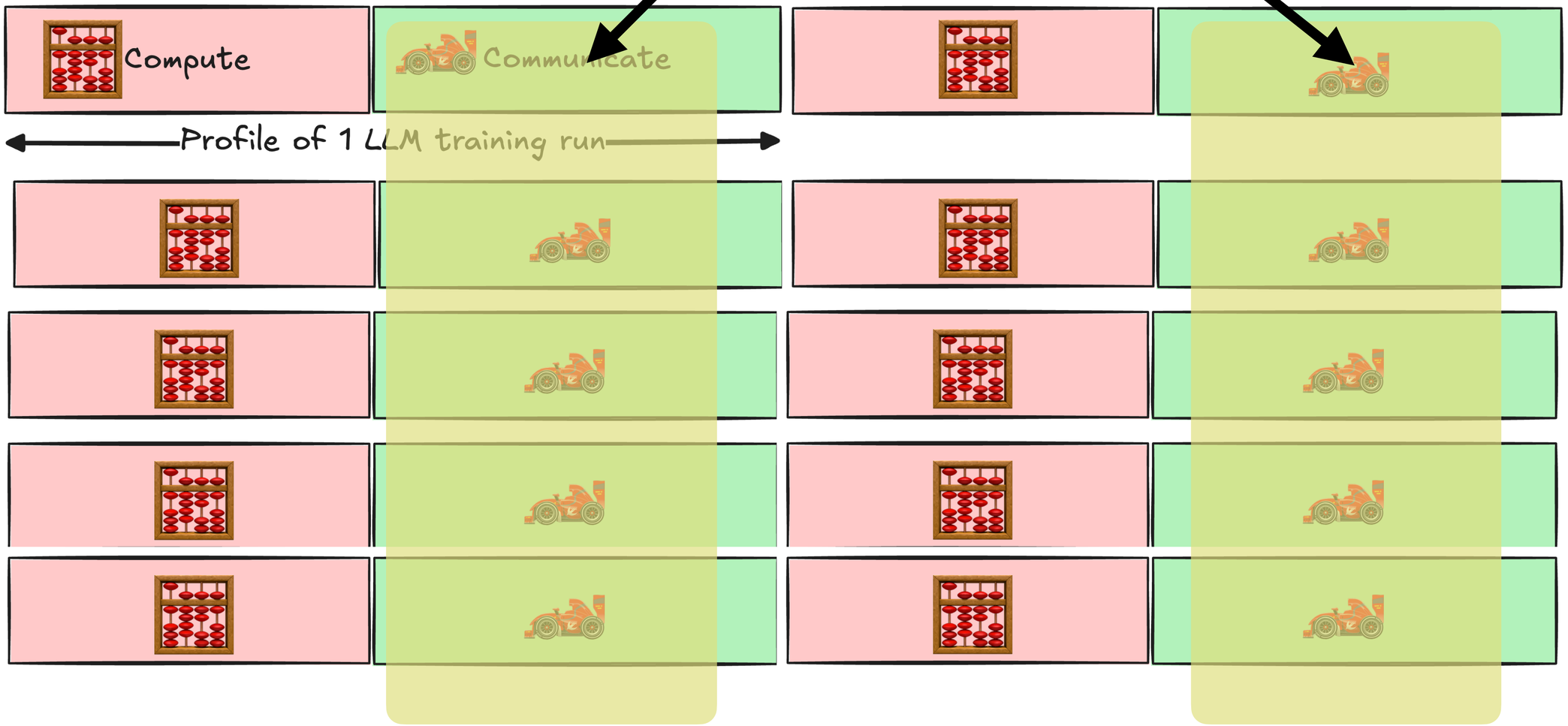
Similar highly correlated apps



🏎️ AI networks

# Profile of an LLM trai

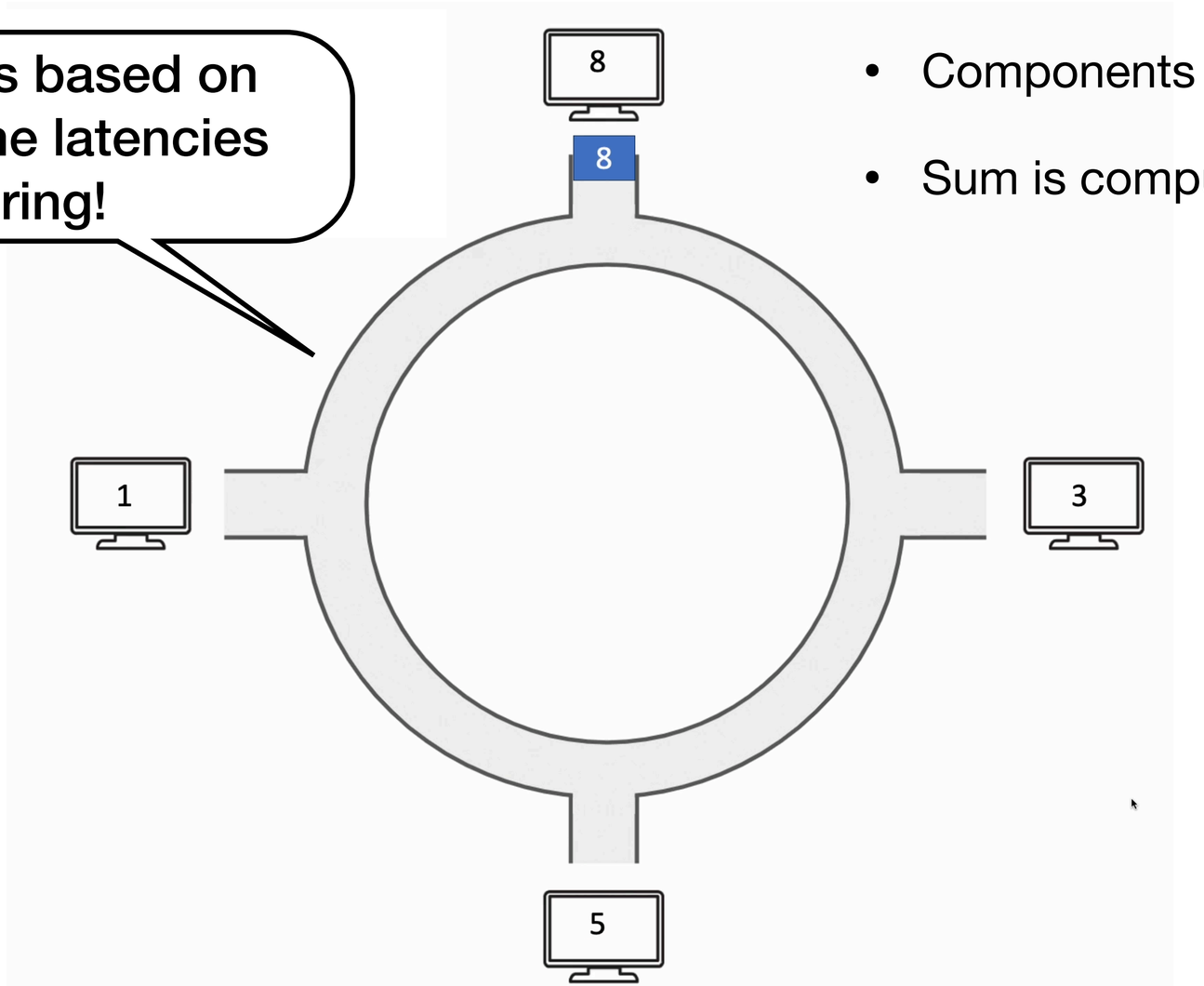
> 30 % of total job completion time is spent on networking!




