







Resilience for AI Workloads at Scale

Lerna Ekmekcioglu
Sr. Solutions Engineer
Clockwork.io



Resilience for AI Workloads at Scale

Lerna Ekmekcioglu
Sr. Solutions Engineer
Clockwork.io



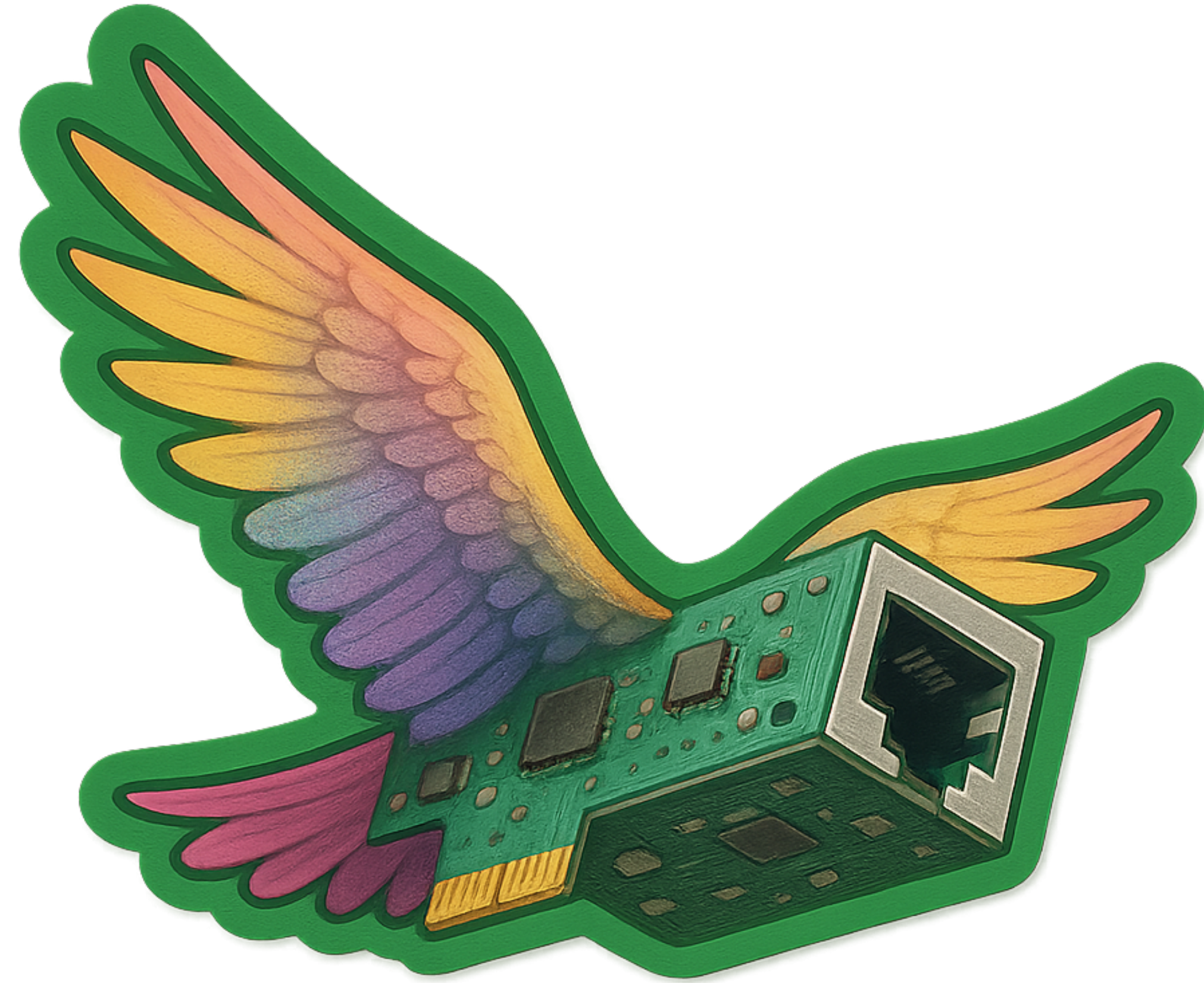


AI/ML Eng/
Researcher

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



AI-generated

AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus





DEEP SPACE

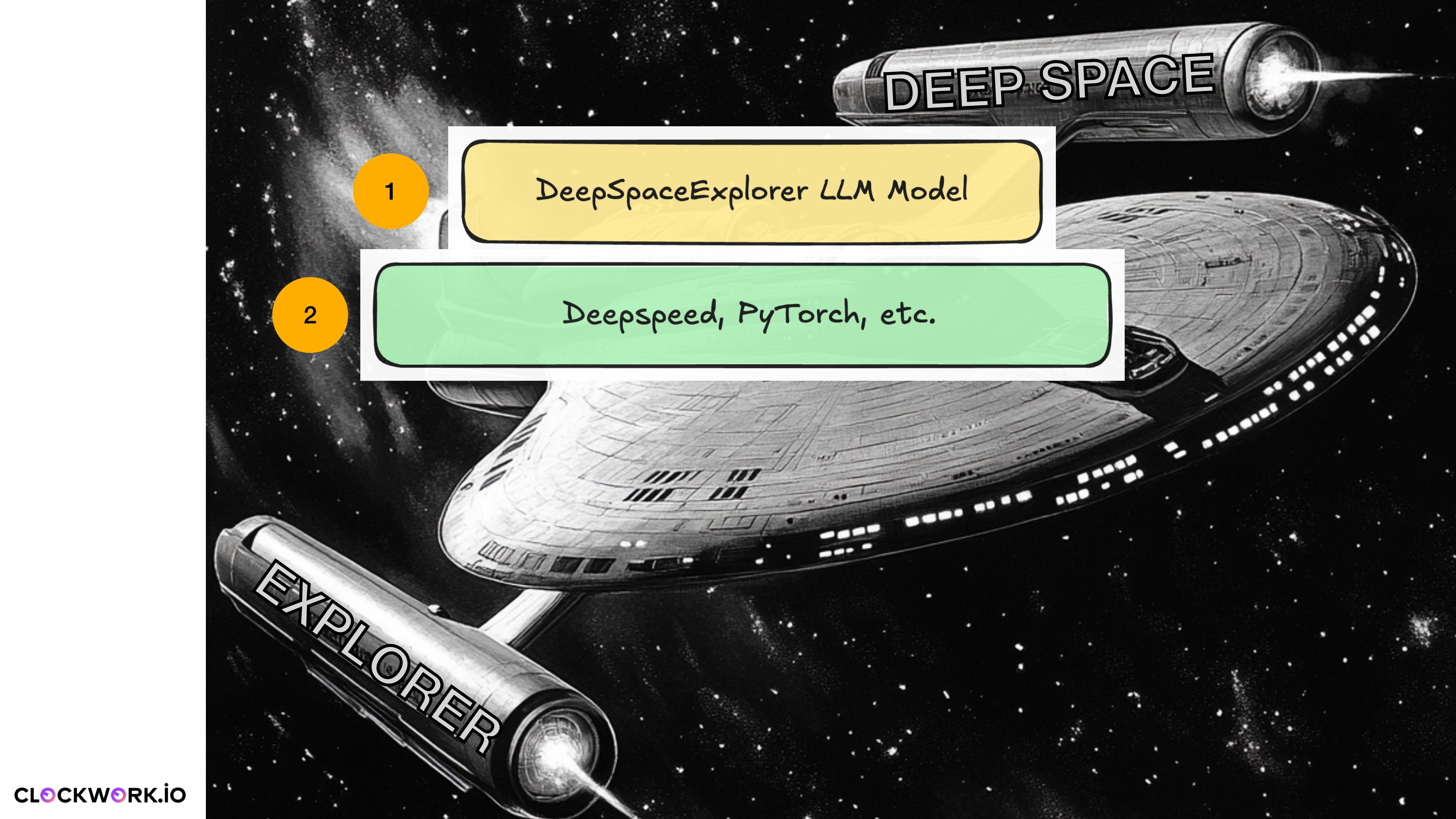
EXPLORER

DEEP SPACE

1

DeepSpaceExplorer LLM Model

EXPLORER



DEEP SPACE

1

DeepSpaceExplorer LLM Model

2

Deepspeed, PyTorch, etc.

EXPLORER



Large scale jobs

1

DeepSpaceExplorer LLM Model

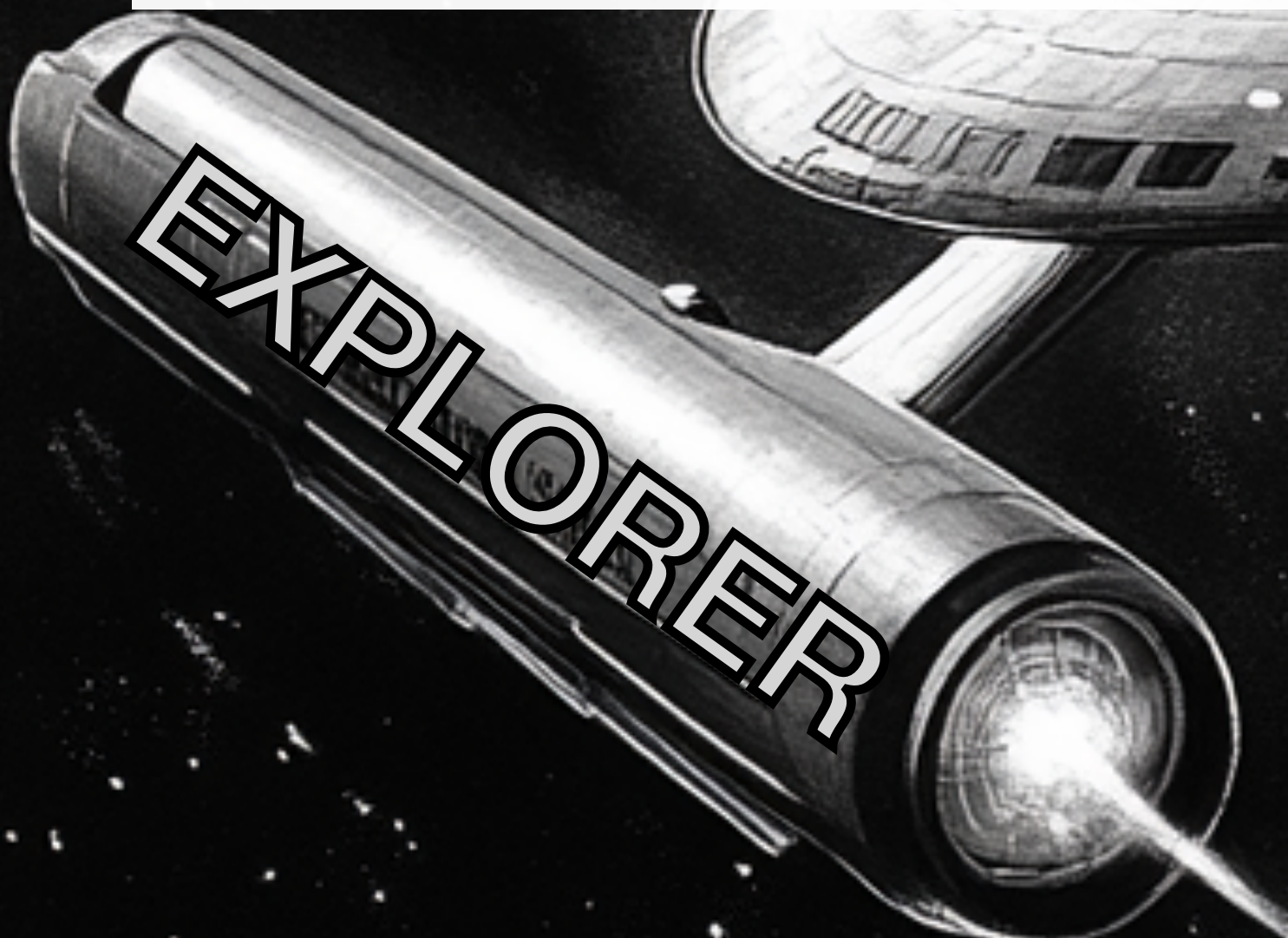
2

Deepspeed, PyTorch, etc.

3

 NCCL (Nvidia Collective Communications Library) 

Inter-GPU communication



DEEP SPACE

1

DeepSpaceExplorer LLM Model

2

Deepspeed, PyTorch, etc.

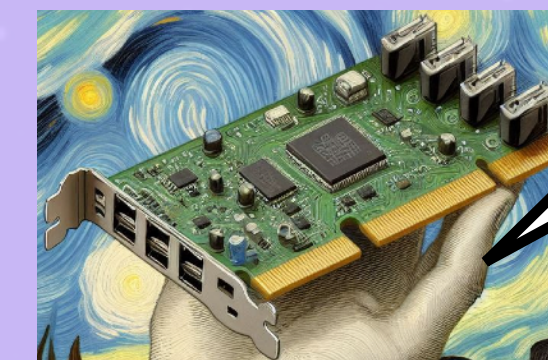
3

 NCCL (Nvidia Collective Communications Library) 

RDMA
capable!

