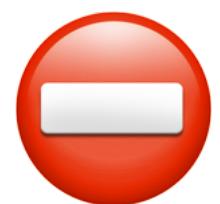


Lack of fine-grained  
visibility  
at queue pair level



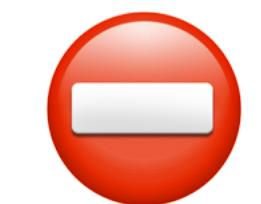
Pre-race & race time!



Visibility



NIC/link flapping



Reliability

# 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



# DEEP SPACE

Large scale  
jobs

1

DeepSpaceExplorer LLM Model

2

Deepspeed, PyTorch, etc.

3

NCCL (Nvidia Collective Communications Library)

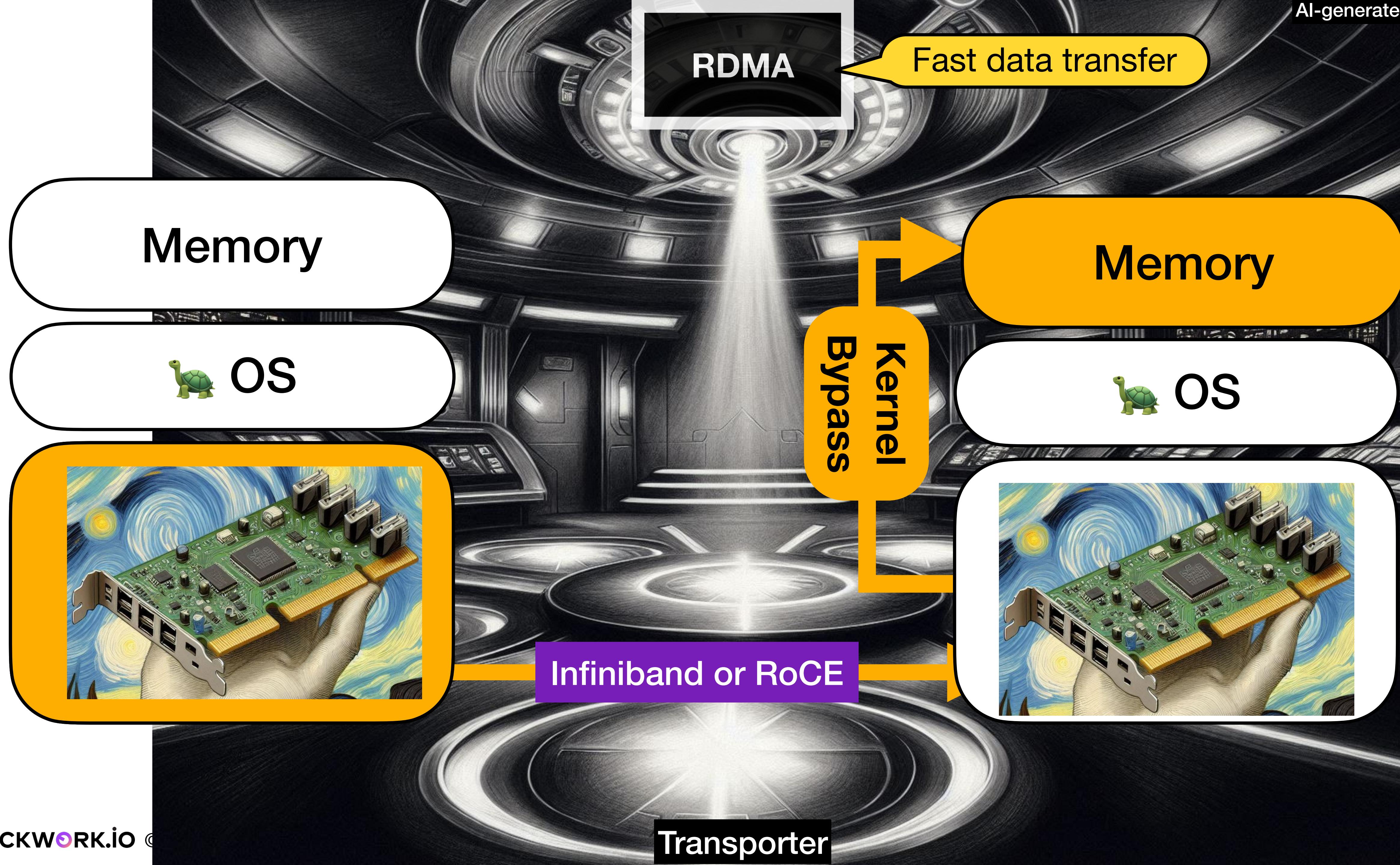
4

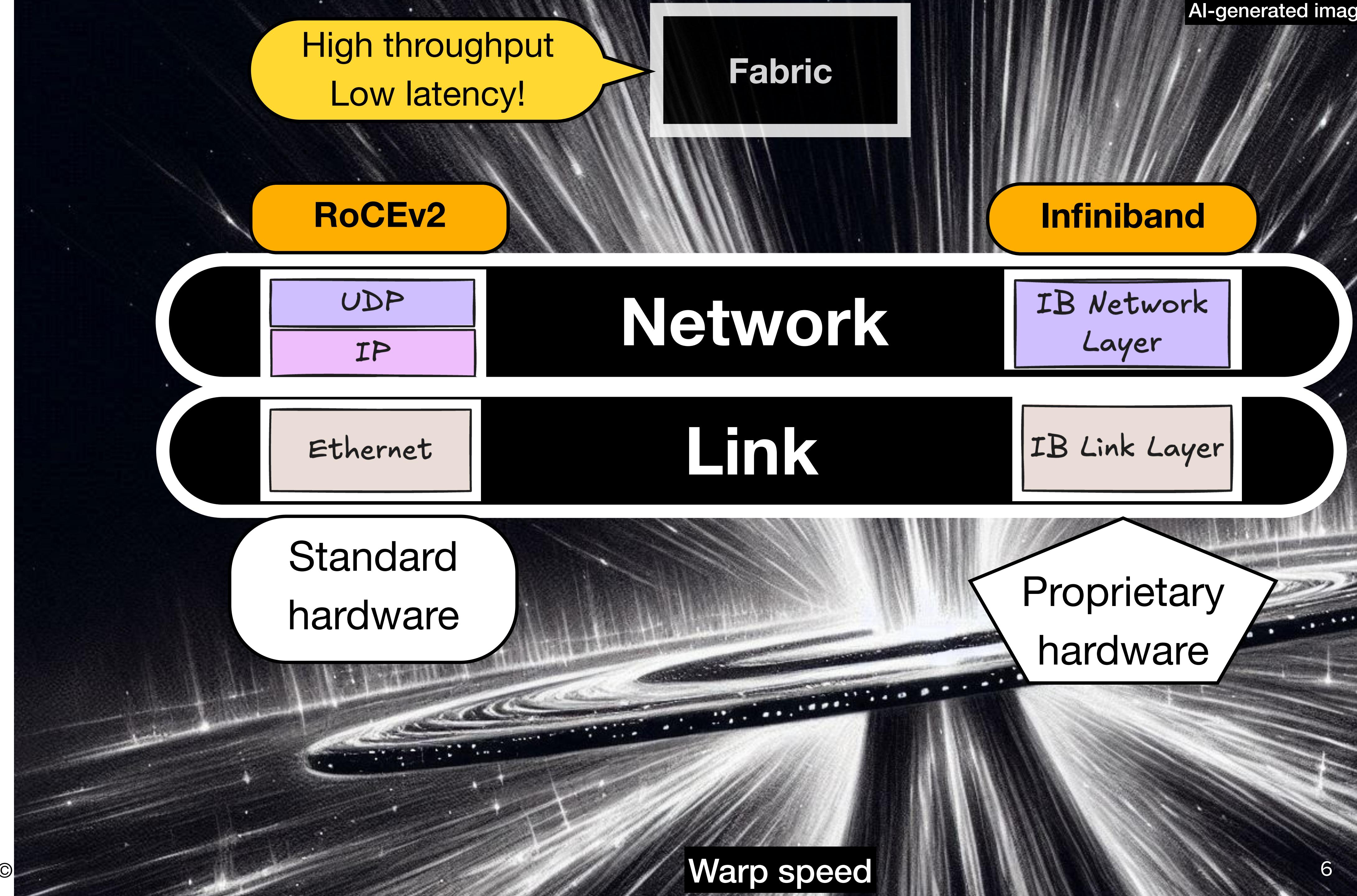
Network Device (Infiniband or RoCE)

Inter-GPU  
communication

RDMA  
capable!







