
rise to the
RISE
TO THE
CHALLENGE
challenge
the challenege
AI-generated


NAS

Resilience for AI Workloads at Scale

Lerna Ekmekcioglu
Sr. Solutions Engineer
Clockwork.io

AI-generated

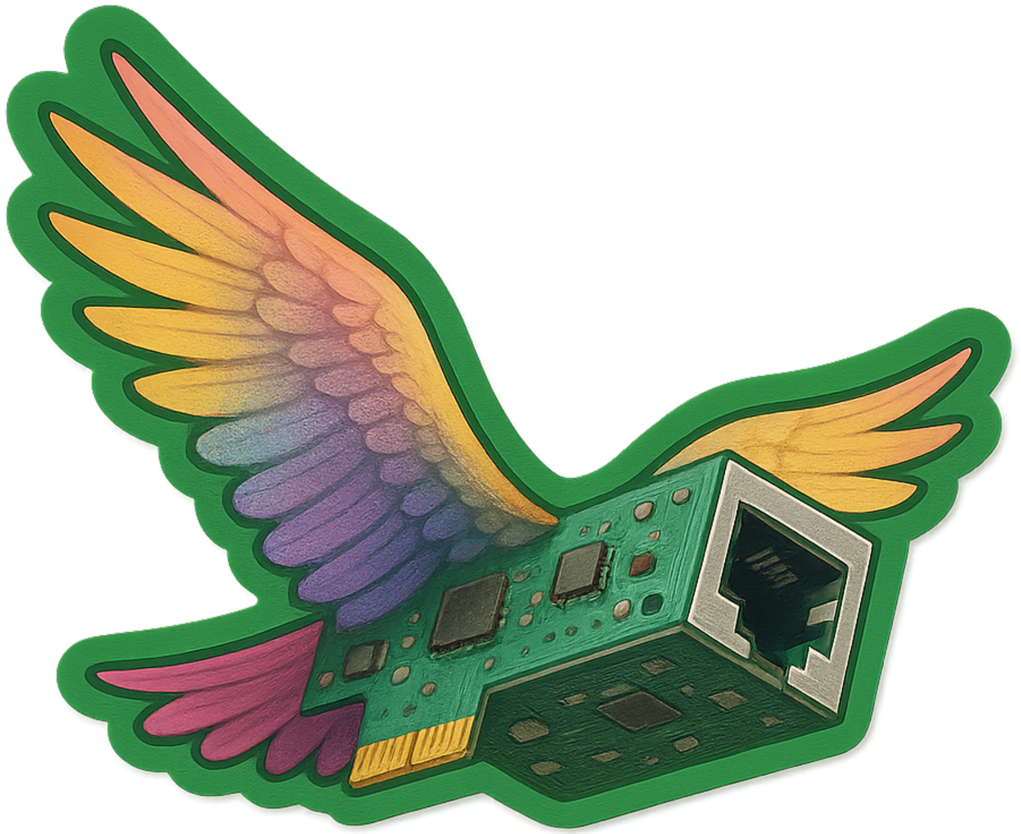
Resilience for AI Workloads at Scale

Lerna Ekmekcioglu
Sr. Solutions Engineer
Clockwork.io

AI-generated


AI/ML Eng/
Researcher

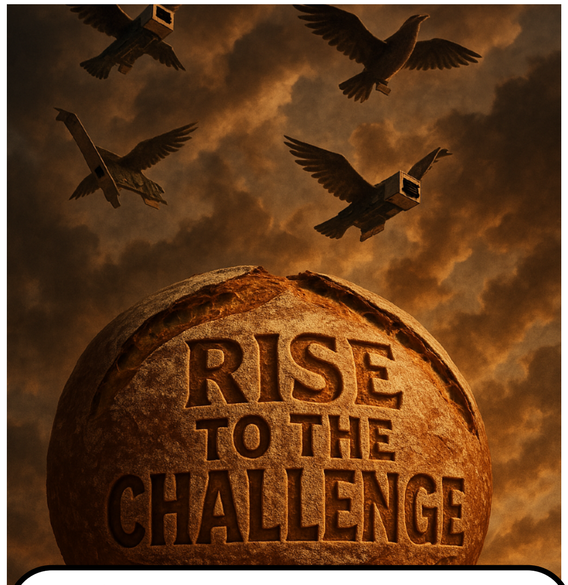

SRE/
Infra eng
AI/ML Eng/
Researcher



Flappy

Challenges for AI workloads

🙈Visibility



🔥 Reliability

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus

CLOCKWORK.IO

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

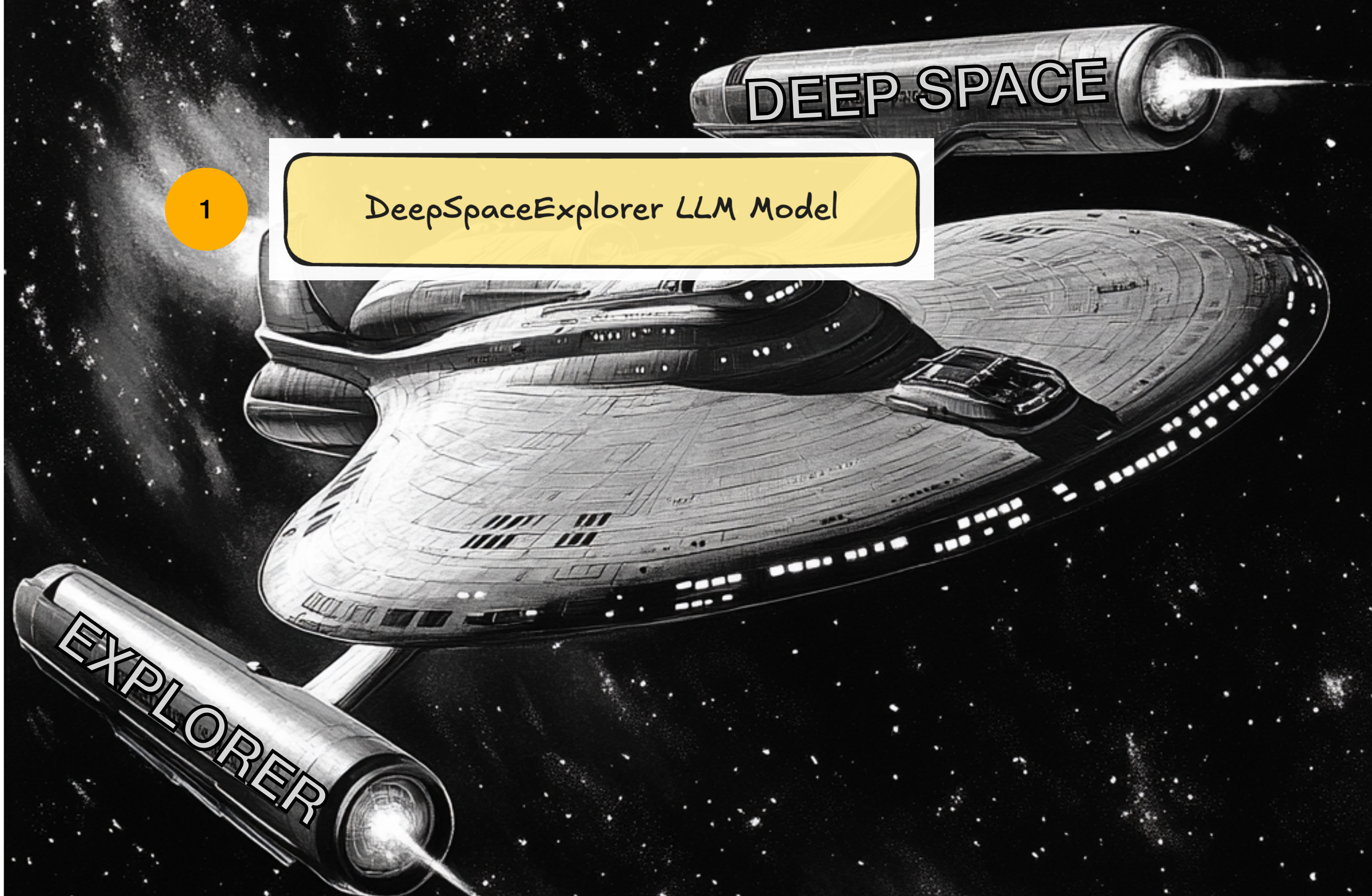
Department of Temporal Affairs

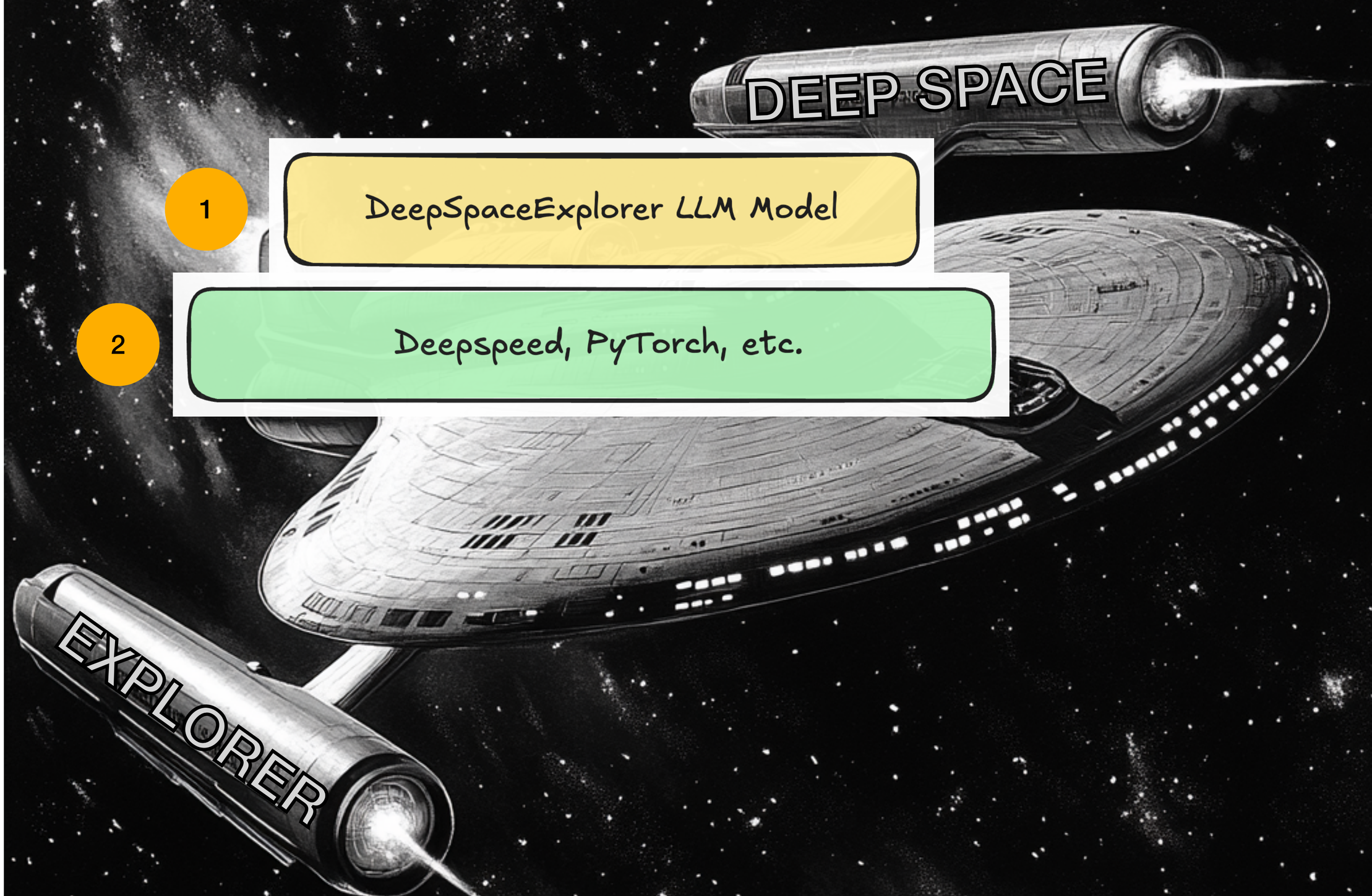
Tempus Mundi Servamus

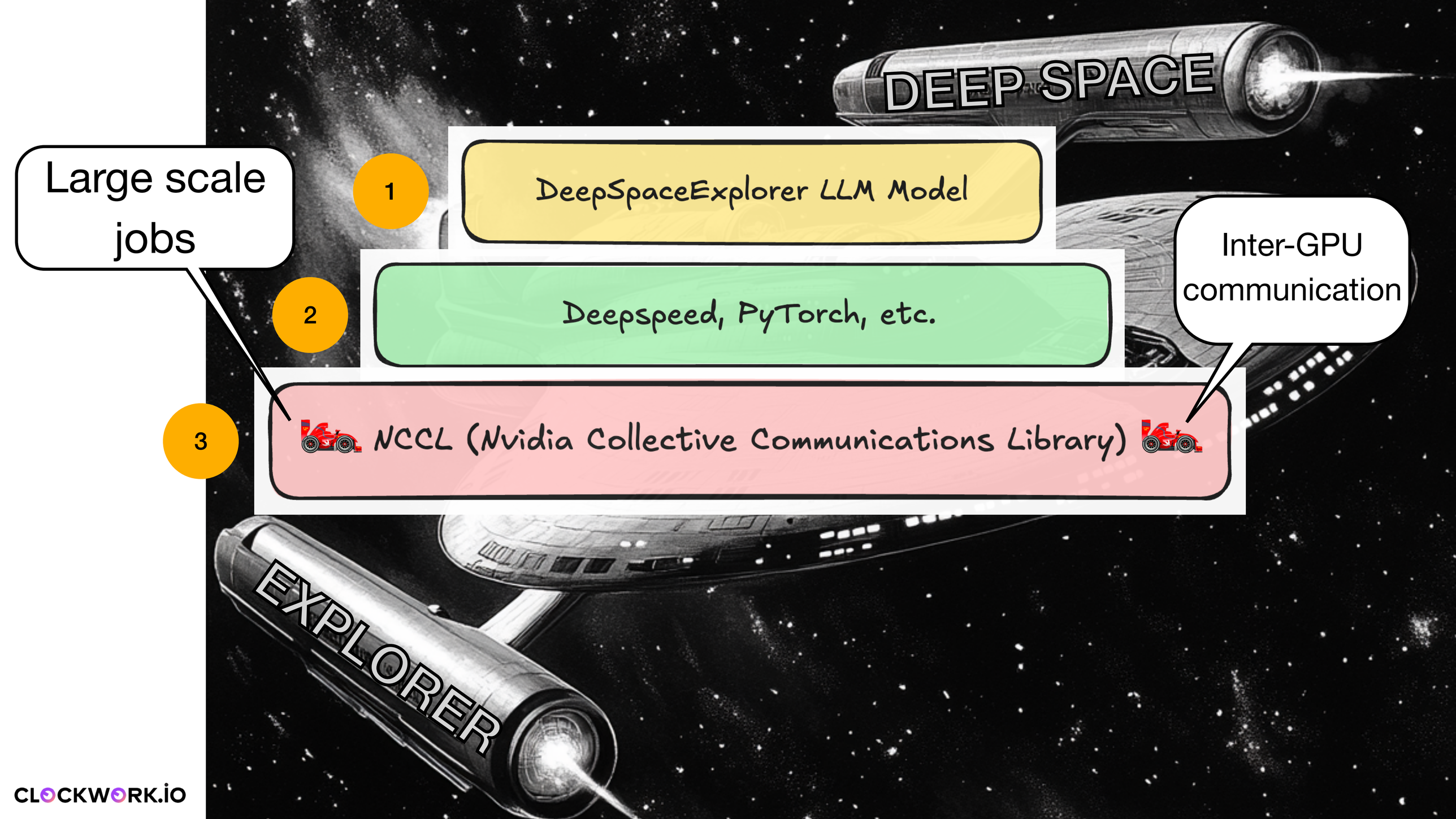
CLOCKWORK.IO

AI-generated


DEEP SPACE
EXPLORER


DEEP SPACE
1
DeepSpaceExplorer LLM Model
EXPLORER


DEEP SPACE
1
DeepSpaceExplorer LLM Model
2
Deepspeed, PyTorch, etc.
EXPLORER


DEEP SPACE
Large scale jobs
1
DeepSpaceExplorer LLM Model
2
Deepspeed, PyTorch, etc.
3
NCCL (Nvidia Collective Communications Library)
Inter-GPU communication
EXPLORER

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

RDMA capable!

Remote Direct Memory Access (RDMA)

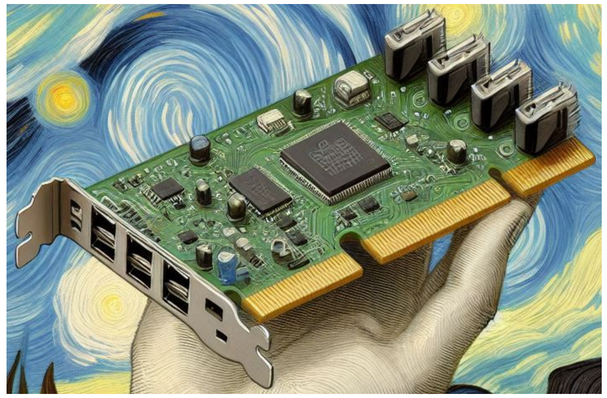
Fast data transfer

Memory

🐢 OS

Memory

🐢 OS



Remote Direct Memory Access (RDMA)

Fast data transfer

Memory

🐢 OS

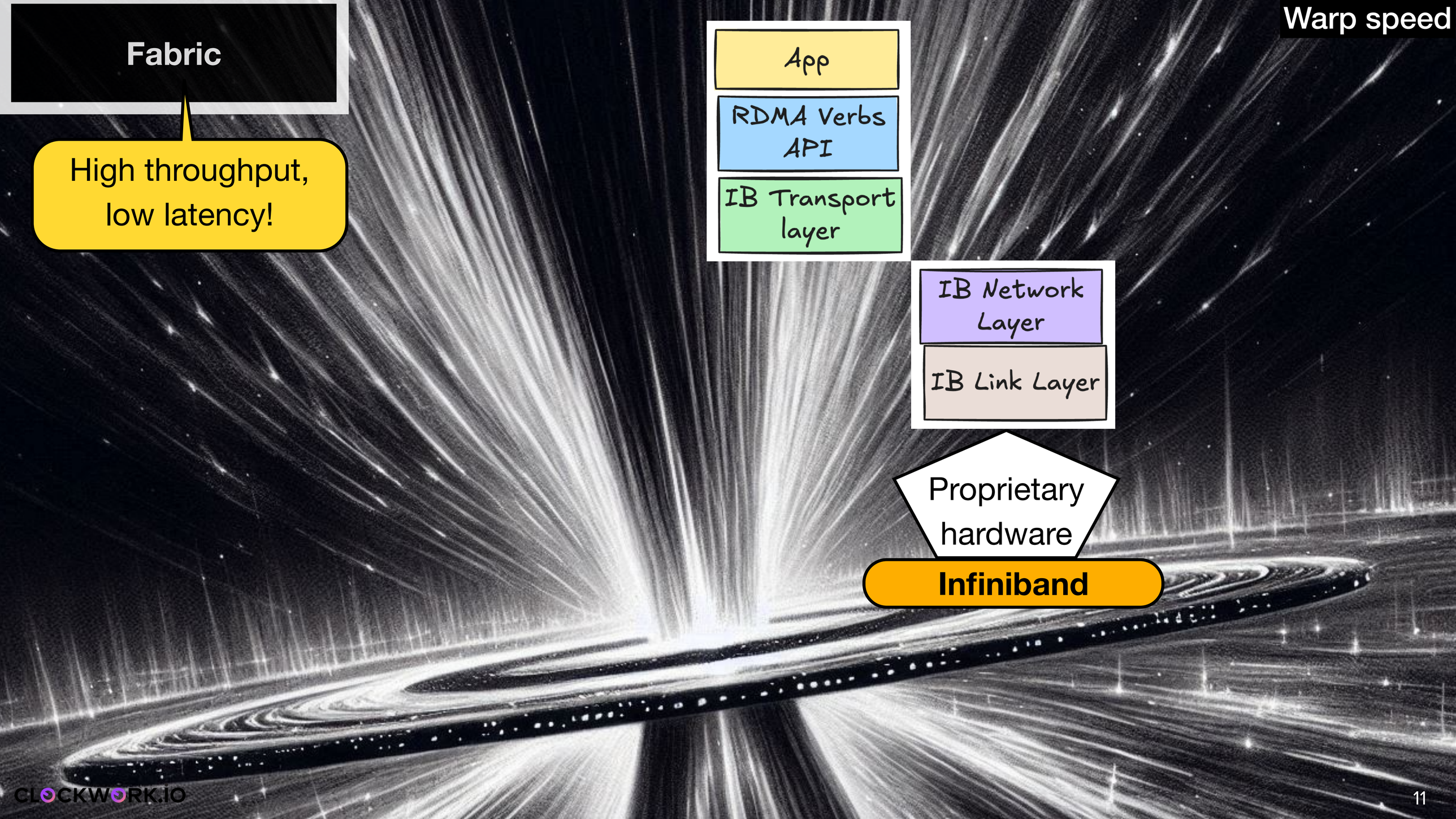
Kernel Bypass

Memory

🐢 OS

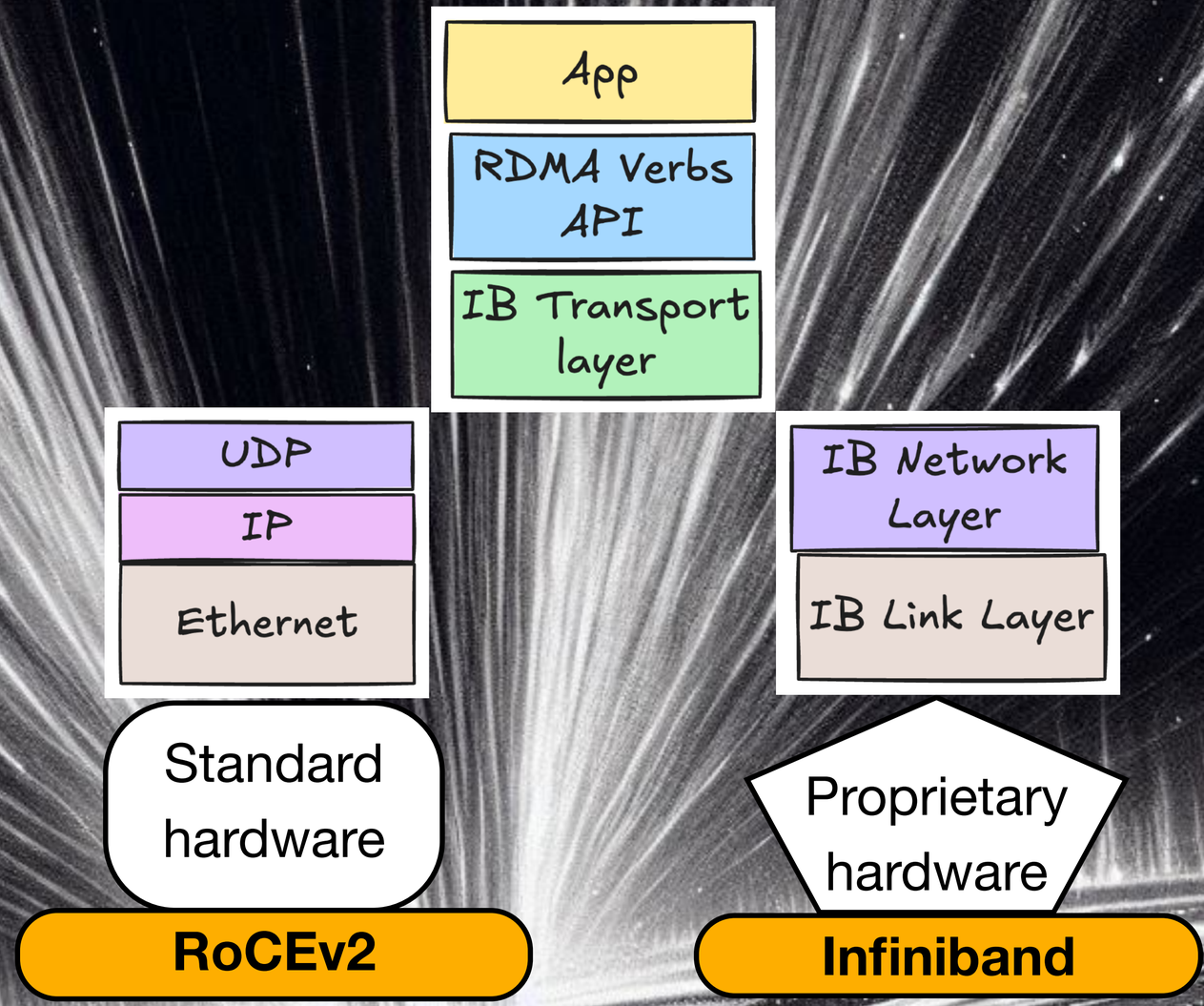
Fabric


Fabric
High throughput, low latency!
App
RDMA Verbs API
IB Transport layer


Fabric
High throughput, low latency!
App
RDMA Verbs API
IB Transport layer
IB Network Layer
IB Link Layer
Proprietary hardware
Infiniband

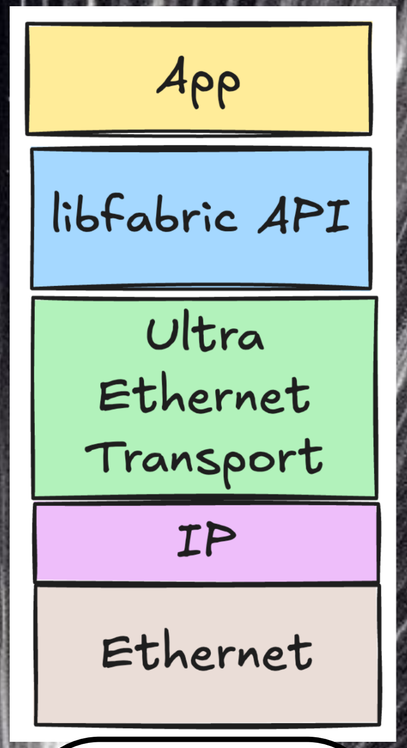
Fabric

High throughput, low latency!