Source: AMD
Source: Nvidia

# All-reduce (NCCL ring implementation)

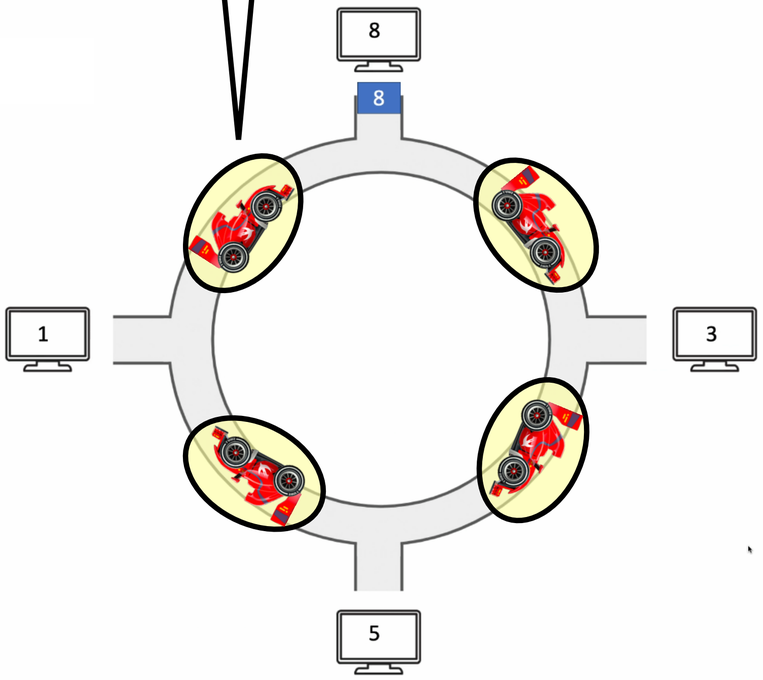
**Finish time is based on the sum of the latencies in the ring!**




- Components of the sum get passed around
- Sum is computed then passed around

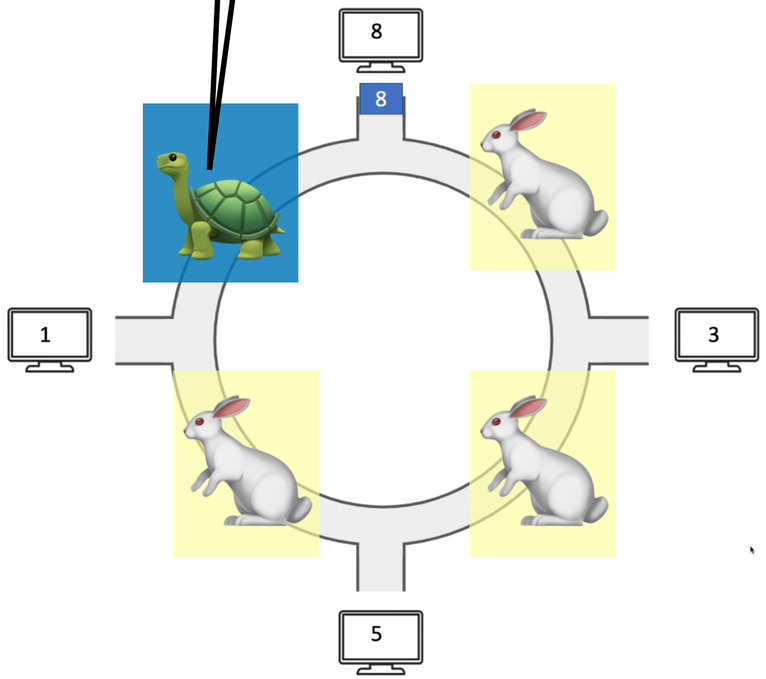
# Communicate

GPU's won't proceed to next step until all synchronize partial results.



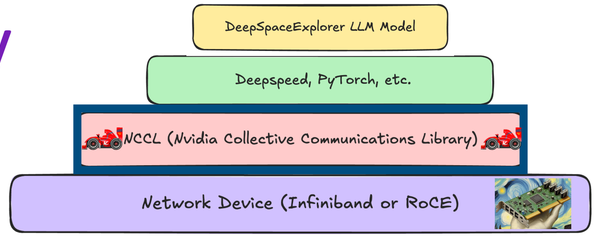

At a pit stop, pit crew is not done until all teammates are done.

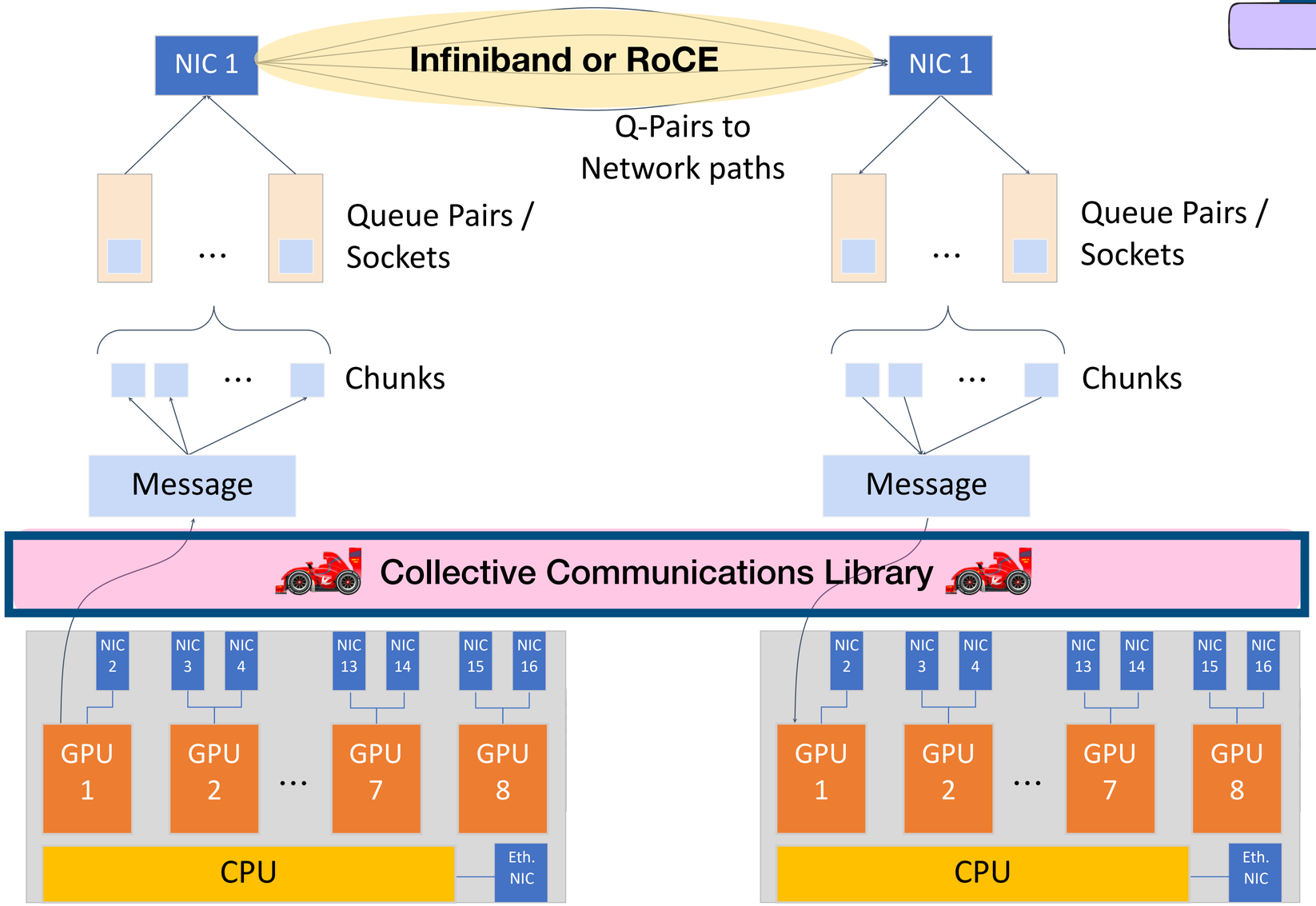
Slowest flow (or node) determines throughput.