# Agenda

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

# Time Force, Department of Temporal Affairs

Tempus Mundi Servamus



## Traditional networks



Delays  
Best effort service

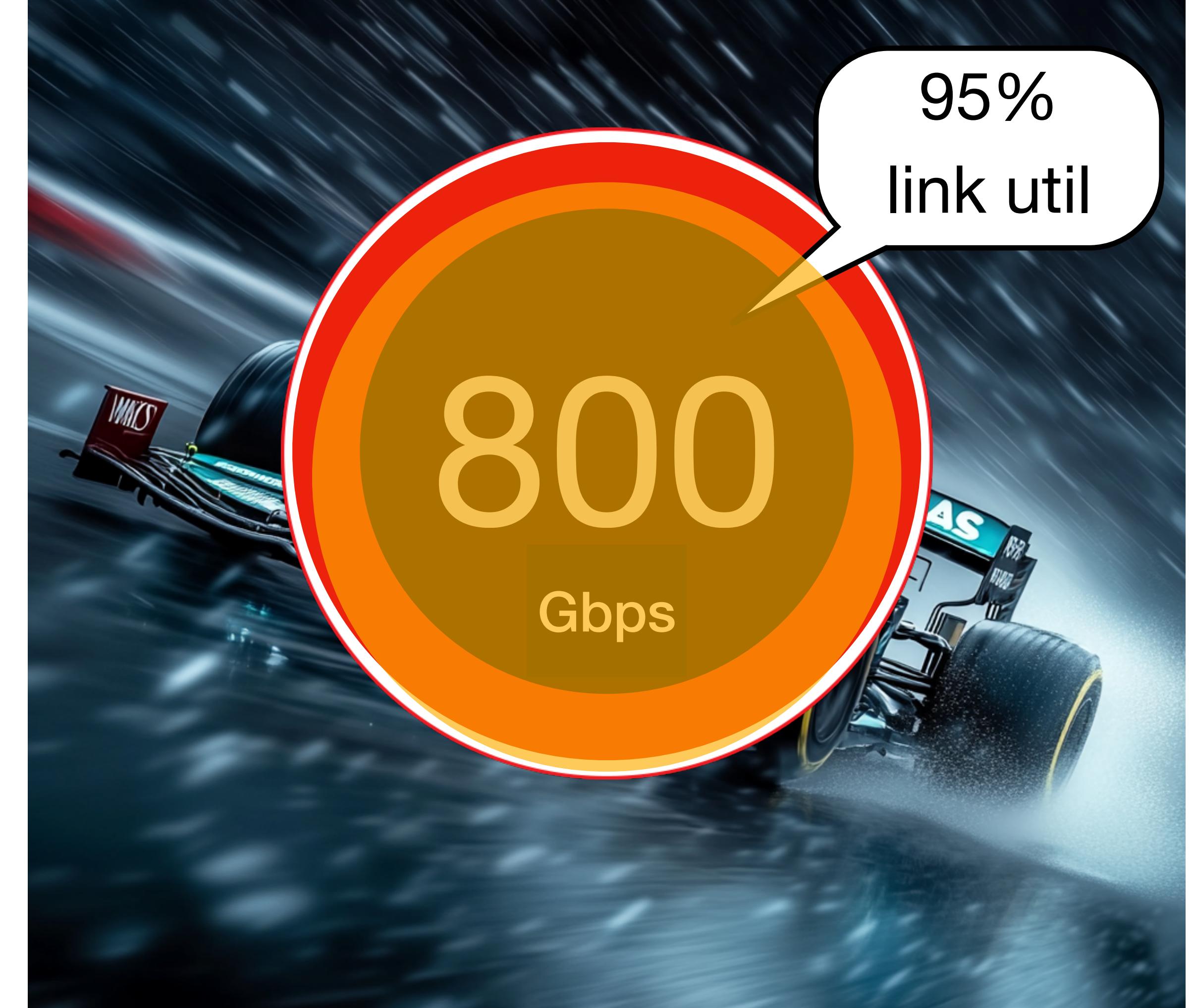
## AI networks



Minimal delays  
Best performance



Traditional networks



AI networks

Speed of 🐭 app doesn't directly affect speed of 🐘 app



Traditional networks

Similar highly correlated apps



AI networks

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

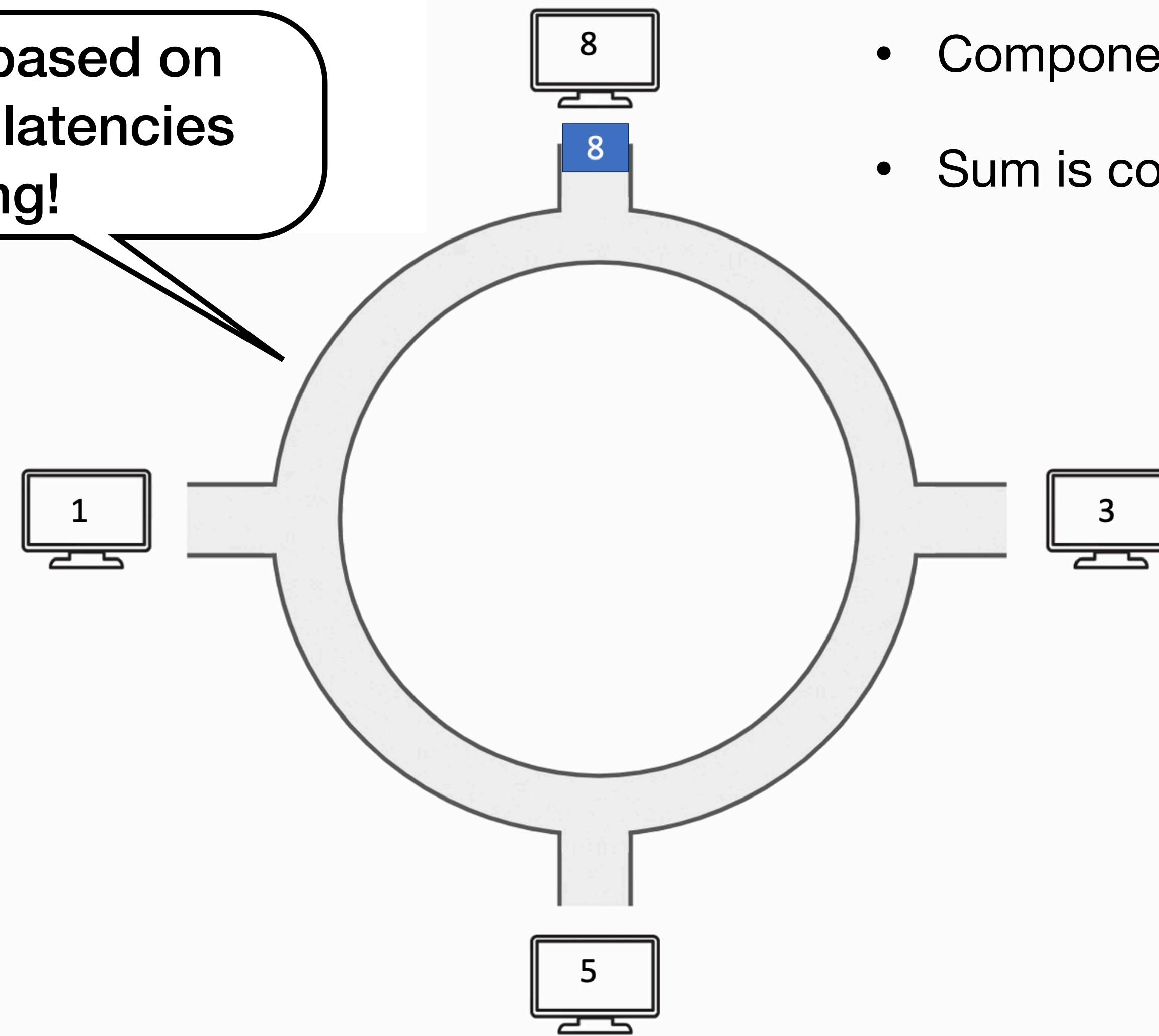
## Profile of an LLM train



Source: AMD  