Fabric

High throughput, low latency!

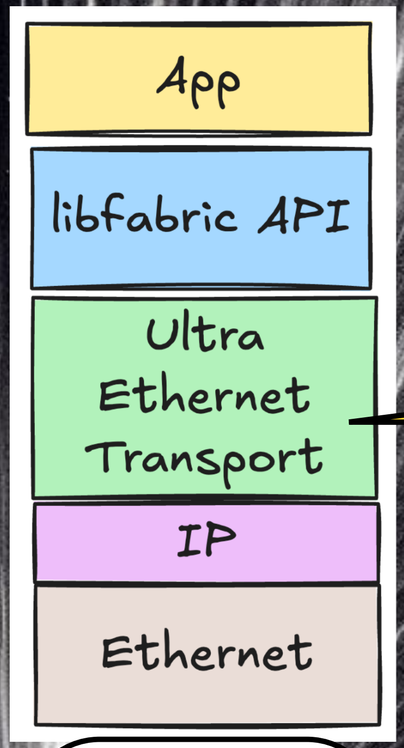


Standard hardware

UltraEthernet

Fabric

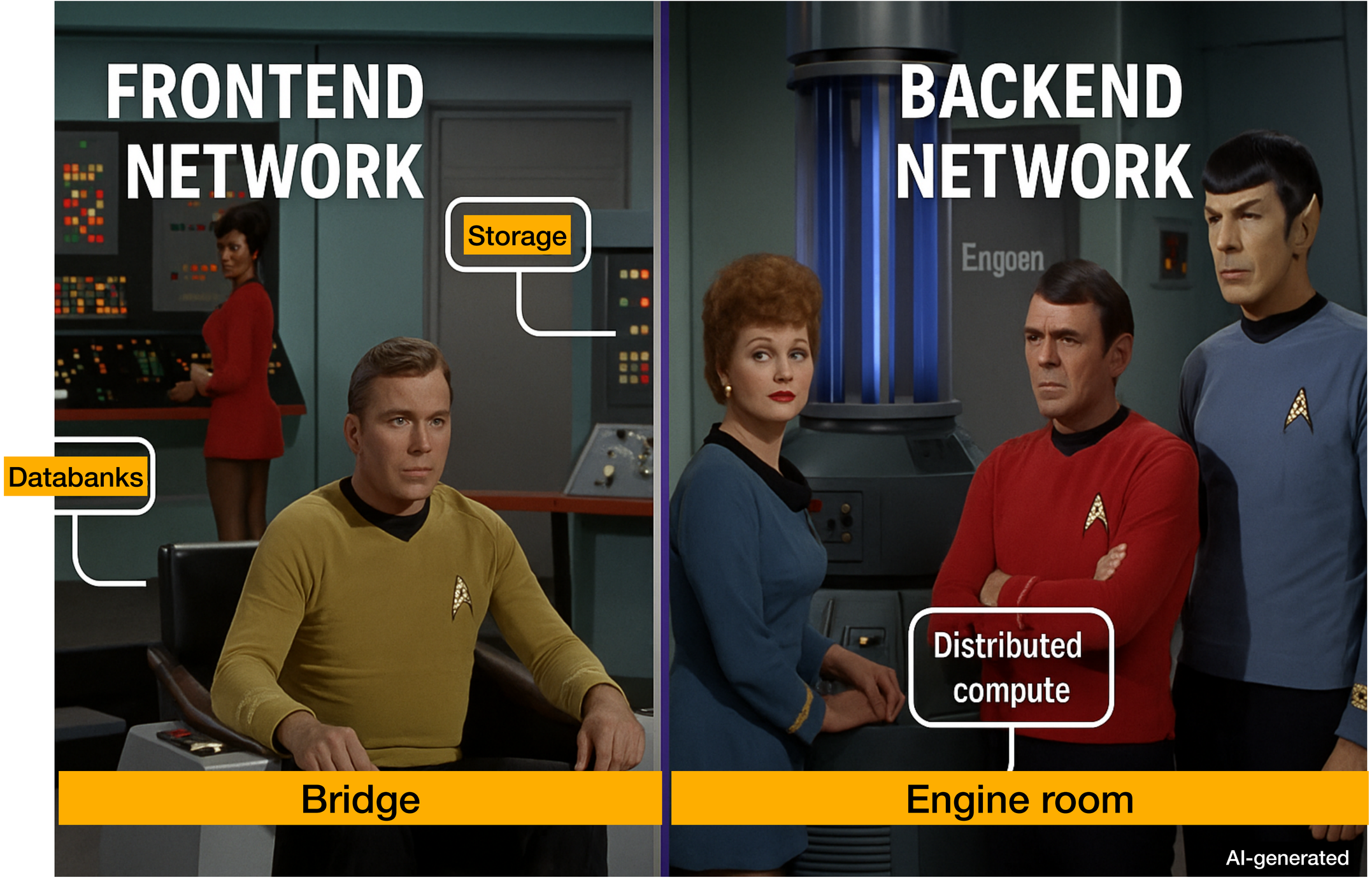
High throughput, low latency!

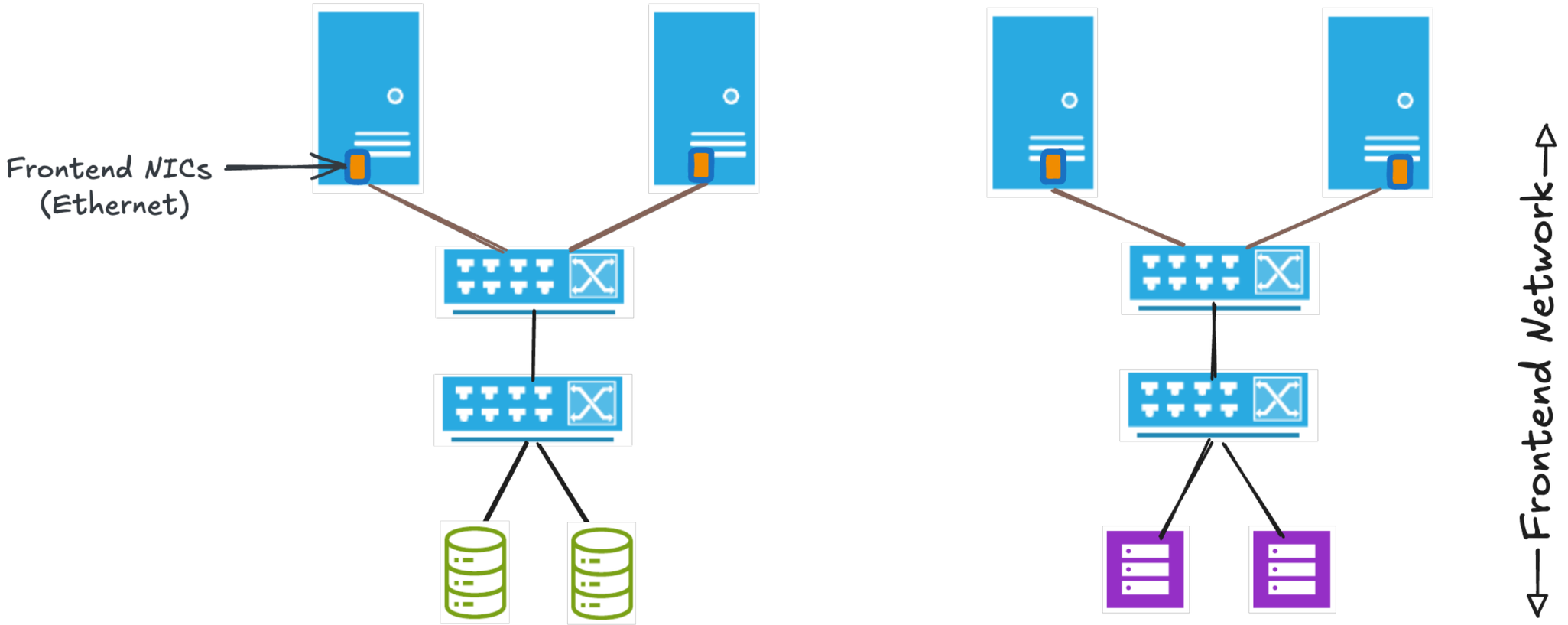


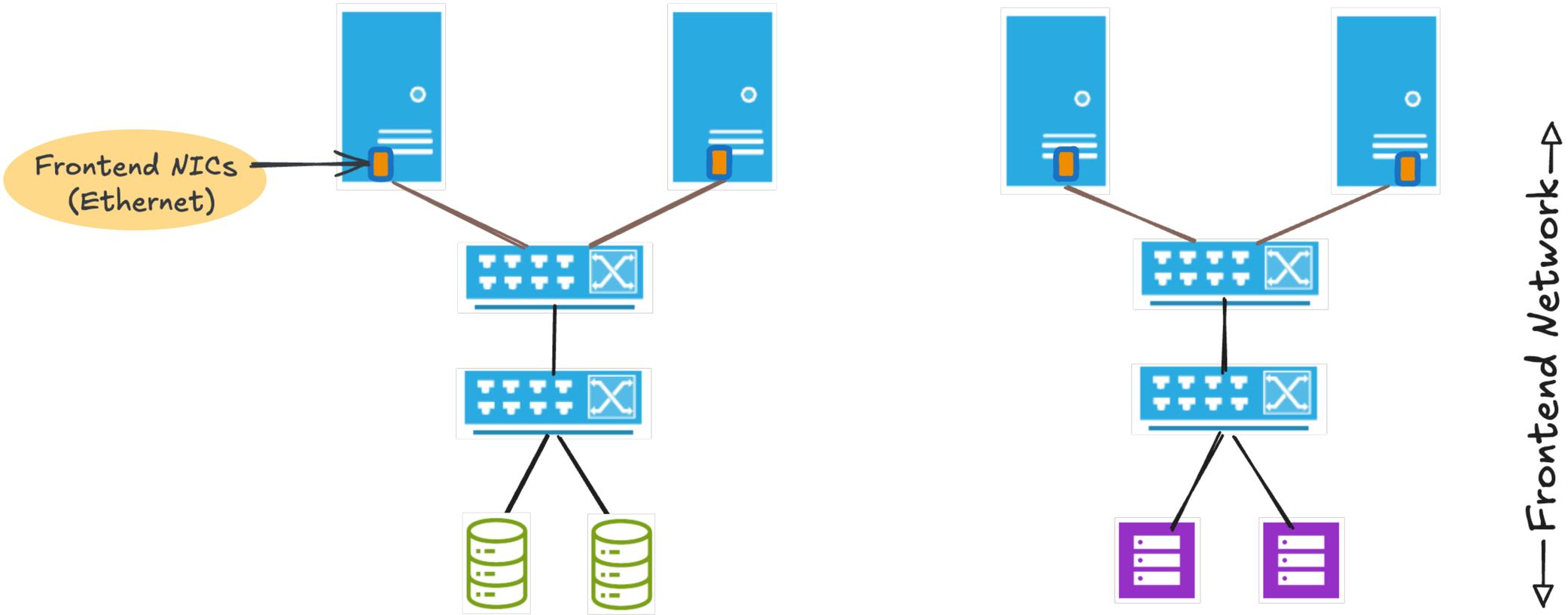
Improved transport!
ultraethernet.org

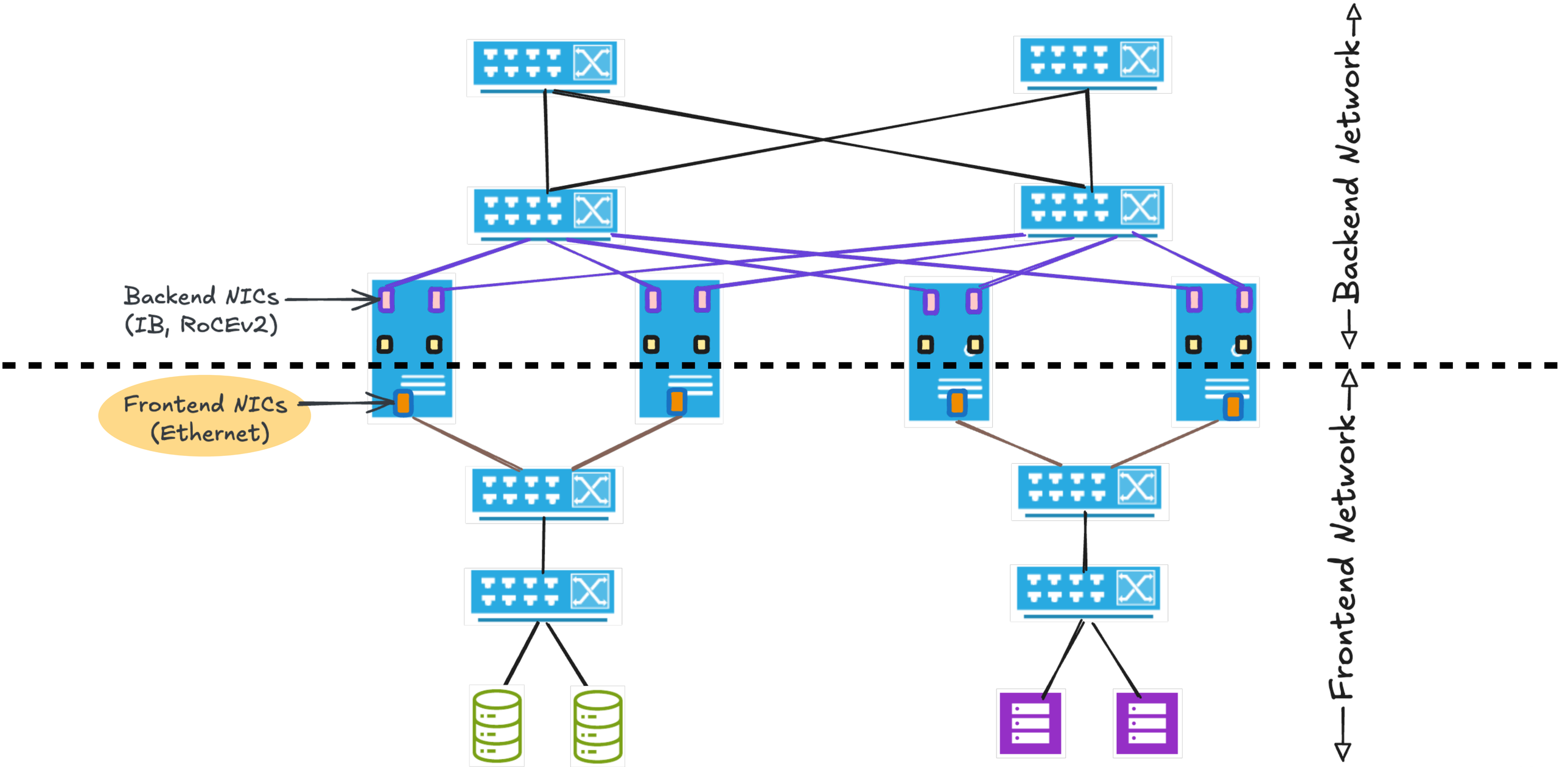
Standard hardware

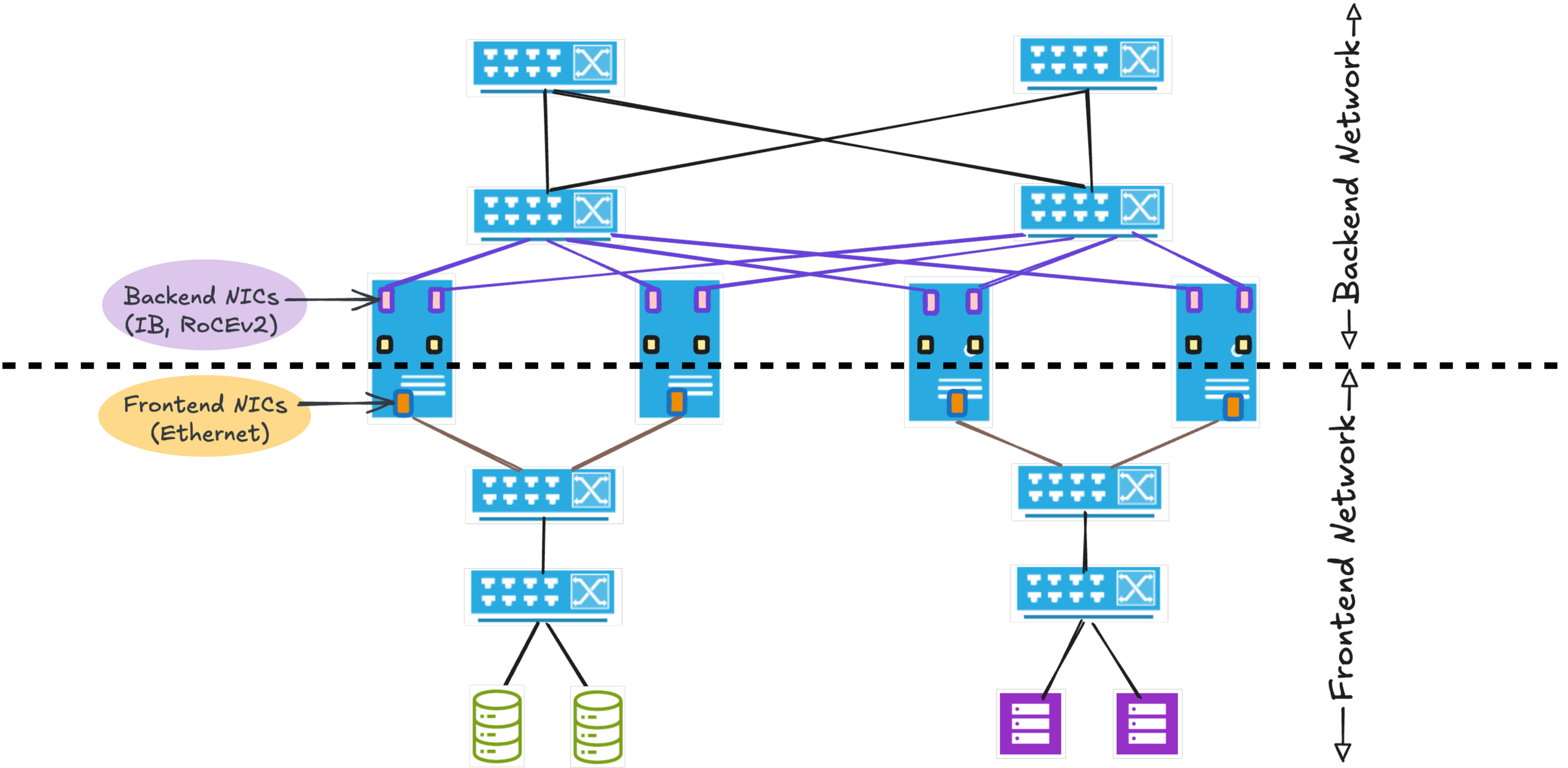
UltraEthernet


FRONTEND NETWORK
Storage
Databanks
Bridge
BACKEND NETWORK
Engoen
Distributed compute
Engine room
AI-generated

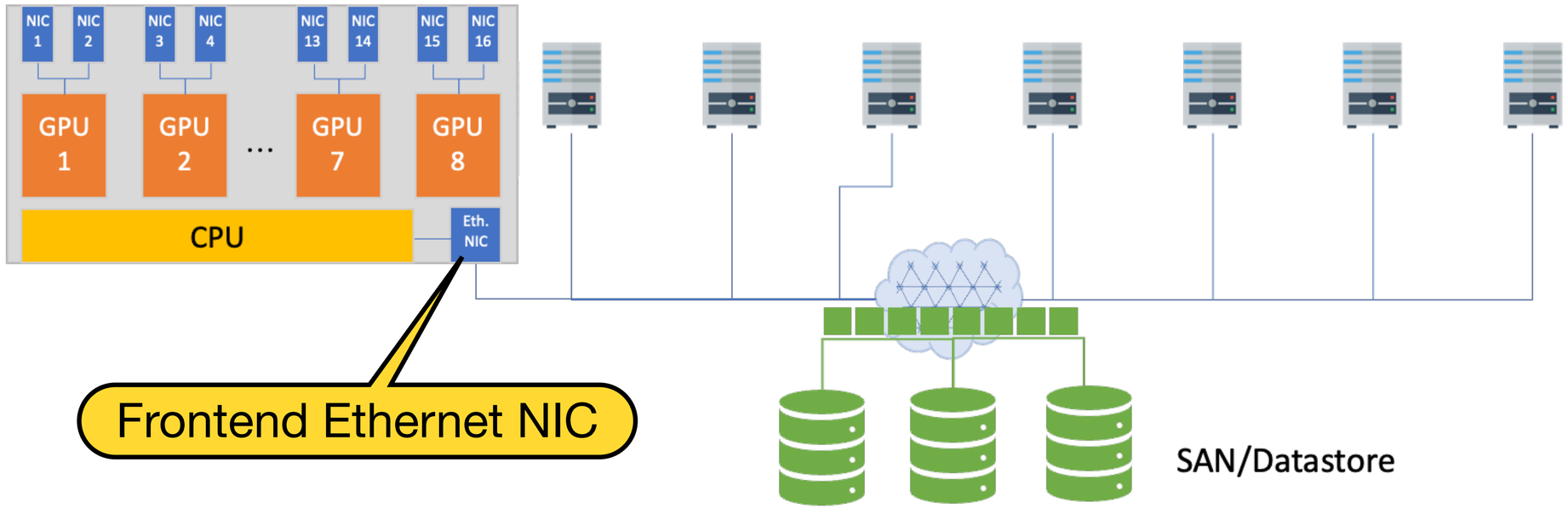

Frontend NICs
(Ethernet)
Frontend Network


Frontend NICs
(Ethernet)
Frontend Network

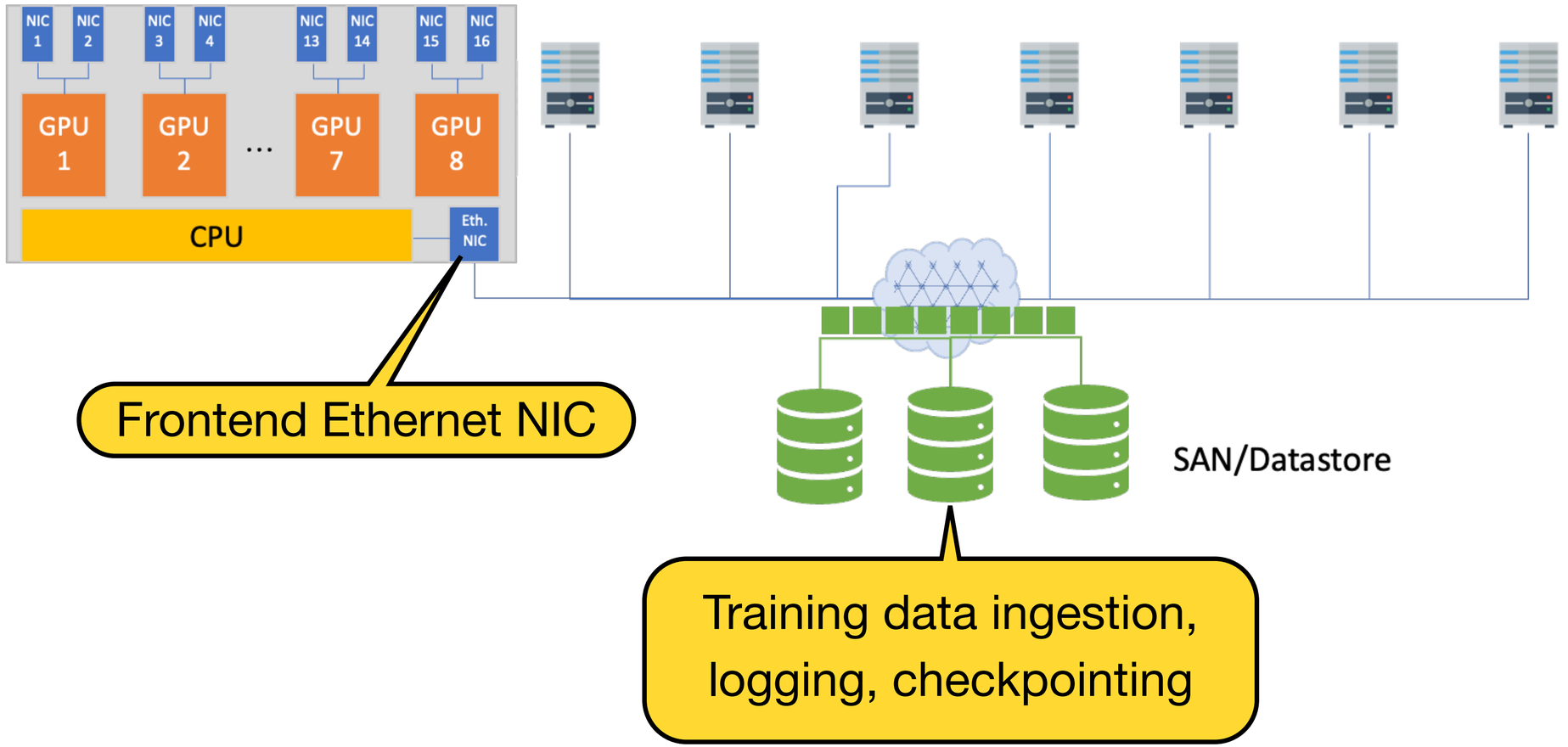

Backend NICs
(IB, RoCEv2)
Frontend NICs
(Ethernet)
Backend Network
Frontend Network


Backend NICs
(IB, RoCEv2)
Frontend NICs
(Ethernet)
Backend Network
Frontend Network