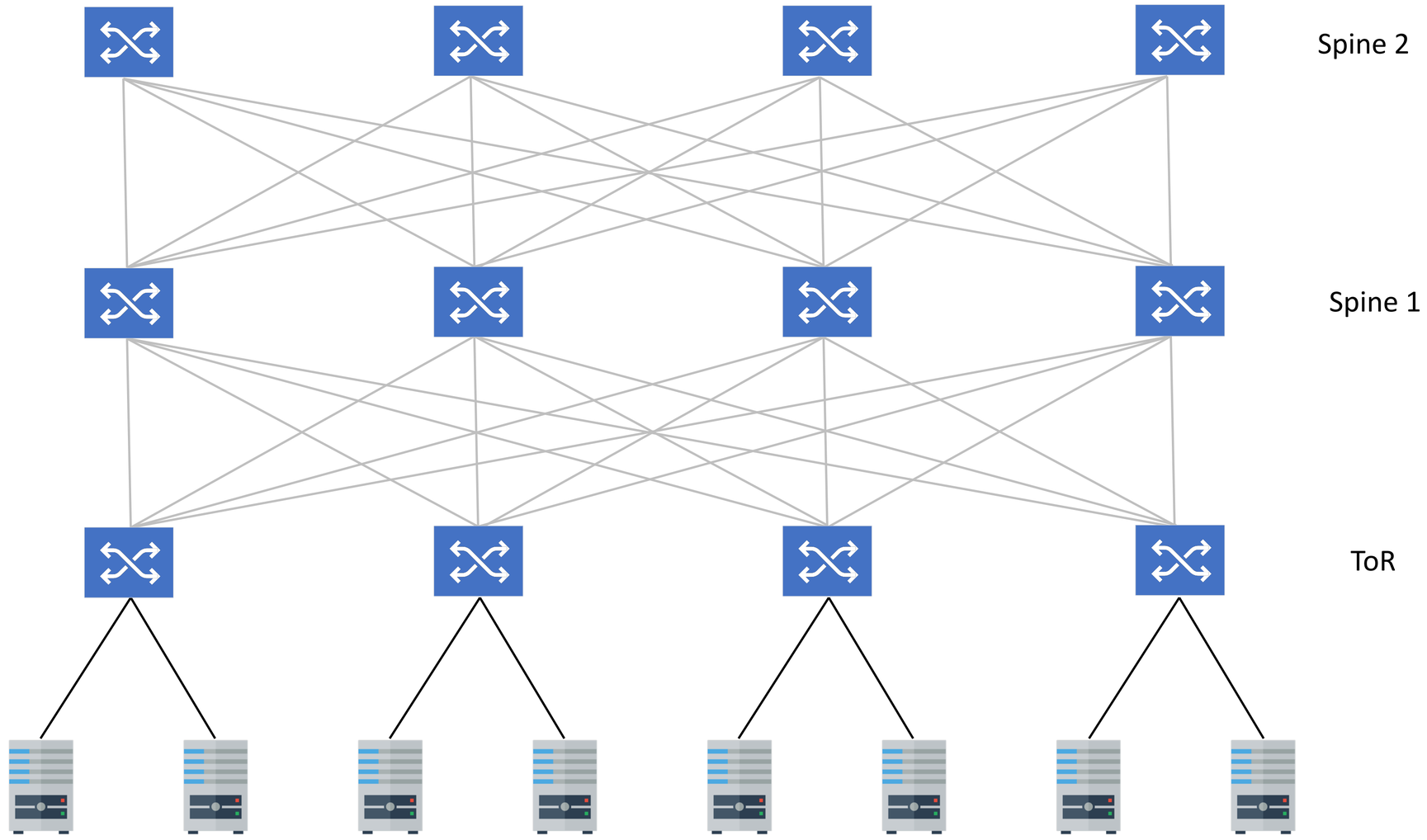

Slowest determines speed.