Source: Nvidia

# All-reduce (NCCL ring implementation)

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

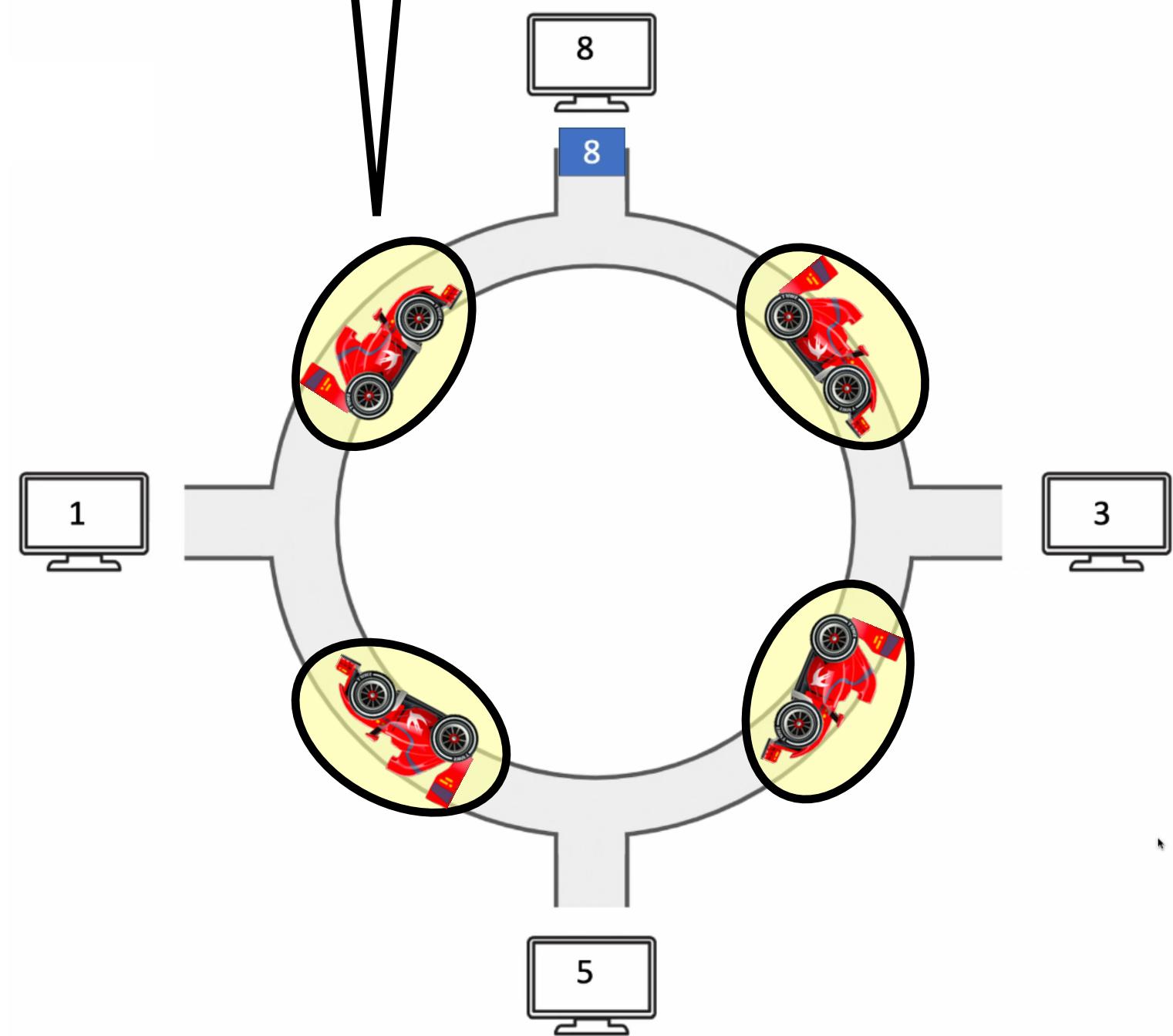


- Components of the sum get passed around
- Sum is computed then passed around



# Communicate

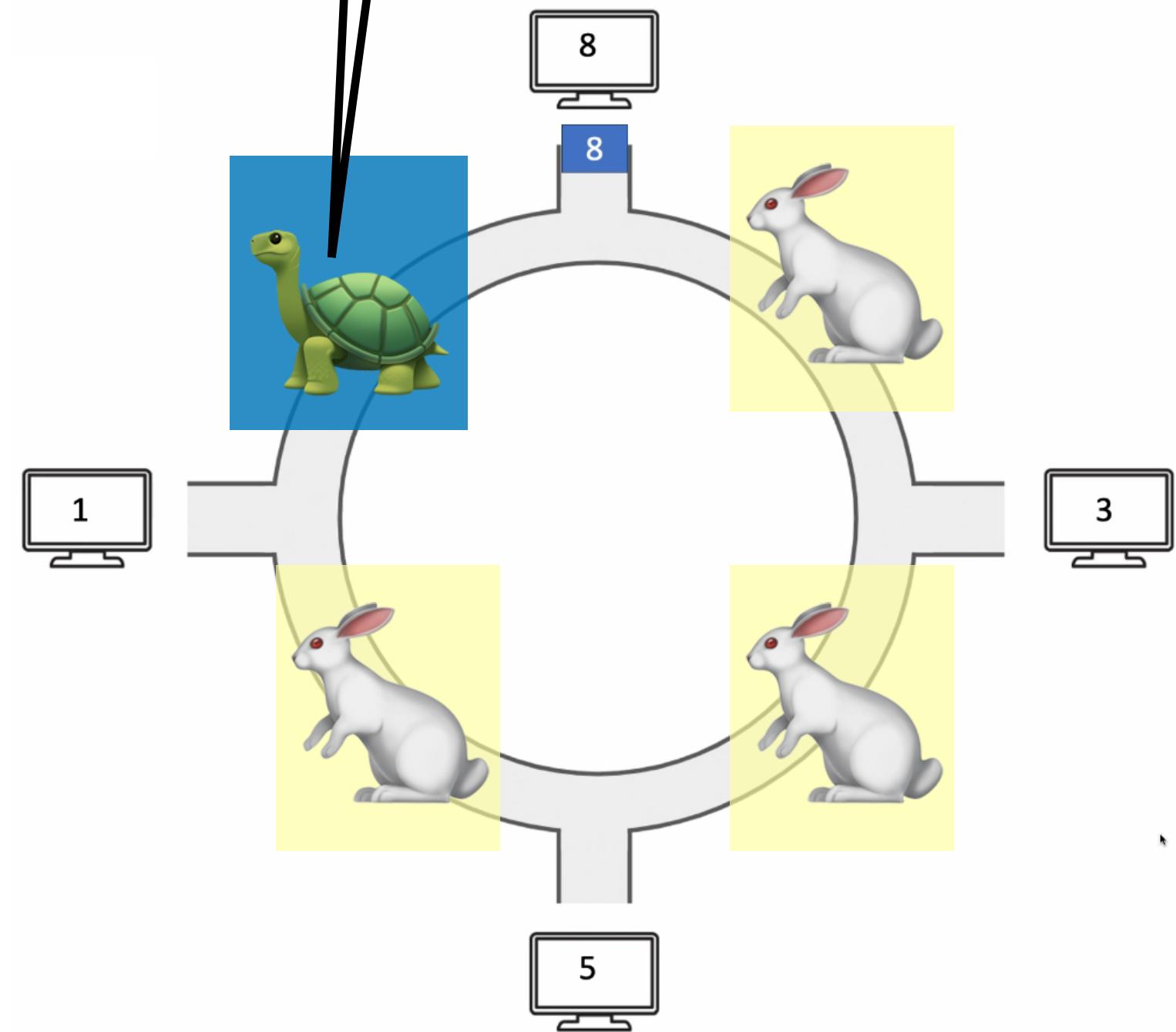
GPU's won't proceed to next step until all synchronize partial results.



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



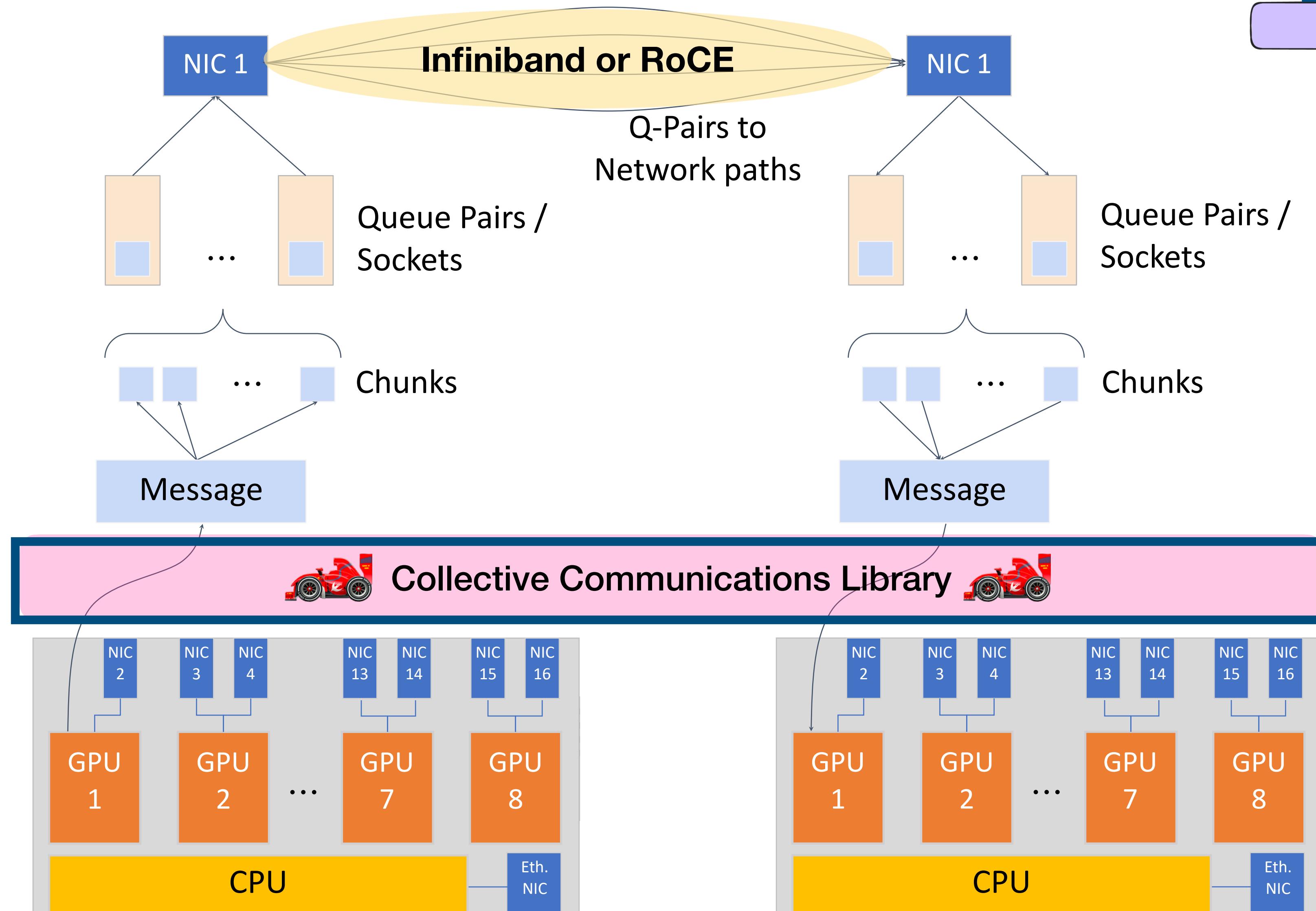
Slowest flow (or node)  
determines throughput.



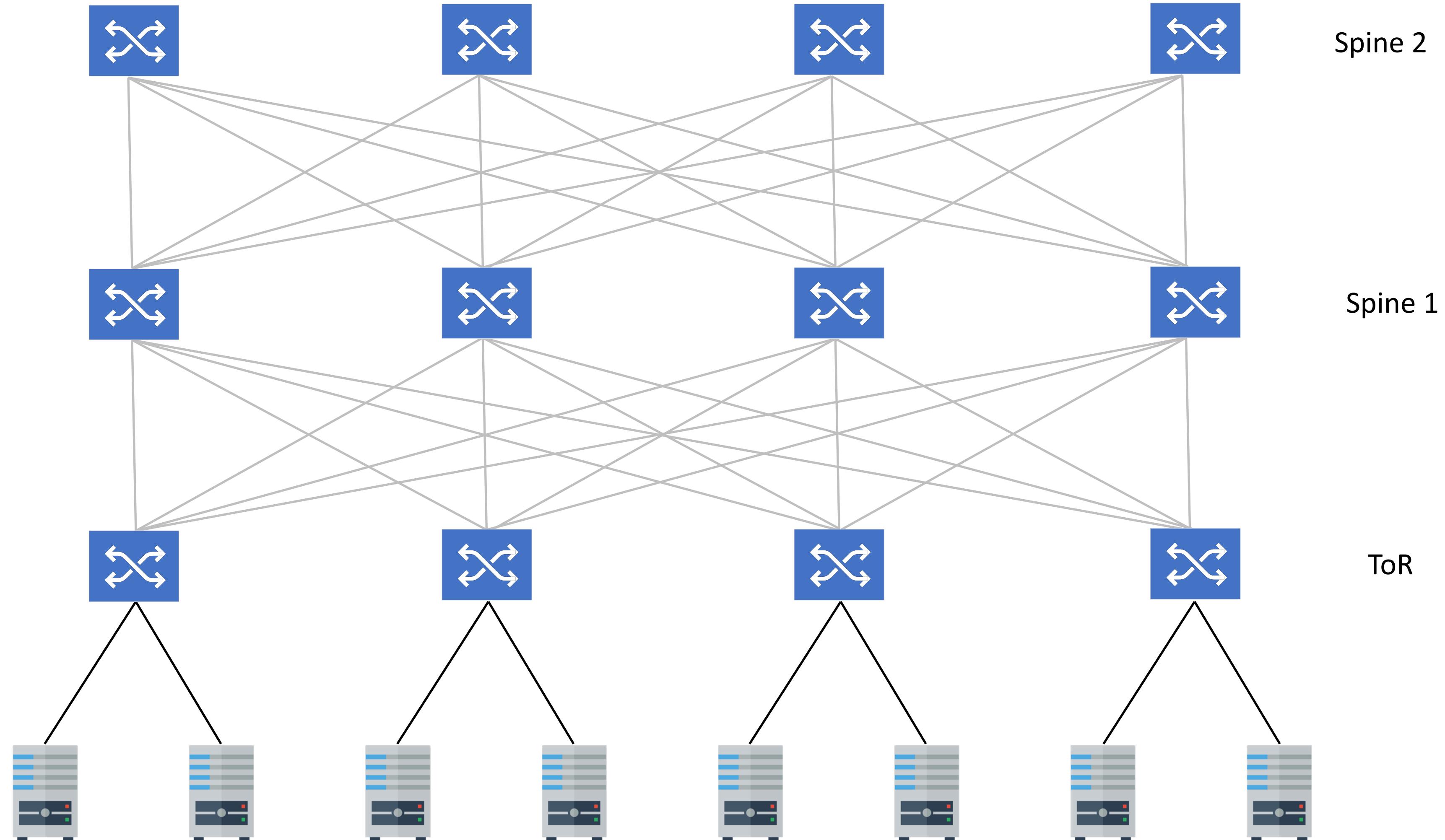
Slowest  
determines speed.