4

Network Device (Infiniband or RoCE)



Remote Direct Memory Access (RDMA)

Fast data transfer

Memory

Memory

 OS

 OS



Remote Direct Memory Access (RDMA)

Fast data transfer

Memory

 OS

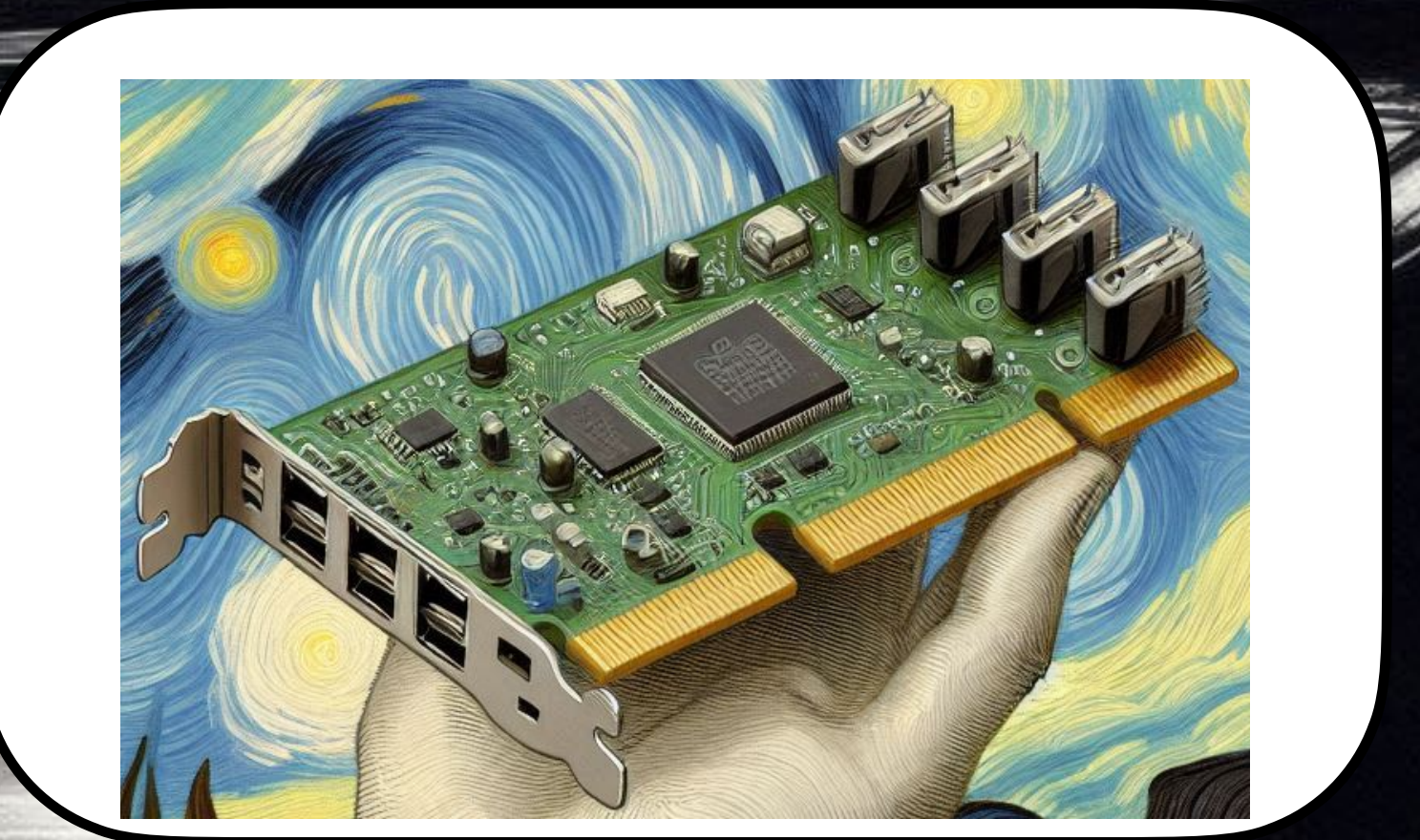


Fabric

Kernel Bypass

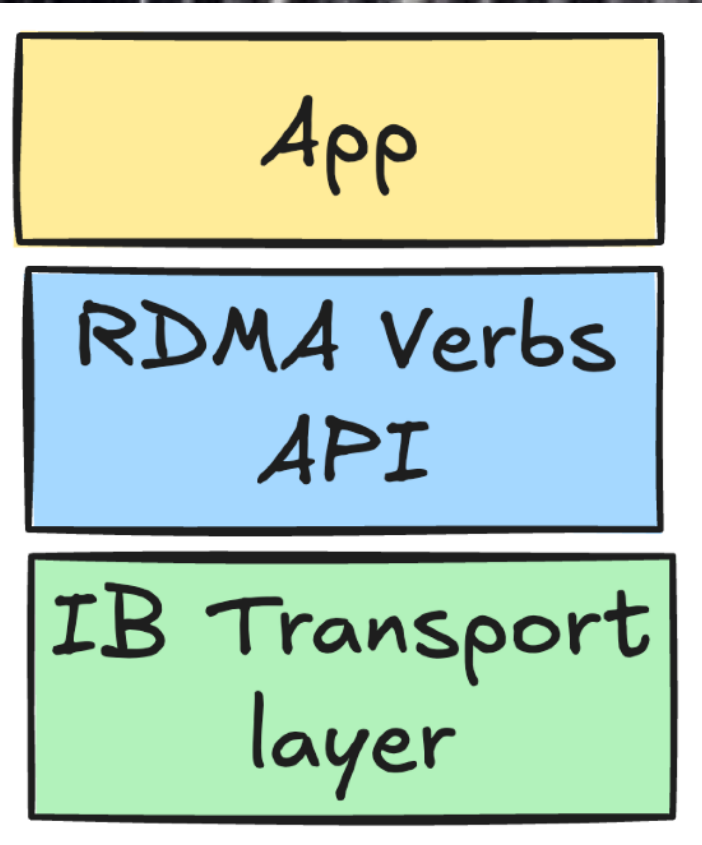
Memory

 OS



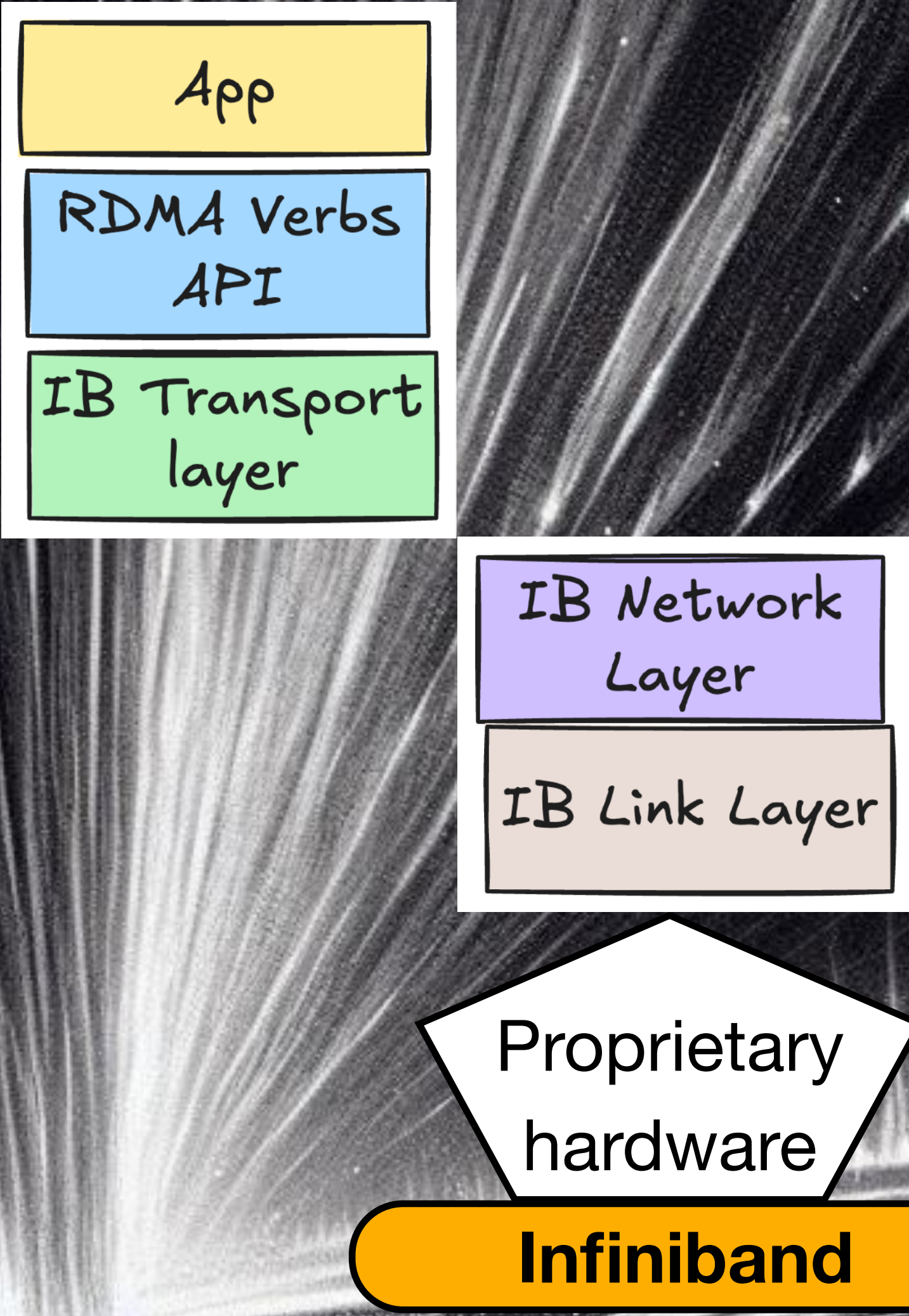
Fabric

High throughput,
low latency!



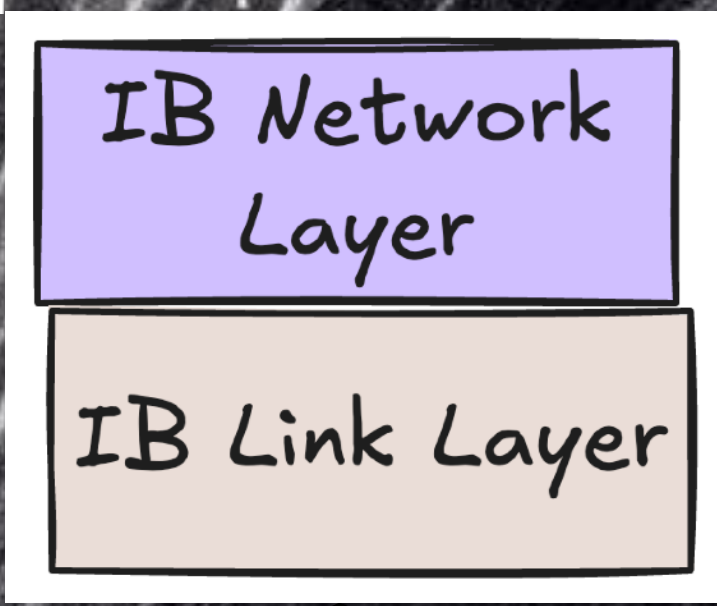
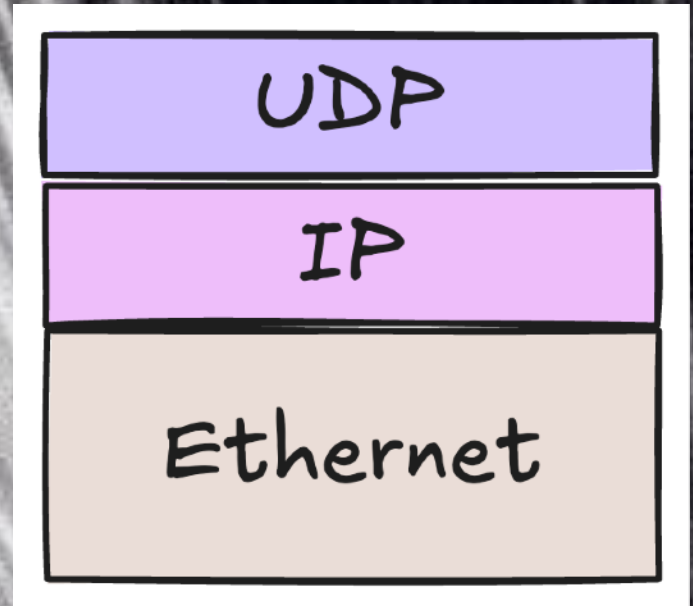
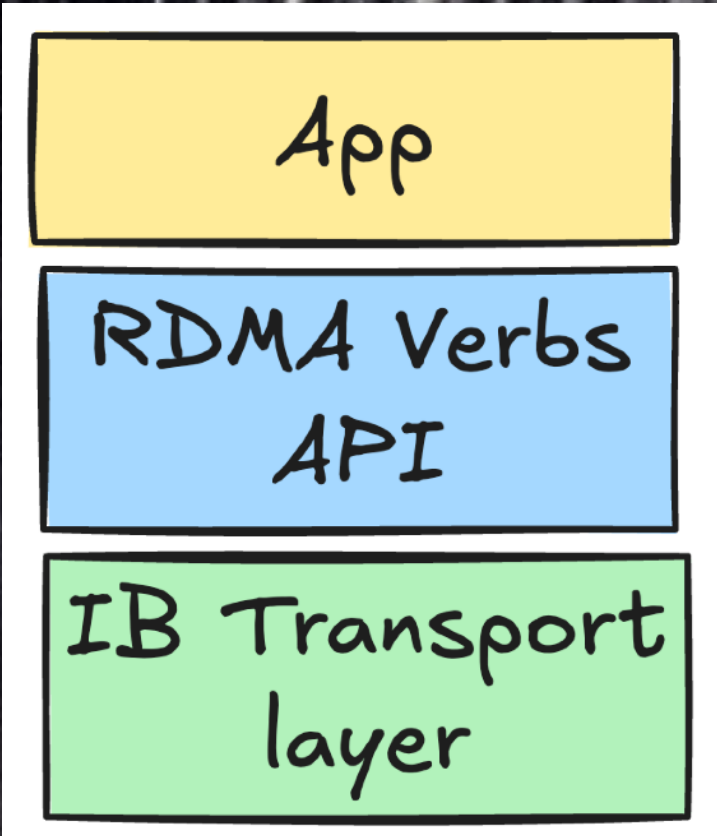
Fabric