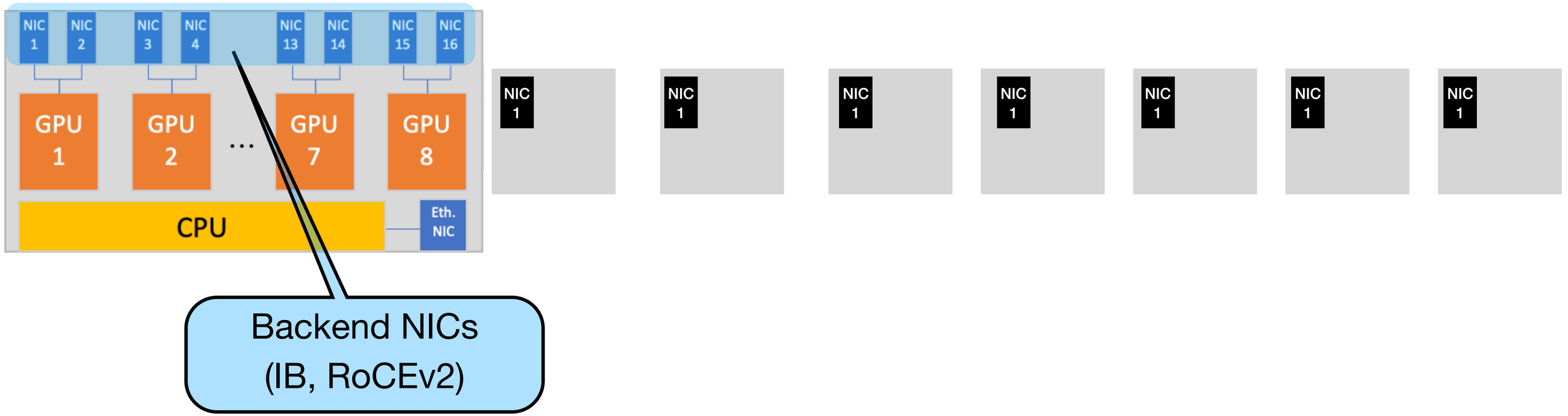
Frontend network



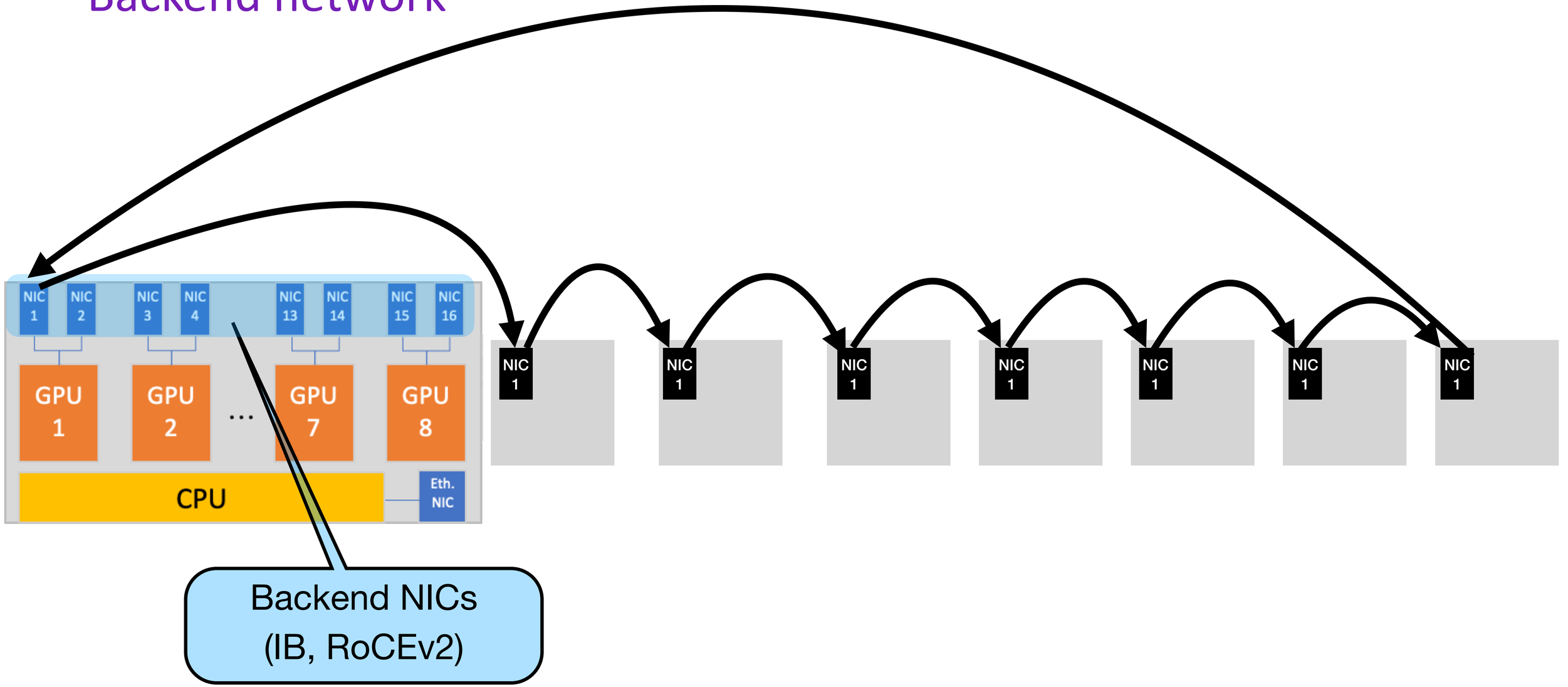
Frontend network



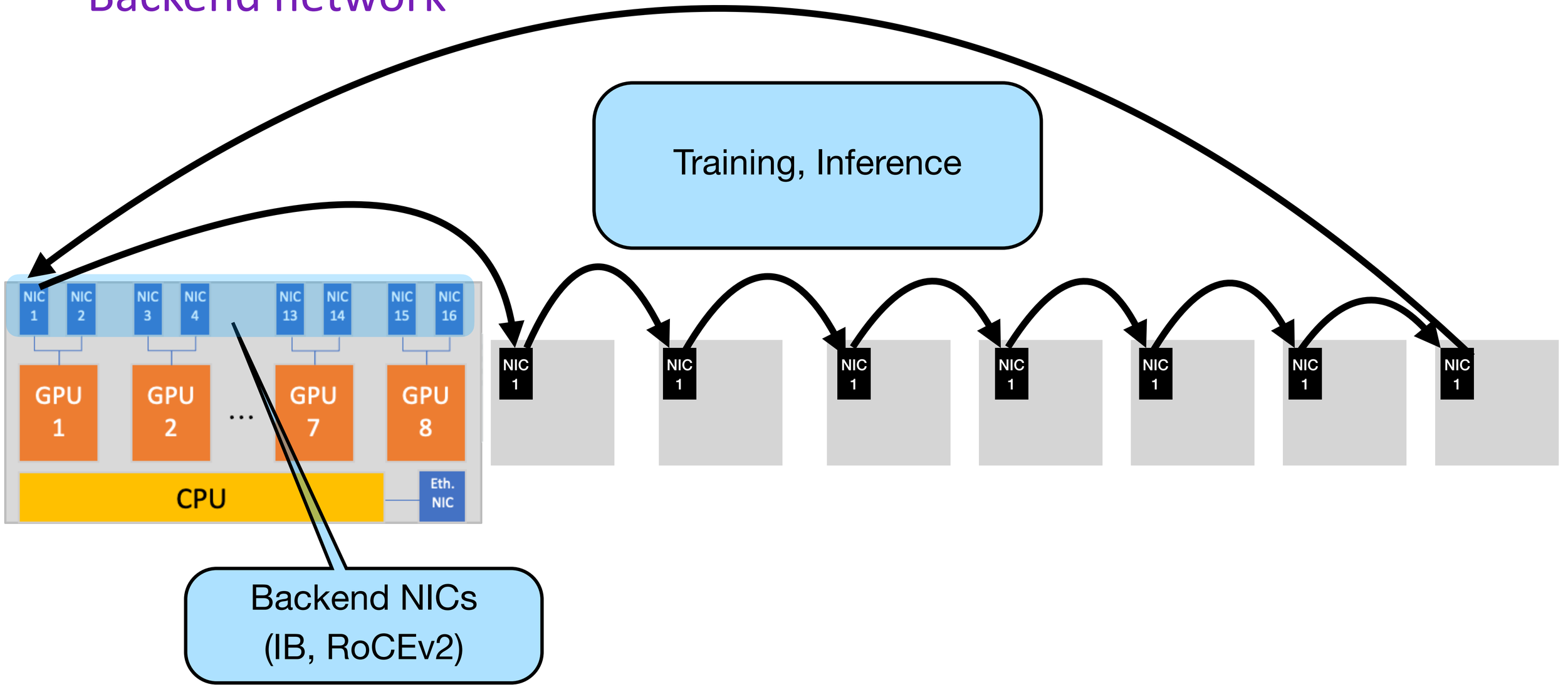
Backend network



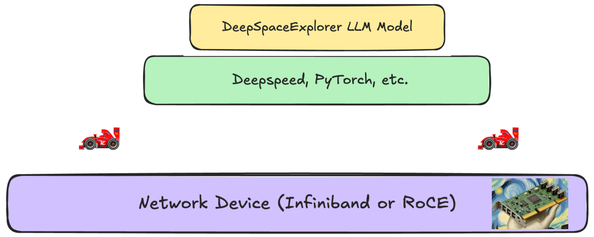
Backend network

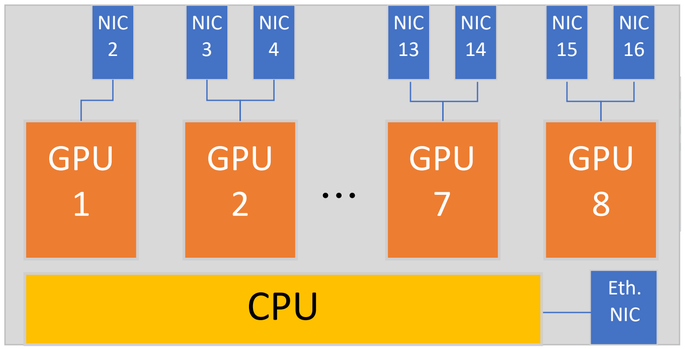


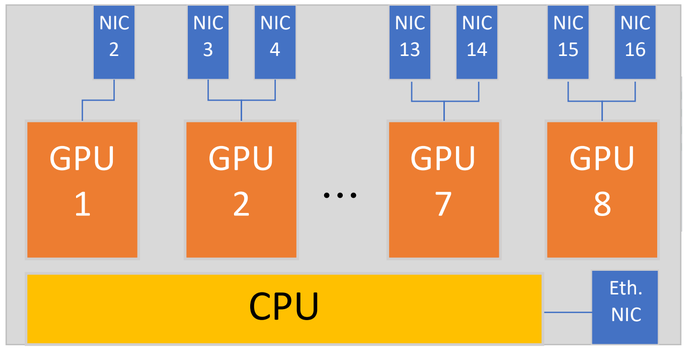
Backend network



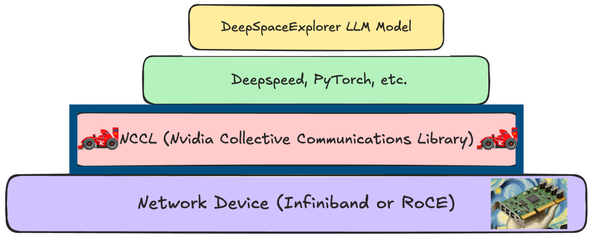
Collective communications

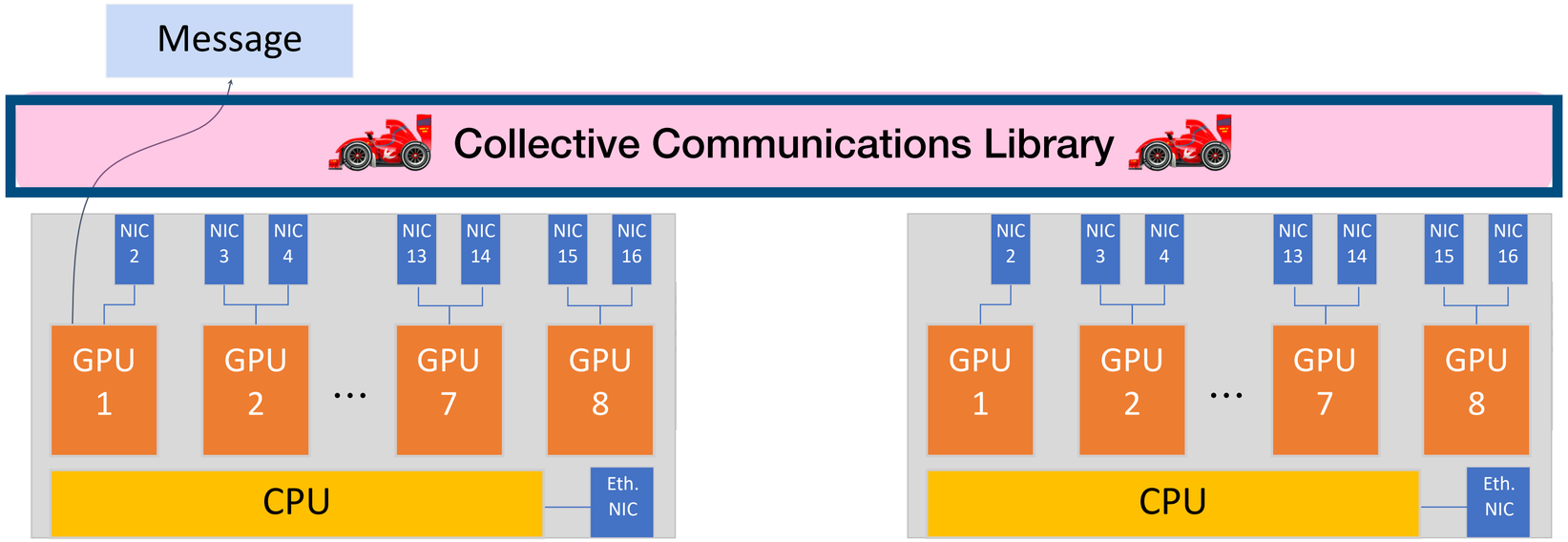




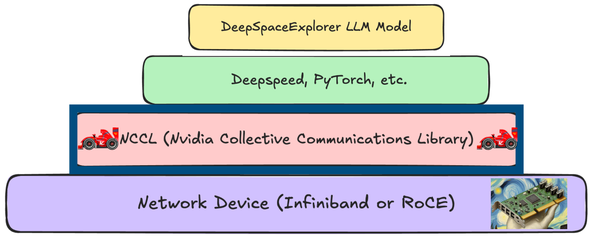


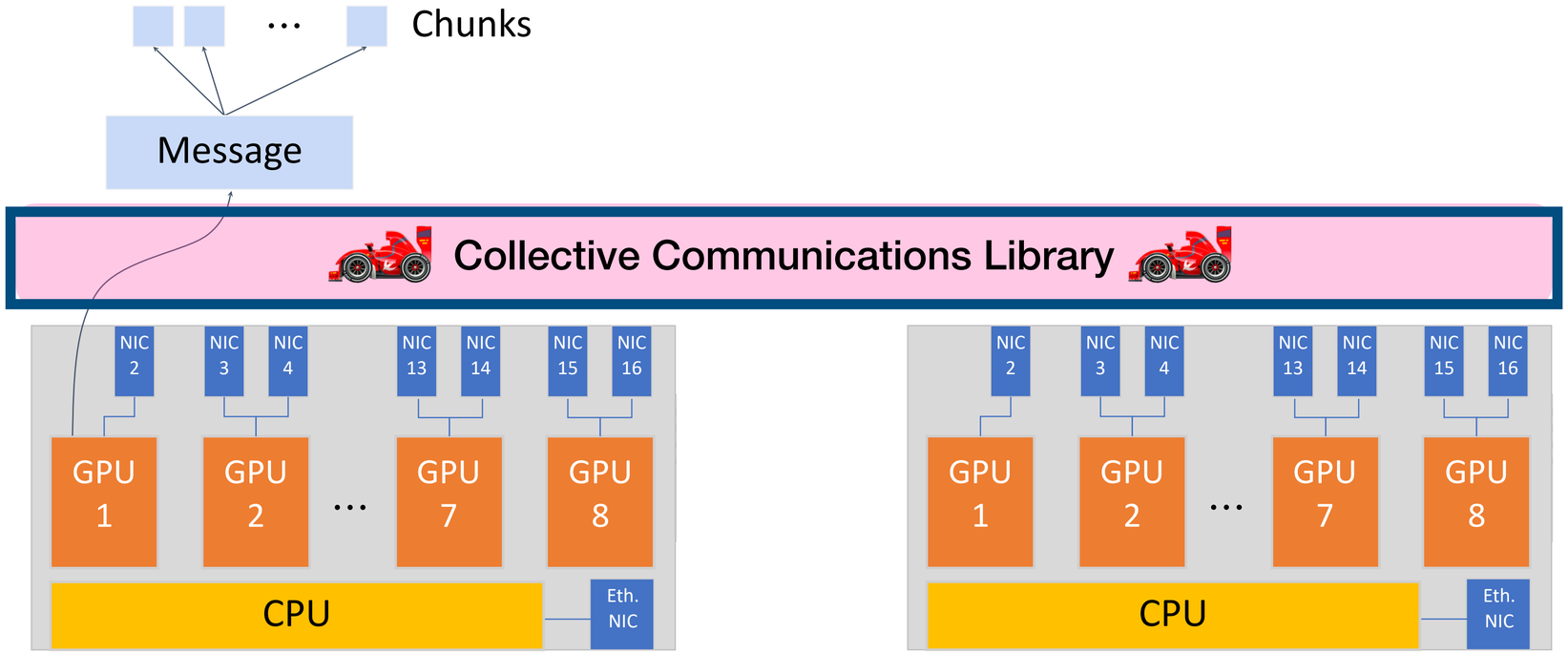
Collective communications



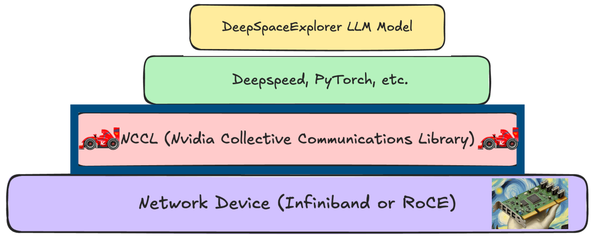


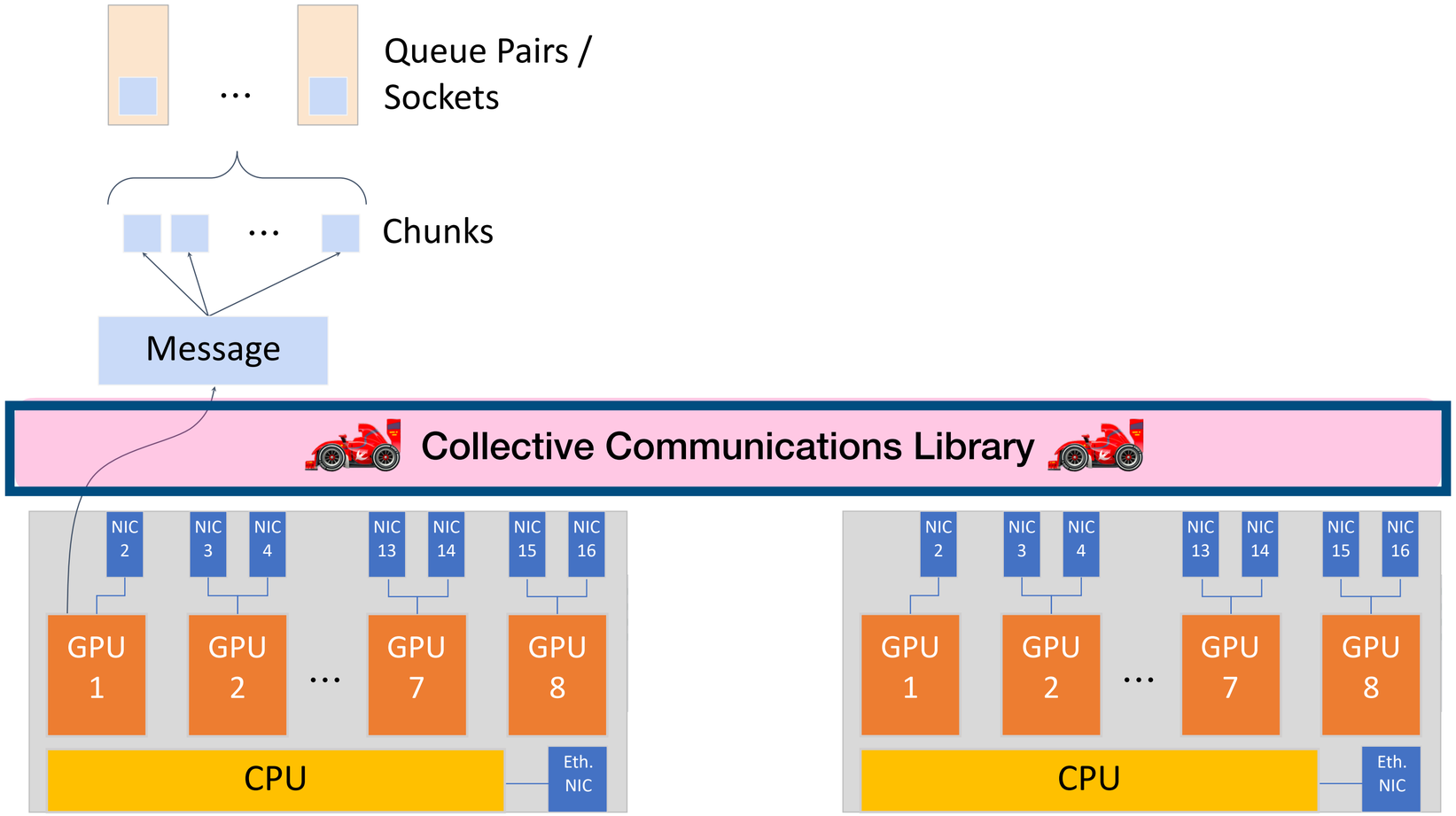
Collective communications



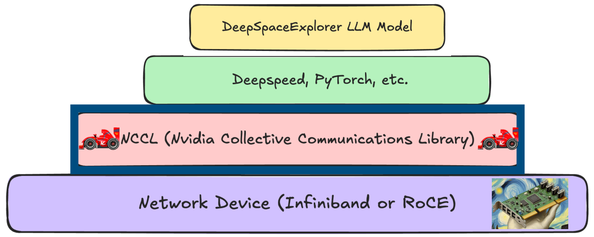


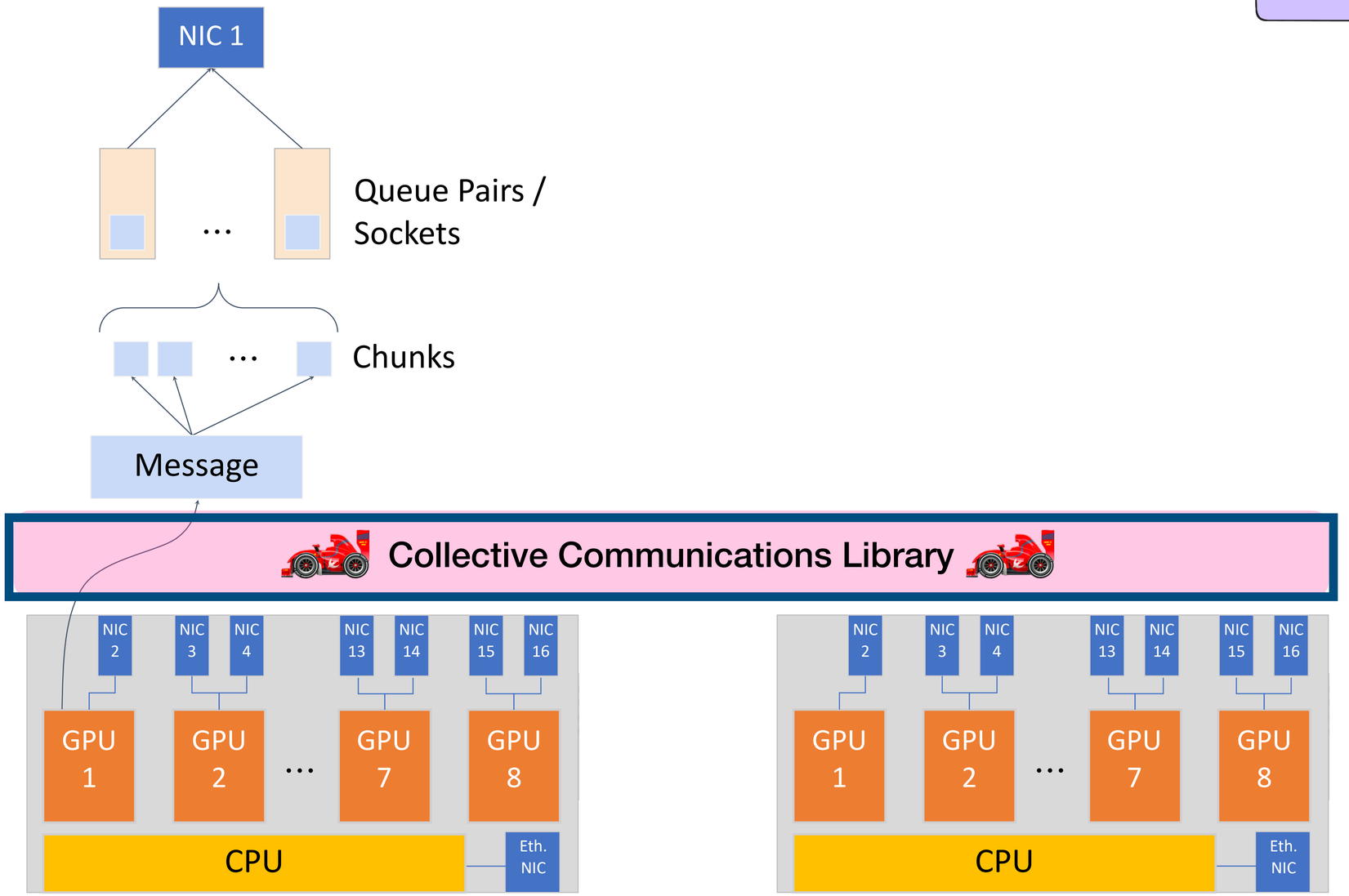
Collective communications



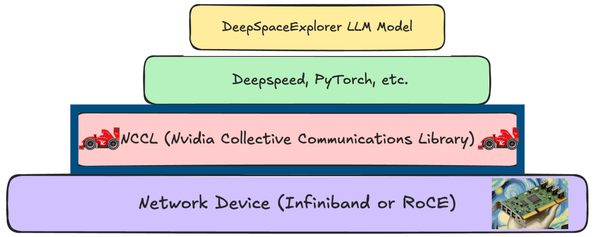


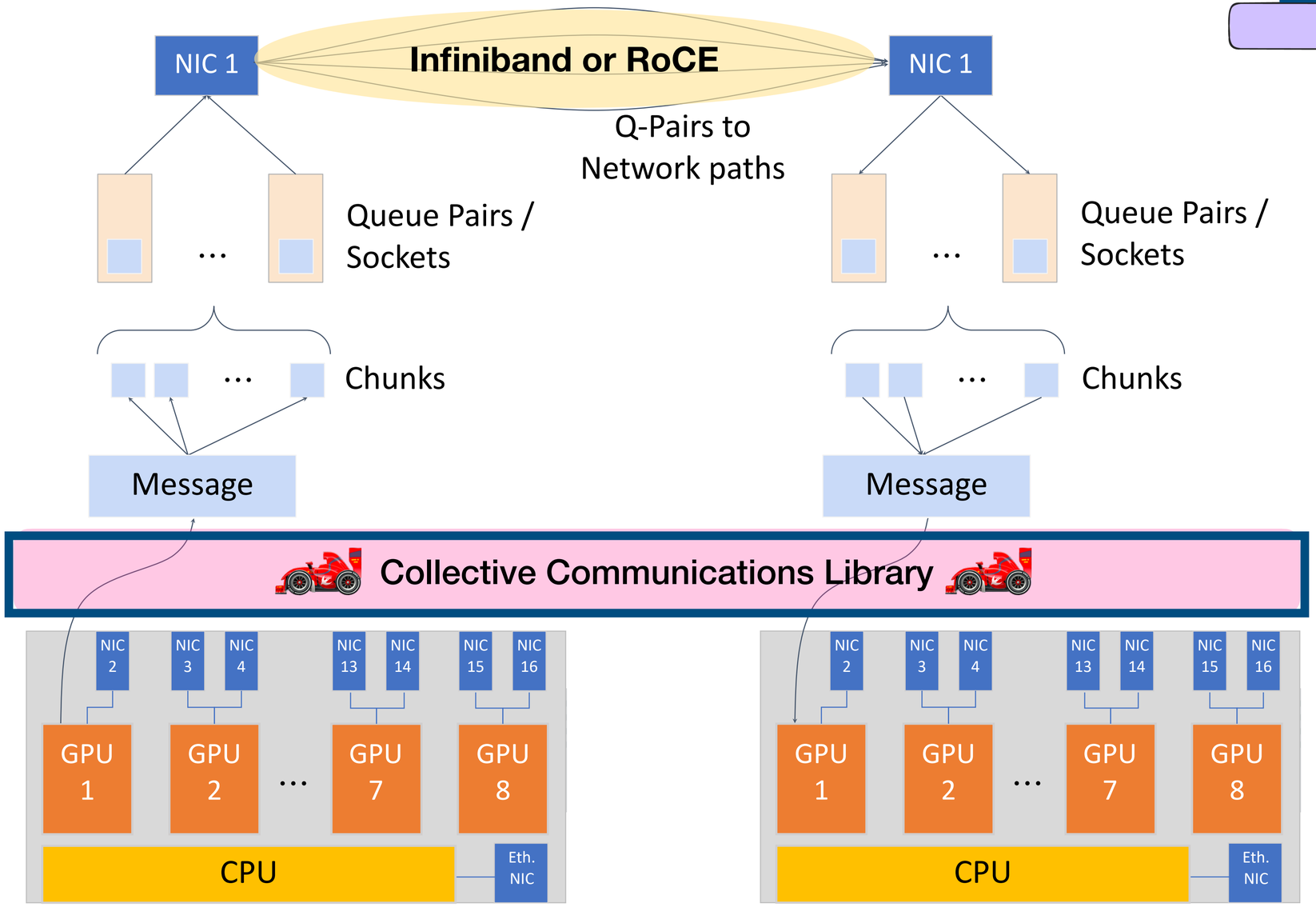
Collective communications





Collective communications





AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus

CLOCKWORK.IO

AI-generated

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus

AI-generated

CLOCKWORK.IO

Traditional networks

AI networks

Traditional networks

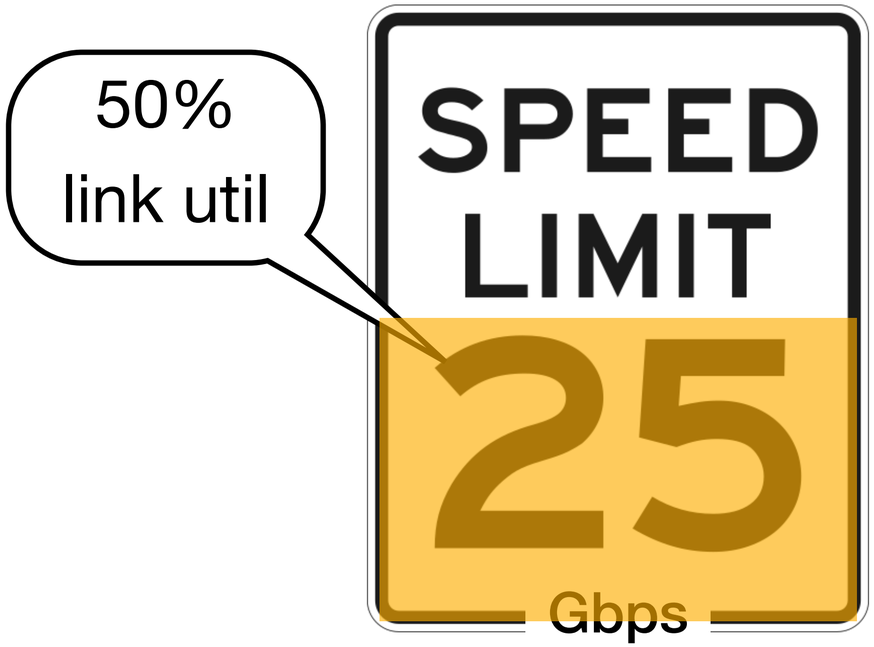
AI networks

Traditional networks

AI networks

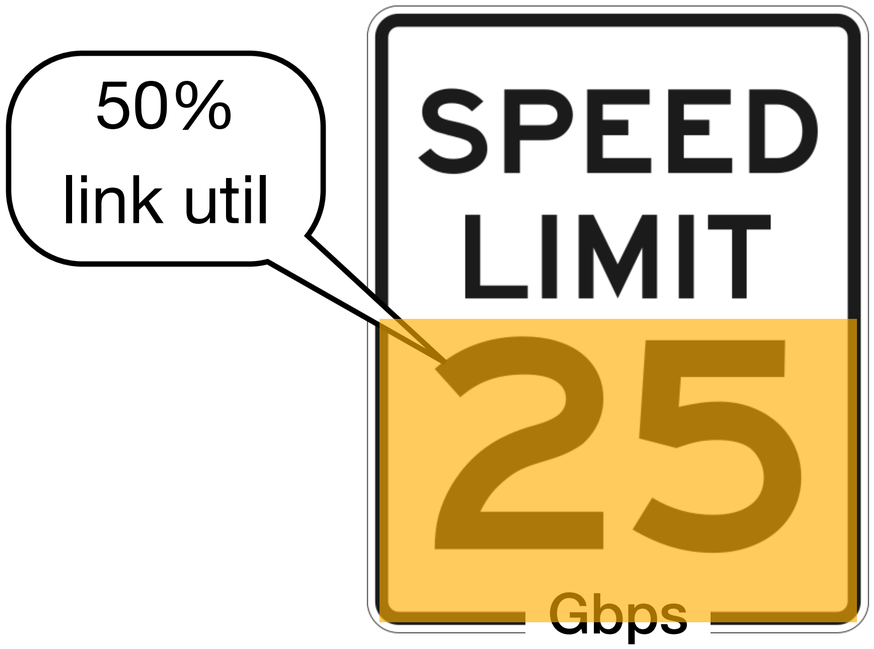
Traditional networks

AI networks



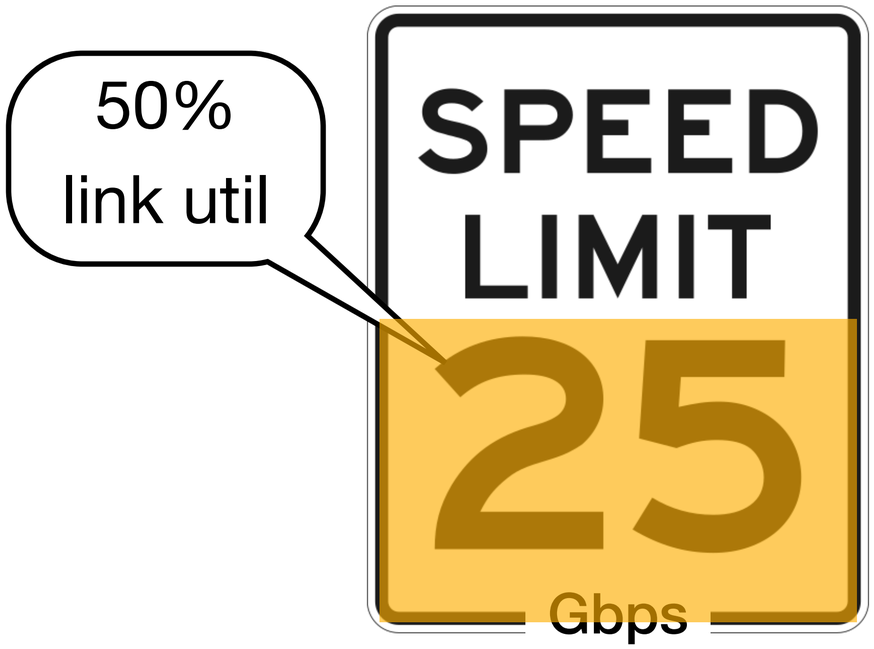
Traditional networks

AI networks



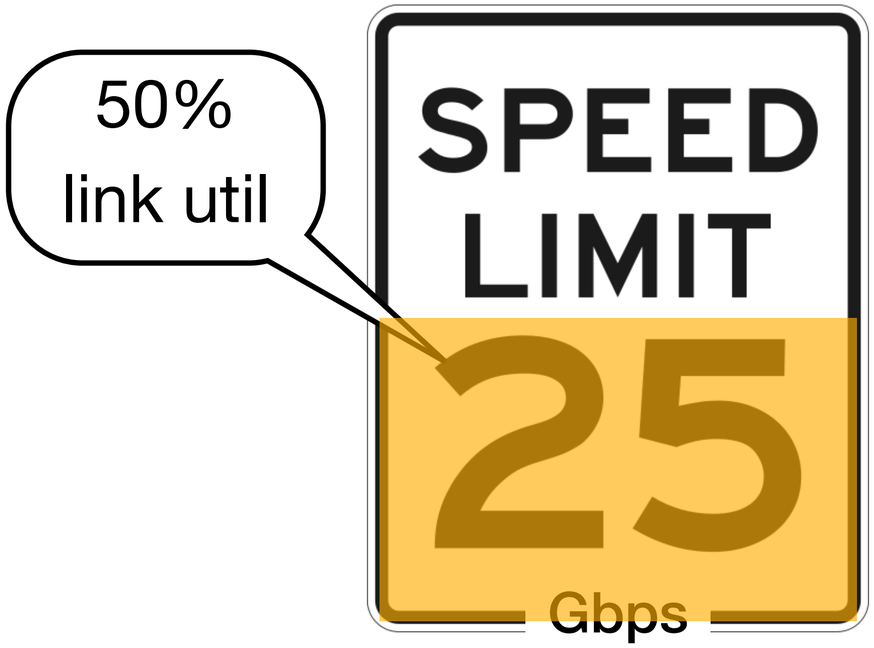
Traditional networks

AI networks

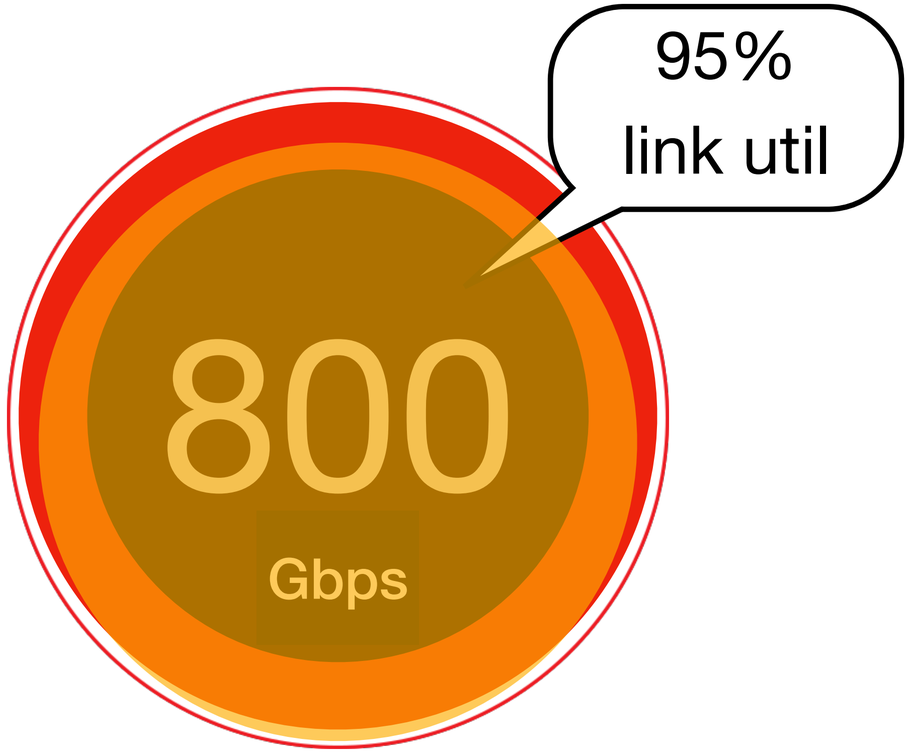


Traditional networks

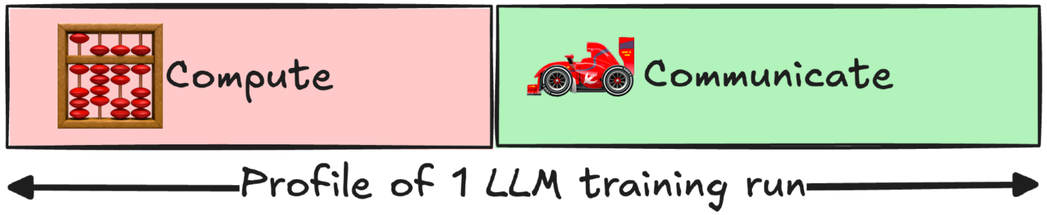
AI networks



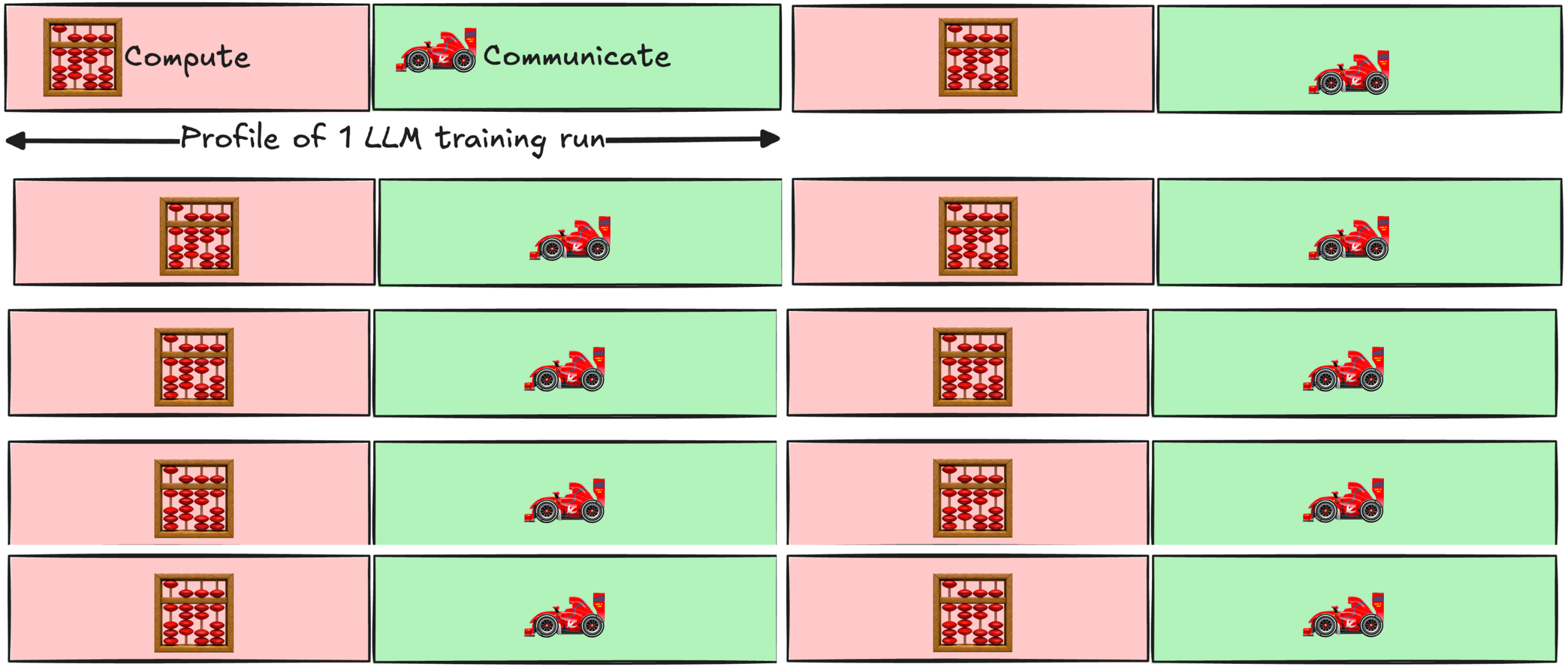
Traditional networks



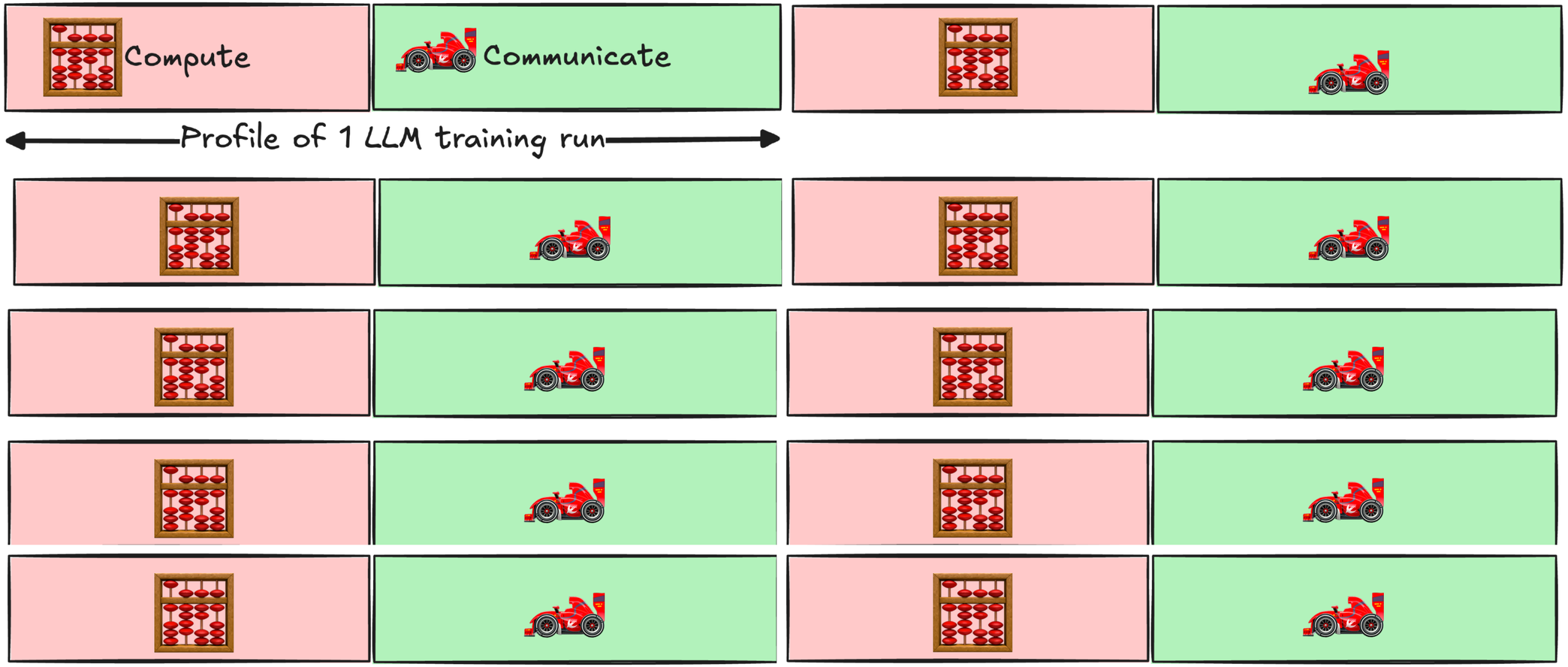
AI networks

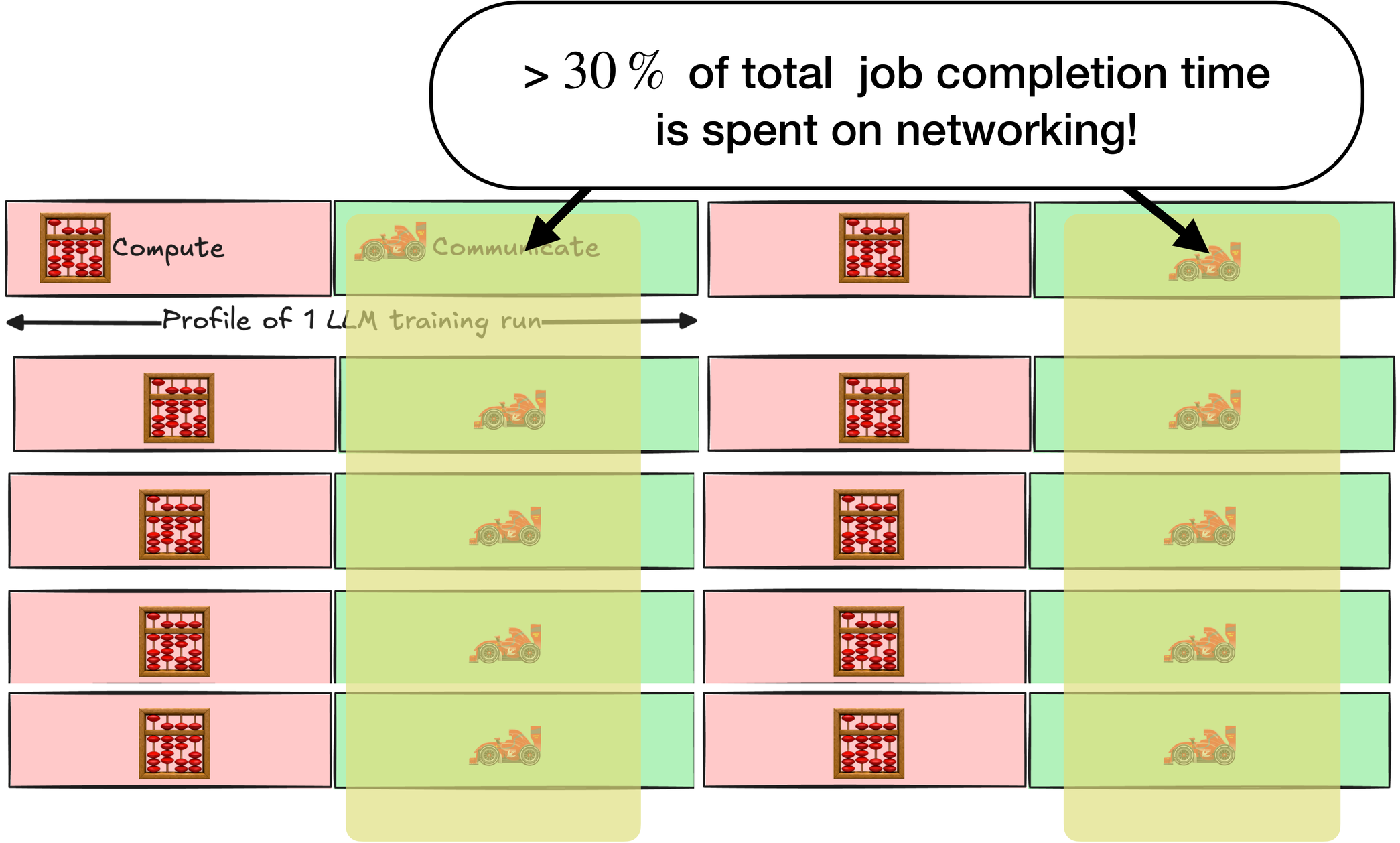


Source: Nvidia



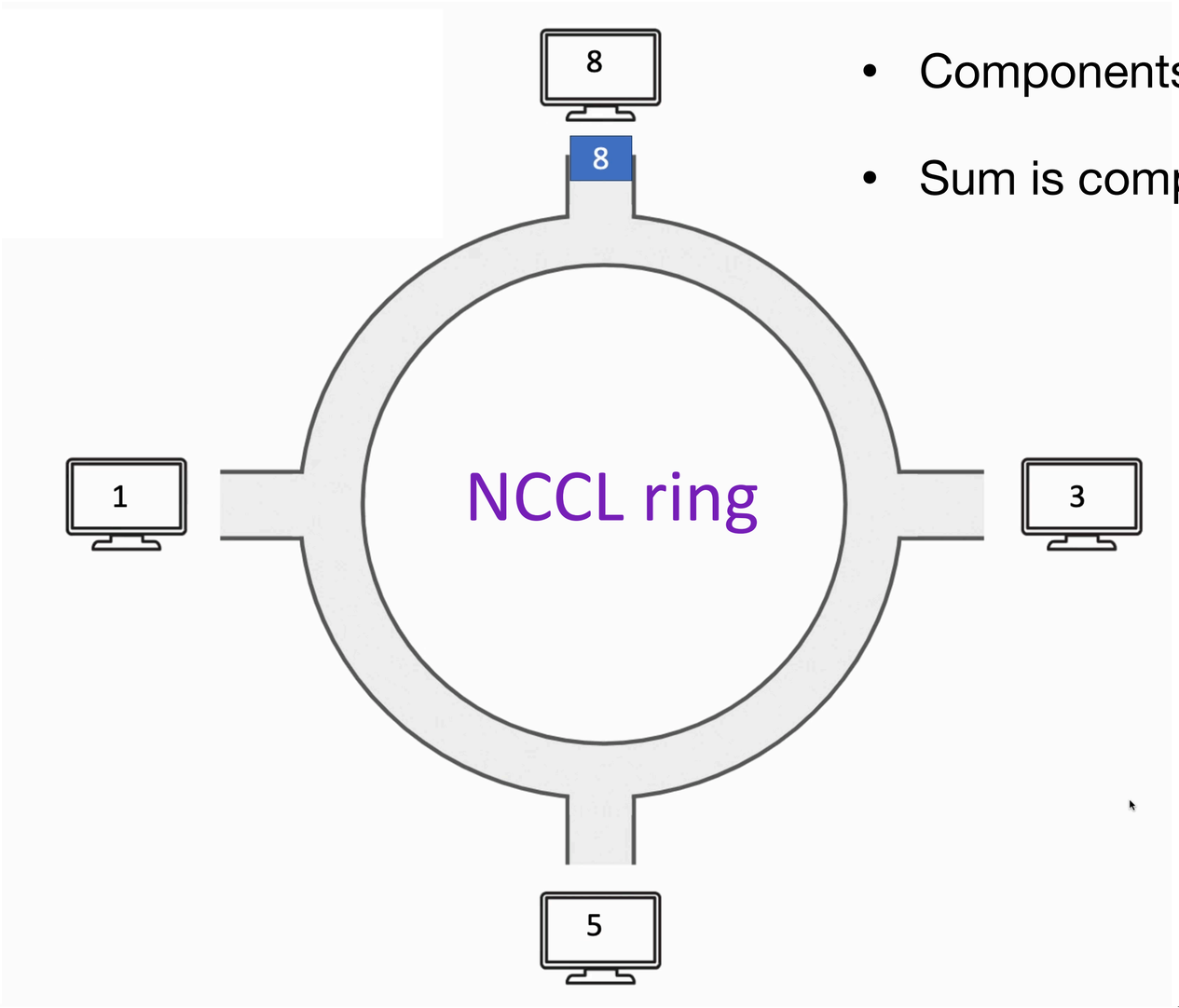
Source: Nvidia


Compute
Communicate
Profile of 1 LLM training run


> 30 % of total job completion time is spent on networking!
Compute
Communicate
Profile of 1 LLM training run