# Under the hood: NCCL or other communication library



# The Interconnection Fabric: A Fat-tree Network (or Rail-Optimized Topology)

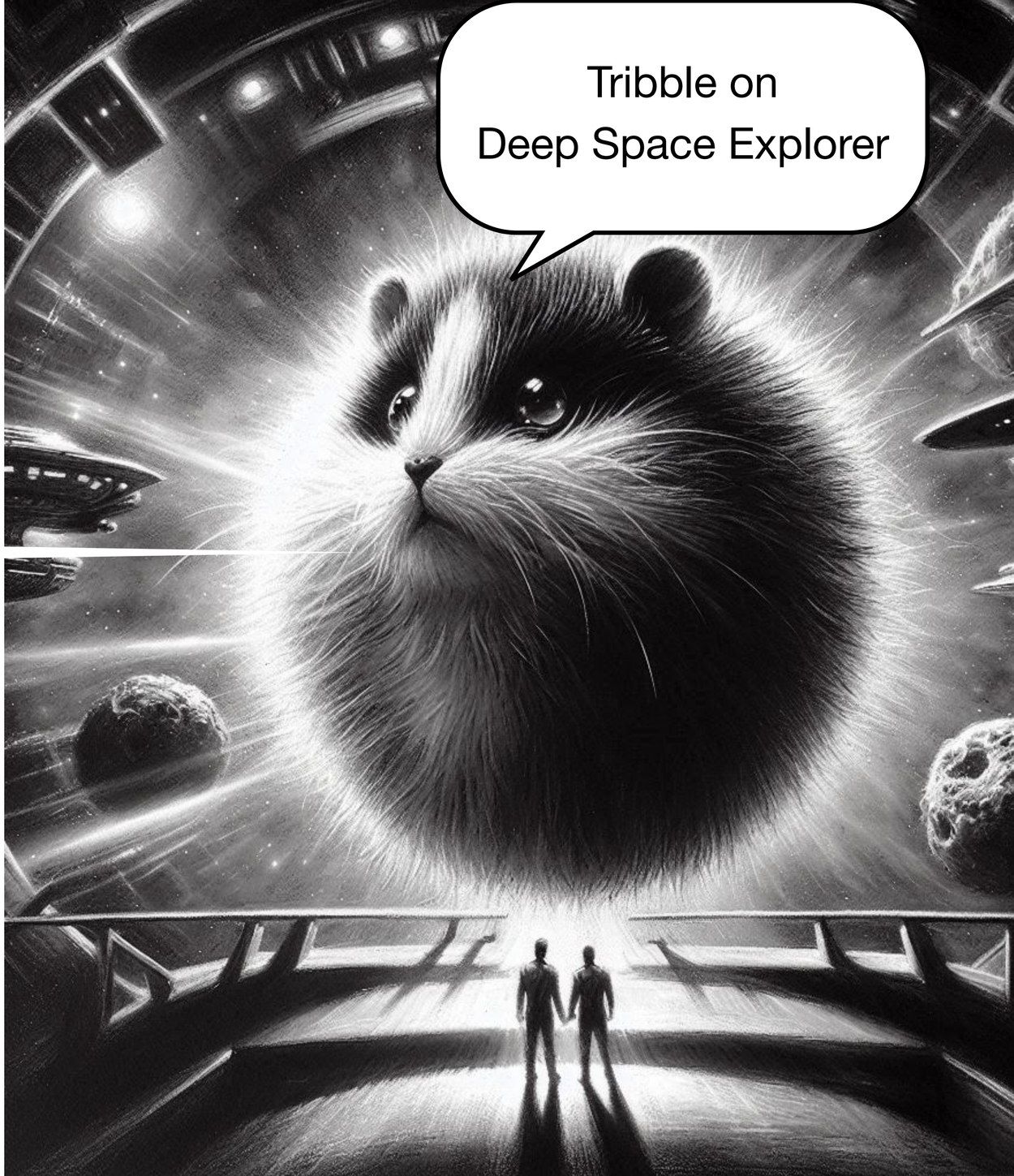


# Agenda

1. AI workload layers
2. Demands from AI networks
3. Challenges in AI networks
4. Key Takeaways

Time Force,
Department of Temporal Affairs

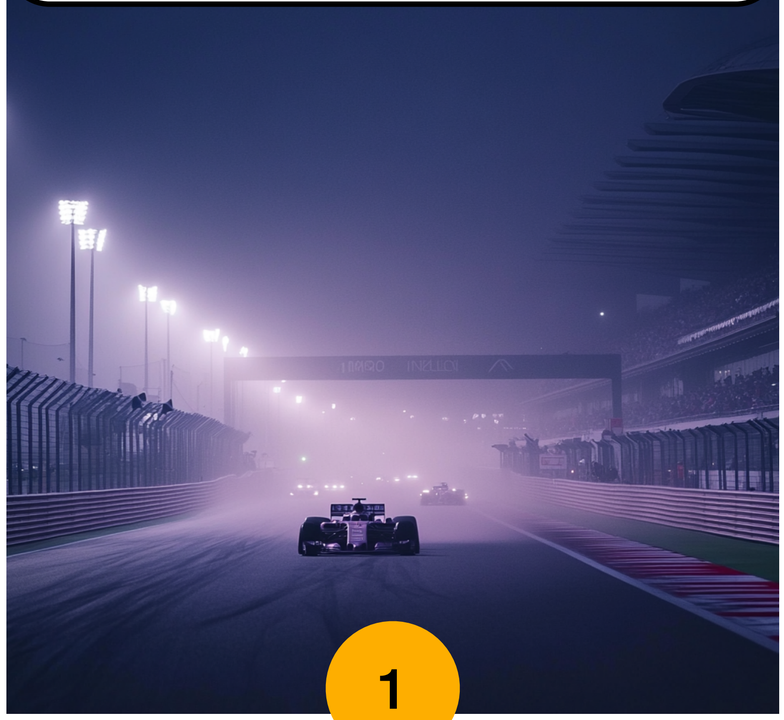
Tempus Mundi Servamus




CLOCKWORK.IO

# Networking challenges of AI/ML workloads

⛔ Visibility

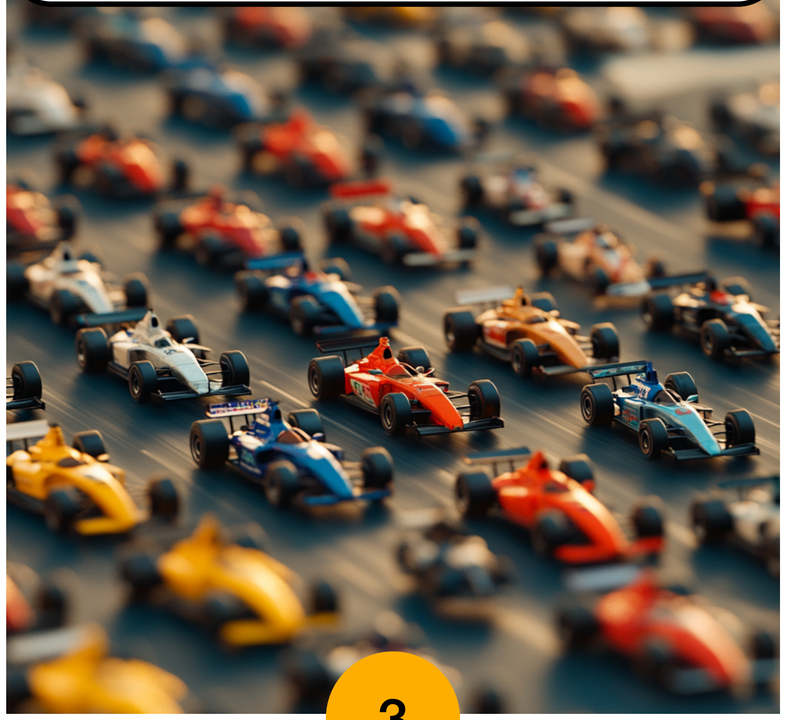


1

⛔ Reliability

2

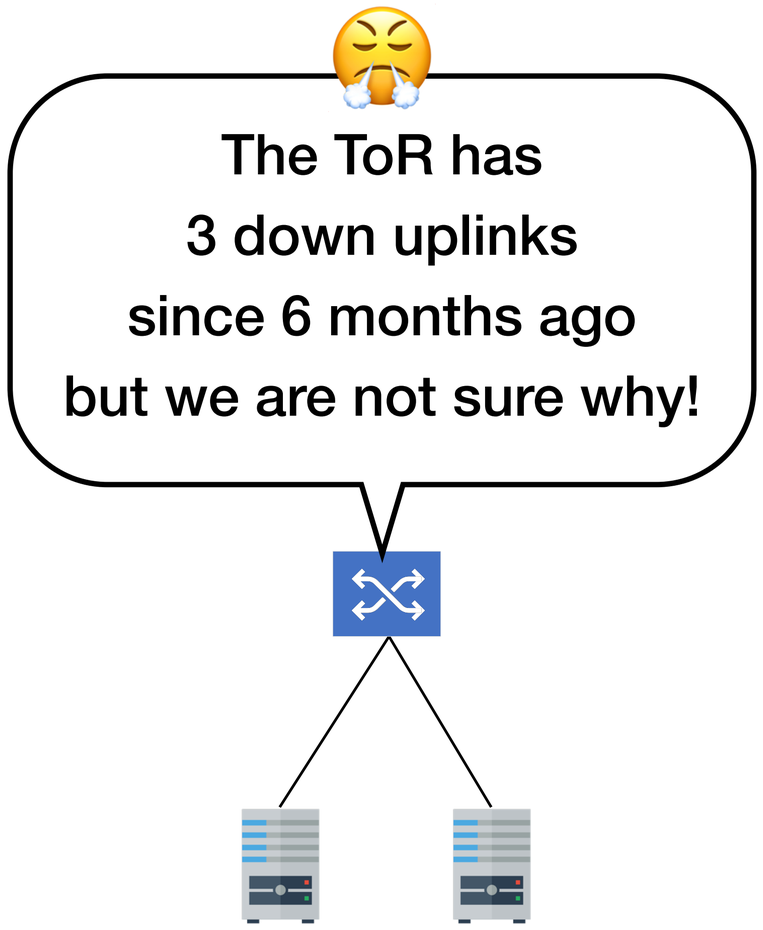
⛔ Performance



3

# 1 Lack of visibility

🚨Alert




## Why?

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