High throughput,
low latency!



Fabric

High throughput,
low latency!



Standard hardware

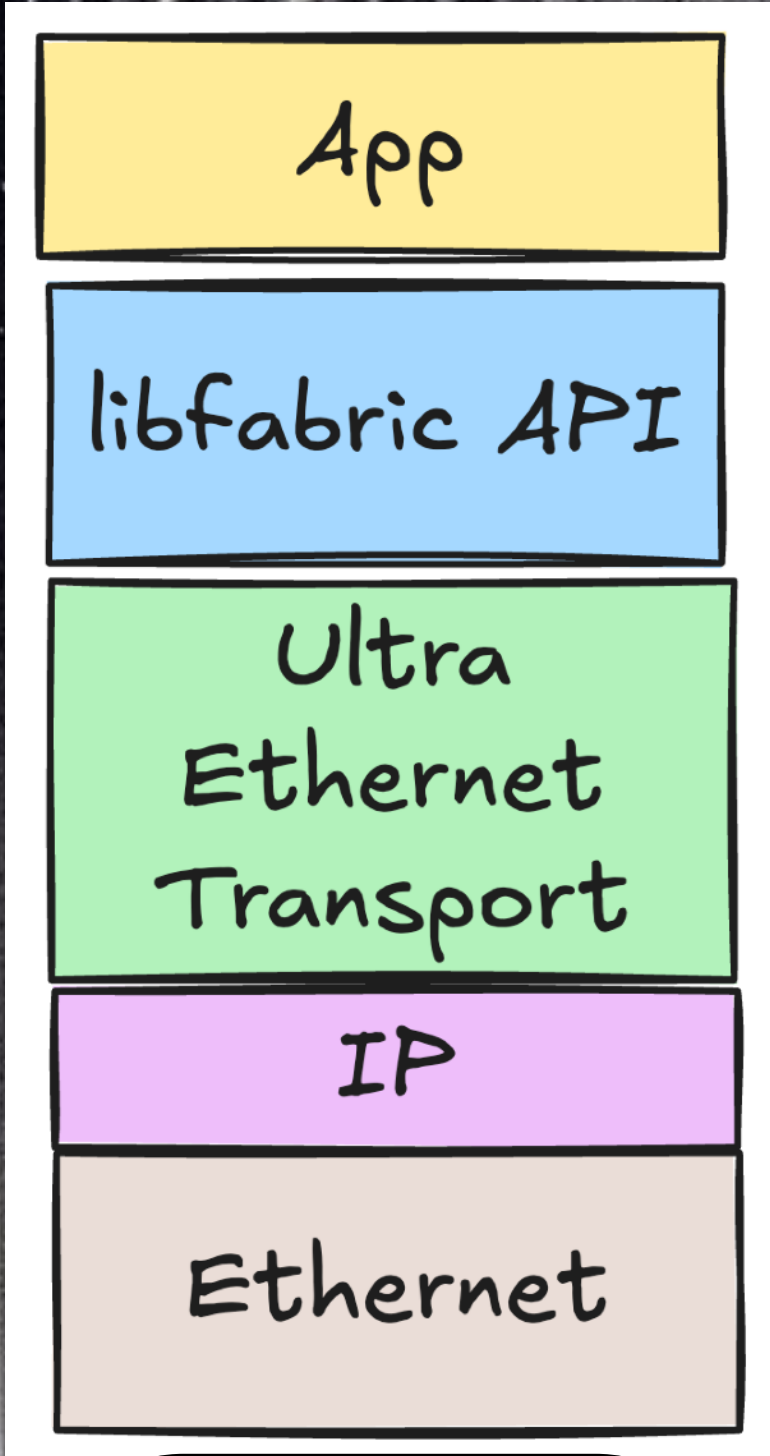
Proprietary hardware

RoCEv2

Infiniband

Fabric

High throughput,
low latency!

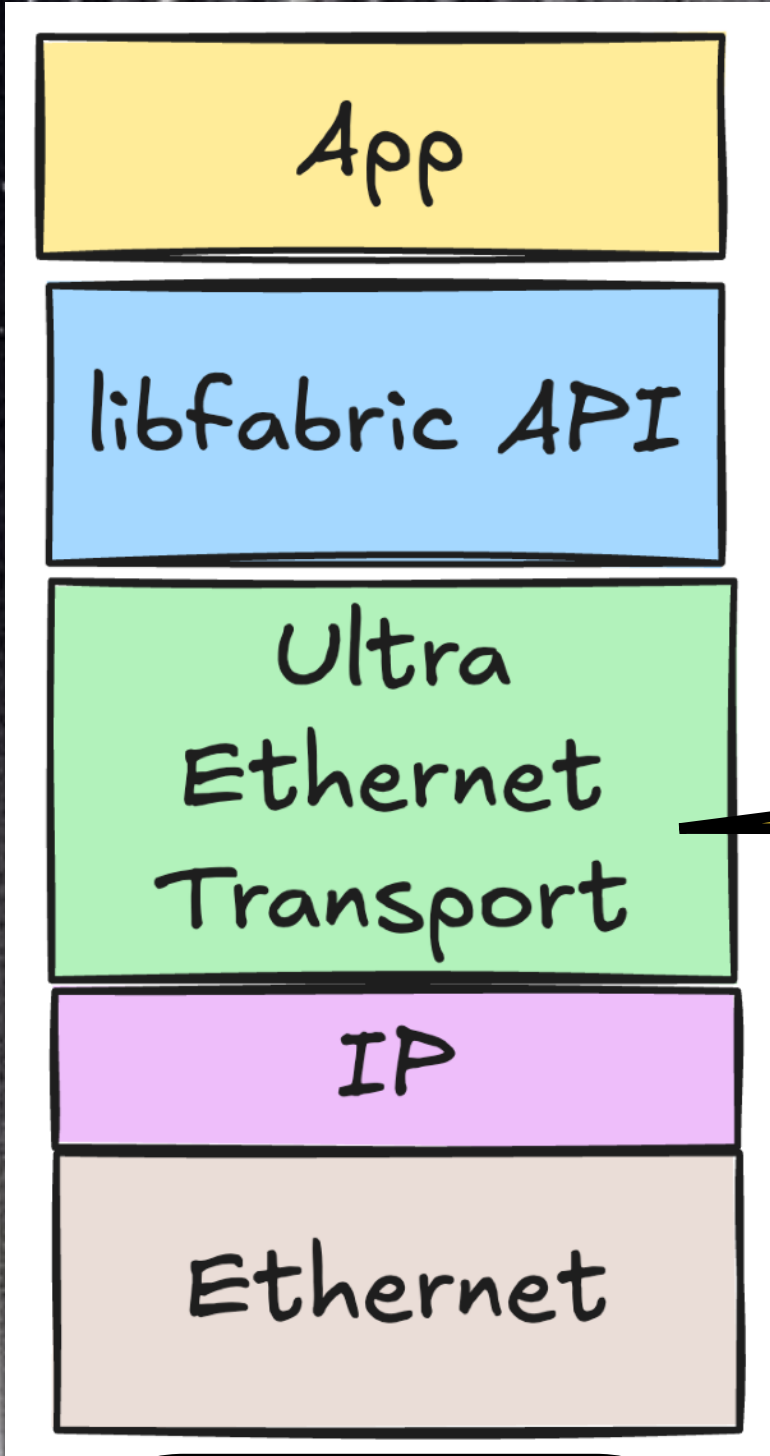


Standard hardware

UltraEthernet

Fabric

High throughput,
low latency!