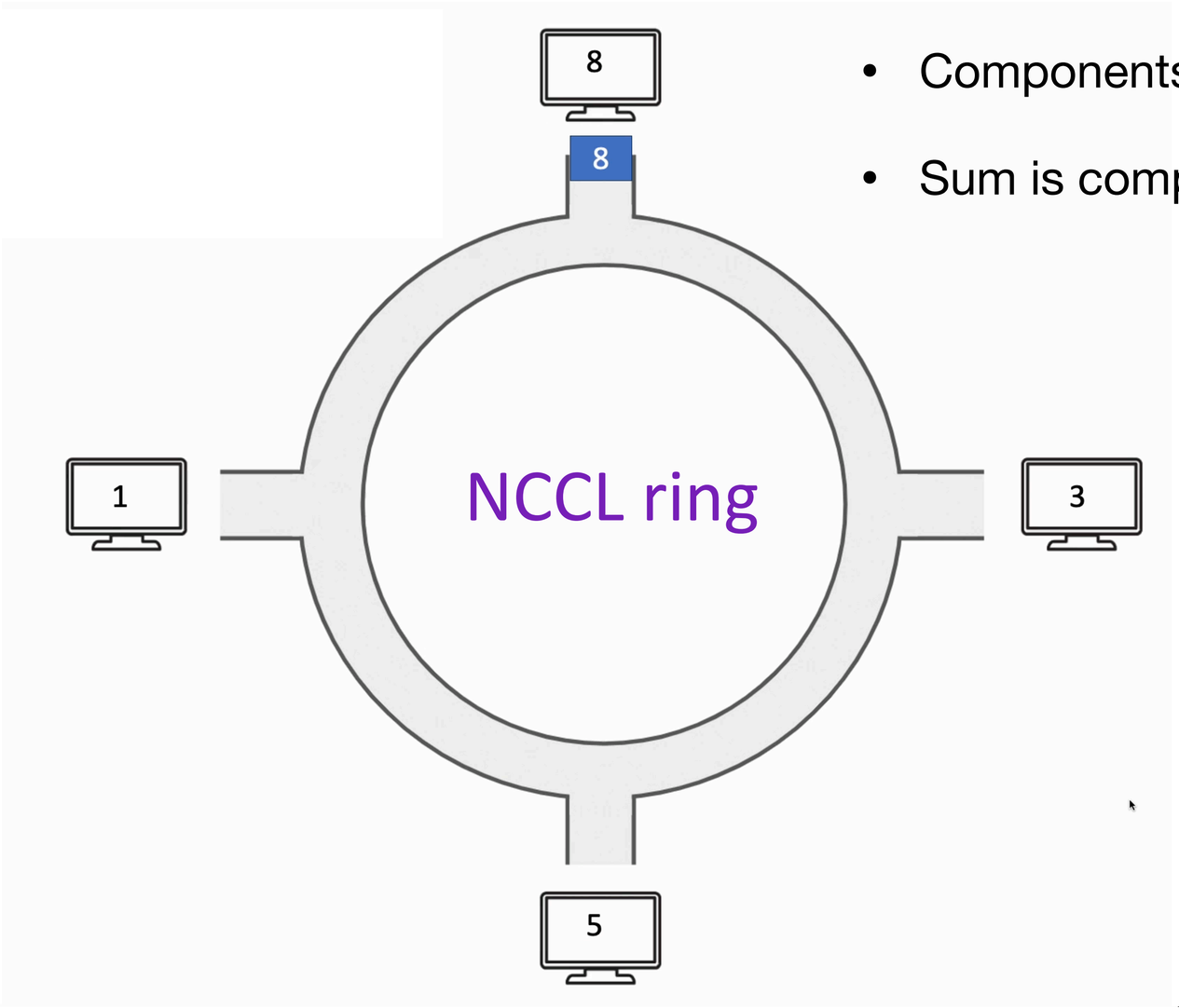
Source: AMD

All-reduce



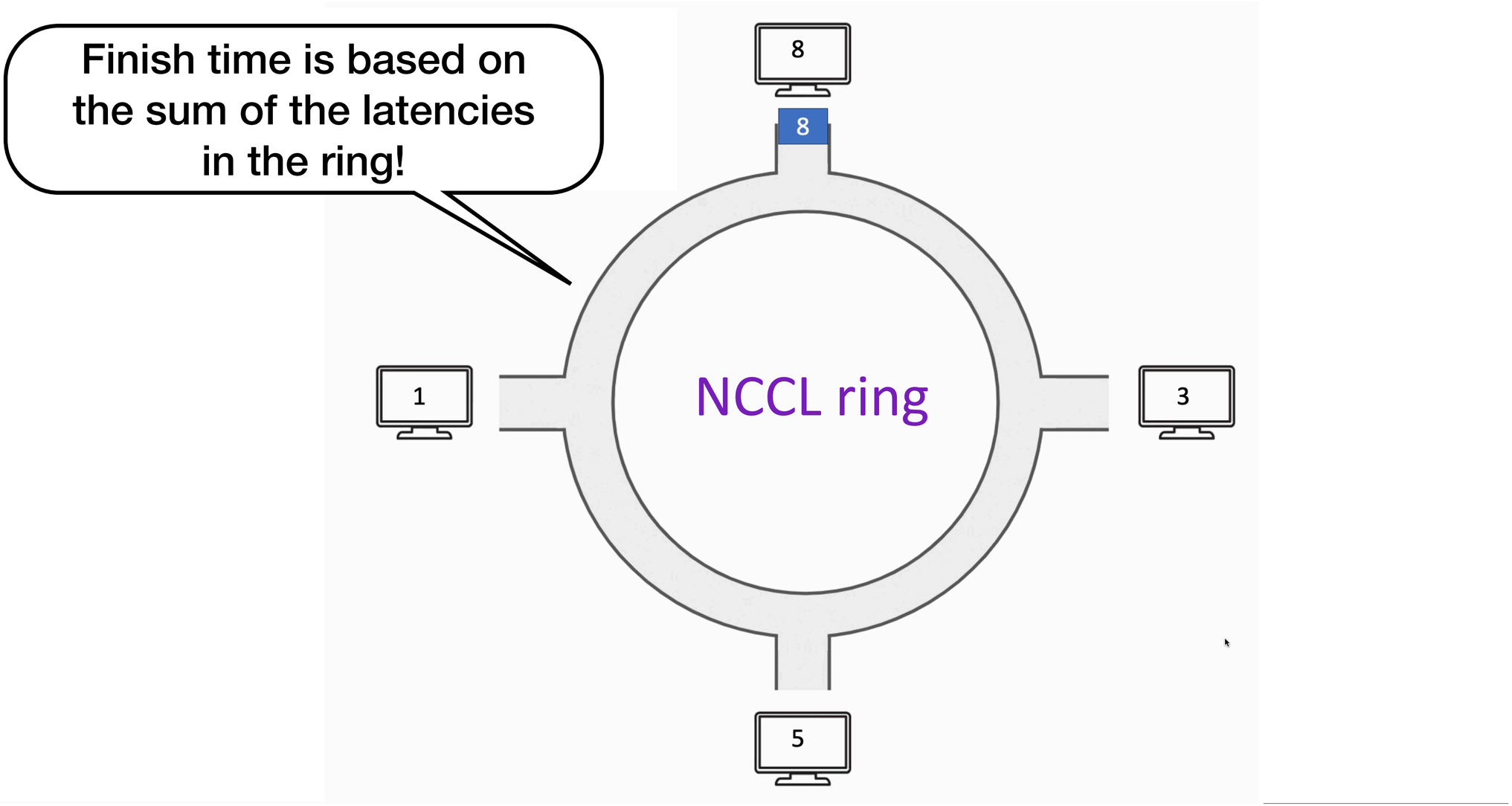
- Components of the sum get passed around
- Sum is computed then passed around

All-reduce

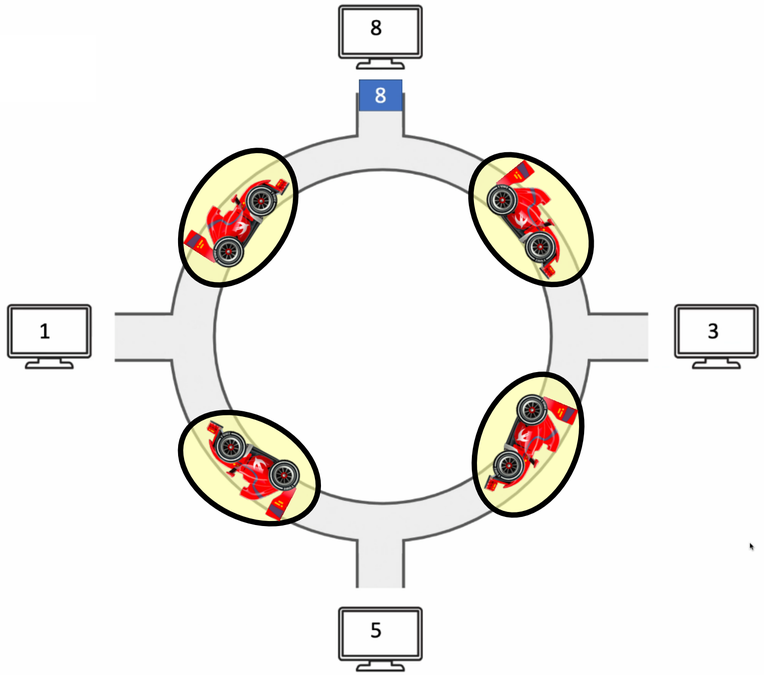


- Components of the sum get passed around
- Sum is computed then passed around

All-reduce

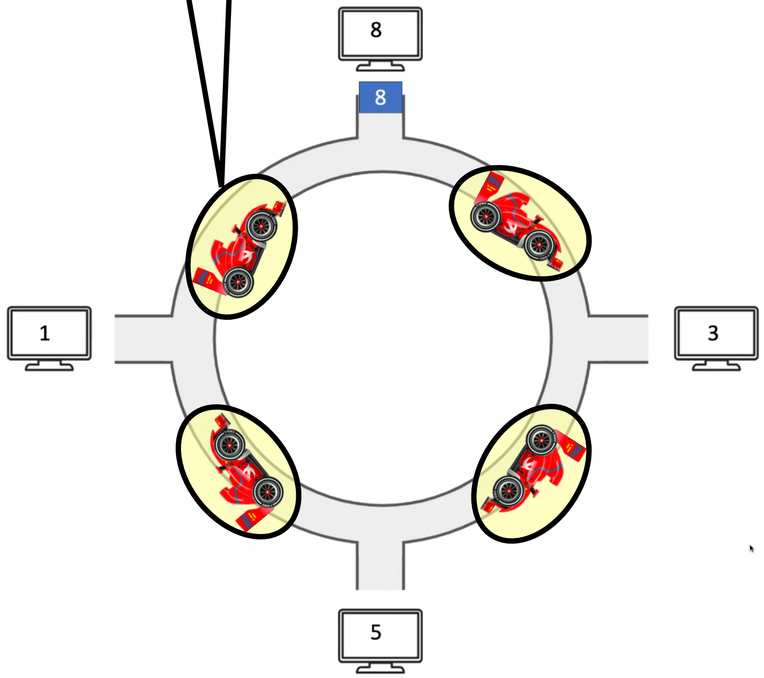

Finish time is based on the sum of the latencies in the ring!
8
8
1
NCCL ring
3
5

Communicate

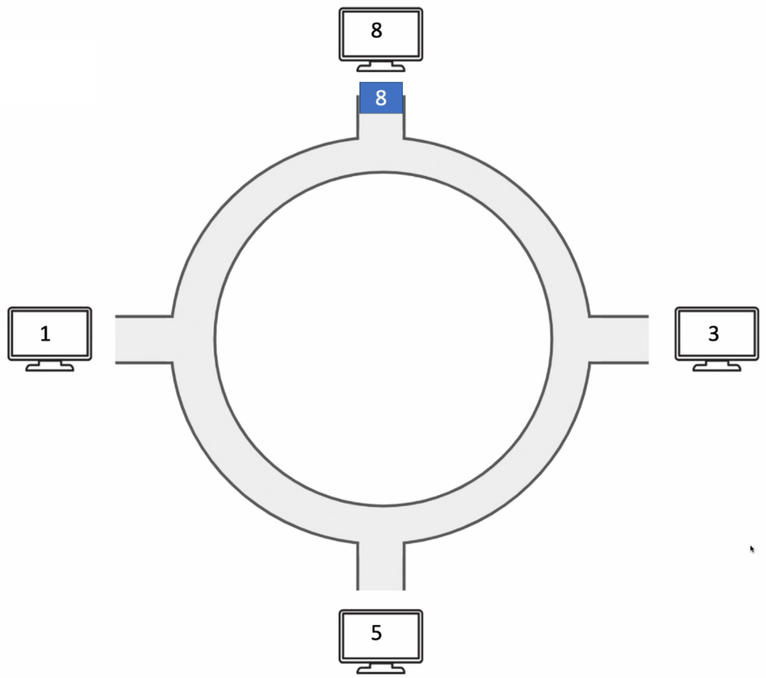


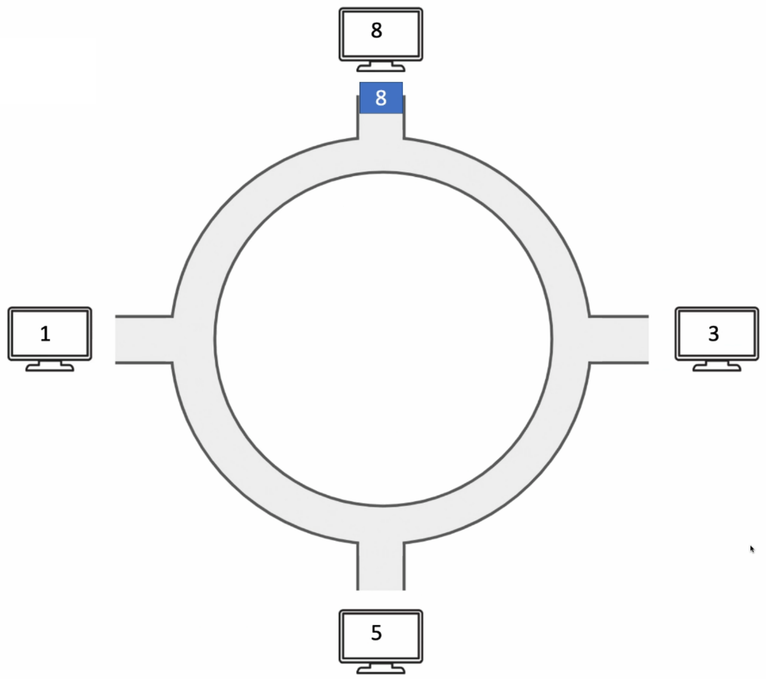
Communicate

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

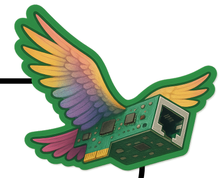


Pit crew is not done until the slowest member is done.

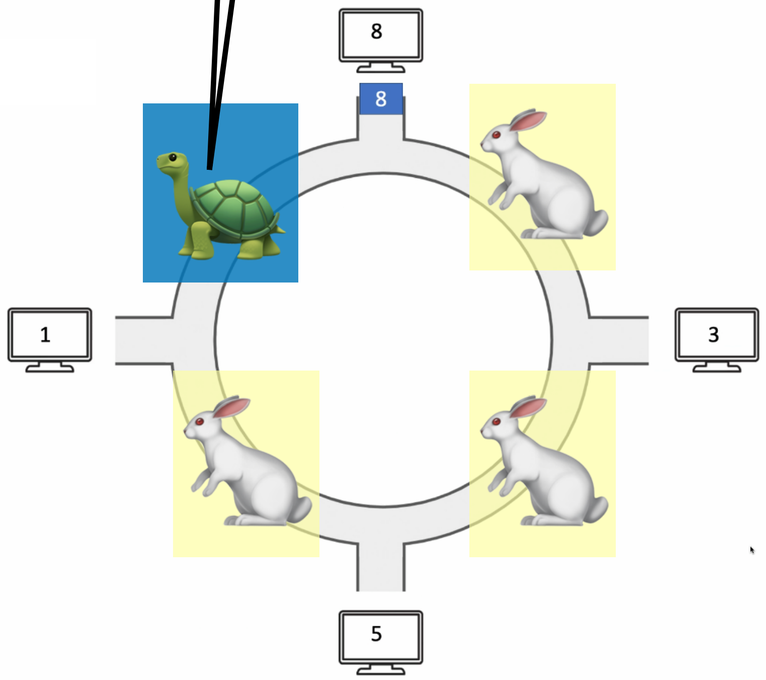

8
8
1
3
5


8
8
1
3
5


Slowest member determines speed.



Slowest flow
determines completion time.



Slowest member
determines speed.