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

# Visibility

# Huygens clock synchronization

- Software based
- High precision at scale 🎯

Measures
one way delays



usenix
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

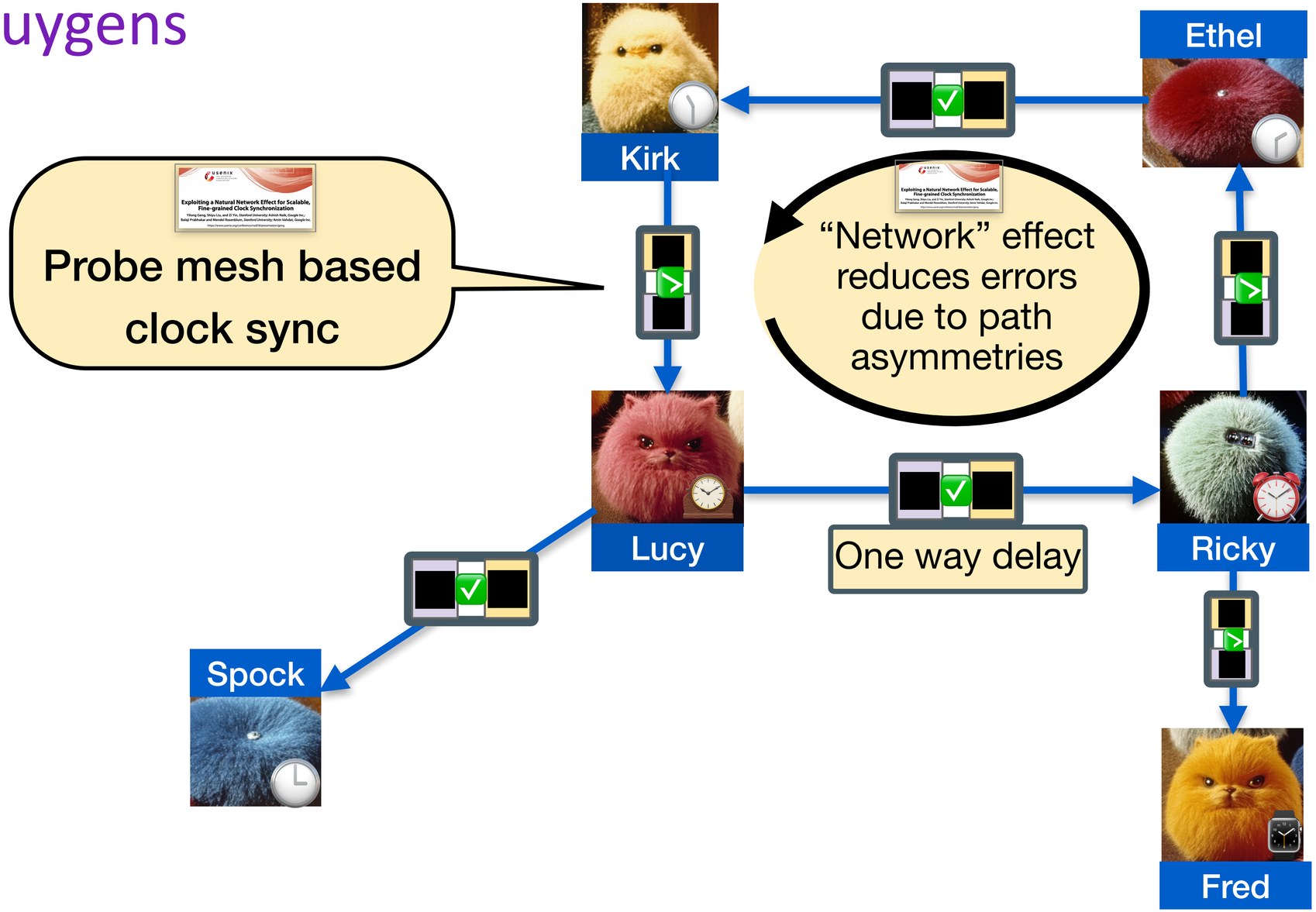
**Exploiting a Natural Network Effect for Scalable, Fine-grained Clock Synchronization**

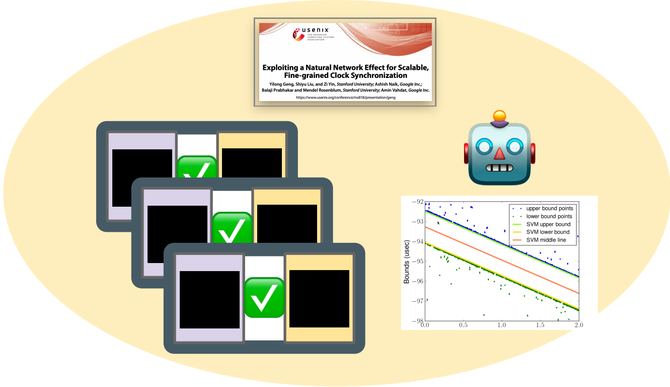
Yilong Geng, Shiyu Liu, and Zi Yin, *Stanford University;* Ashish Naik, *Google Inc.;*
Balaji Prabhakar and Mendel Rosenblum, *Stanford University;* Amin Vahdat, *Google Inc.*

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

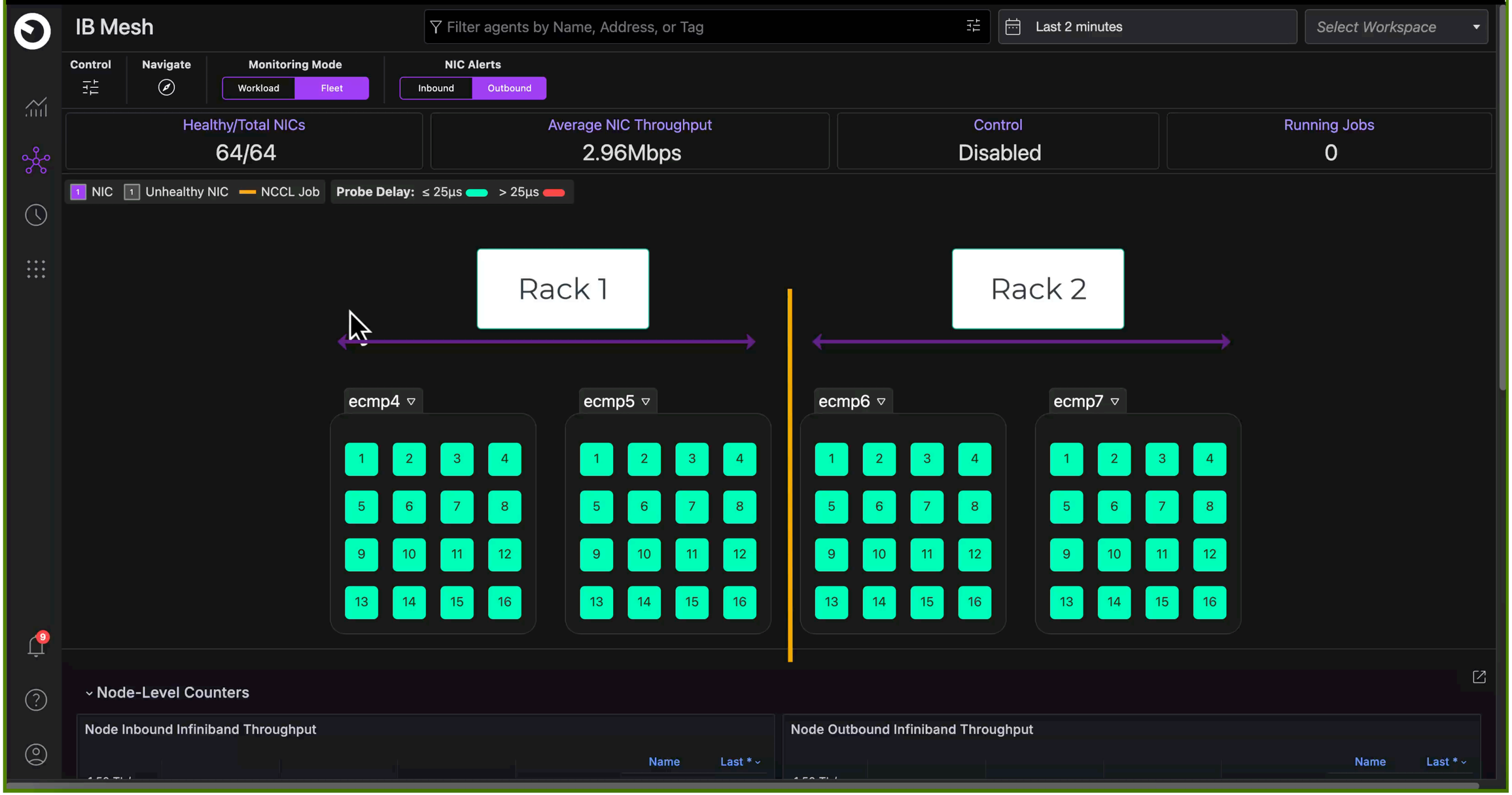
https://www.usenix.org/conference/nsdi18/presentation/geng