Improved transport!
ultraethernet.org

Standard hardware

UltraEthernet

FRONTEND NETWORK

Storage

Databanks

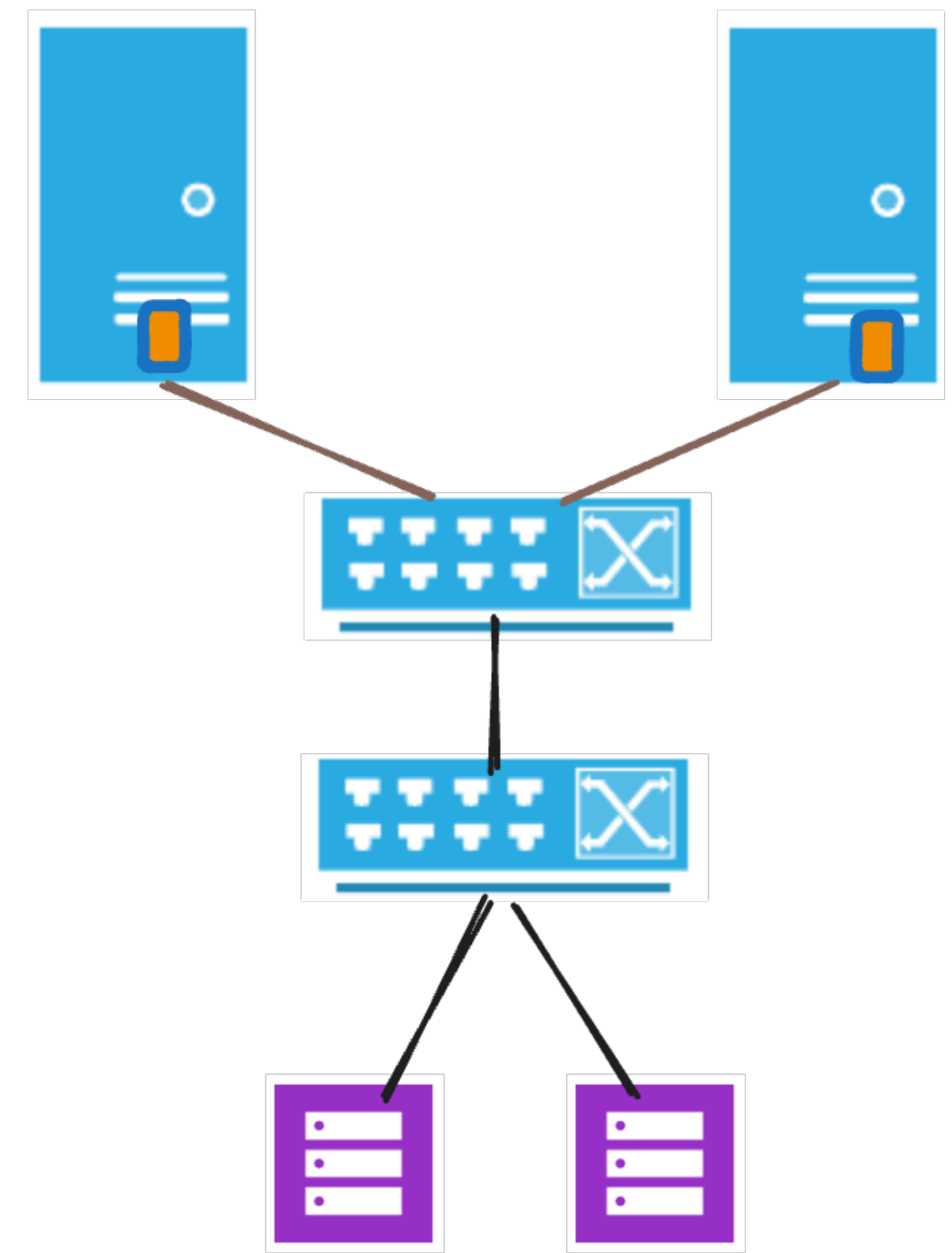
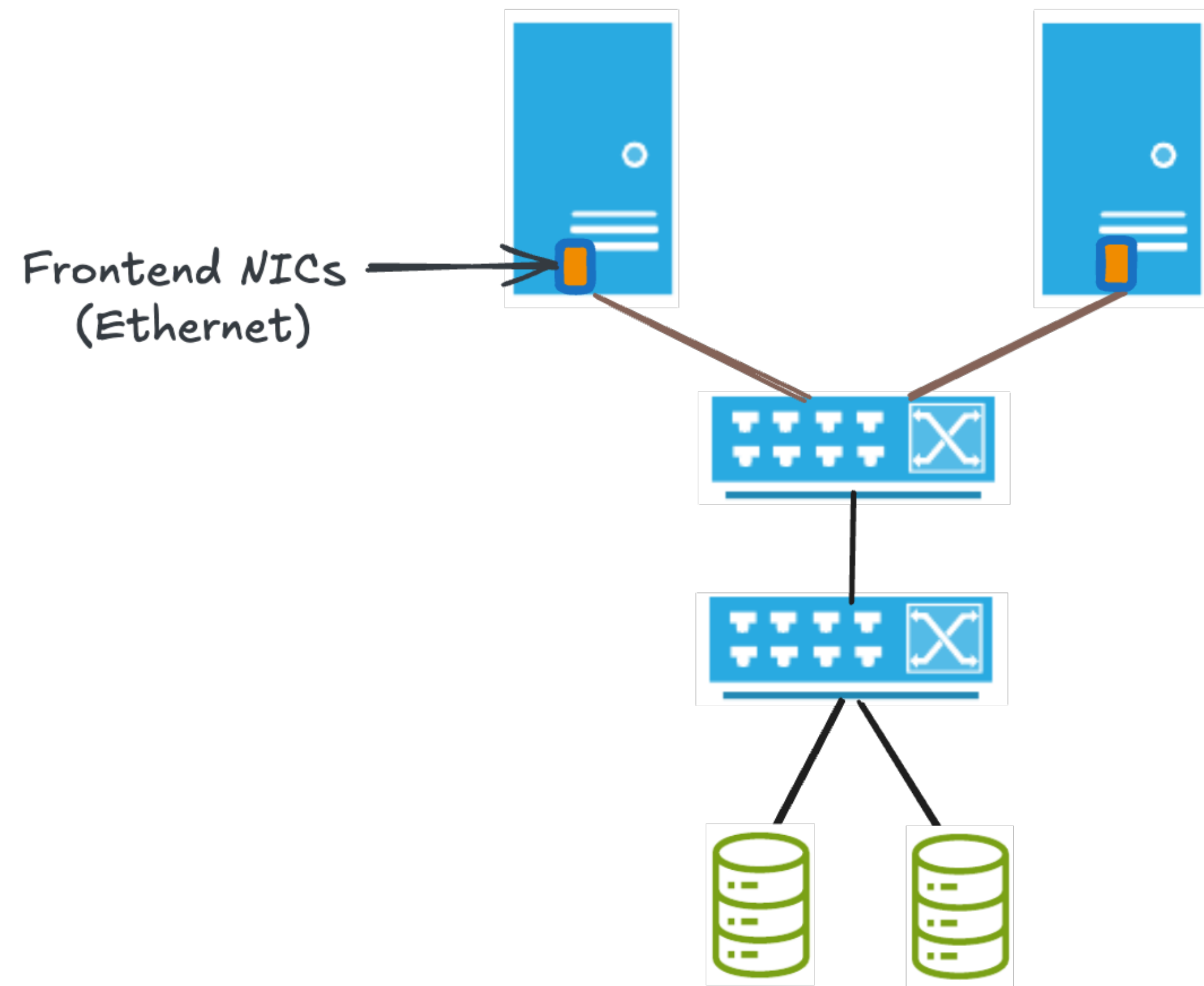
Bridge

BACKEND NETWORK

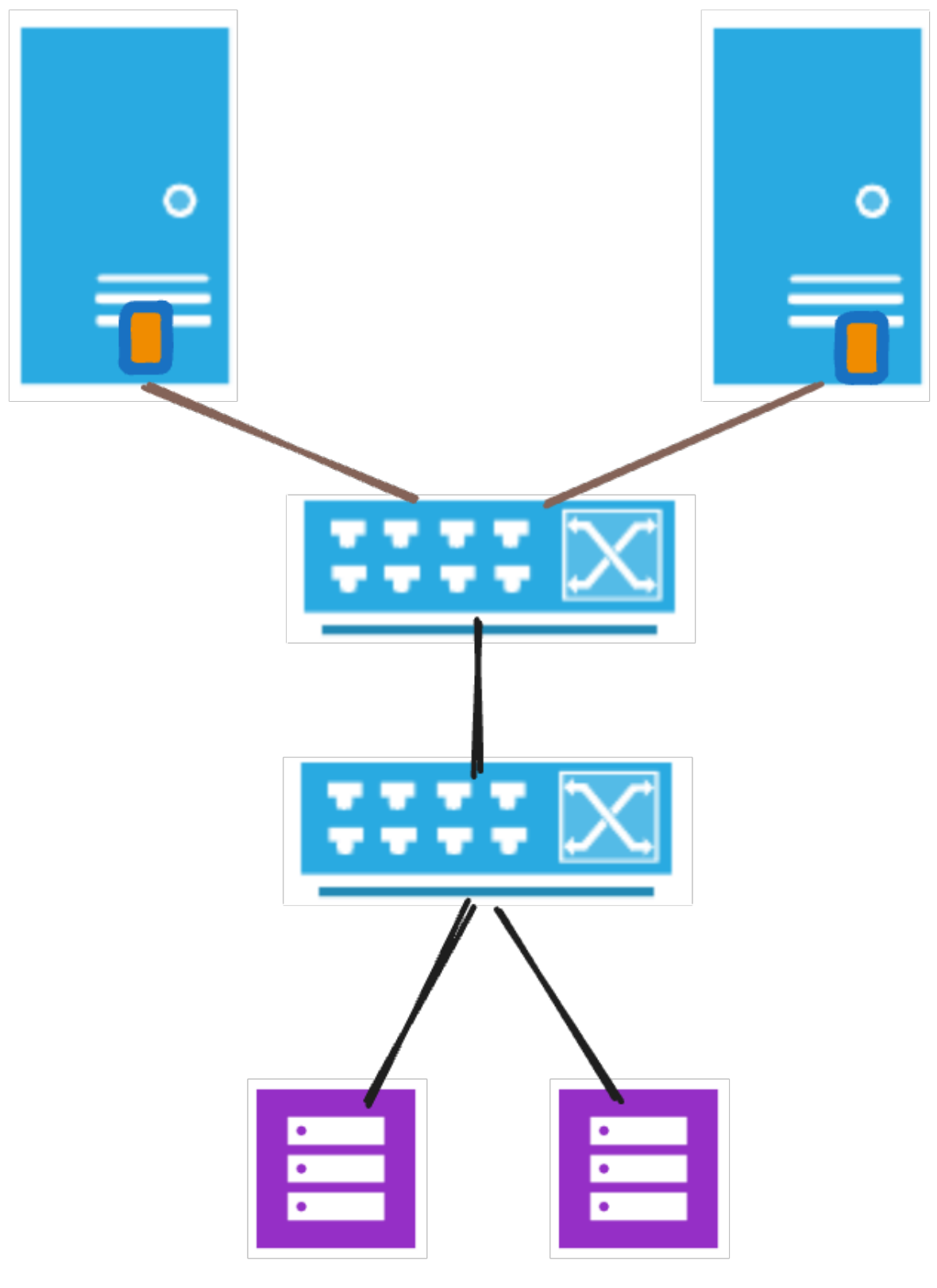
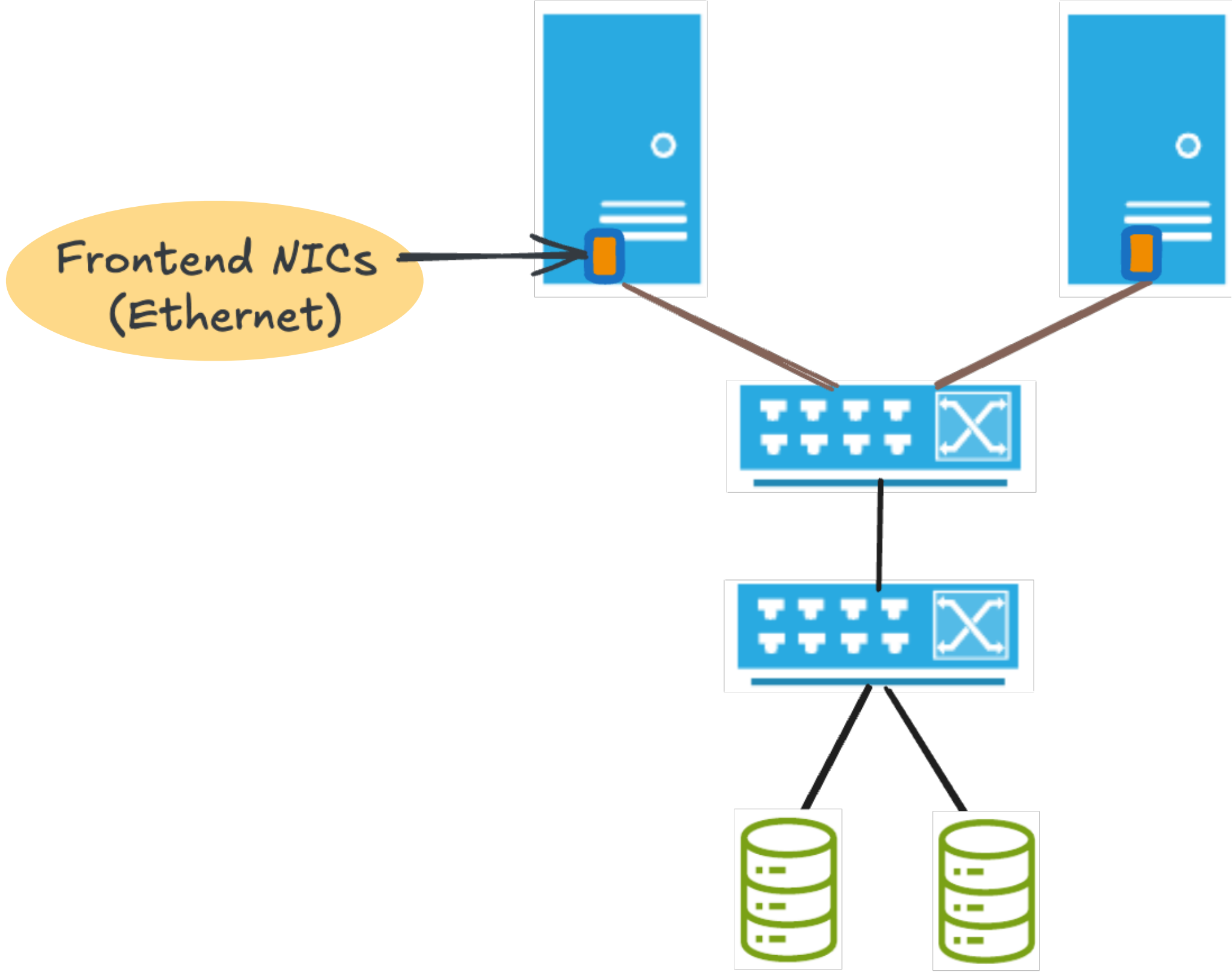
Distributed compute