# 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

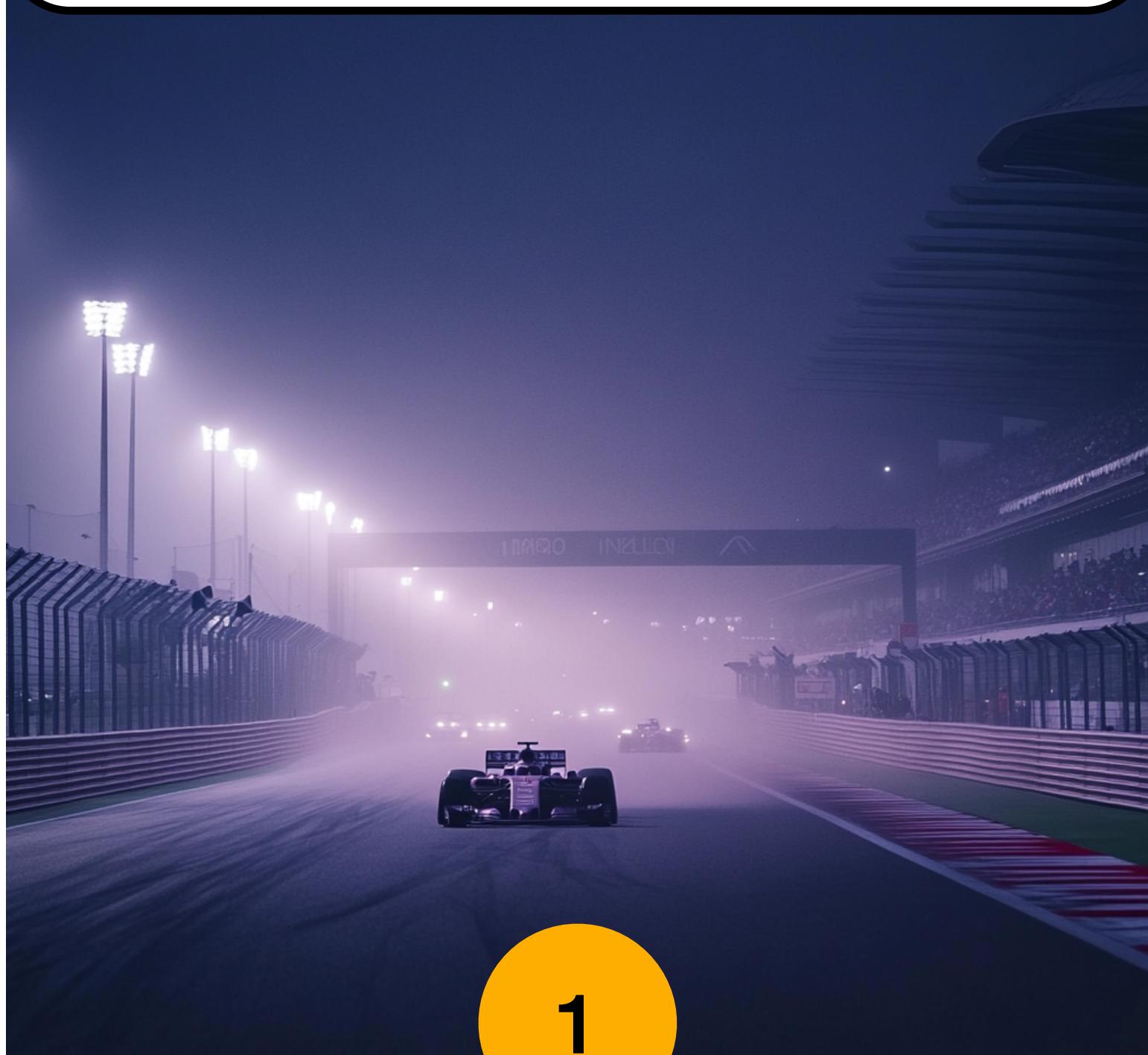
Tempus Mundi Servamus

Tribble on  
Deep Space Explorer

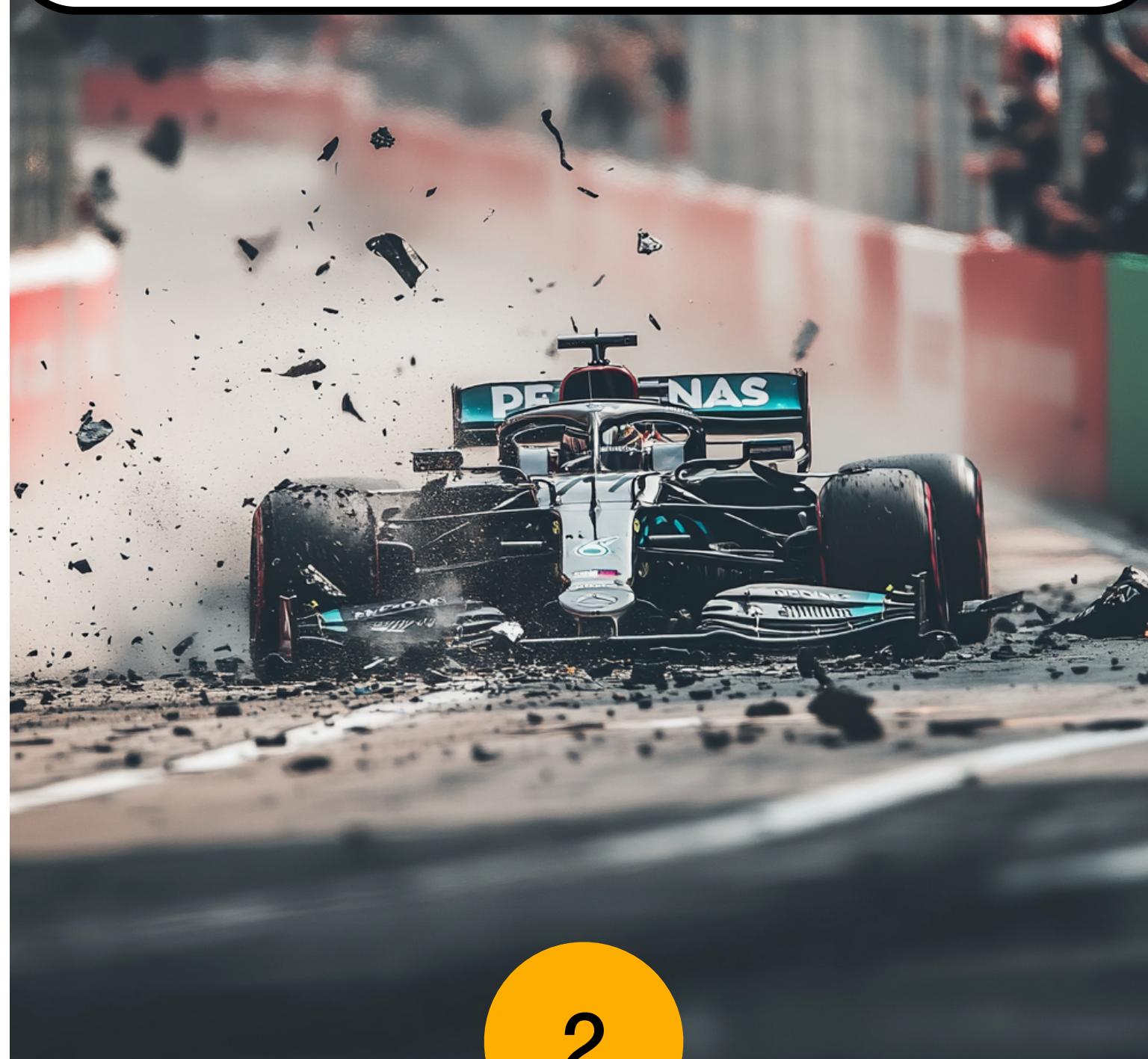
# Networking challenges of AI/ML workloads