# Huygens



# Visibility demo



# 2 Lack of reliability

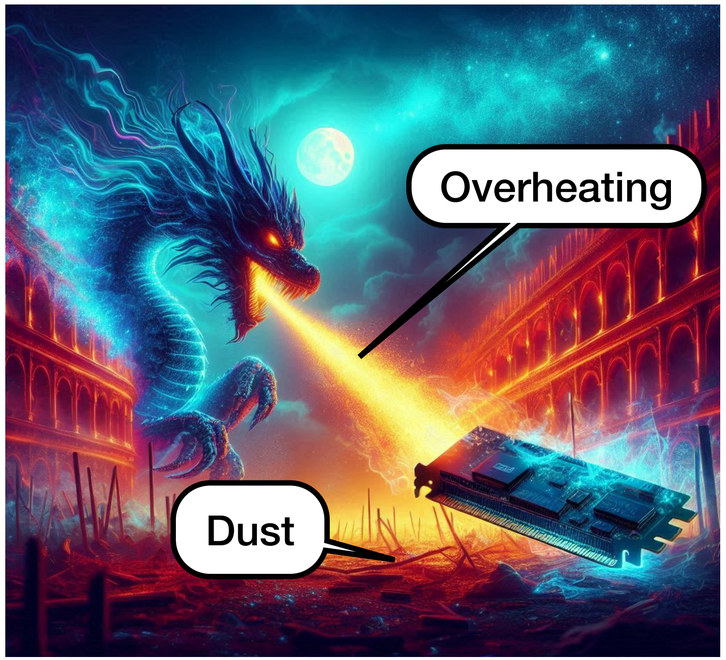
## 🚨 Alert

😤

Link flapped 10 times!
Send technicians
to swap cables!

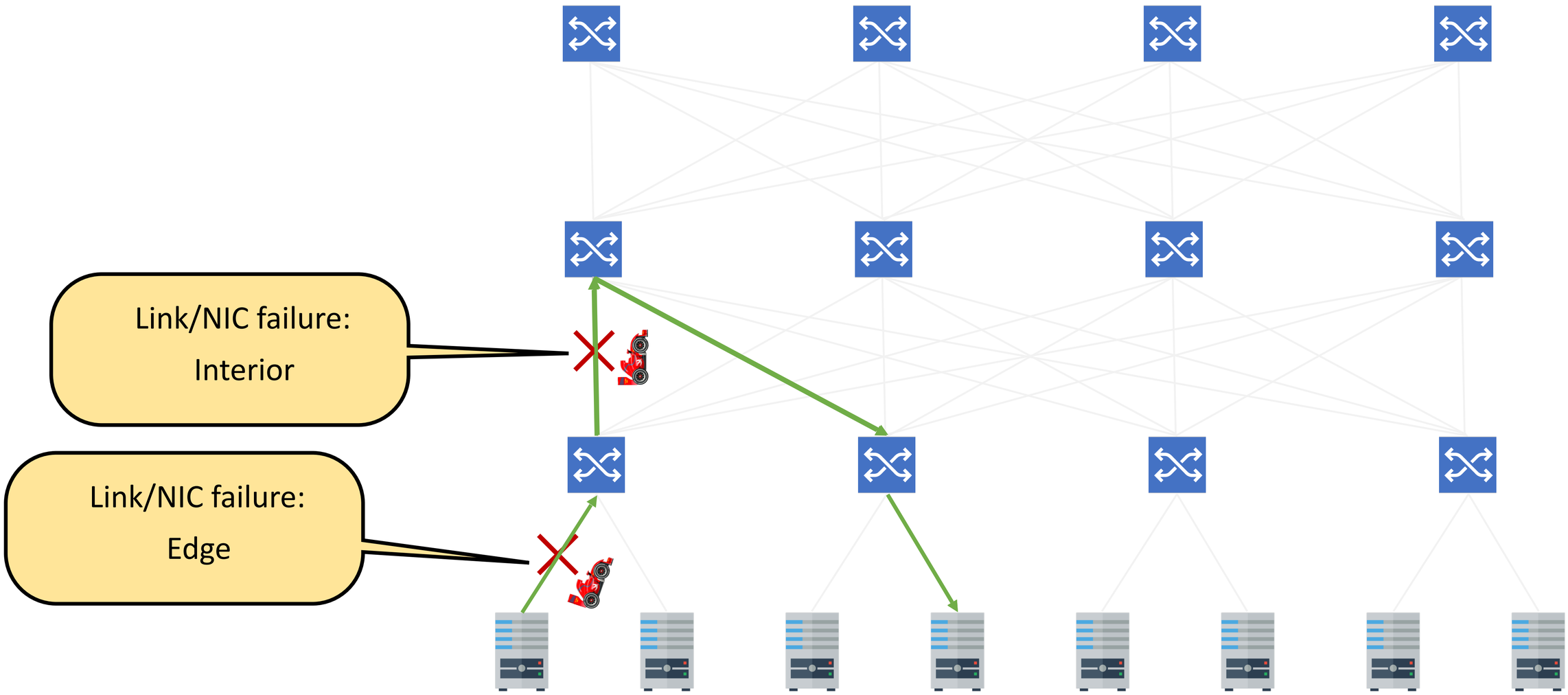
## Why?