Engine room

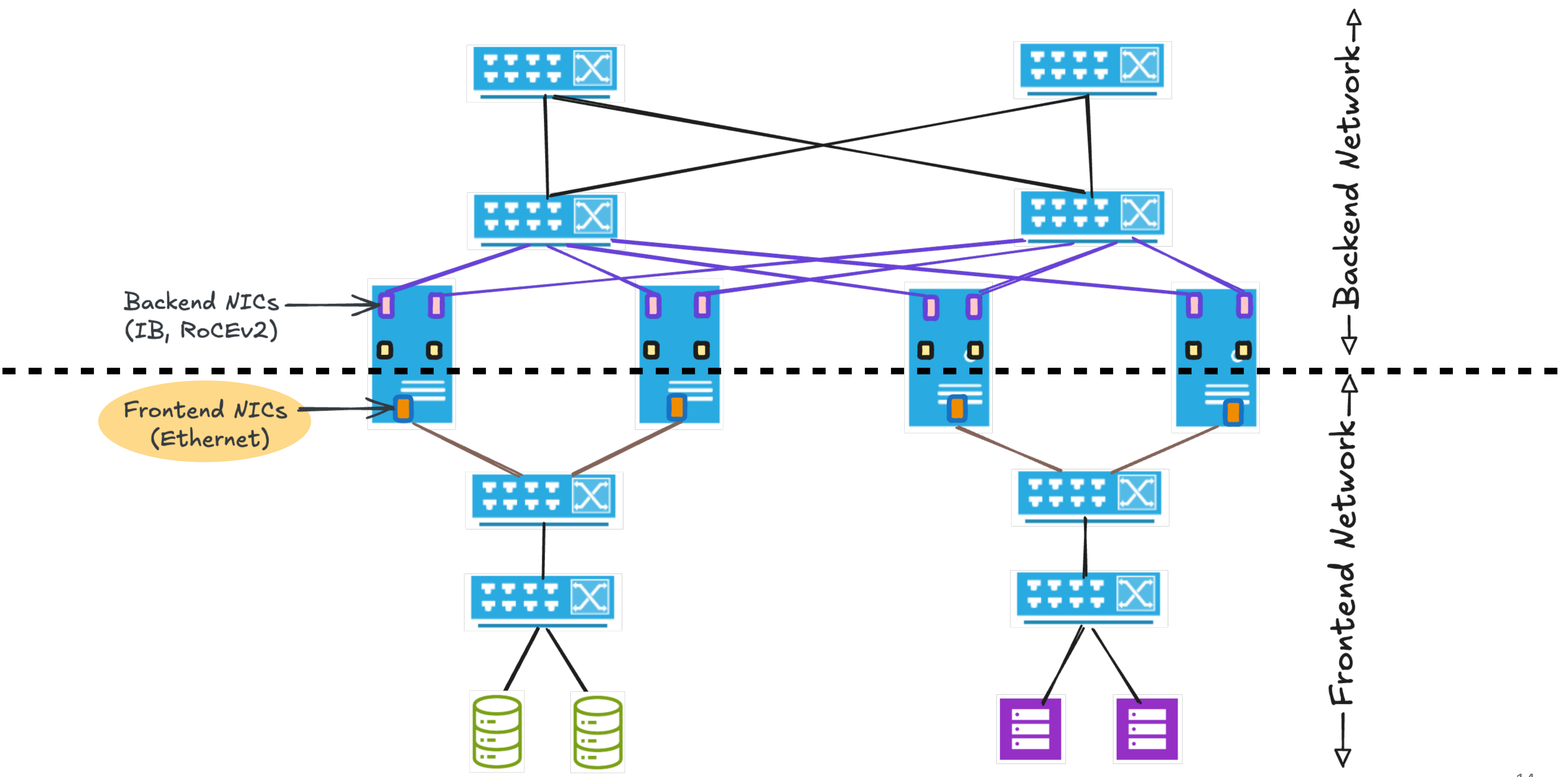
AI-generated

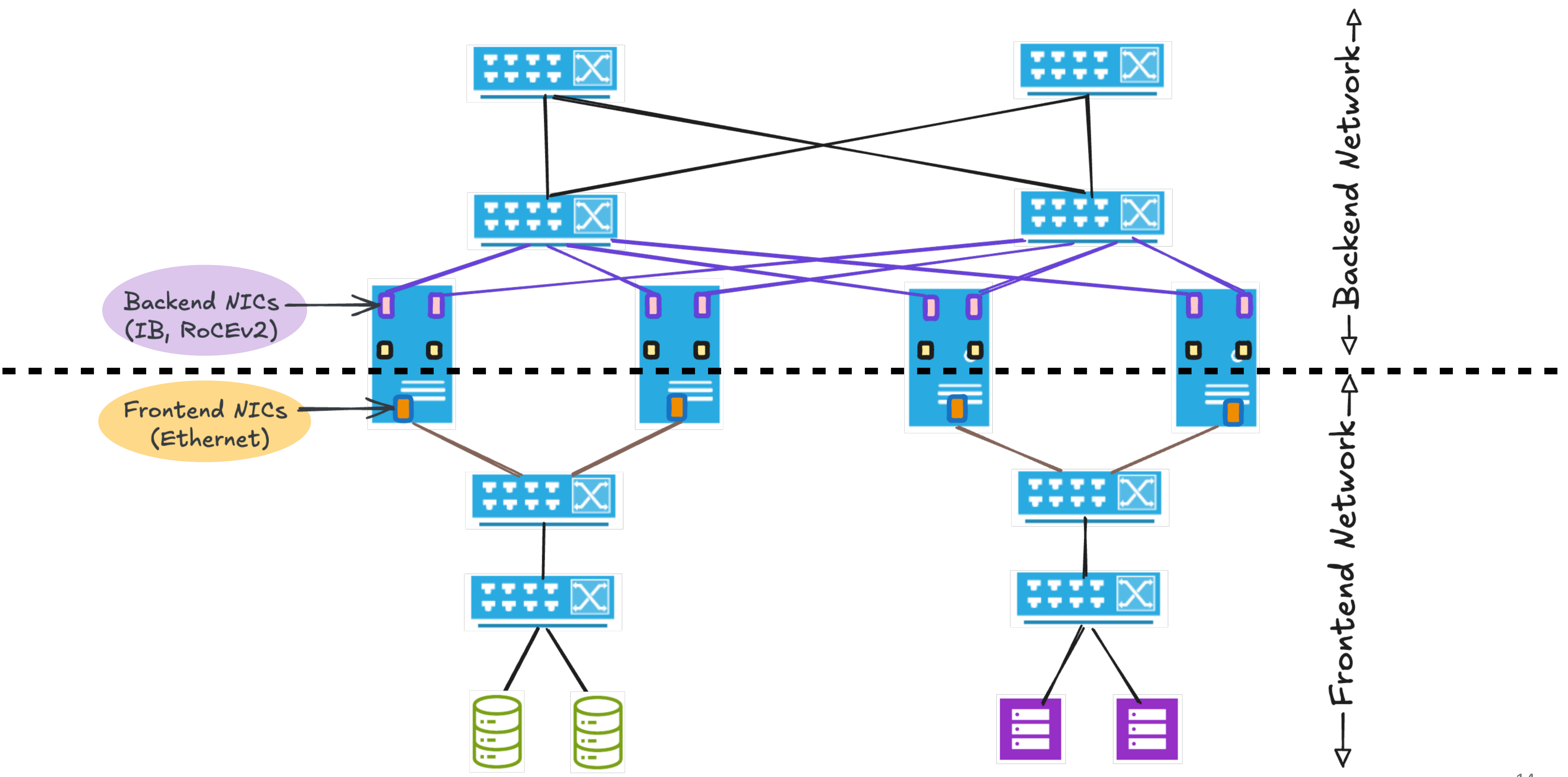


↕ Frontend Network ↕

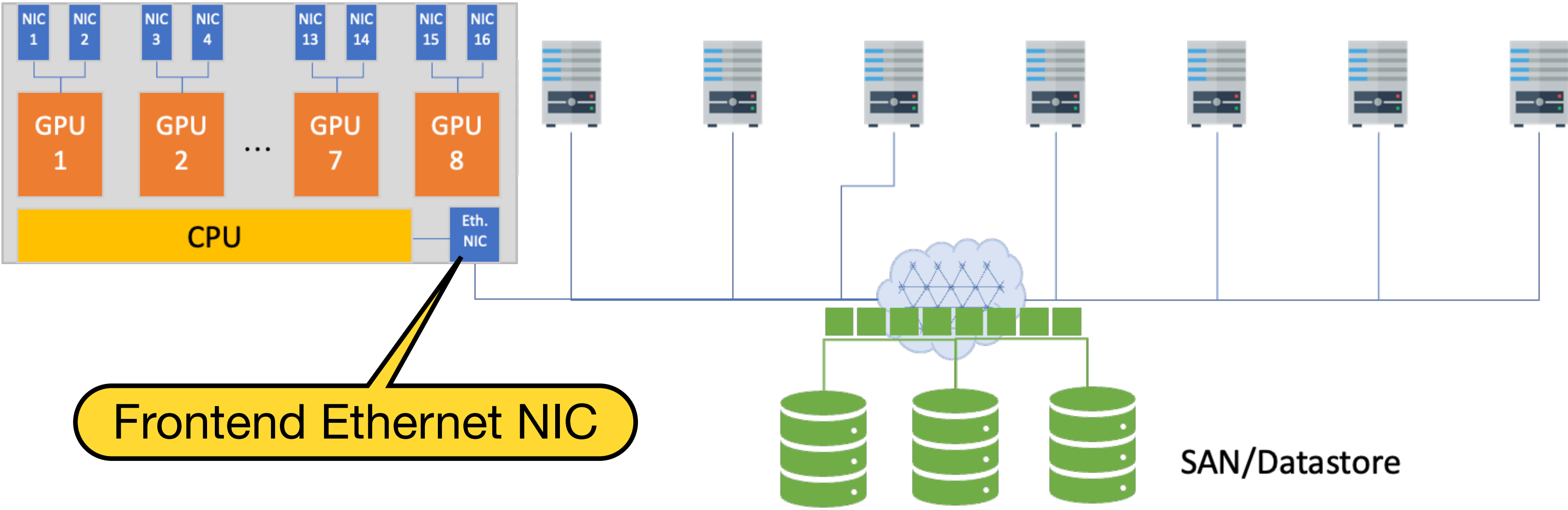


↕ Frontend Network ↗





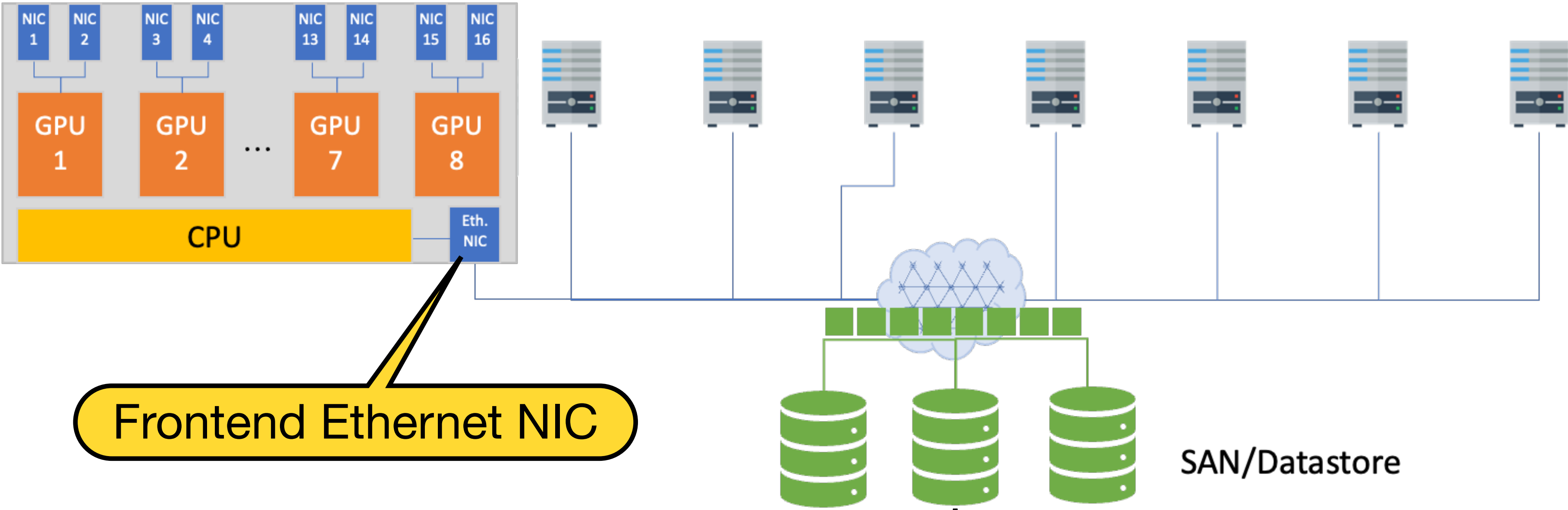
Frontend network



Frontend Ethernet NIC

SAN/Datastore

Frontend network

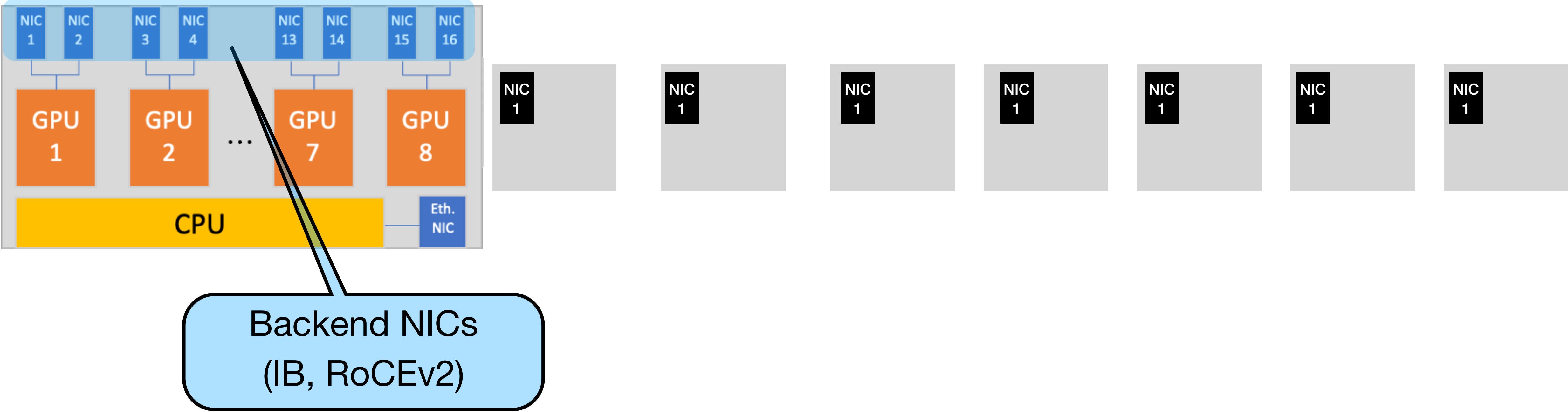