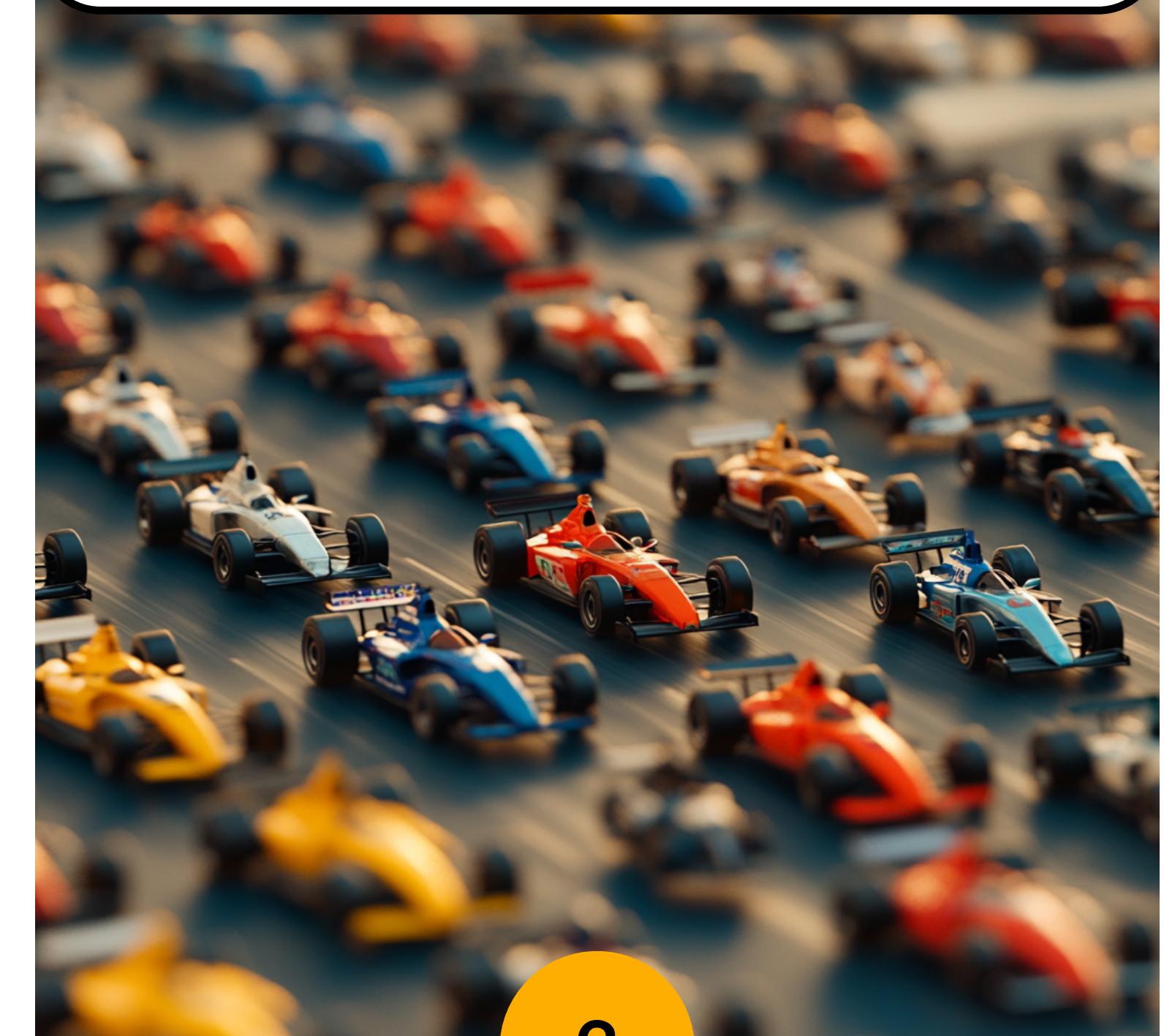
Visibility



Reliability



Performance

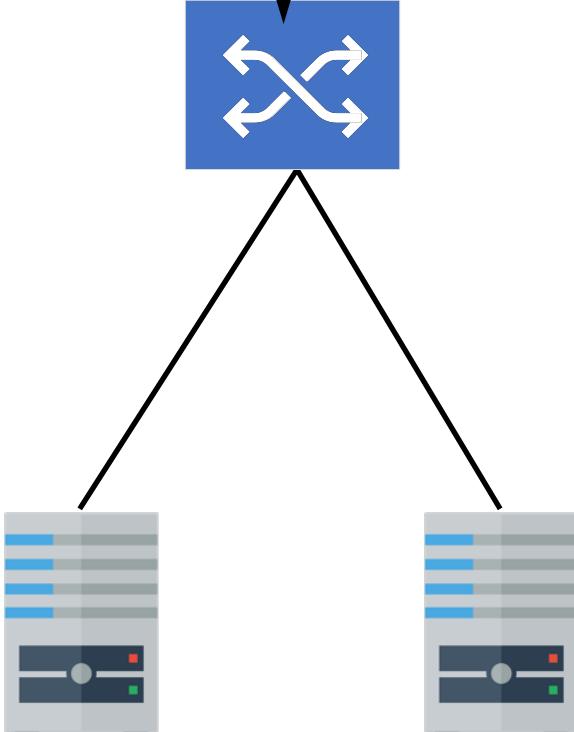


1

# Lack of visibility



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



## 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

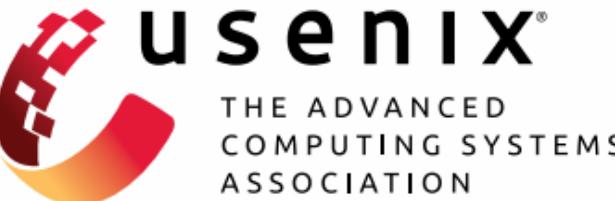
# Visibility



# Huygens clock synchronization

- ▶ Software based
- ▶ High precision at scale 

Measures  
one way delays



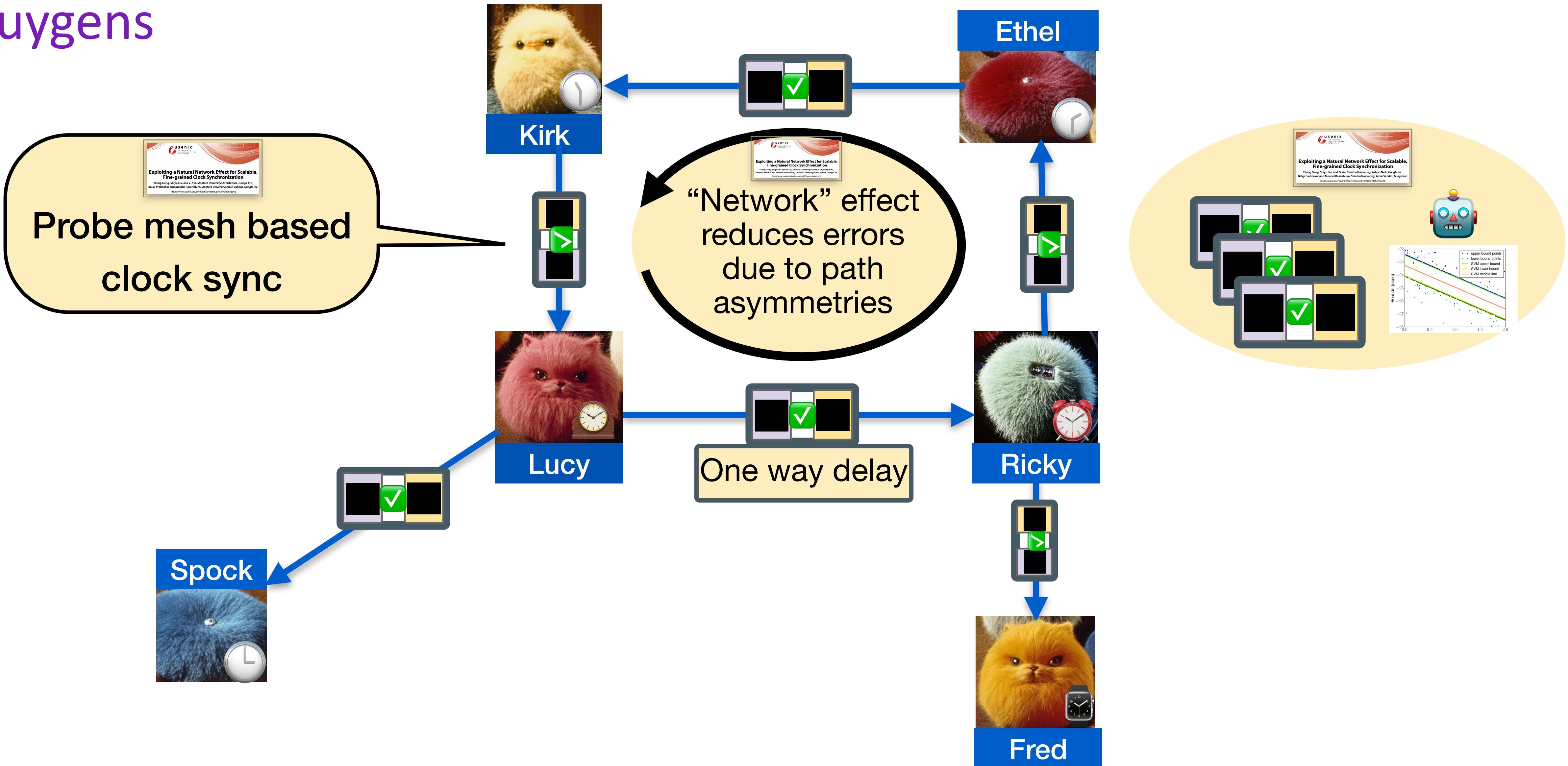
**Exploiting a Natural Network Effect for Scalable, Fine-grained Clock Synchronization**

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

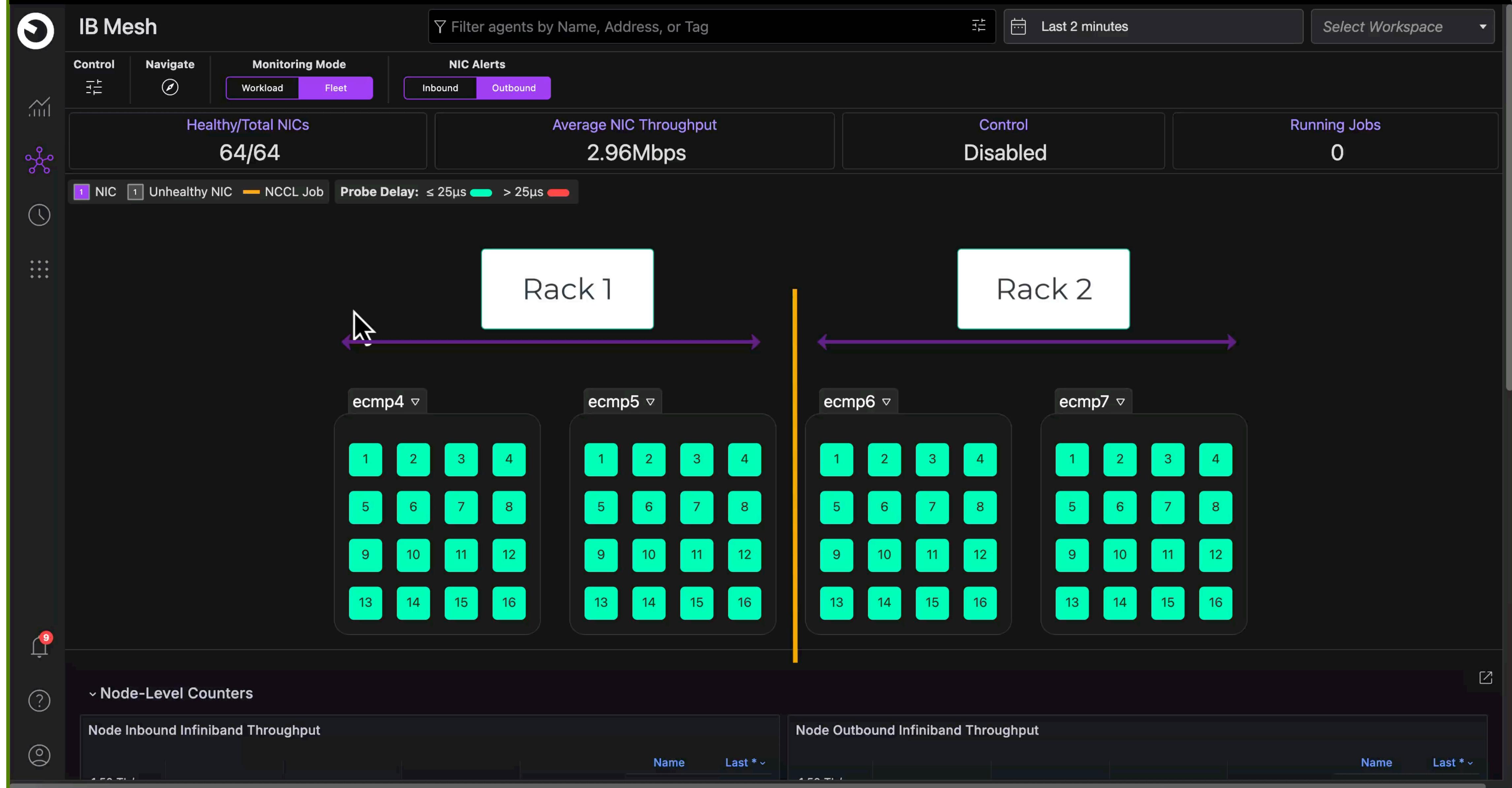
<https://www.usenix.org/conference/nsdi18/presentation/geng>



# Huygens



# Visibility demo



2

## Lack of reliability



Link flapped 10 times!  
Send technicians  
to swap cables!