1. AI workload layers

2. Demands from AI networks

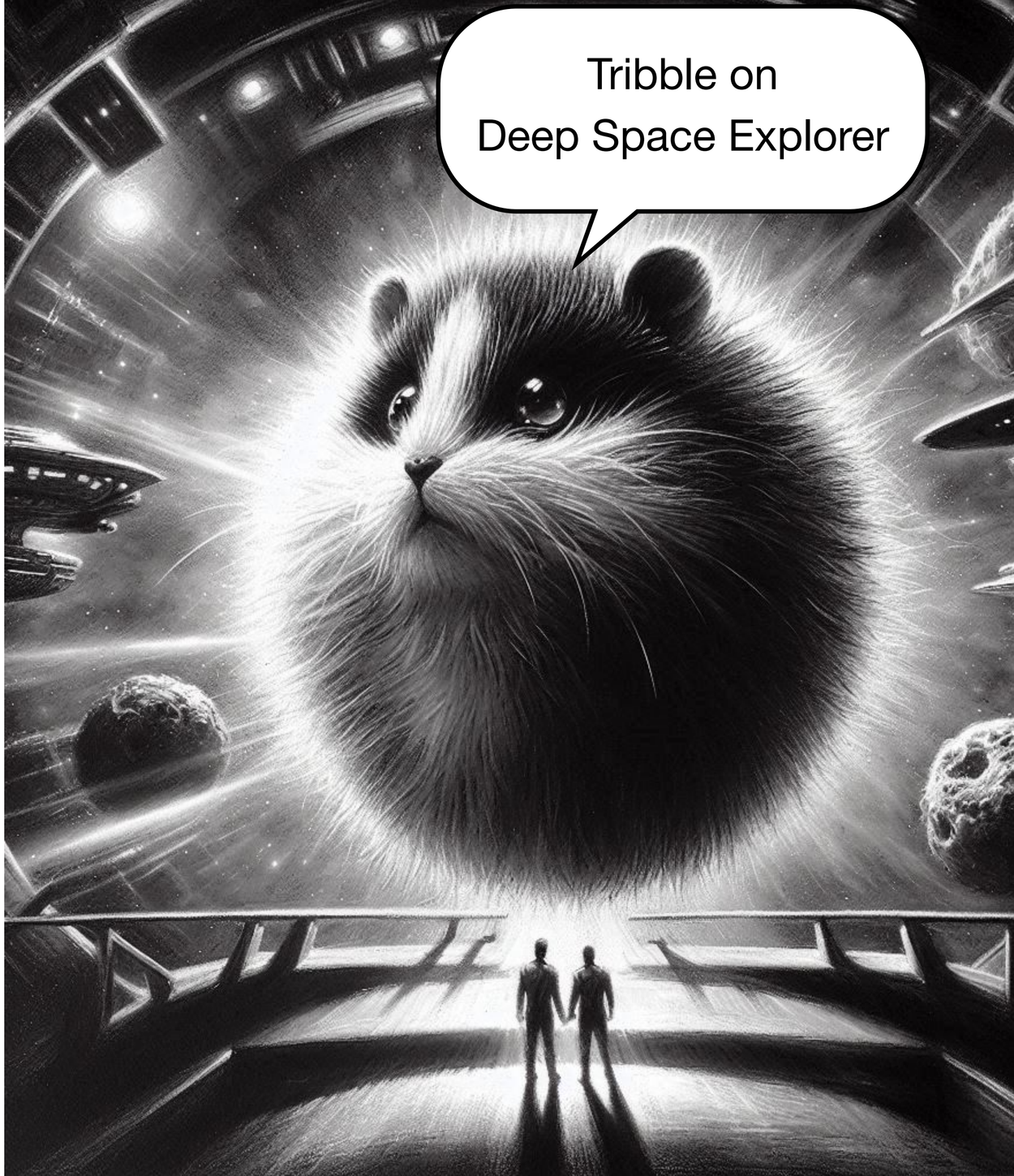
3. Challenges in AI networks

4. Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus



1. AI workload layers

2. Demands from AI networks

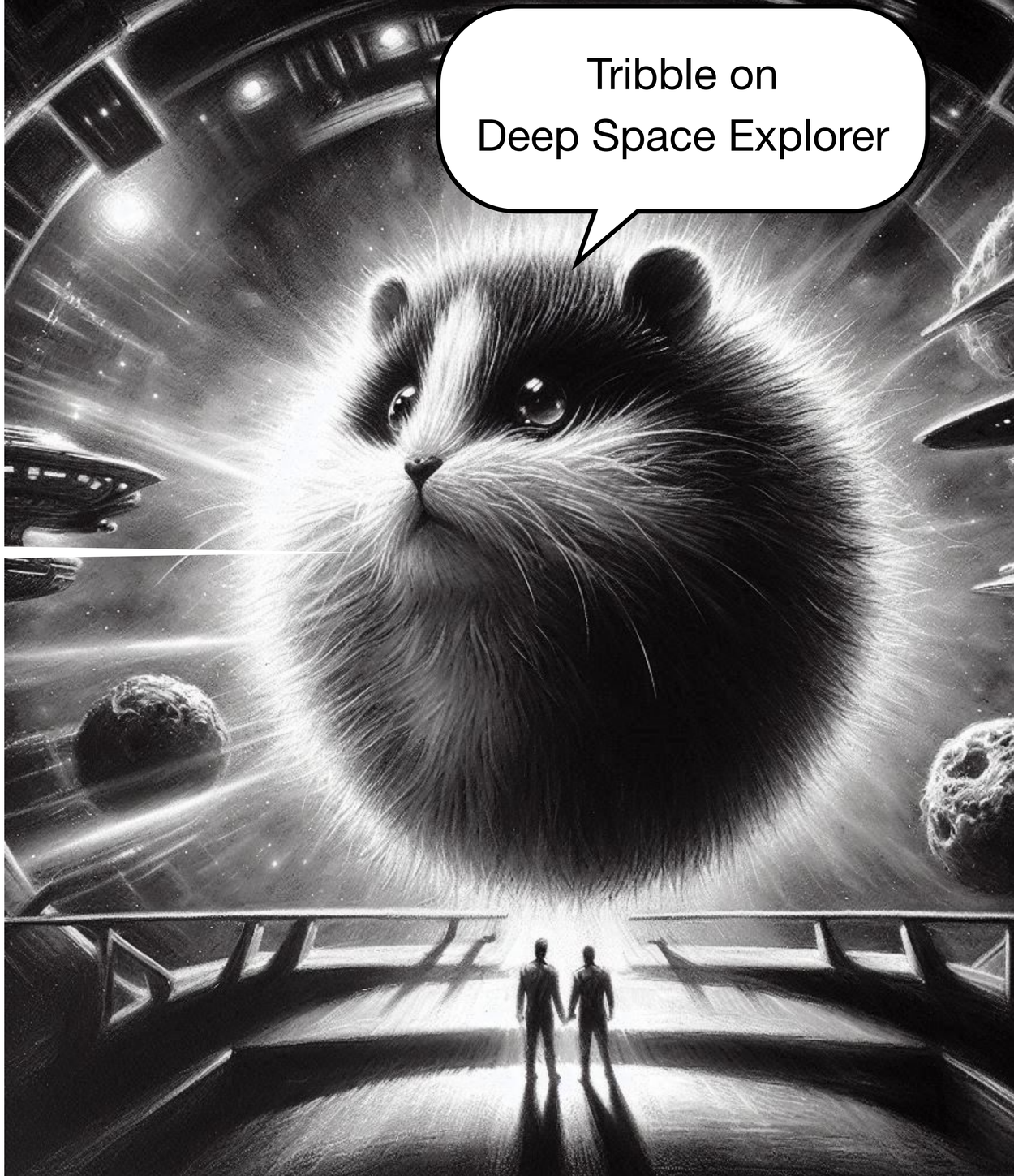
3. Challenges in AI networks

4. Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus

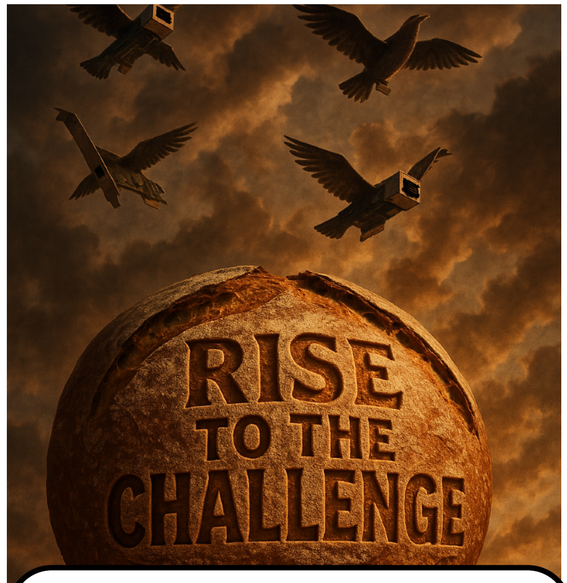


Challenges

🙈 Visibility

Challenges

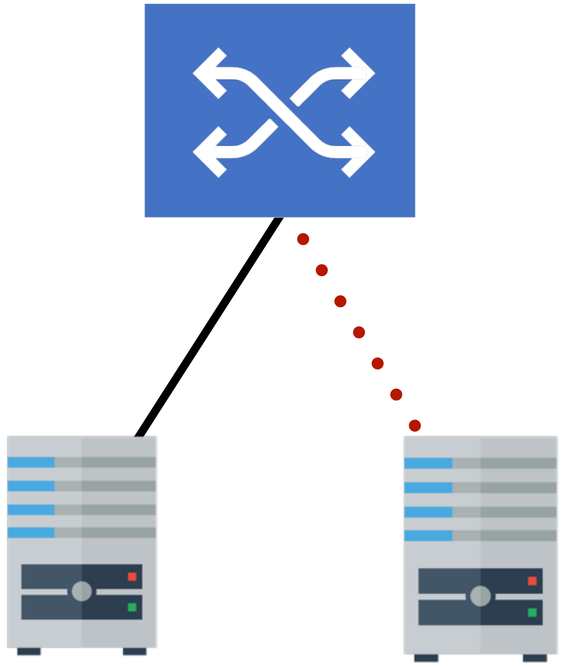
🙈 Visibility



🔥 Reliability

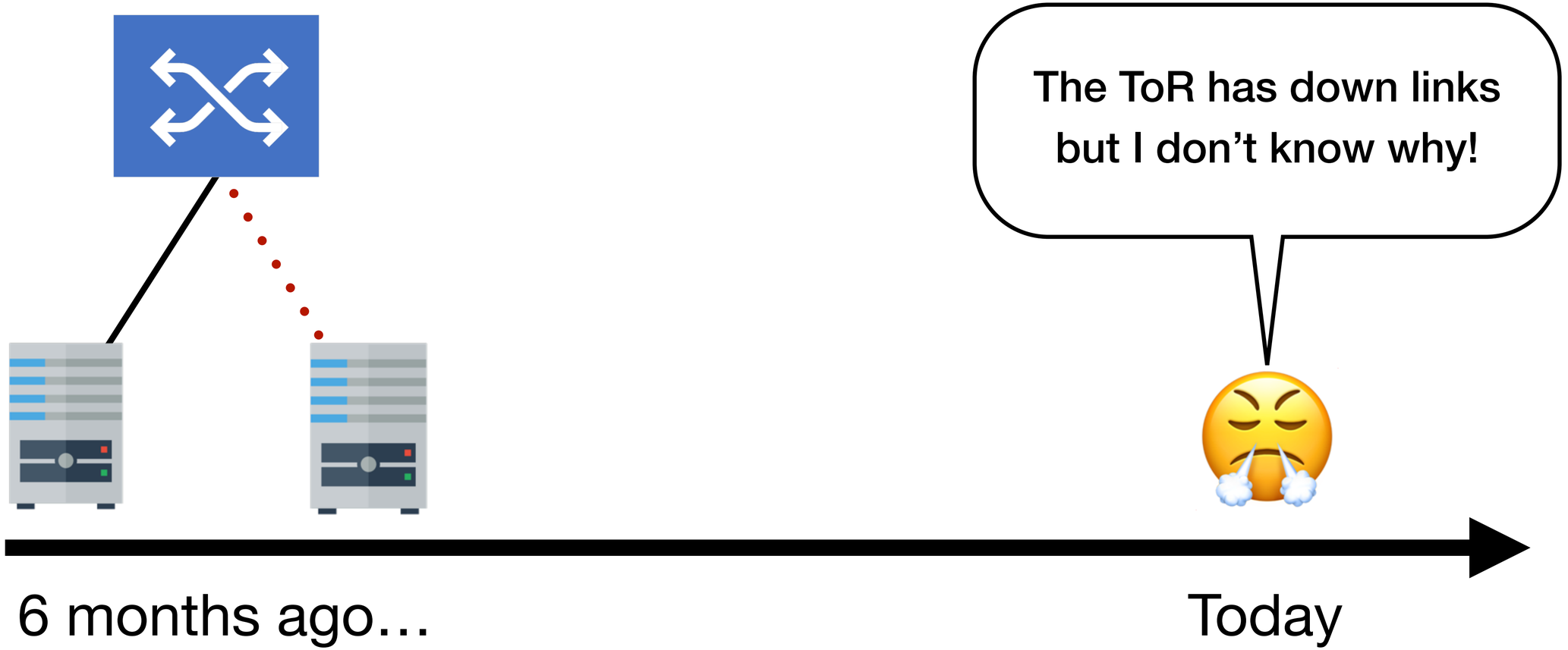
Lack of visibility

Lack of visibility



6 months ago…

Lack of visibility

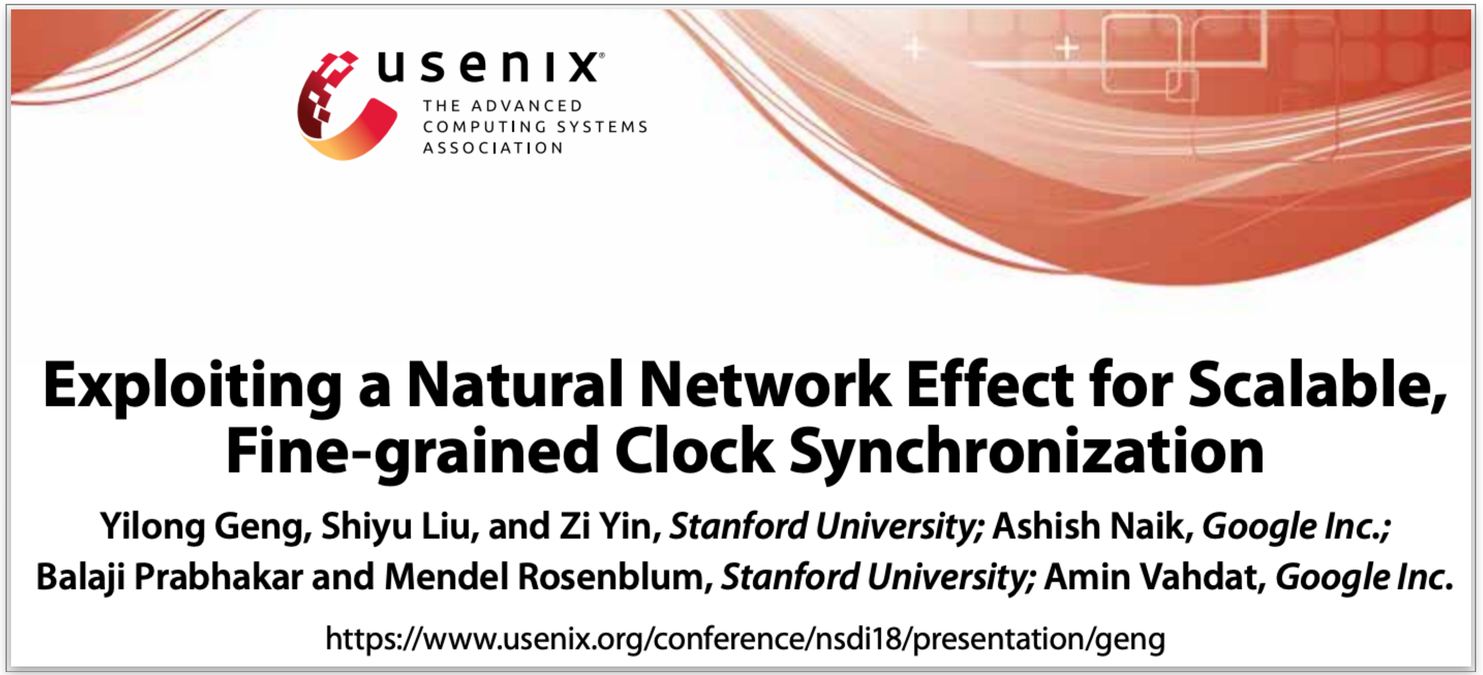


Lack of visibility

- Lack of fine grained metrics
- Inability to identify root cause quickly

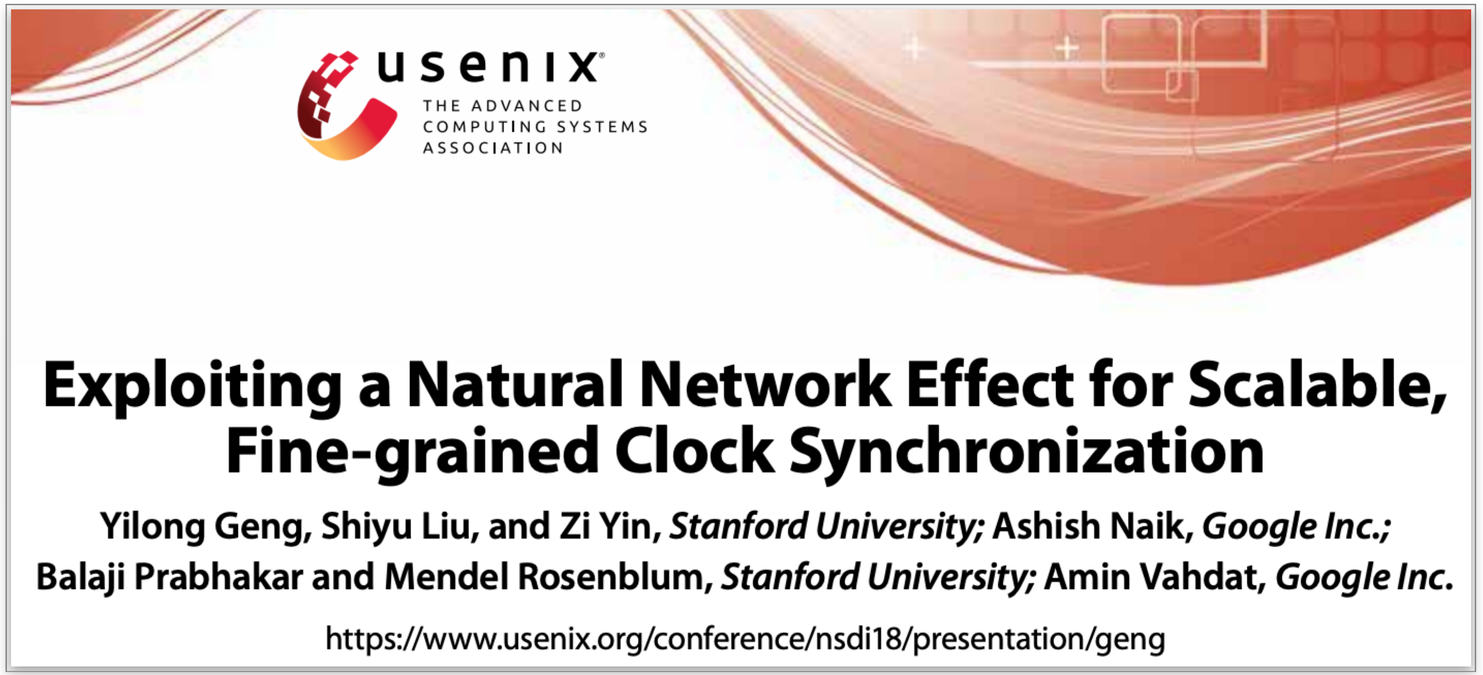
Huygens clock synchronization

- Software, probe mesh based
- High precision at scale



Huygens clock synchronization

- Software, probe mesh based
- High precision at scale

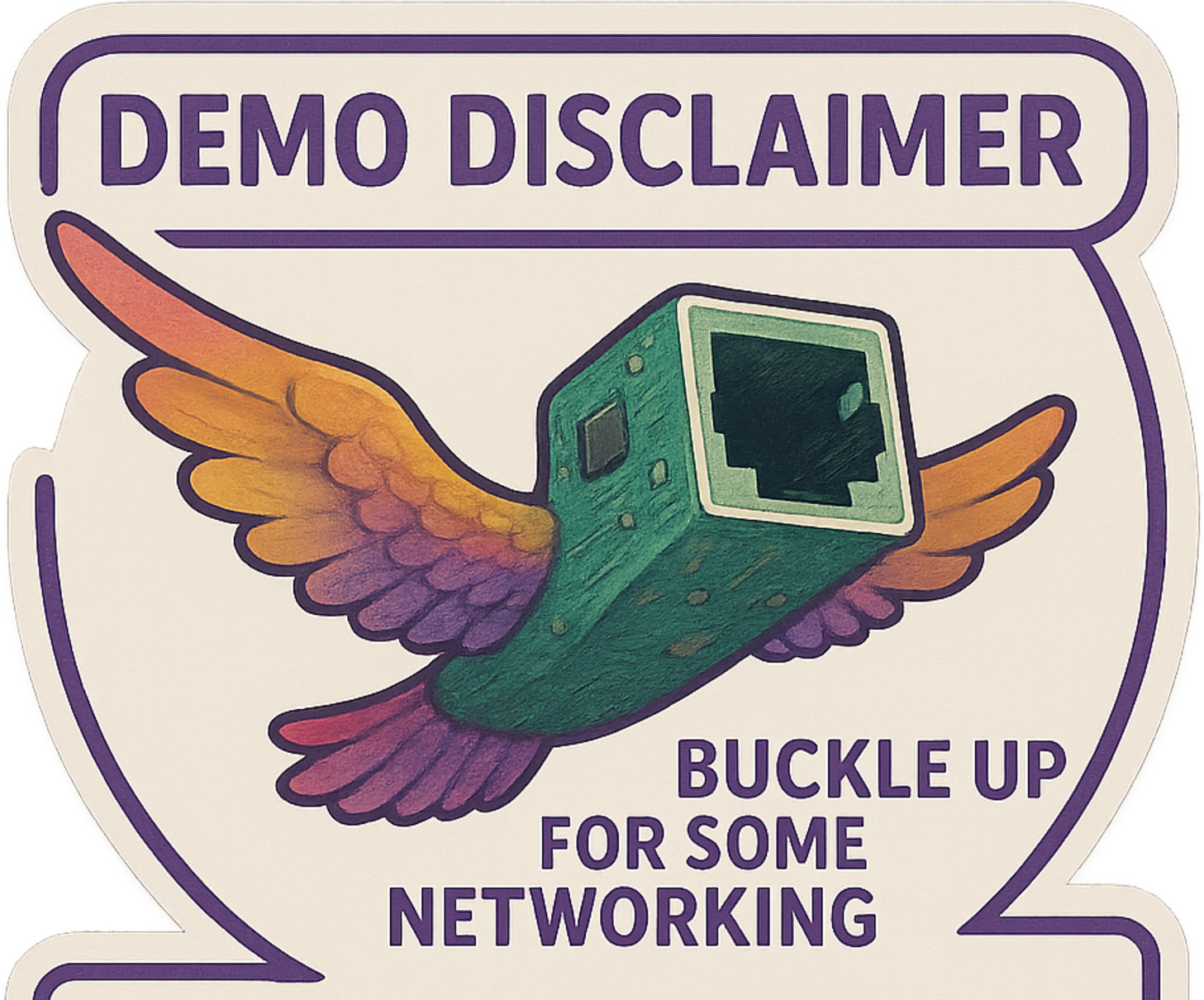


⚠️ Demo Disclaimer ⚠️

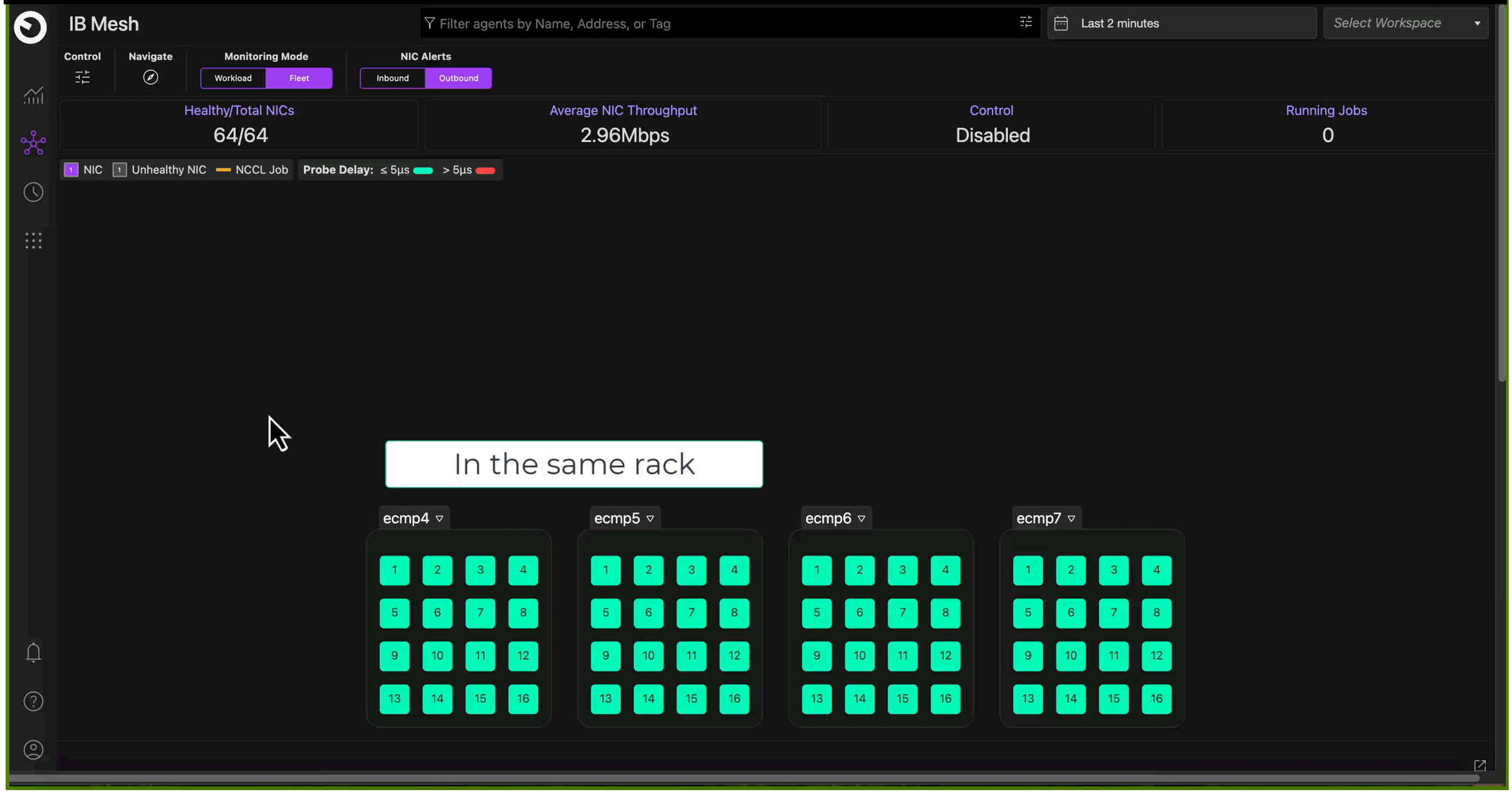
The upcoming demo
may not be
warp-speed thrilling…

unless you really, *really* love networking.

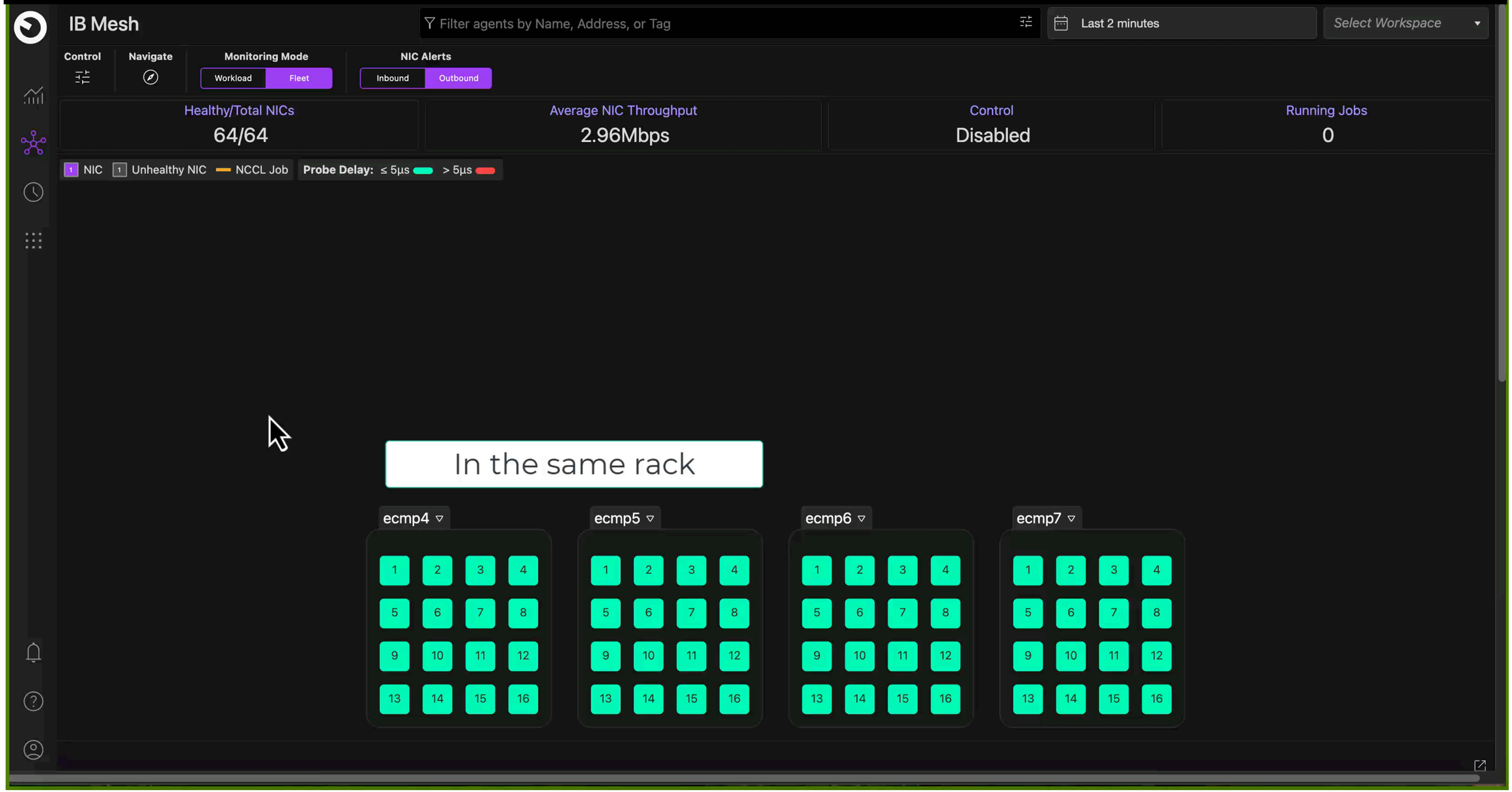
(If you do, you’re in for a treat!)
🚀



Visibility demo



Visibility demo



Lack of reliability

Lack of reliability

One of the most common problems encountered is Infiniband/RoCE link failure. Even if each NIC to leaf switch link had a mean time 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 on a brand new, working cluster.

Estimated Time to First Job Failure (Minutes)				
Mean Time to Failure Per Link	3 years	4 years	5 years	10 years
Number of GPUs				
10,000	157.7	210.2	262.8	525.6
20,000	78.8	105.1	131.4	262.8
50,000	31.5	42.0	52.6	105.1
100,000	15.8	21.0	26.3	52.6

Source: www.semianalysis.com

Lack of reliability

One of the most common problems encountered is Infiniband/RoCE link failure. Even if each NIC to leaf switch link had a mean time 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 on a brand new, working cluster.