Frontend Ethernet NIC

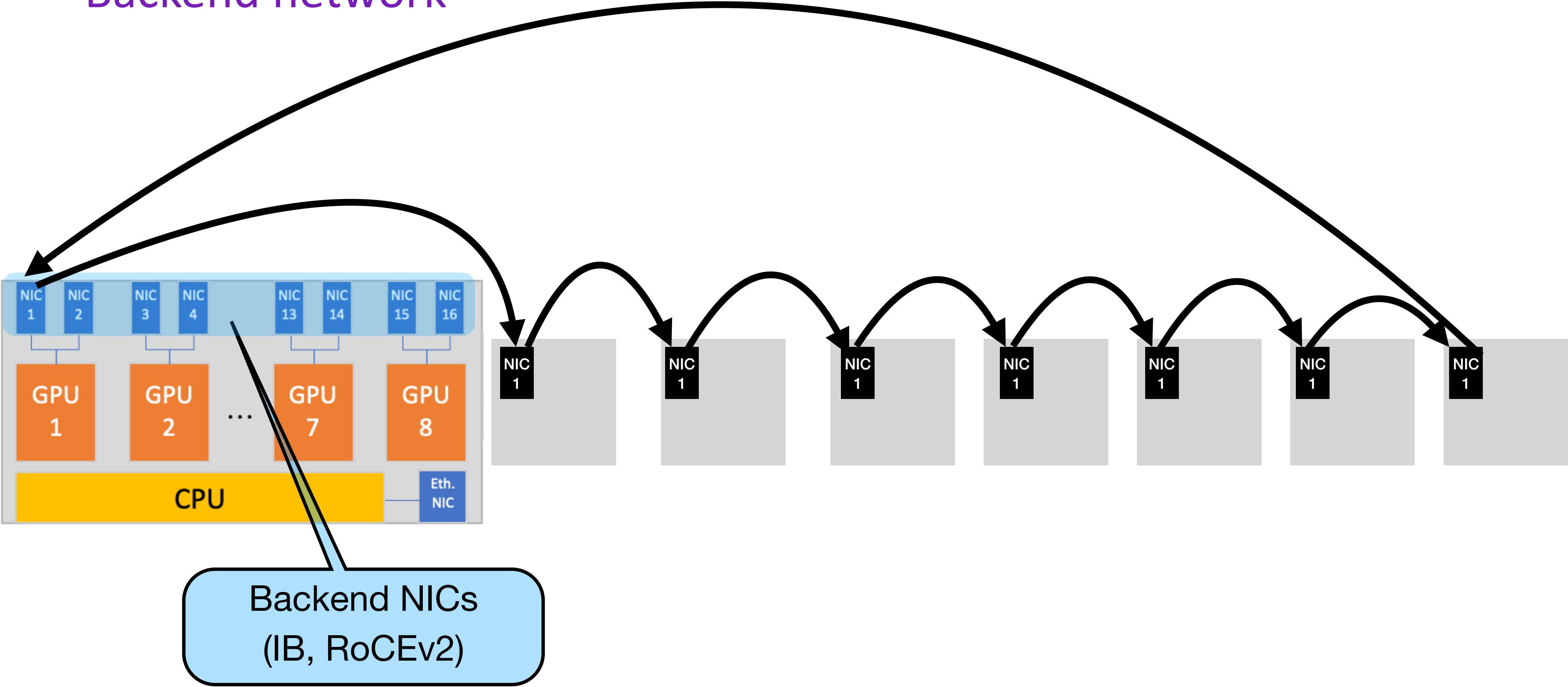
Training data ingestion, logging, checkpointing

SAN/Datastore

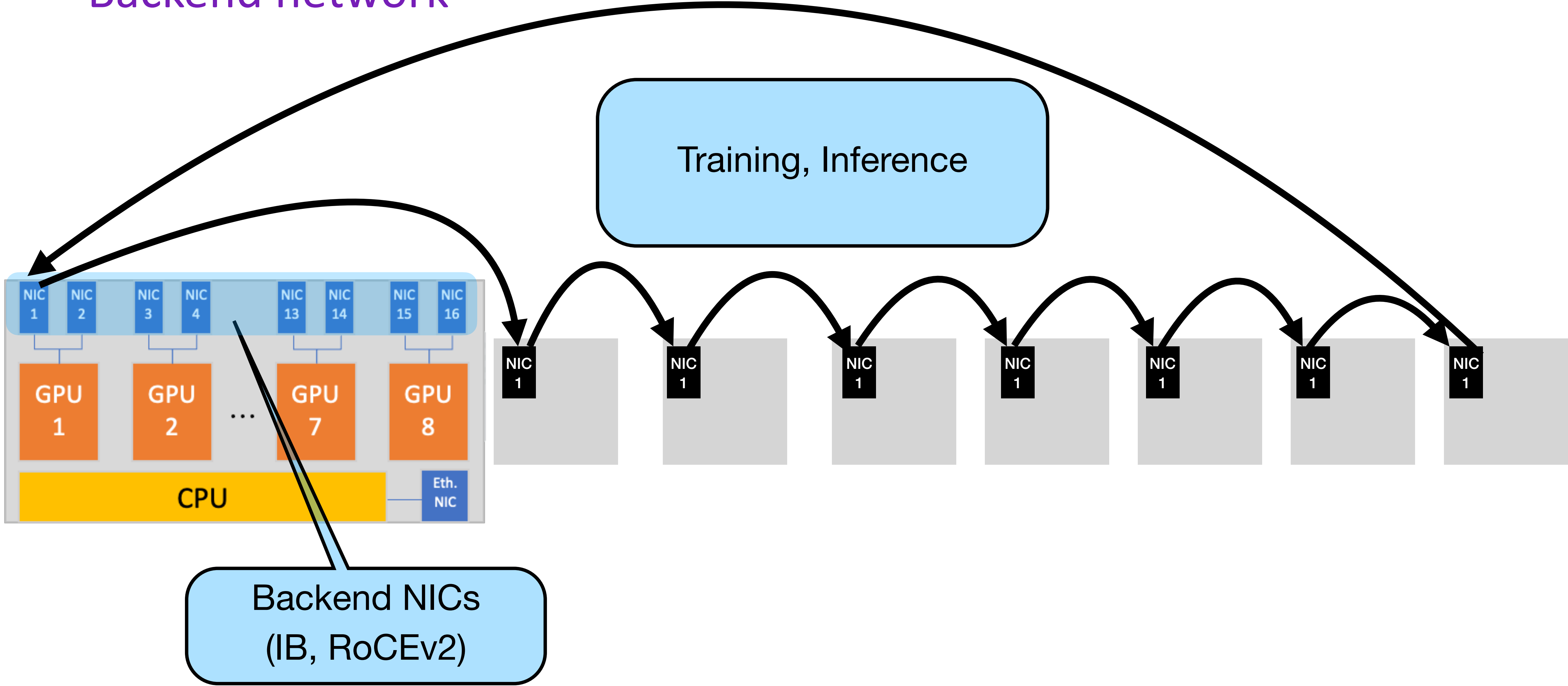
Backend network



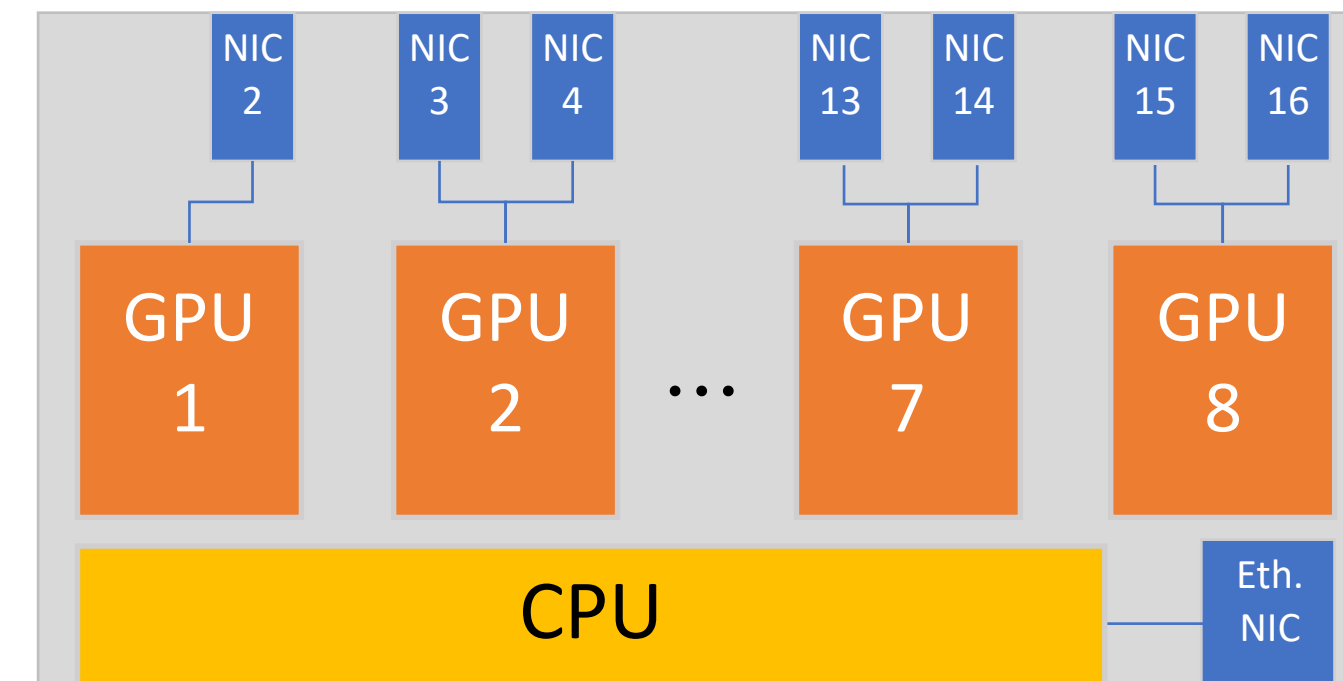
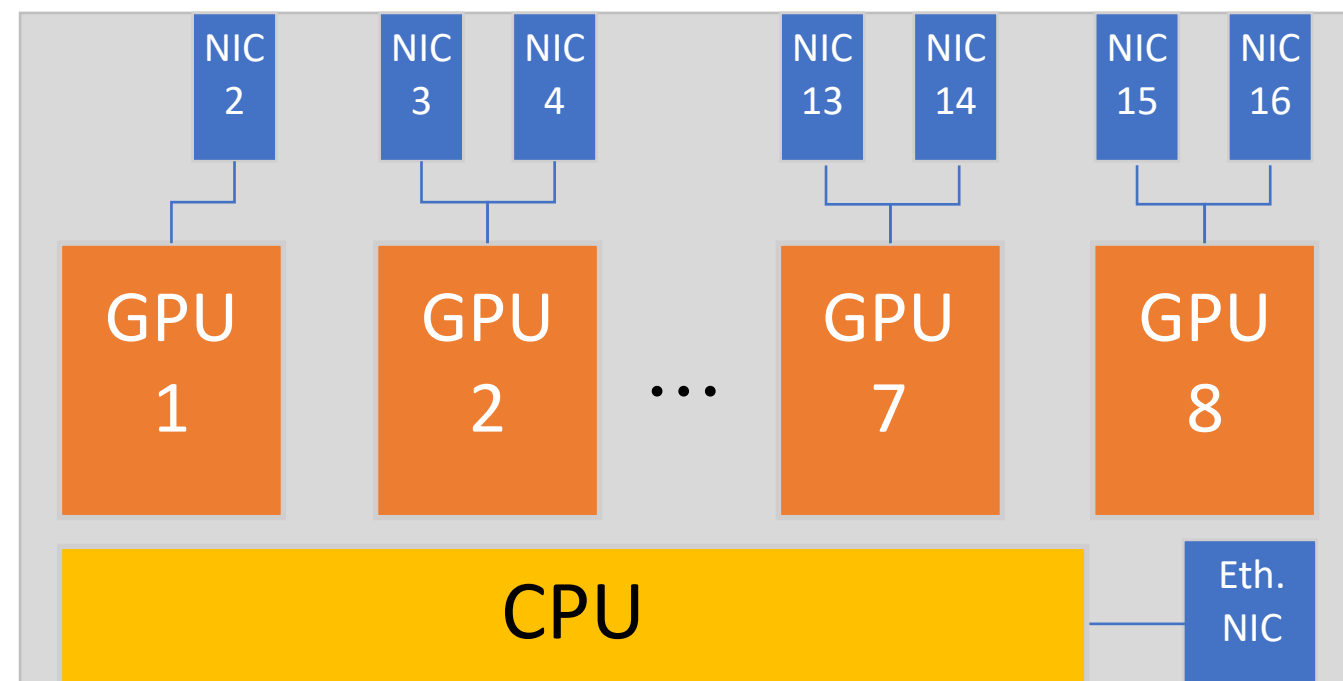
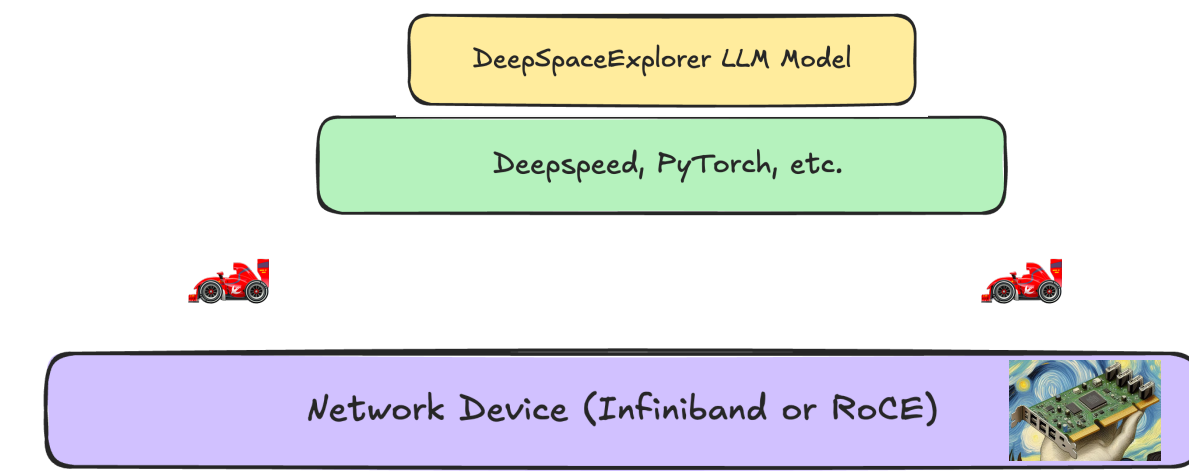
Backend network