Common in large clusters



Source: [semianalysis](#)

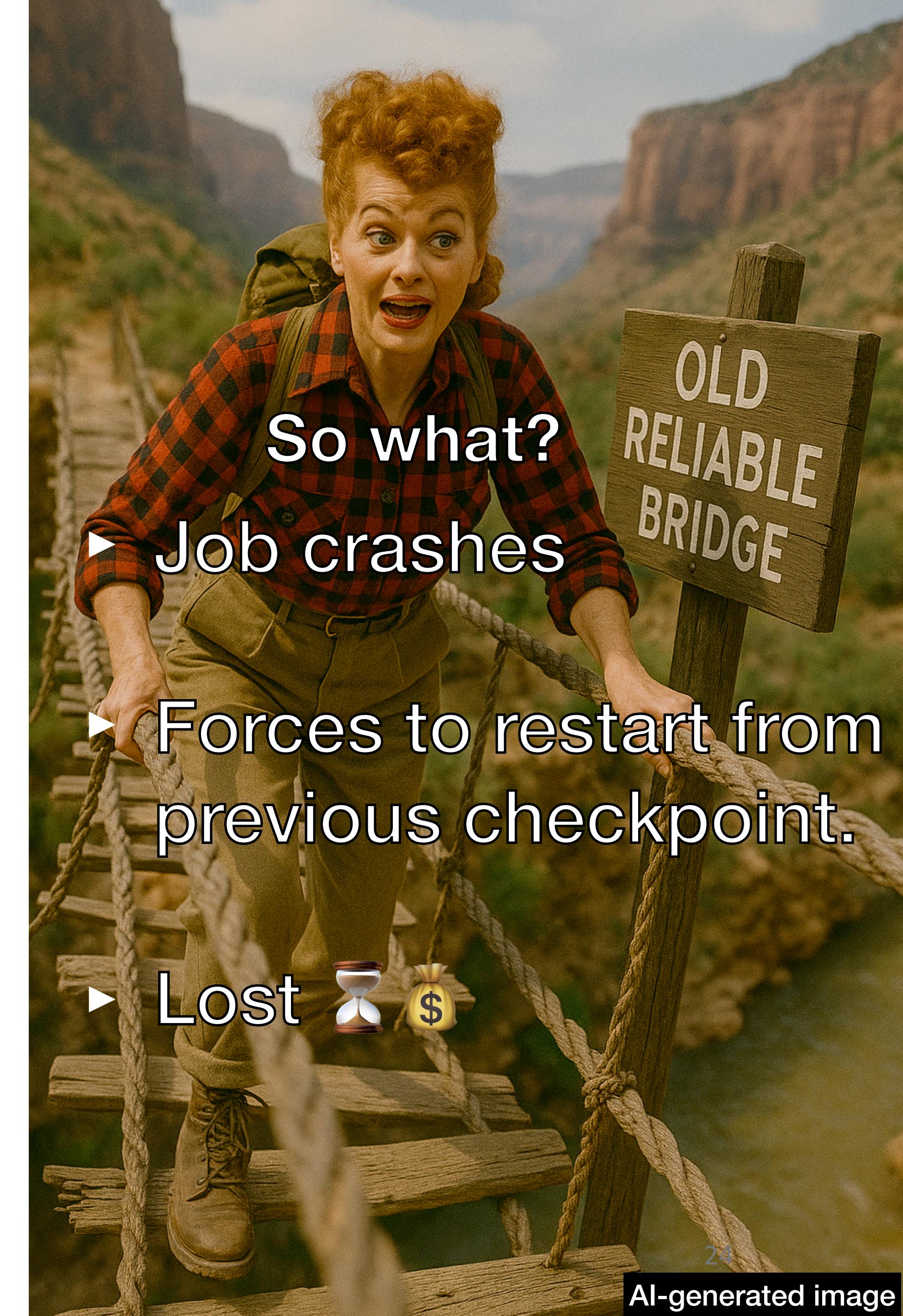
### Why?

So what?

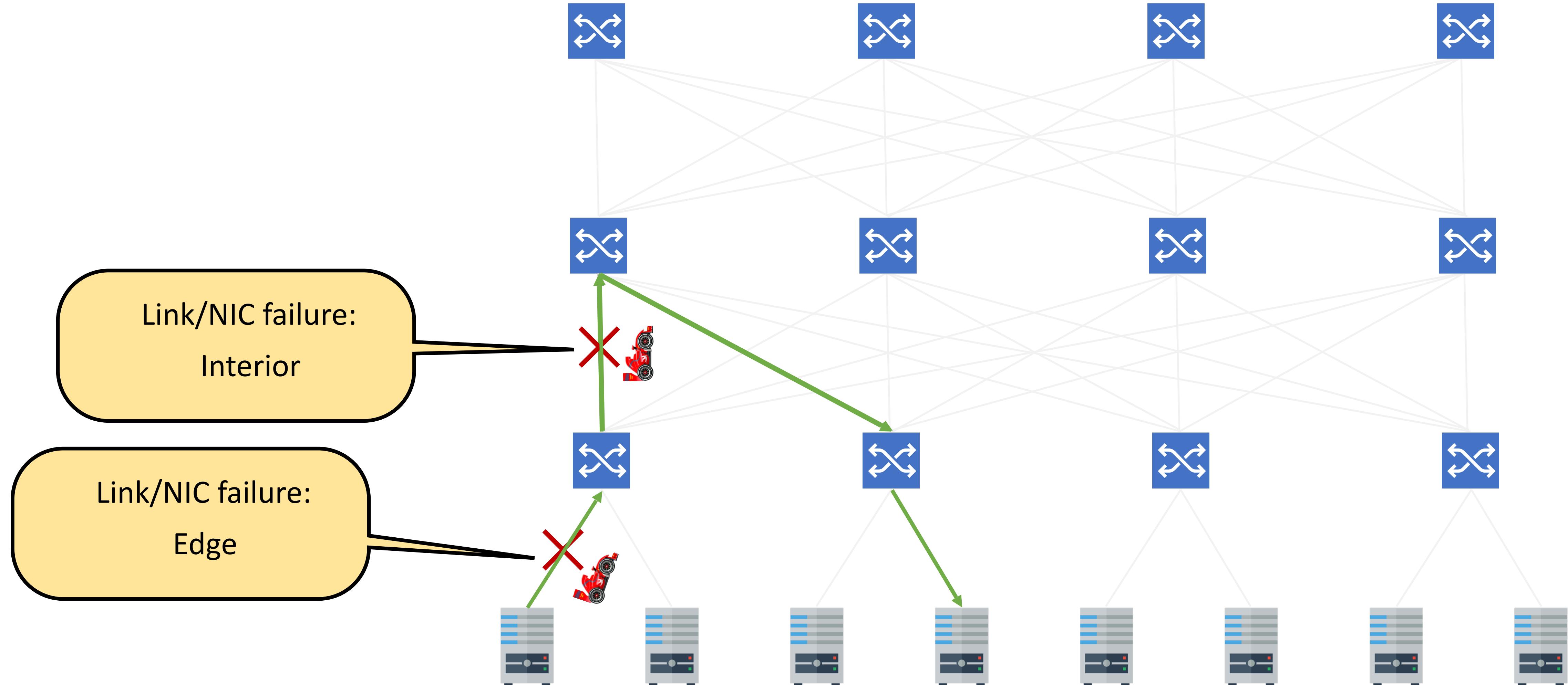
► Job crashes

► Forces to restart from previous checkpoint.