🔥 Common in large clusters

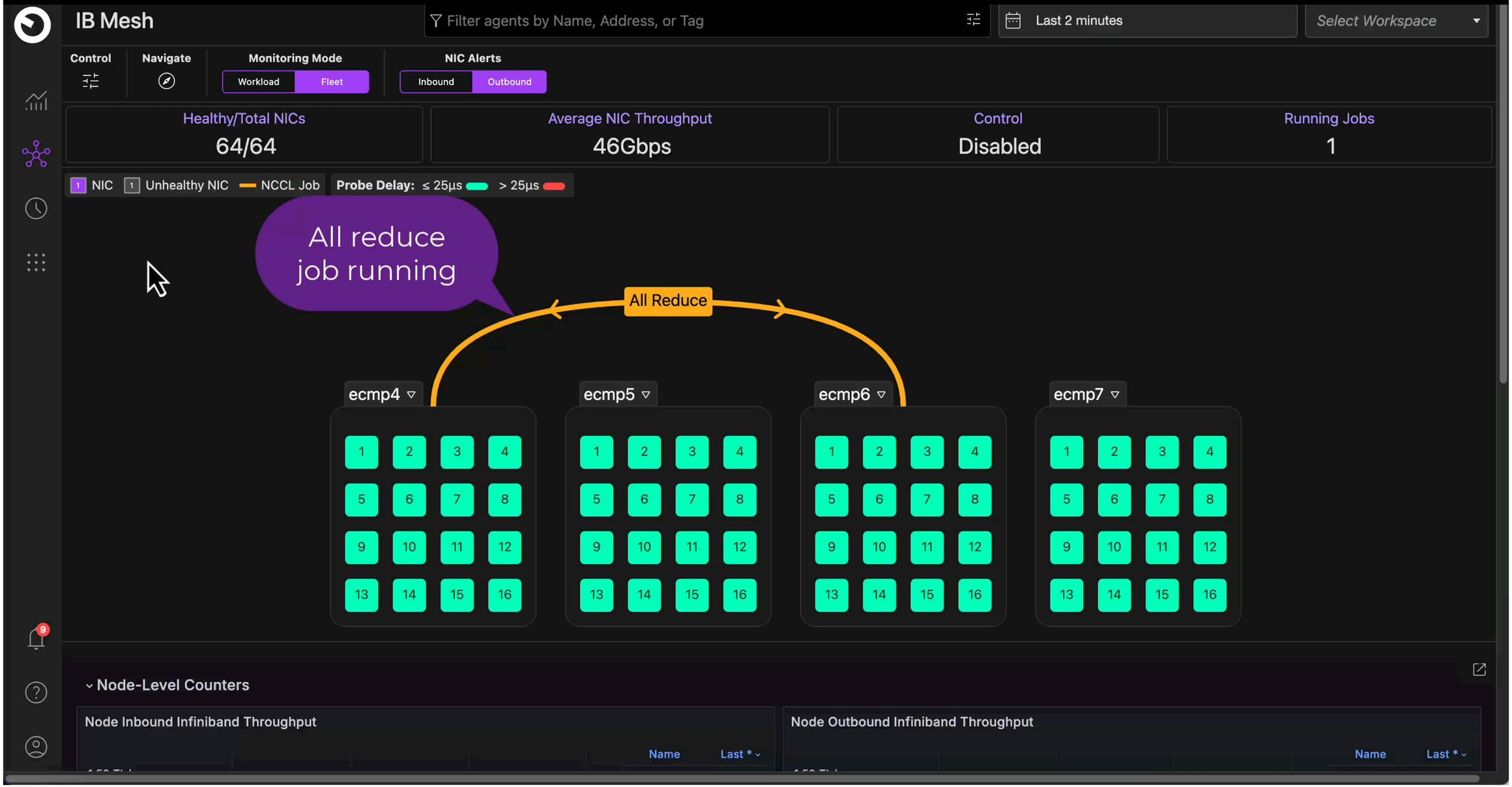



Source: semianalysis

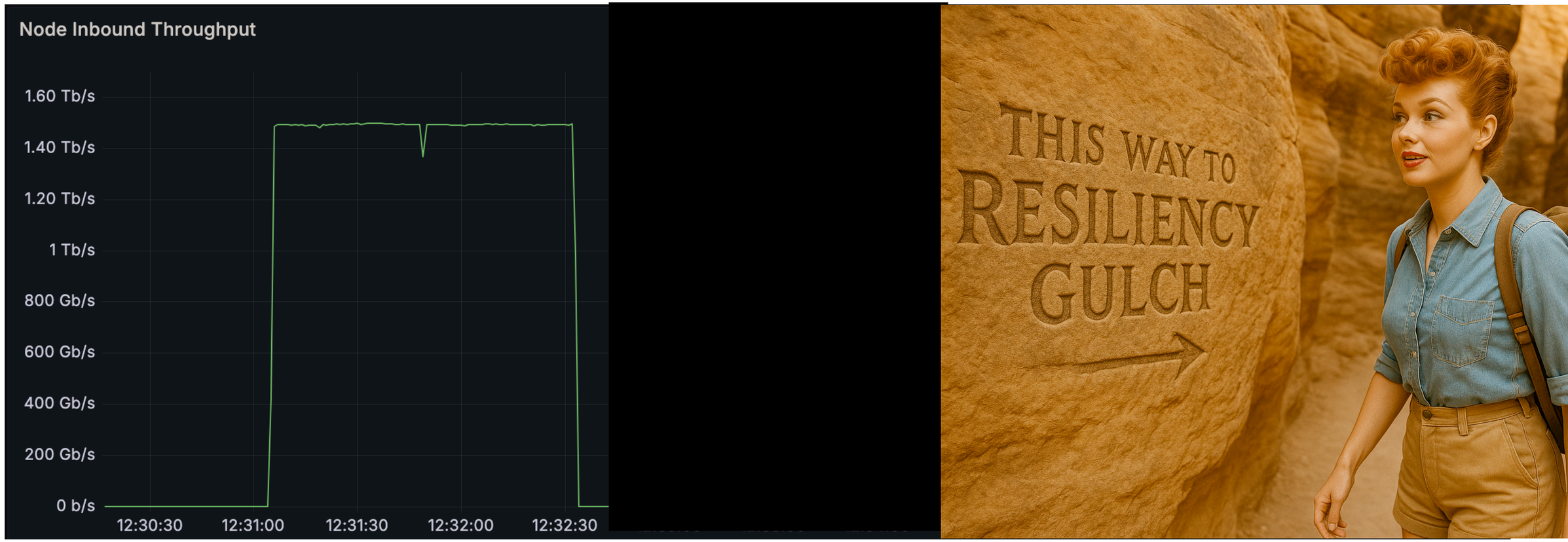
# Link/NIC Flapping



# NIC flapping demo



# Resilience to NIC Flaps in Oracle Cloud



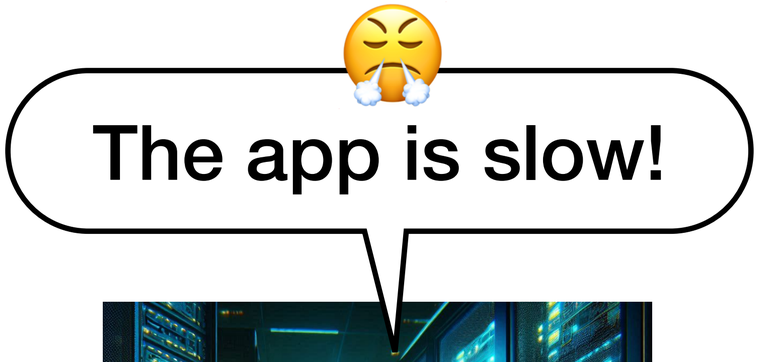

Run 1: No Failure

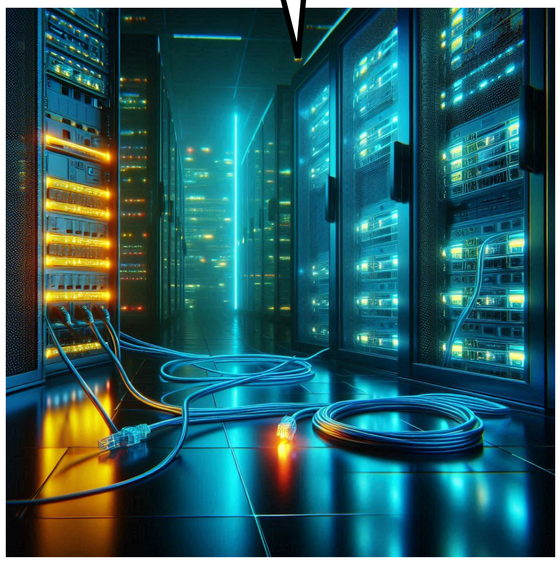
Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