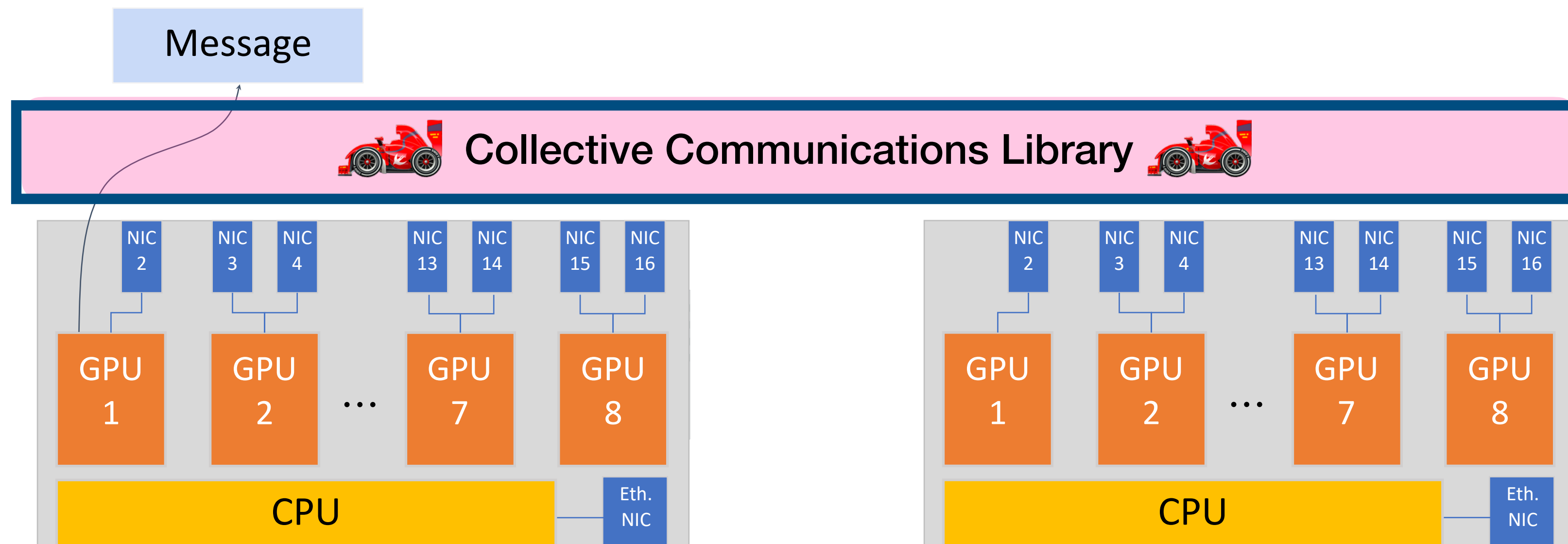
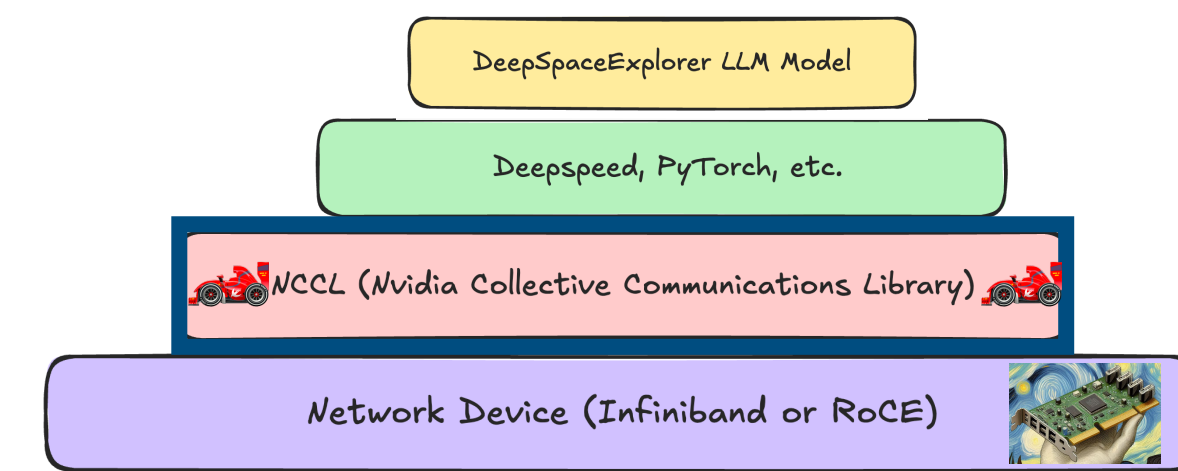
Backend network



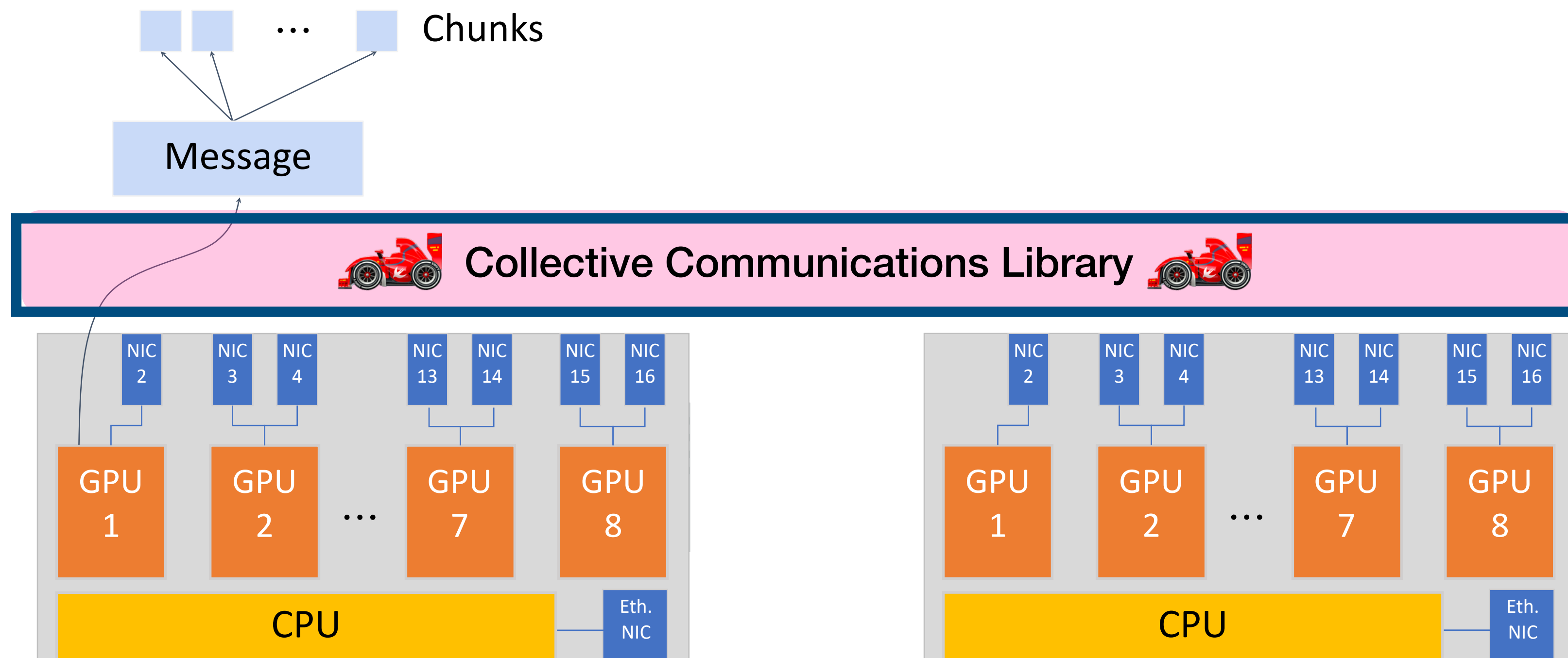
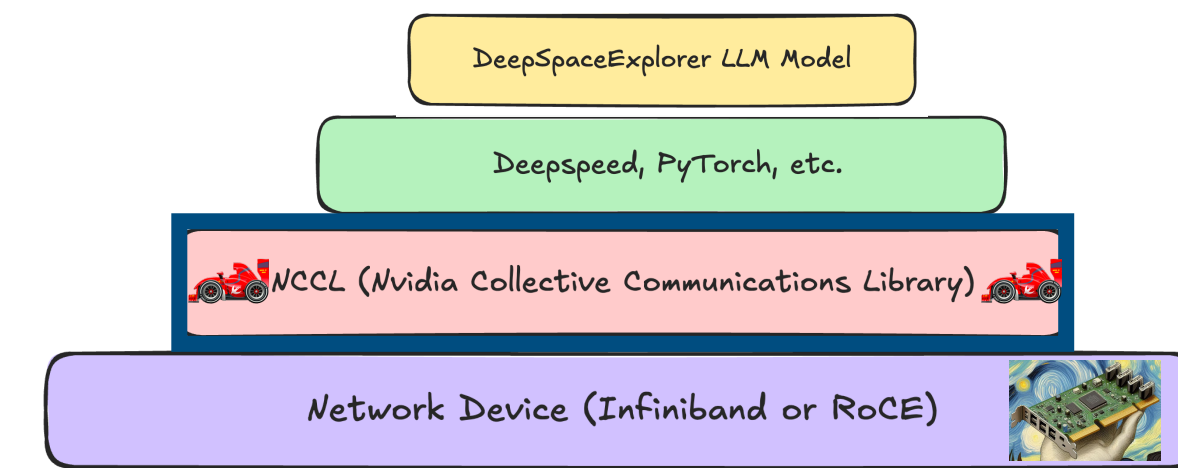
Collective communications