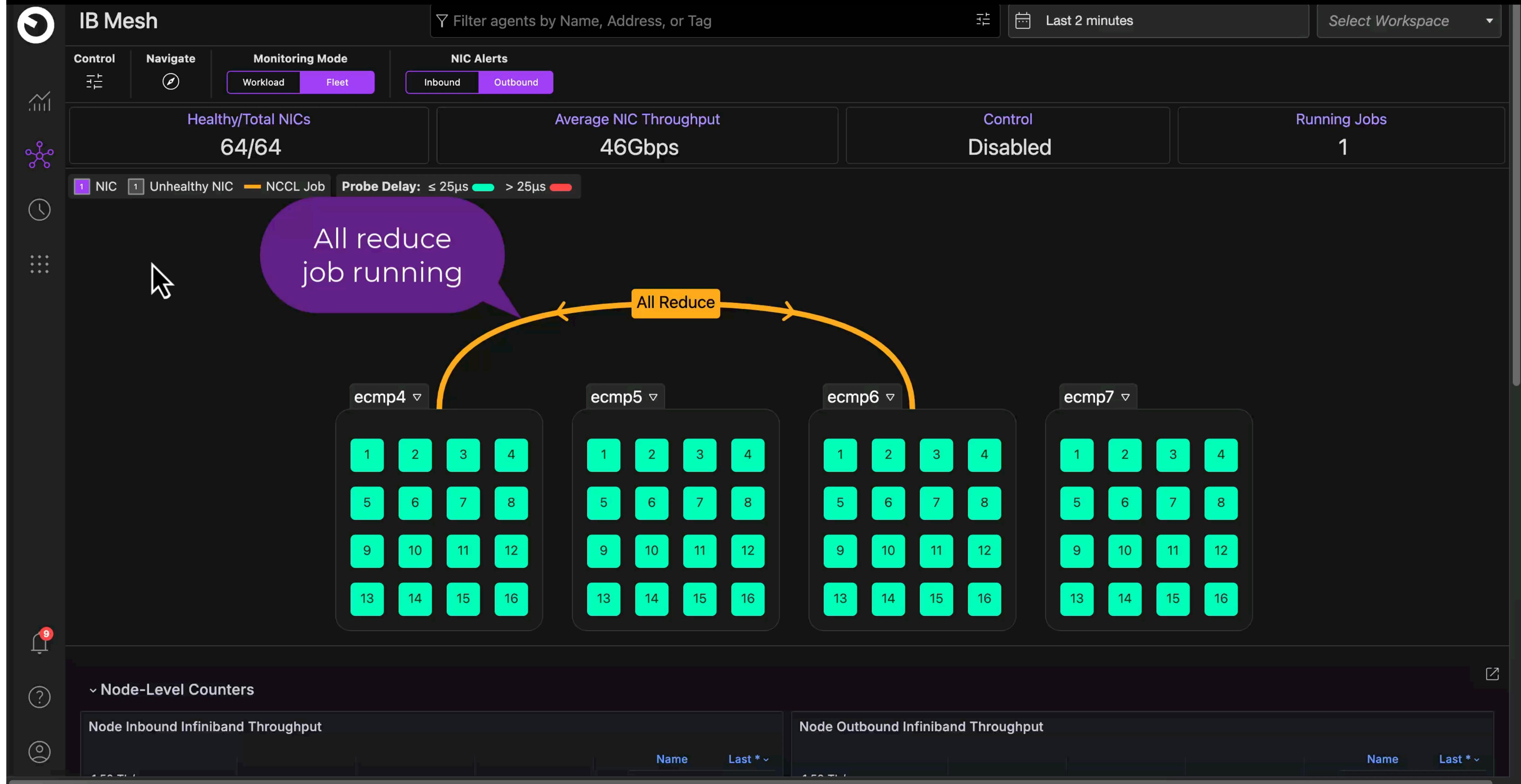
► Lost



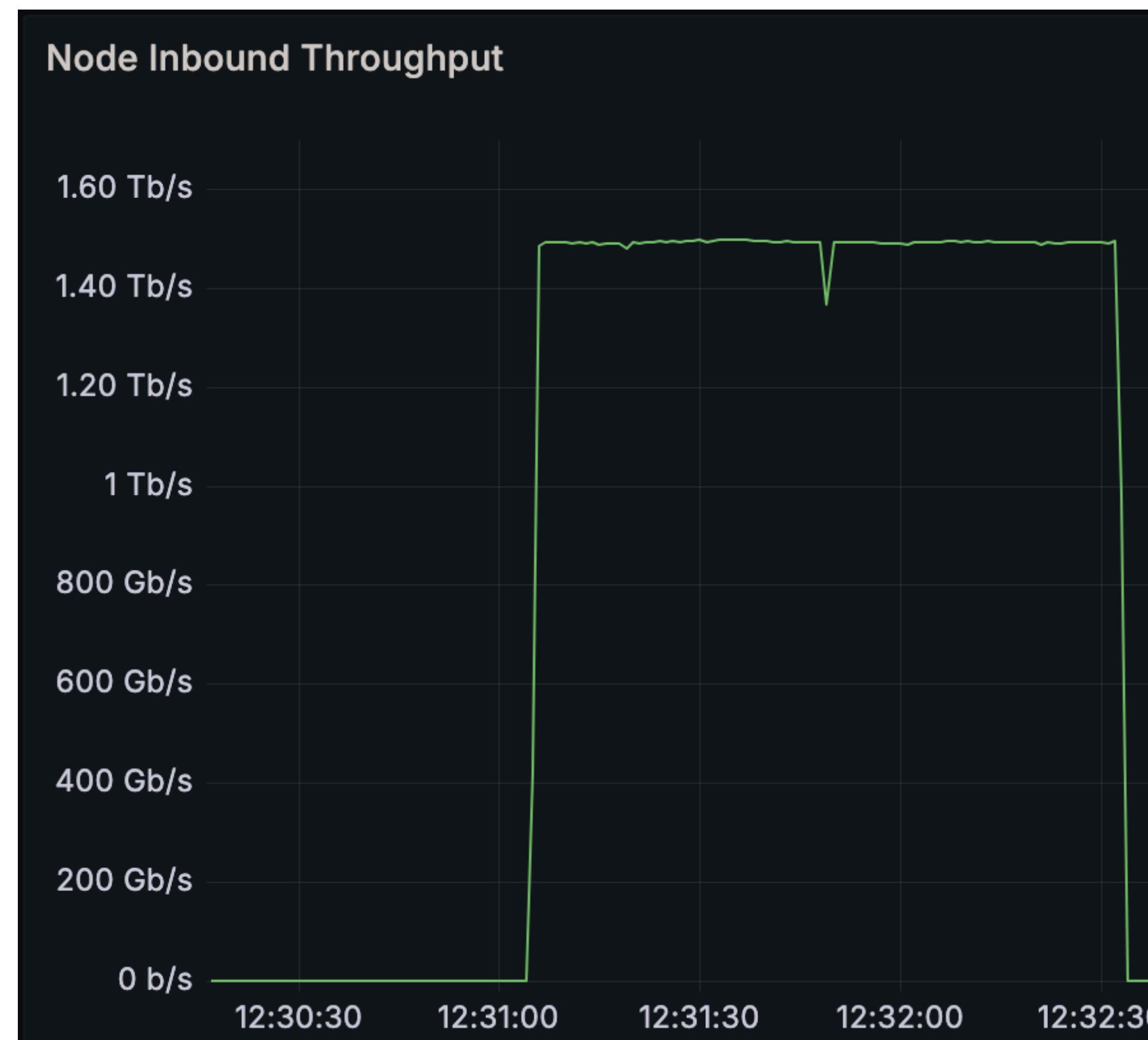
# Link/NIC Flapping



# 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

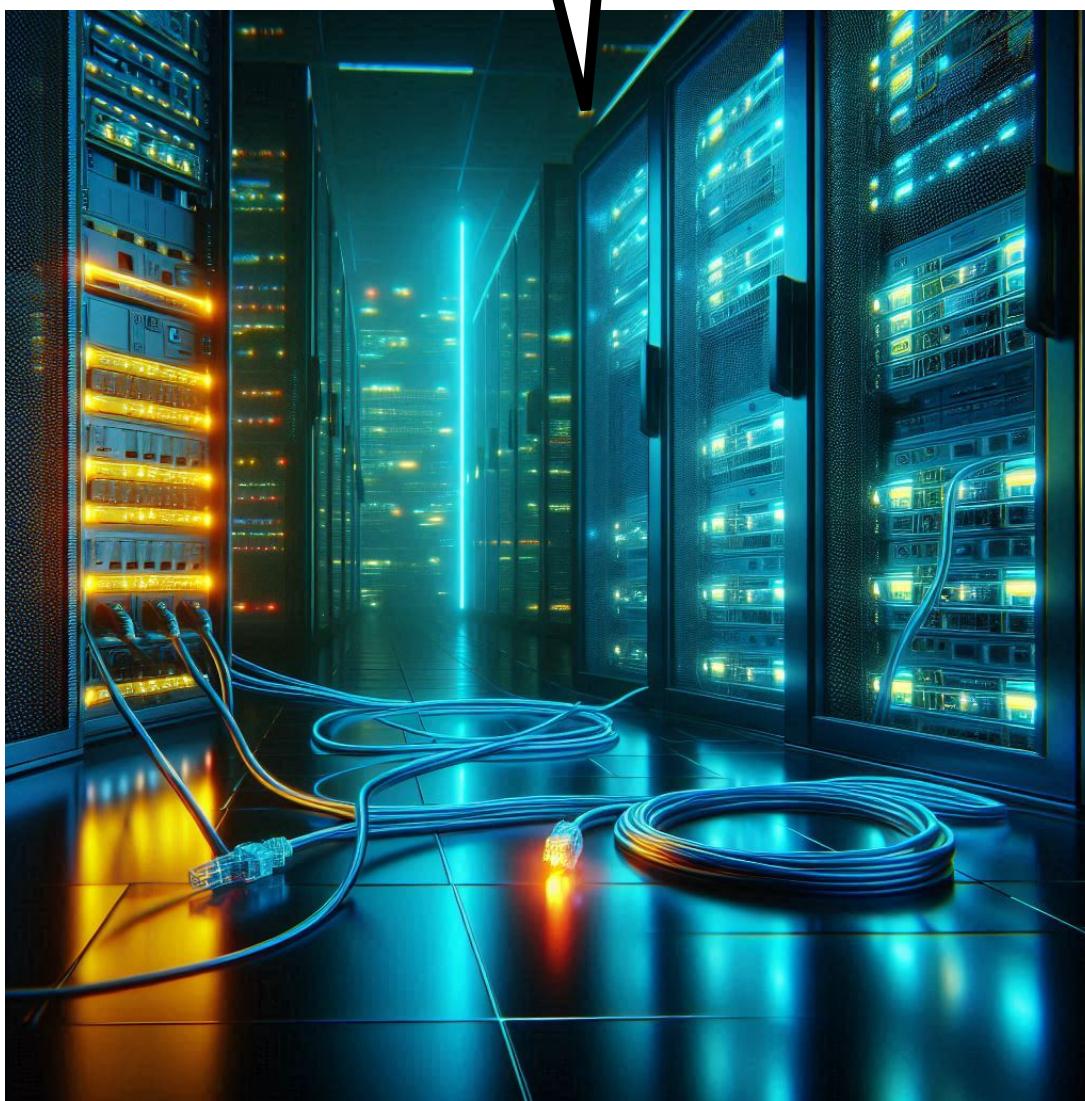


3

## Lack of predictable performance



The app is slow!



Why?