Estimated Time to First Job Failure (Minutes)				
Mean Time to Failure Per Link	3 years	4 years	5 years	10 years
Number of GPUs				
10,000	157.7	210.2	262.8	525.6
20,000	78.8	105.1	131.4	262.8
50,000	31.5	42.0	52.6	105.1
100,000	15.8	21.0	26.3	52.6

Source: www.semianalysis.com

What is link flapping?

- Network link temporarily drops. (up => down)
- It can last in the order of seconds.
  - A few seconds => may self heal and recover from it if < timeout.
  - > a few seconds => collective communication times out => job fails.

What is link flapping?

- Network link temporarily drops. (up => down)
- It can last in the order of seconds.
  - A few seconds => may self heal and recover from it if < timeout.
  - > a few seconds => collective communication times out => job fails.

Lack of reliability

Lack of reliability

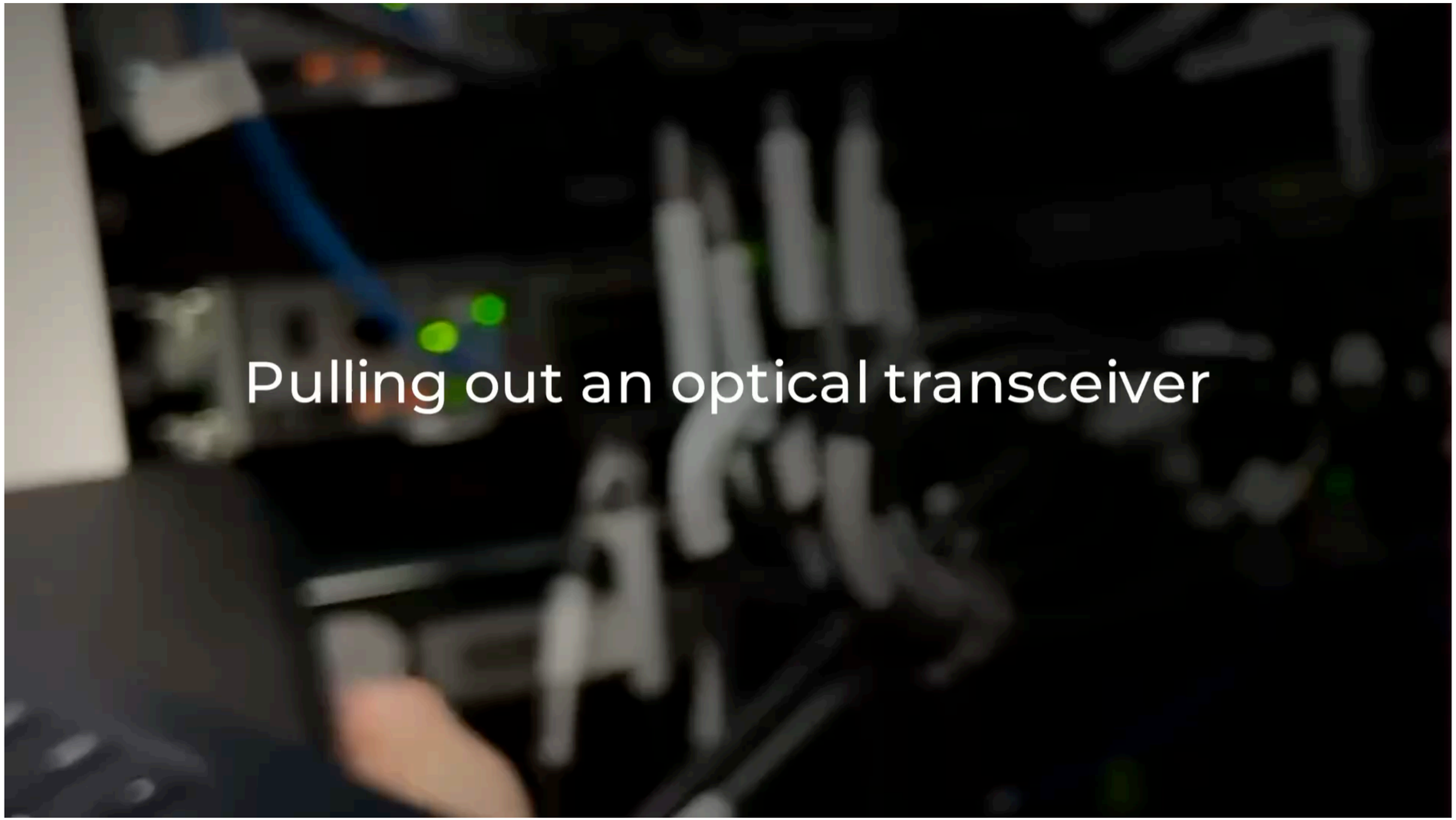
- Fall back to previous checkpoint
- Lost ⌛💰

Poll

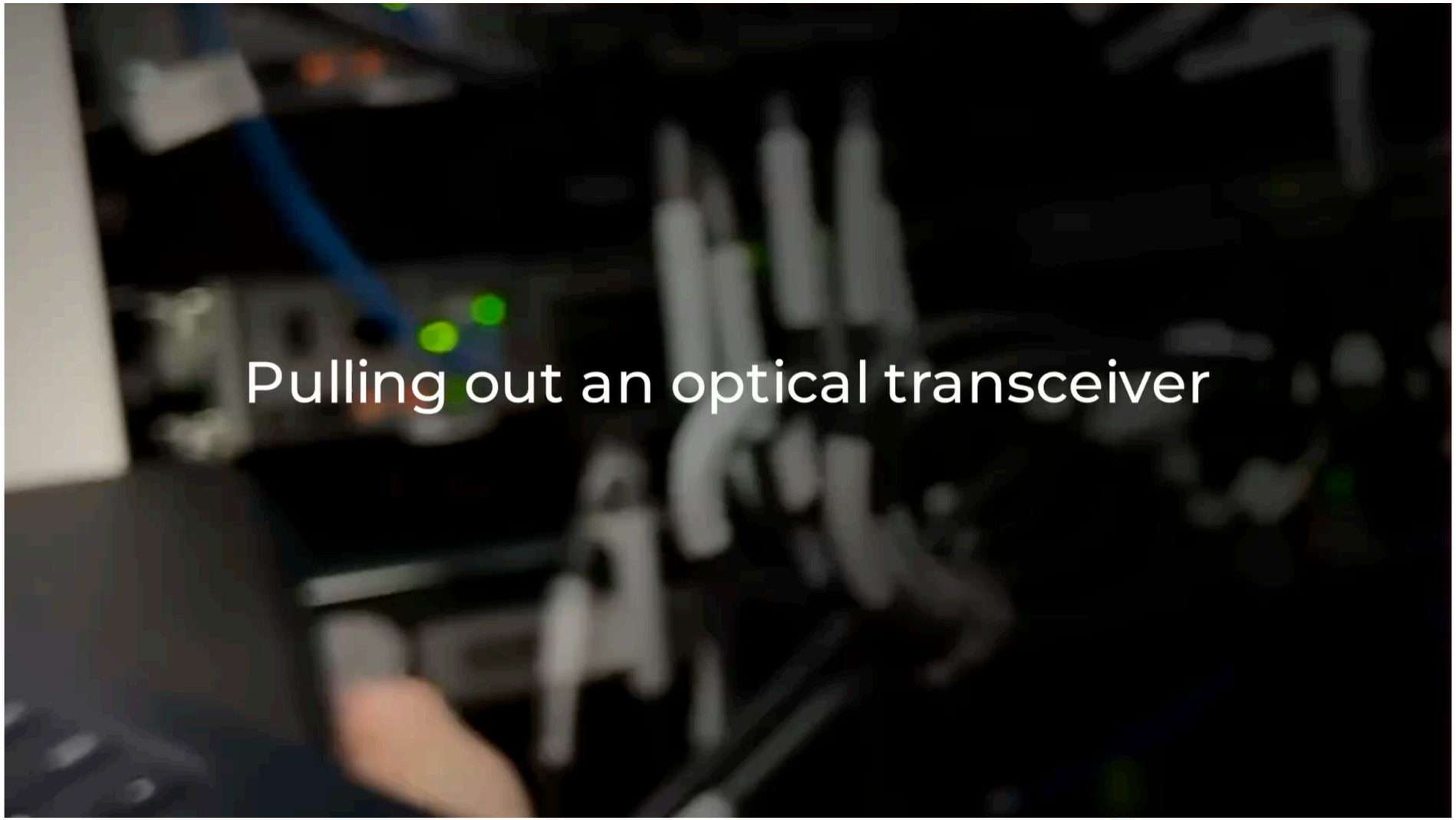
👉 *If your cluster had a personality, which would it be?*

- **The Ghoster** 👻 (drops connections without warning)
- **The Zen Master** 🧘🏾‍♀️(always stable)