# 3 Lack of predictable performance

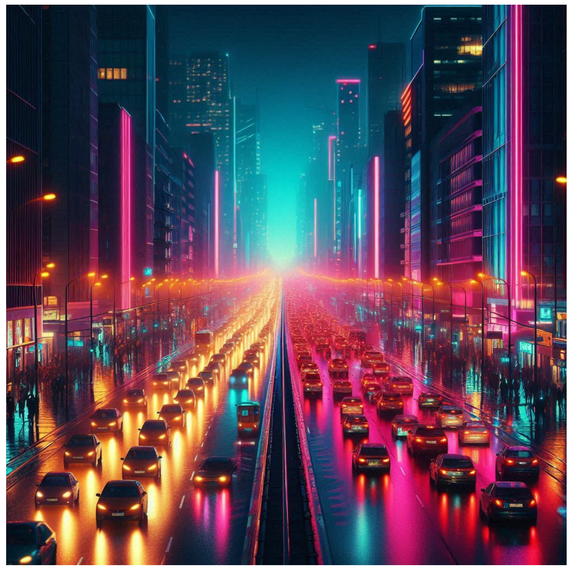
🚨 **Alert**



**Why?**

Flows contending for bandwidth




**So what?**

Low throughput, high latency

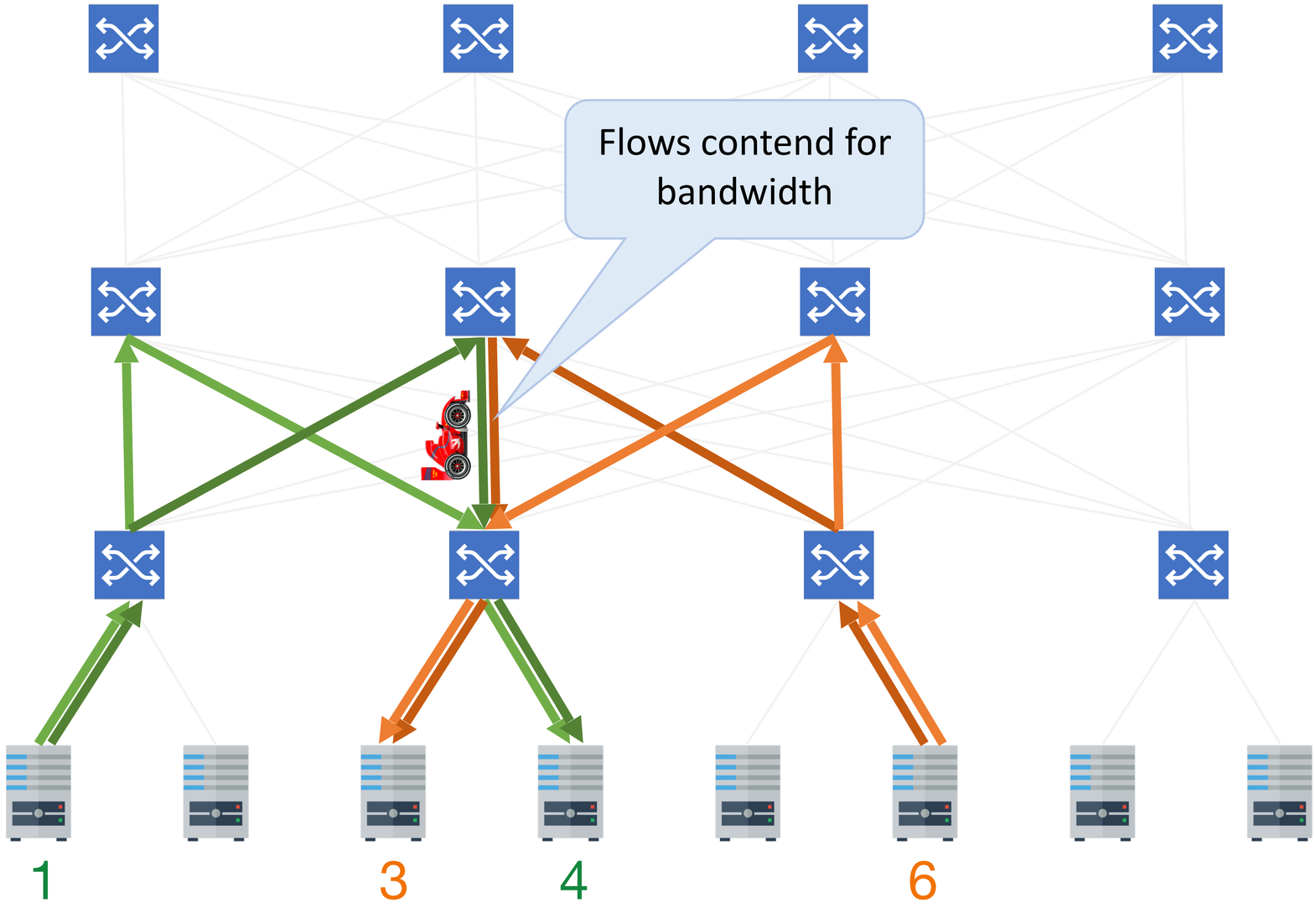


AI-generated images

# Fabric Contention

## Contention

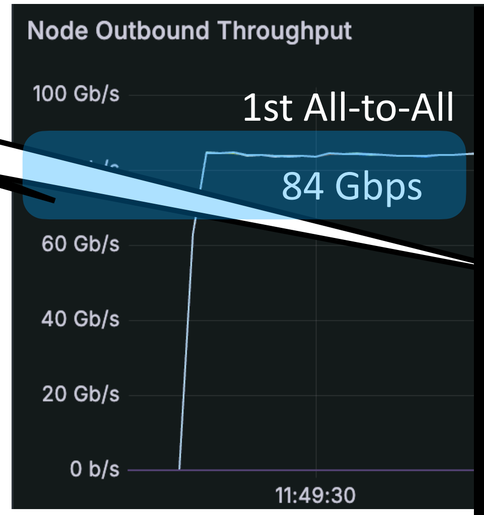
- Queue pairs collide contending for bandwidth
- Senders are not aware which paths are used vs which are free to use



# Detecting and eliminating contention on ECMP Test Bed

⬇️ Lower Throughput!

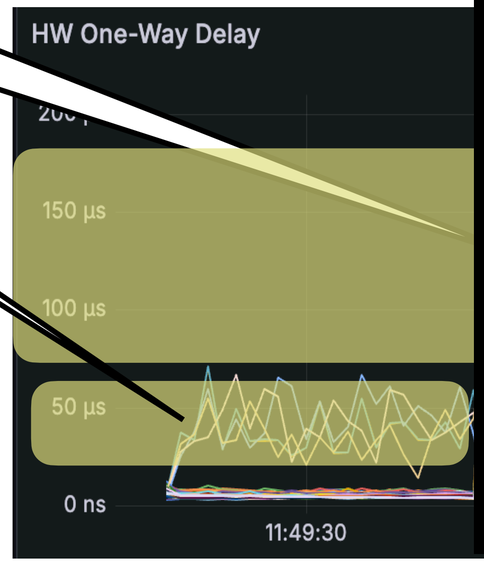
Throughput




🕵🏽 Contention!

🐢 Queue pairs with high one way delays!

One-way Delays




# Key Takeaways

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

Visibility

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


Reliability

💨 Delays in slow queue pairs reduce overall job throughput.

Performance



AI-generated images

# One **Unified** Circuit, Many Cars, Mixed Parts!

- Training, inference and now storage workloads converging on one network.

*Ah yes… I see…*

*less partitioned, and*

*more UNIFIED AI networks*

*in your future!*


AI-generated image

# The Road Ahead

🔥 Efficiency > Speed + Scale

👋🏽 Bye Monolithic Infra,

🙋🏽‍♀️ Hello Heterogenous Envs

*Cross-vendor, mixed generation GPUs, NIC speeds!*

🧠 Smart Cars 🏎️

👁️ Visibility, 💪🏽 reliability, and 💨 performance


AI-generated image

# Thank You

Huygens Paper
& Slides


🖖🏽 lerna@clockwork.io