Flows contending  
for bandwidth



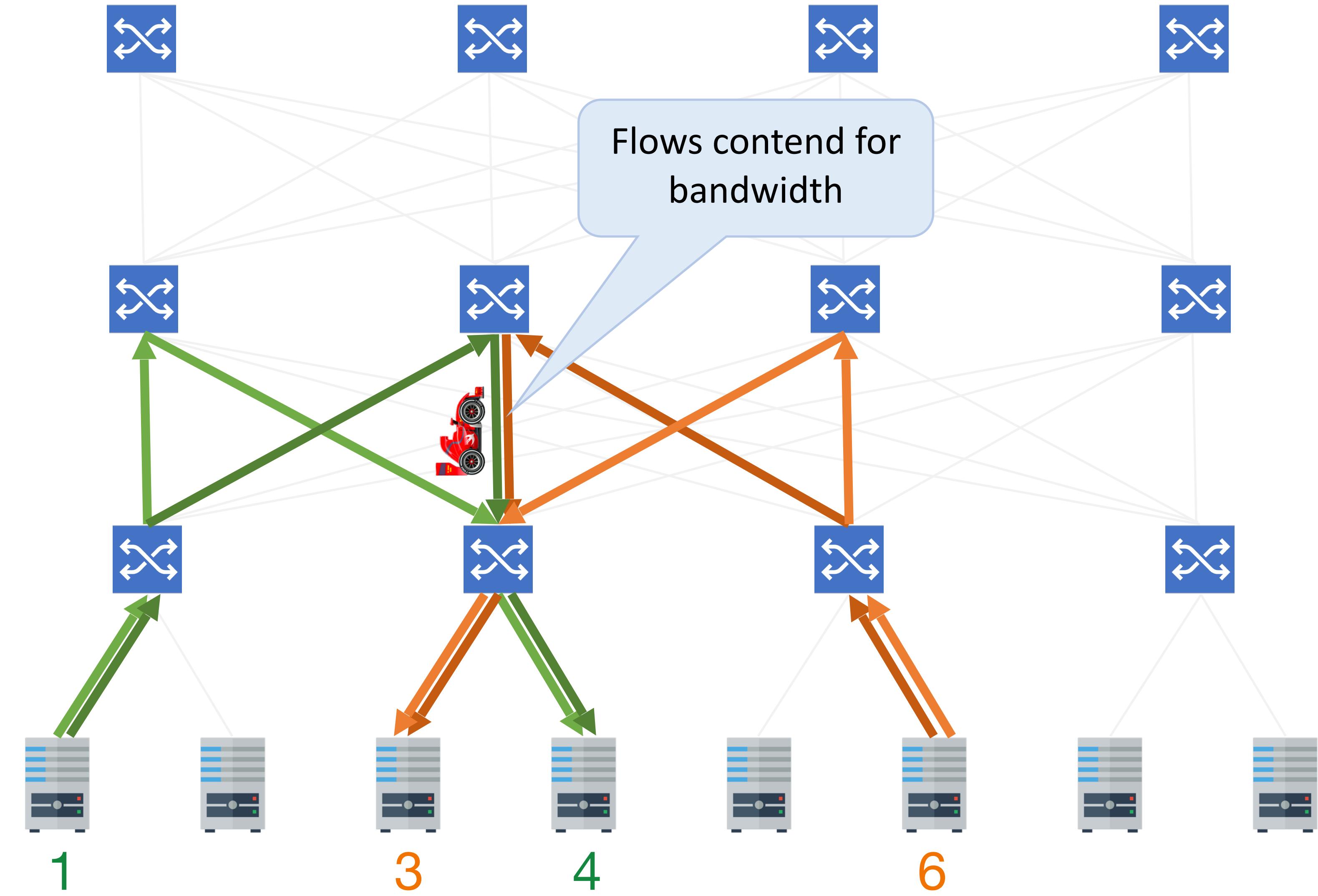
So what?

Low throughput,  
high latency

# Fabric Contention

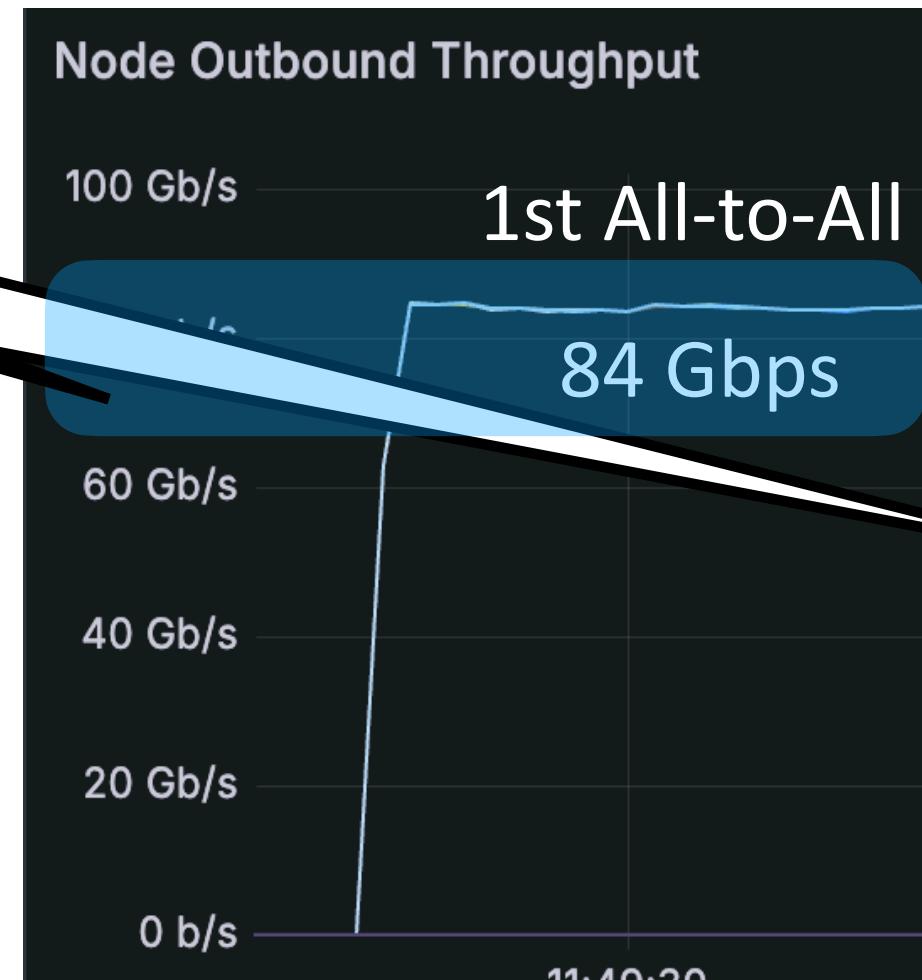
## Contention

- 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

T  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

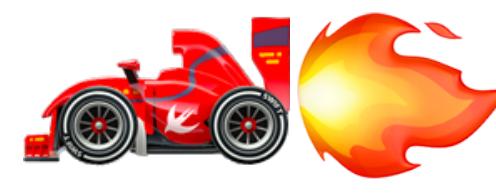


# One Unified Circuit, Many Cars 🚗, 🏎️, Mixed Parts!

- ▶ Training 🏎️, inference 🚗 and now storage 🚛 workloads converging on one network.

*Ah yes... I see...  
less partitioned, and  
more UNIFIED AI networks  
in your future!*





# The Road Ahead

🔥 Efficiency > Speed + Scale

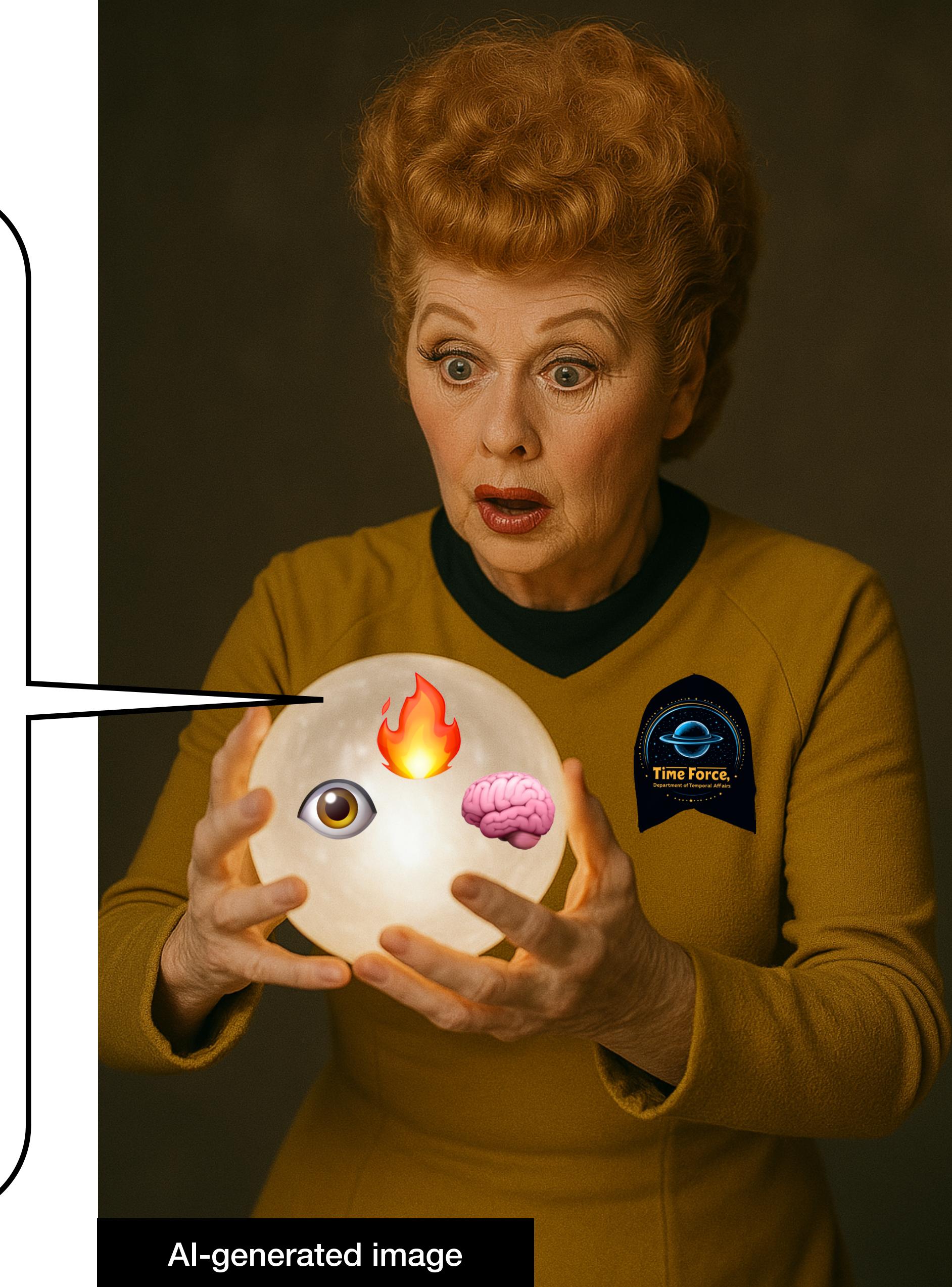
👋 Bye Monolithic Infra,

🙋 Hello Heterogenous Envs

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

🧠 Smart Cars 🏎️

👁️ Visibility, 💪 reliability, and 💨 performance



AI-generated image

# Thank You

Huygens Usenix Paper  
& Slides



[lerna@clockwork.io](mailto:lerna@clockwork.io)