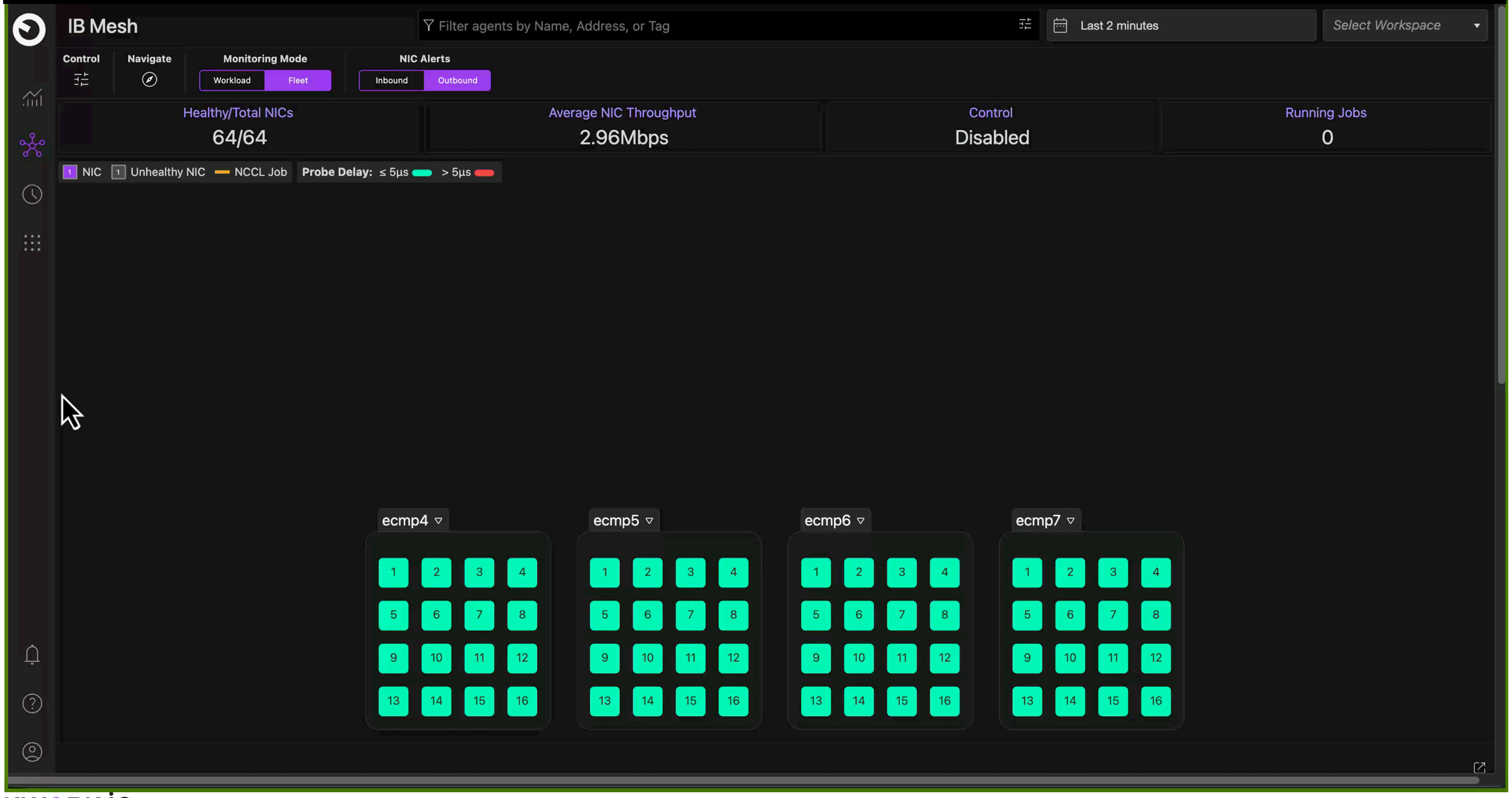
Lack of reliability



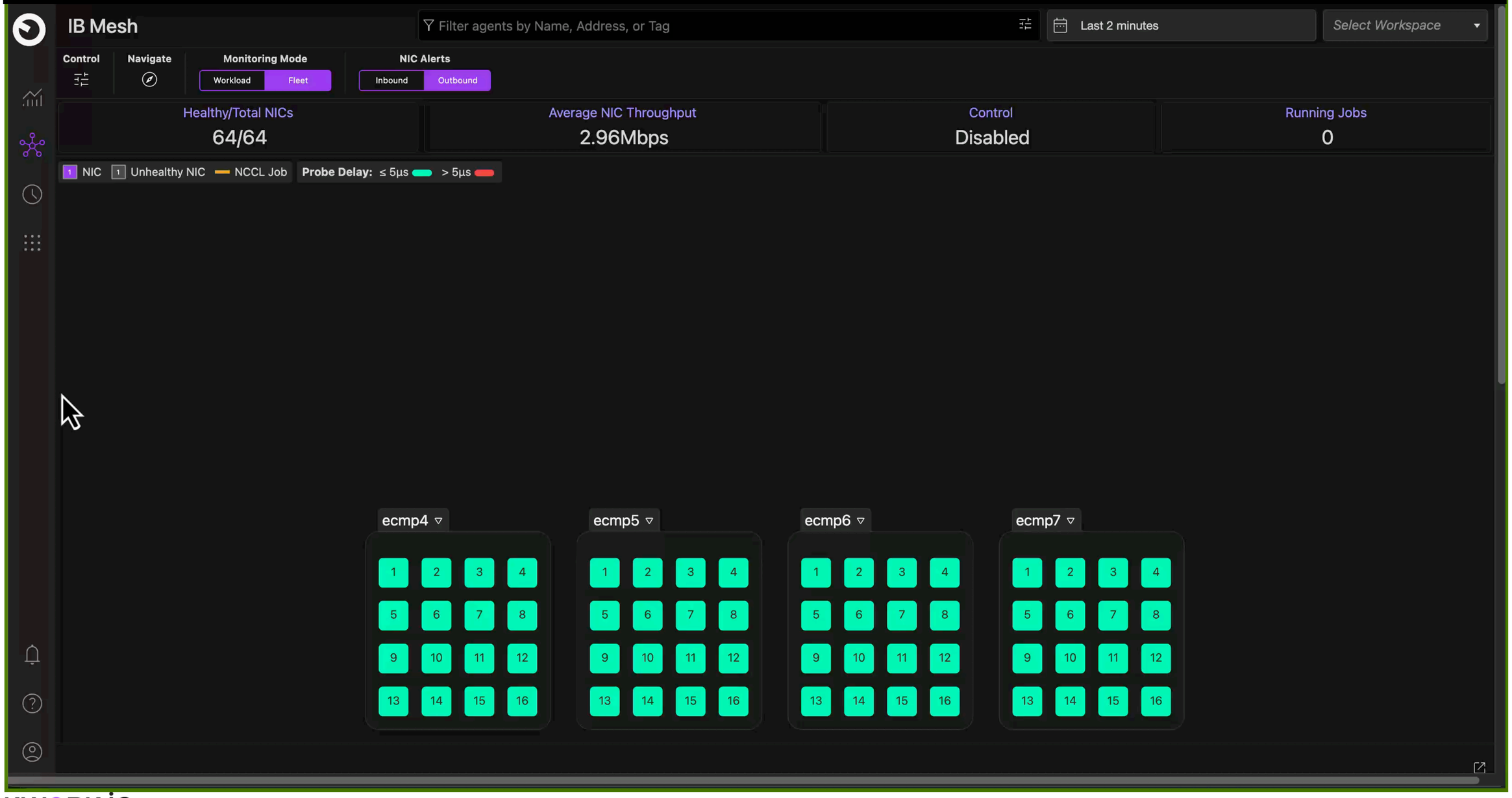
Lack of reliability



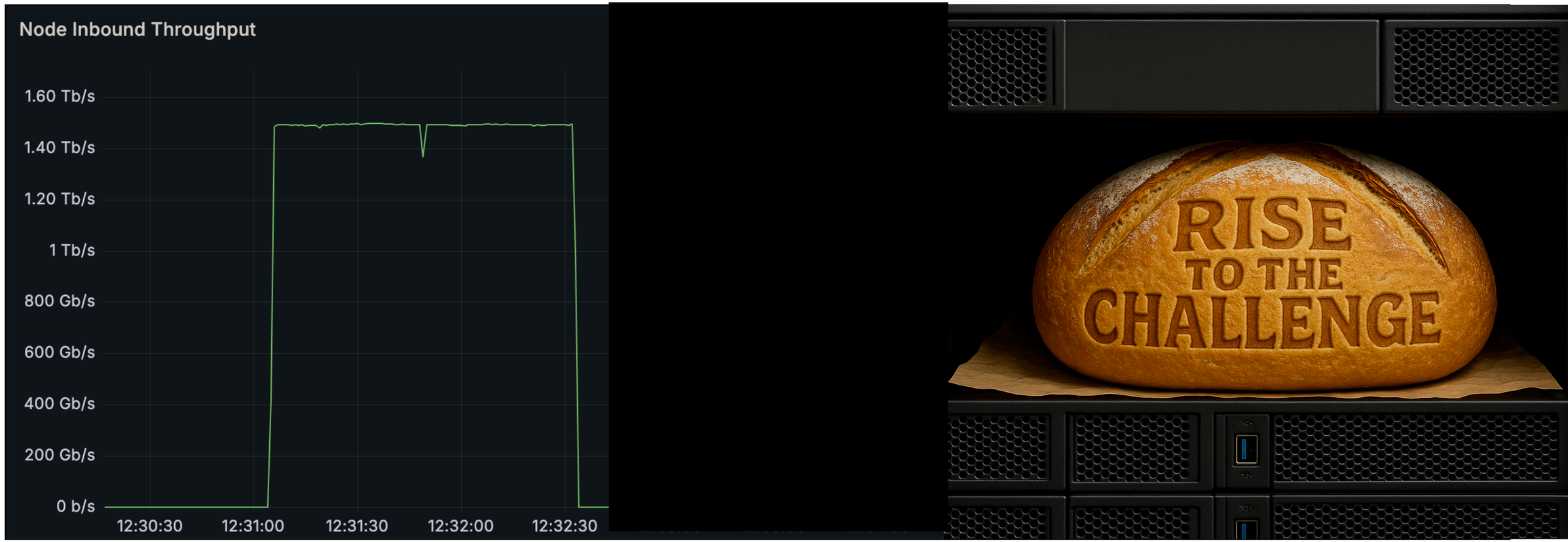
Flapping demo



Flapping demo



NIC Flaps in a cluster on Oracle Cloud

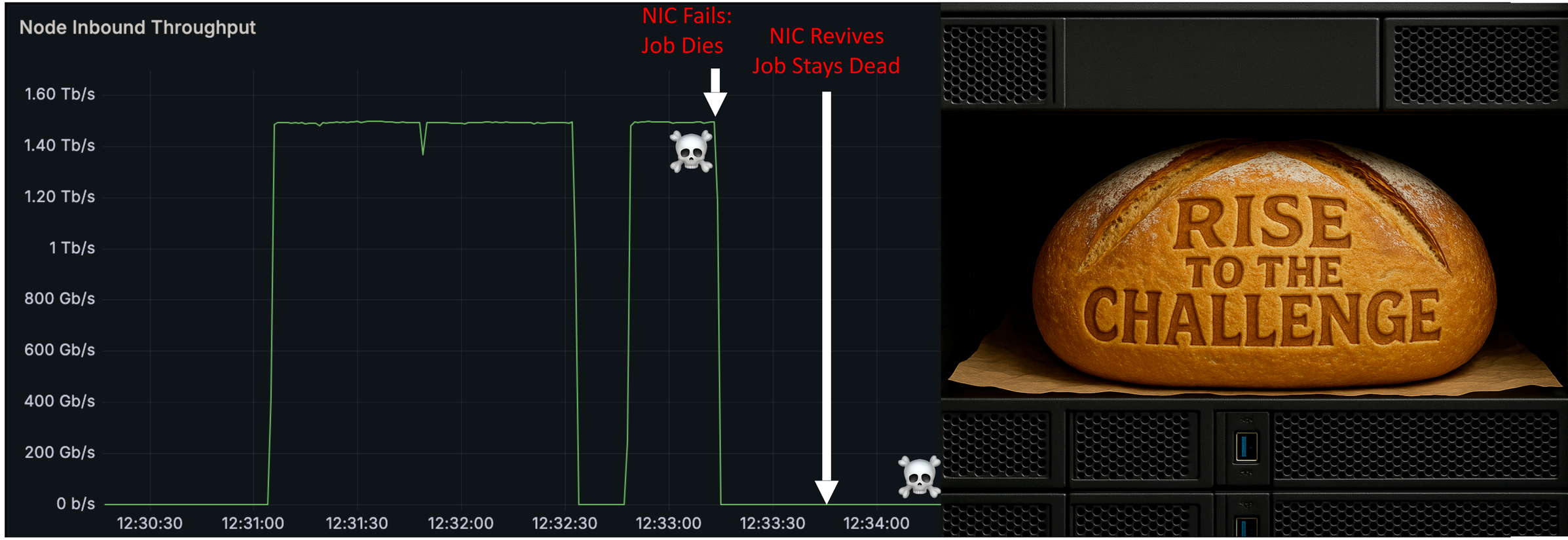


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

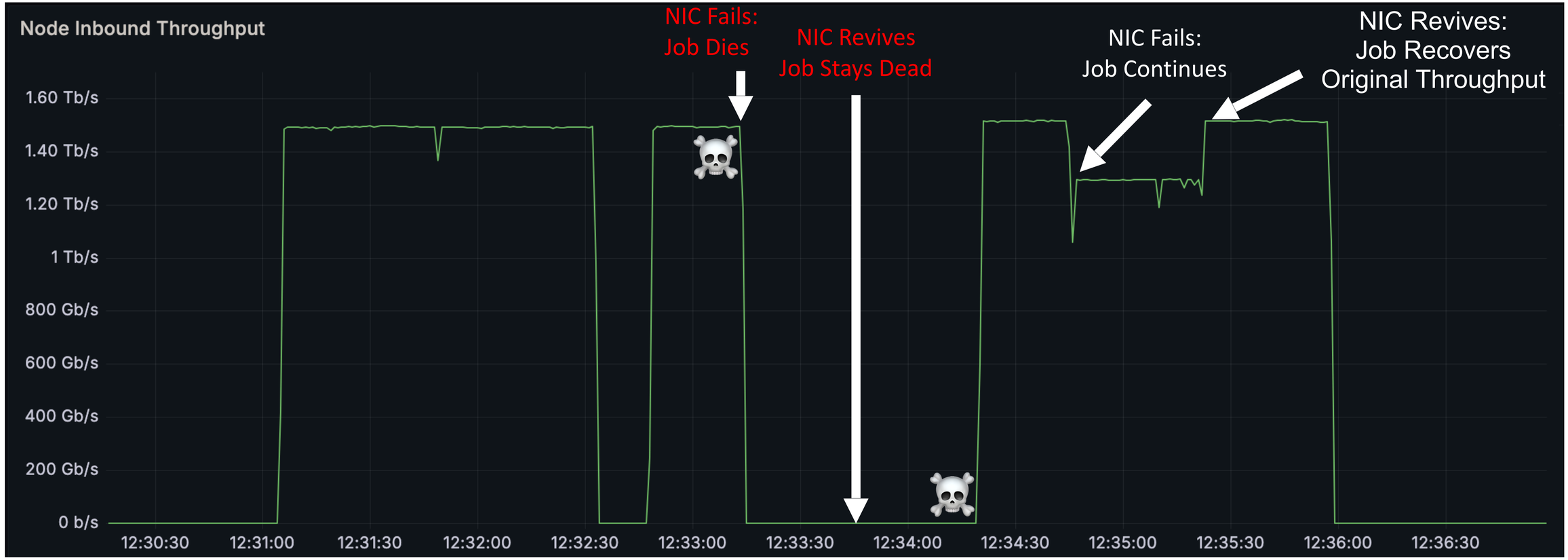


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

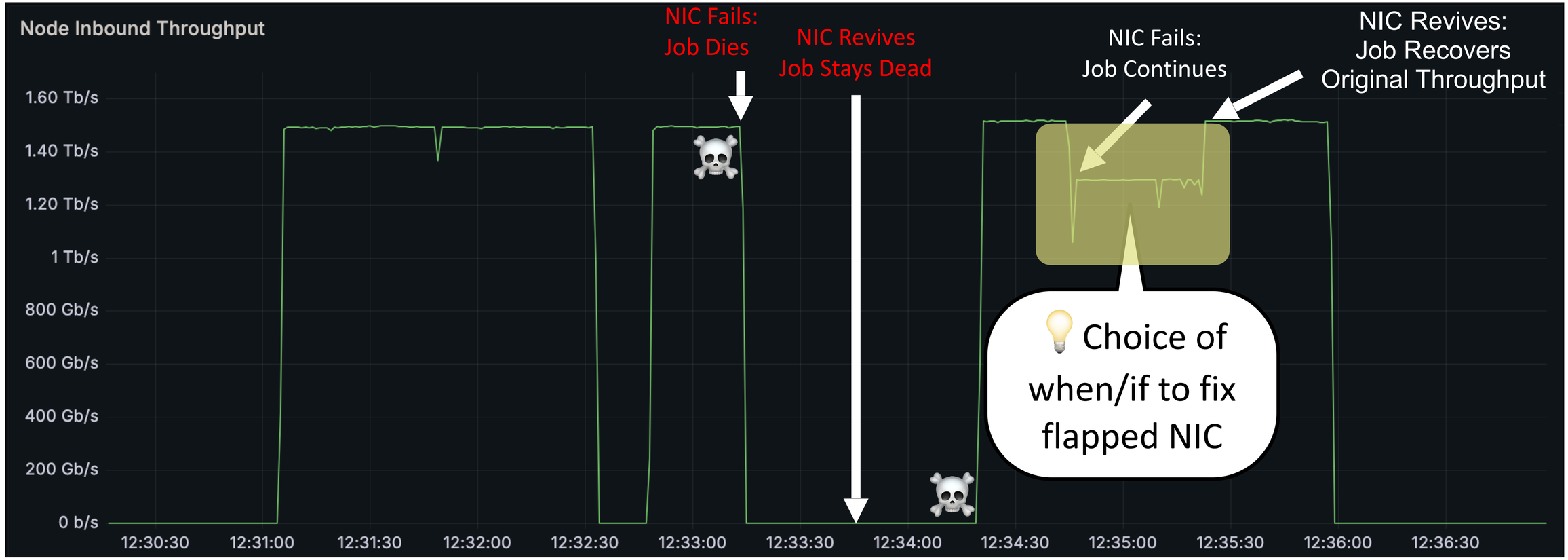


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

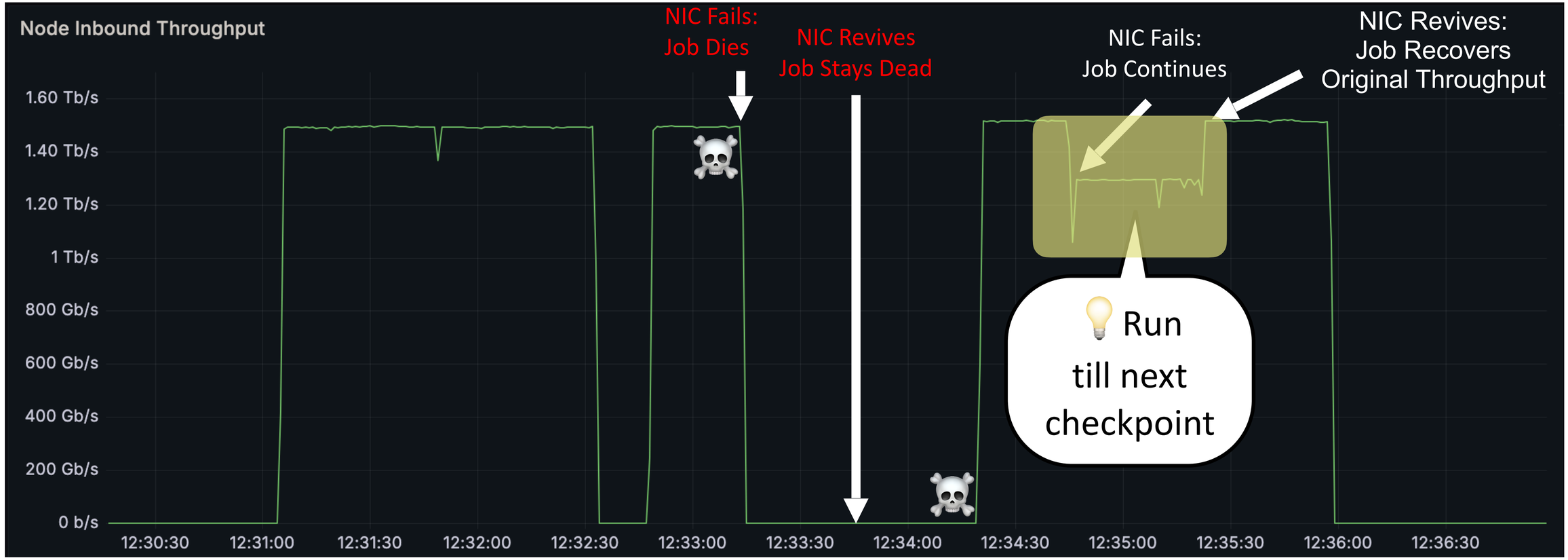


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

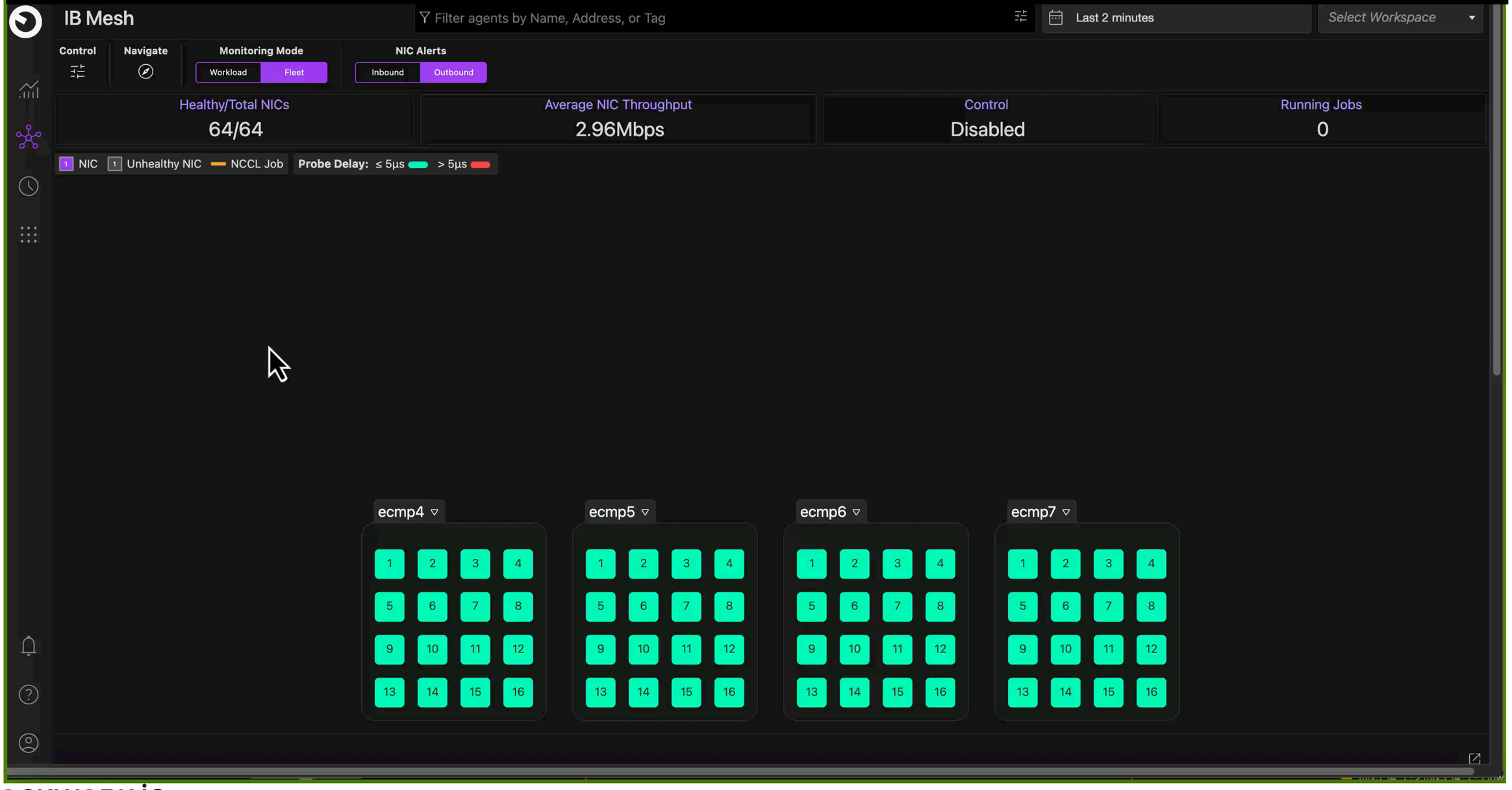


Run 1: No Failure

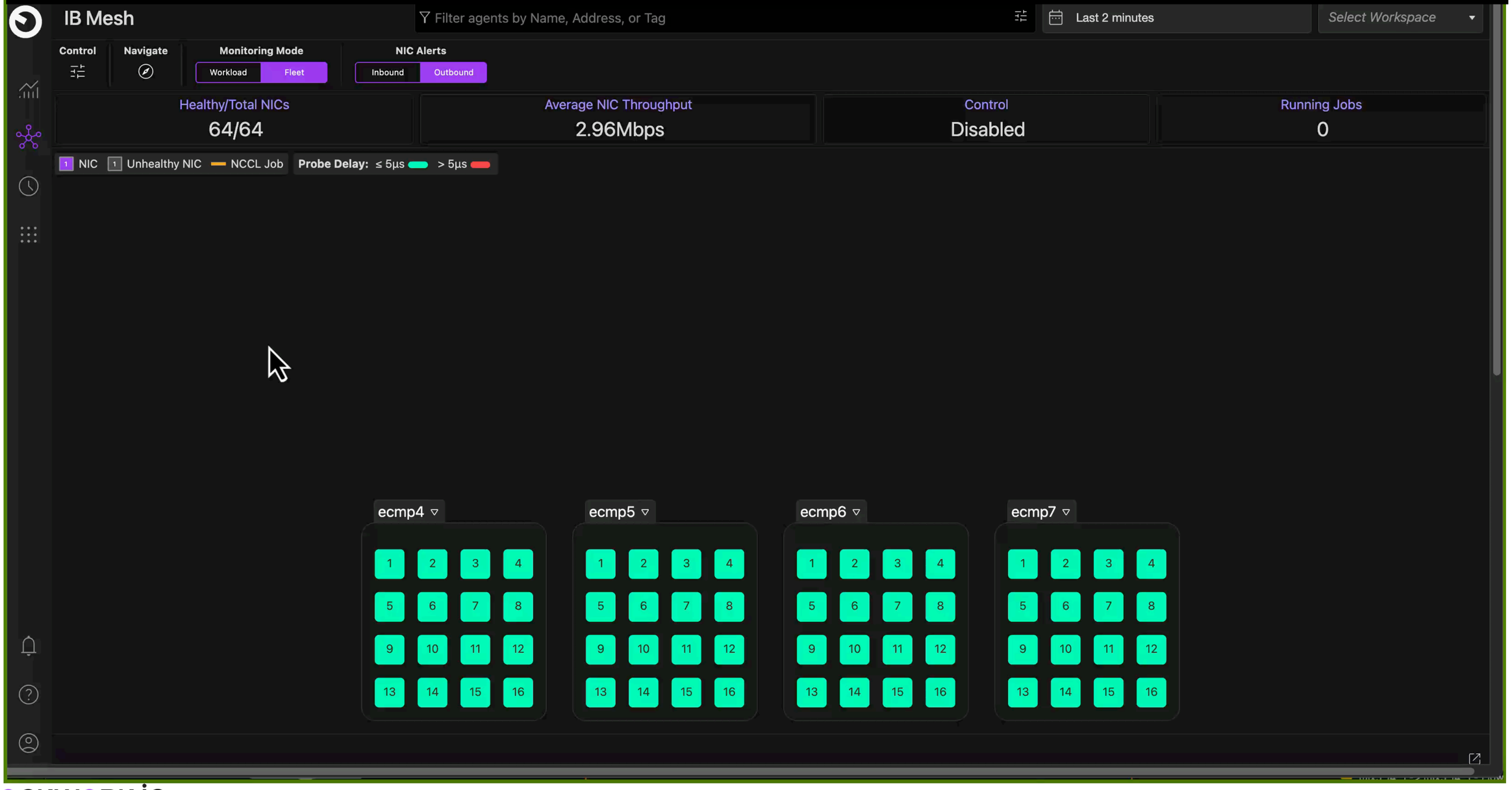
Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

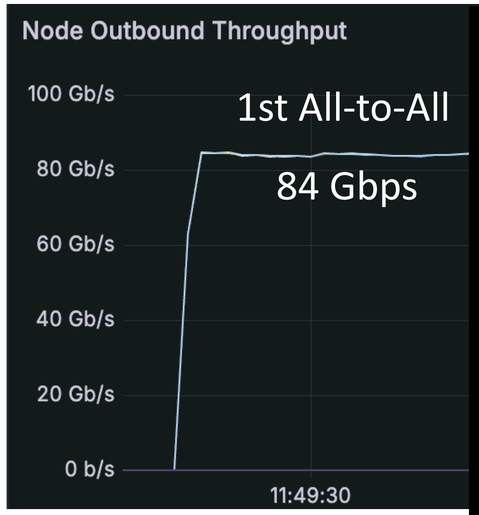
Contention demo

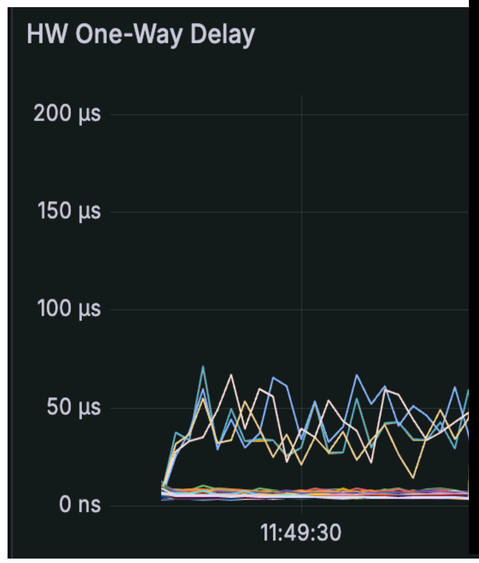


Contention demo



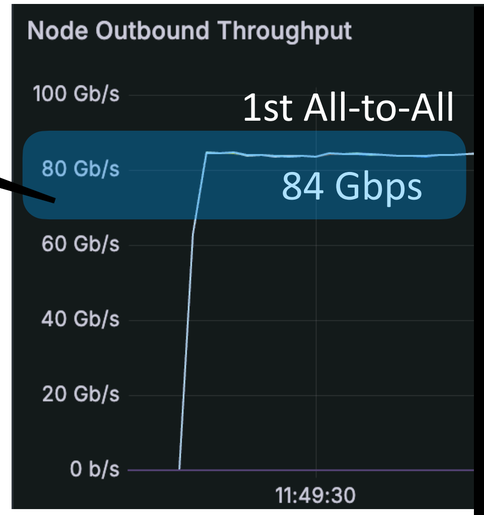
Contention on ECMP Test Bed



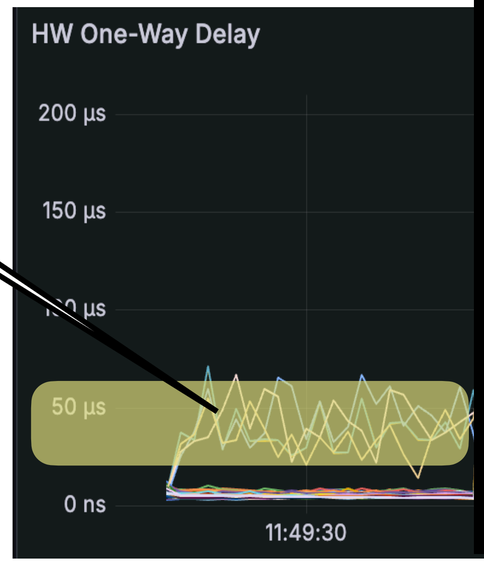


Contention on ECMP Test Bed

⬇️ Lower Throughput!

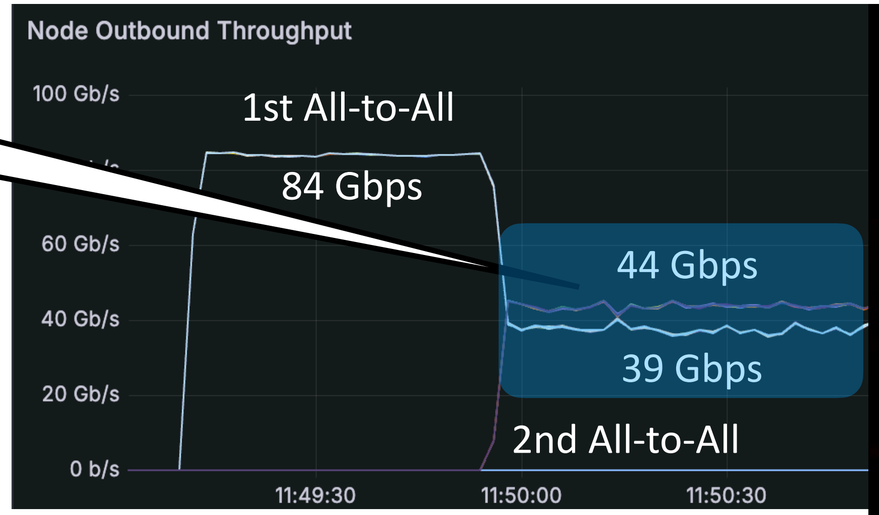


🐢 Queue pairs with high one way delays!

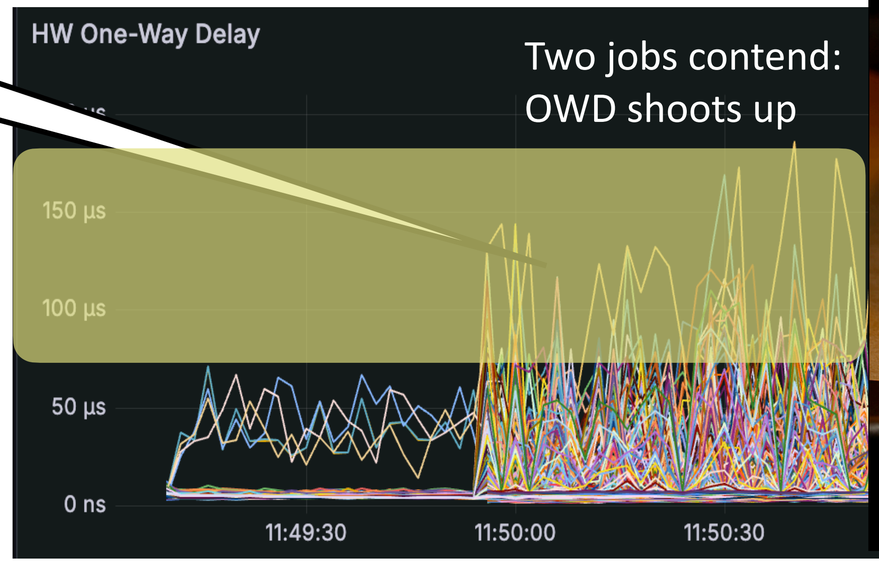


Contention on ECMP Test Bed

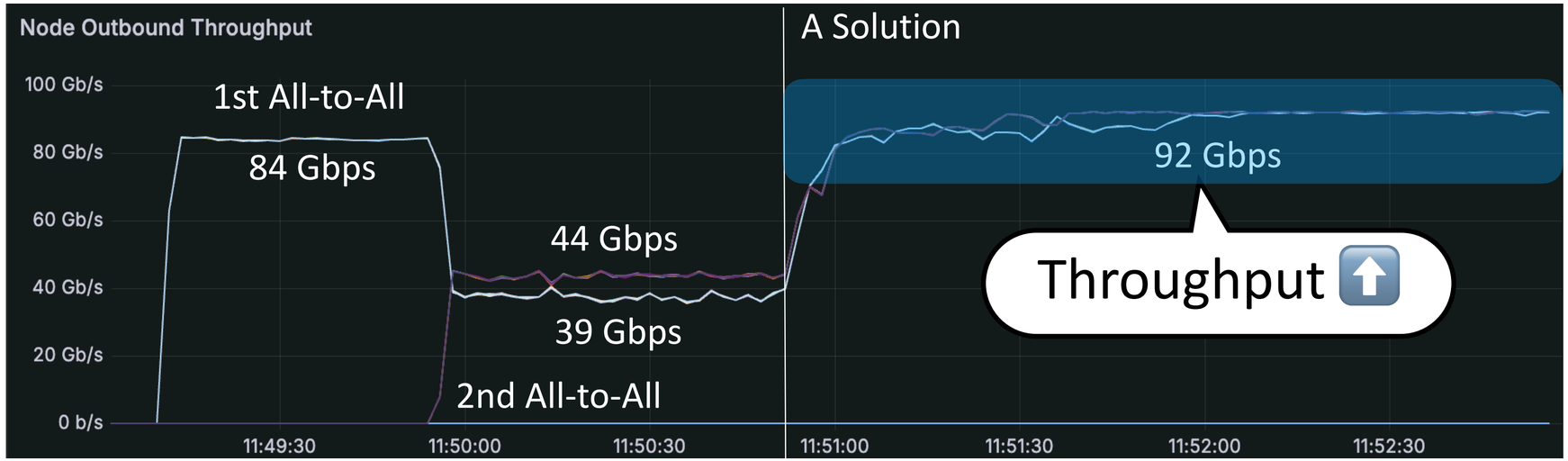
⬇️ Lower Throughput!

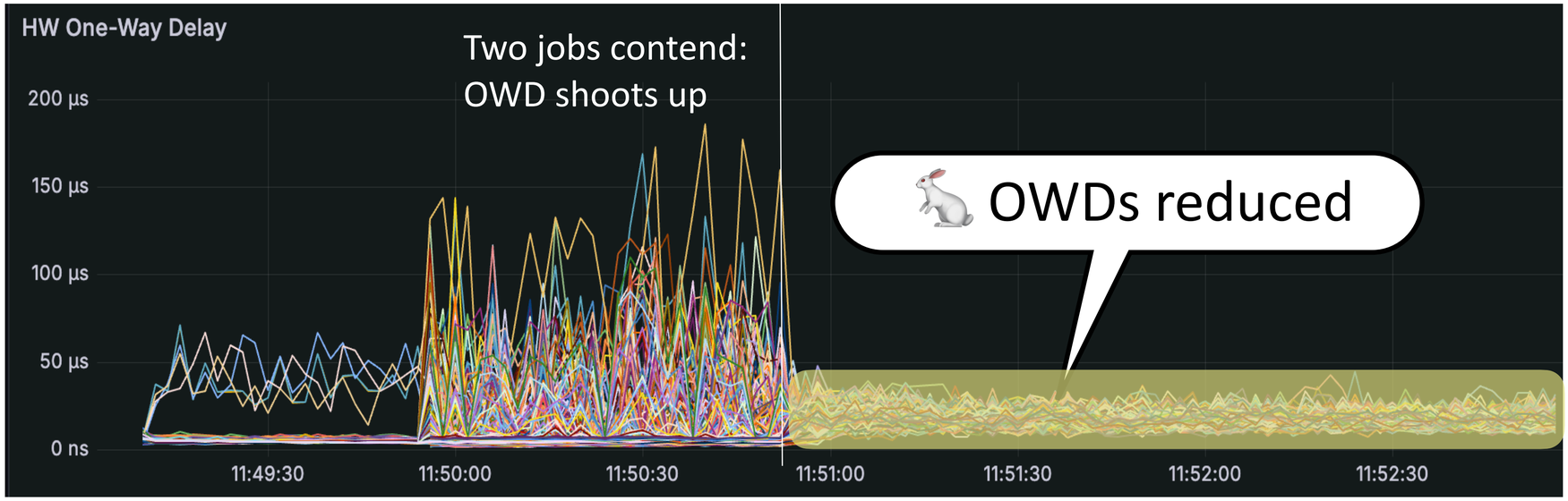


🐢 Queue pairs with high one way delays



Contention on ECMP Test Bed





Poll

👉 *If you could wave a magic wand 🪄 and get rid of a networking pain forever, which would you choose?*

- **No more link flaps**
- **Self-healing NICs**

Key Takeaways

🔎 Lack of visibility into queue pairs slows down diagnosis.

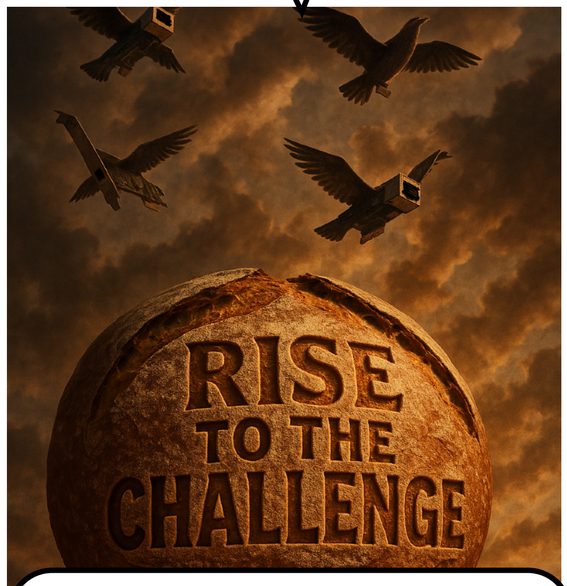
🙈 Visibility

Key Takeaways

🔎 Lack of visibility into queue pairs slows down diagnosis.

🙈 Visibility

NIC/link flapping cause jobs to crash.



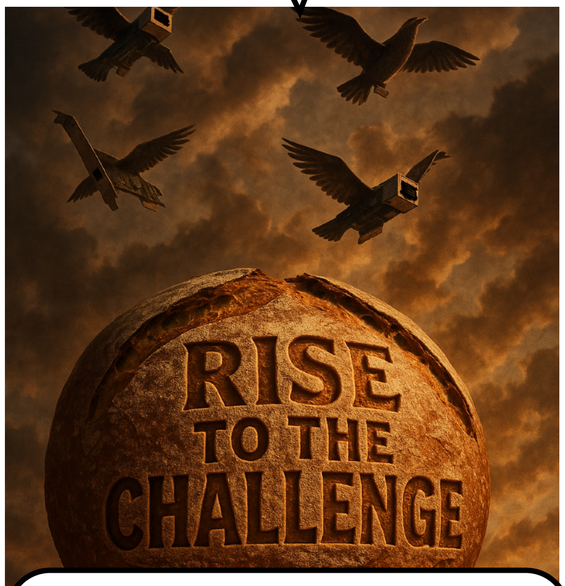
🔥 Reliability

Key Takeaways

🔎 Lack of visibility into queue pairs slows down diagnosis.

🙈 Visibility

NIC/link flapping cause jobs to crash.



🔥 Reliability

💨 Delays in slowest flows degrade throughput.

🐌 Performance

Thank You

linkedin.com/in/lerna

We’re hiring!
careers@clockwork.io
clockwork.io/careers

Thank You

linkedin.com/in/lerna

We’re hiring!
careers@clockwork.io
clockwork.io/careers

Lucille Ball => Star Trek