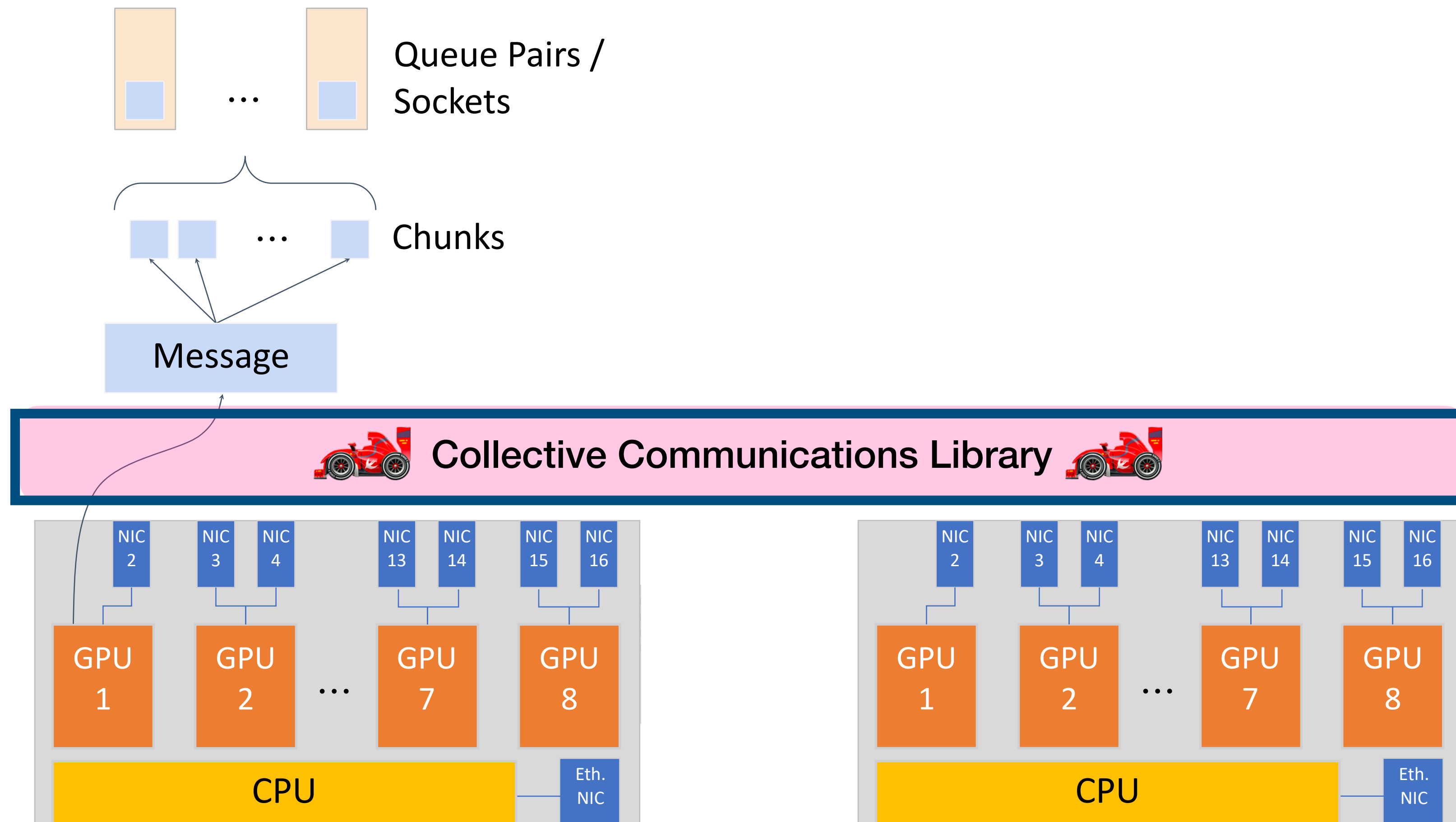
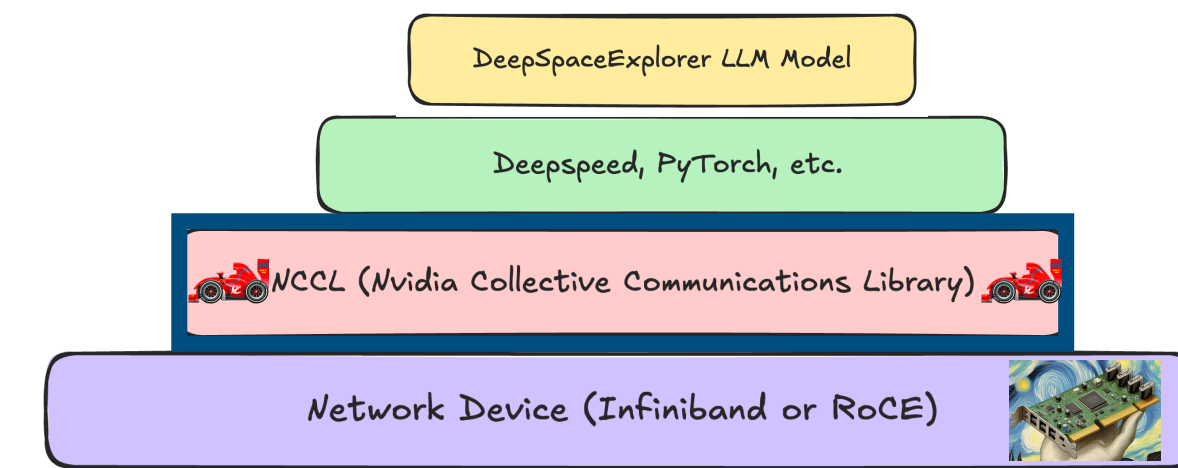
Collective communications



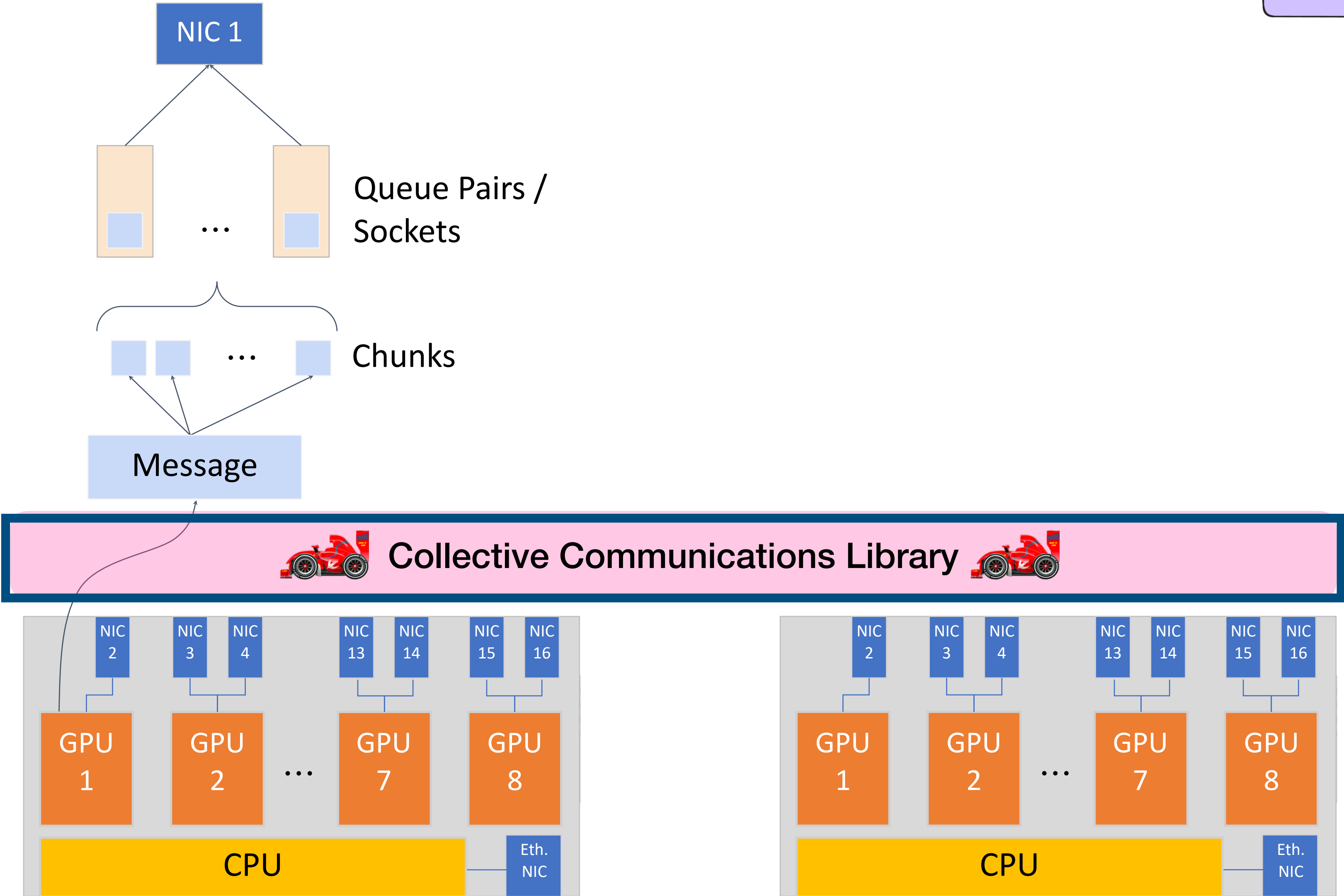
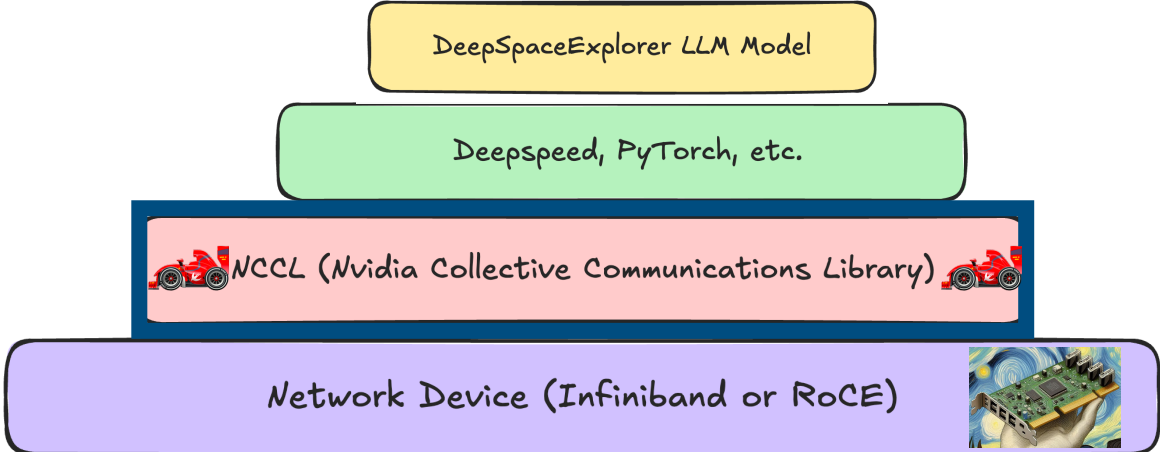
Collective communications



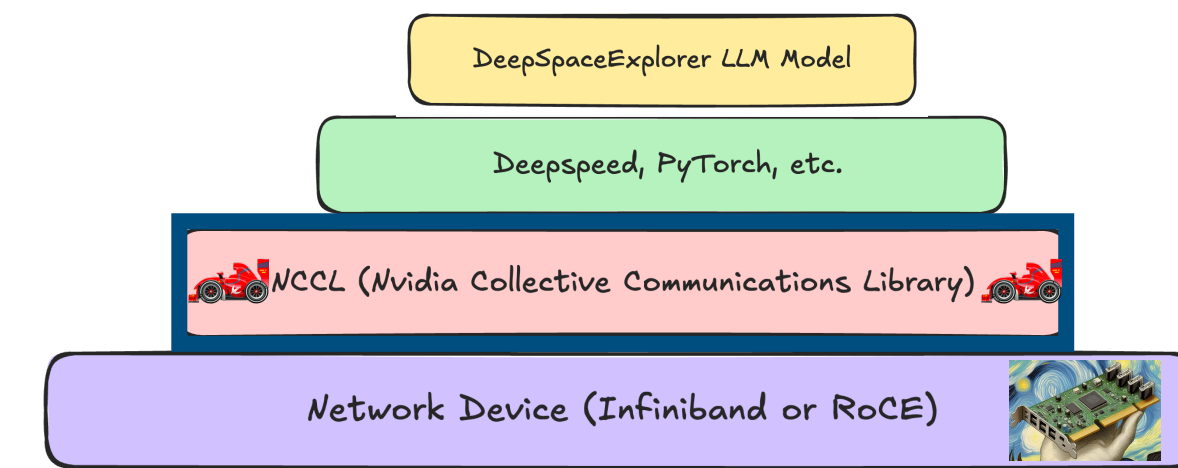
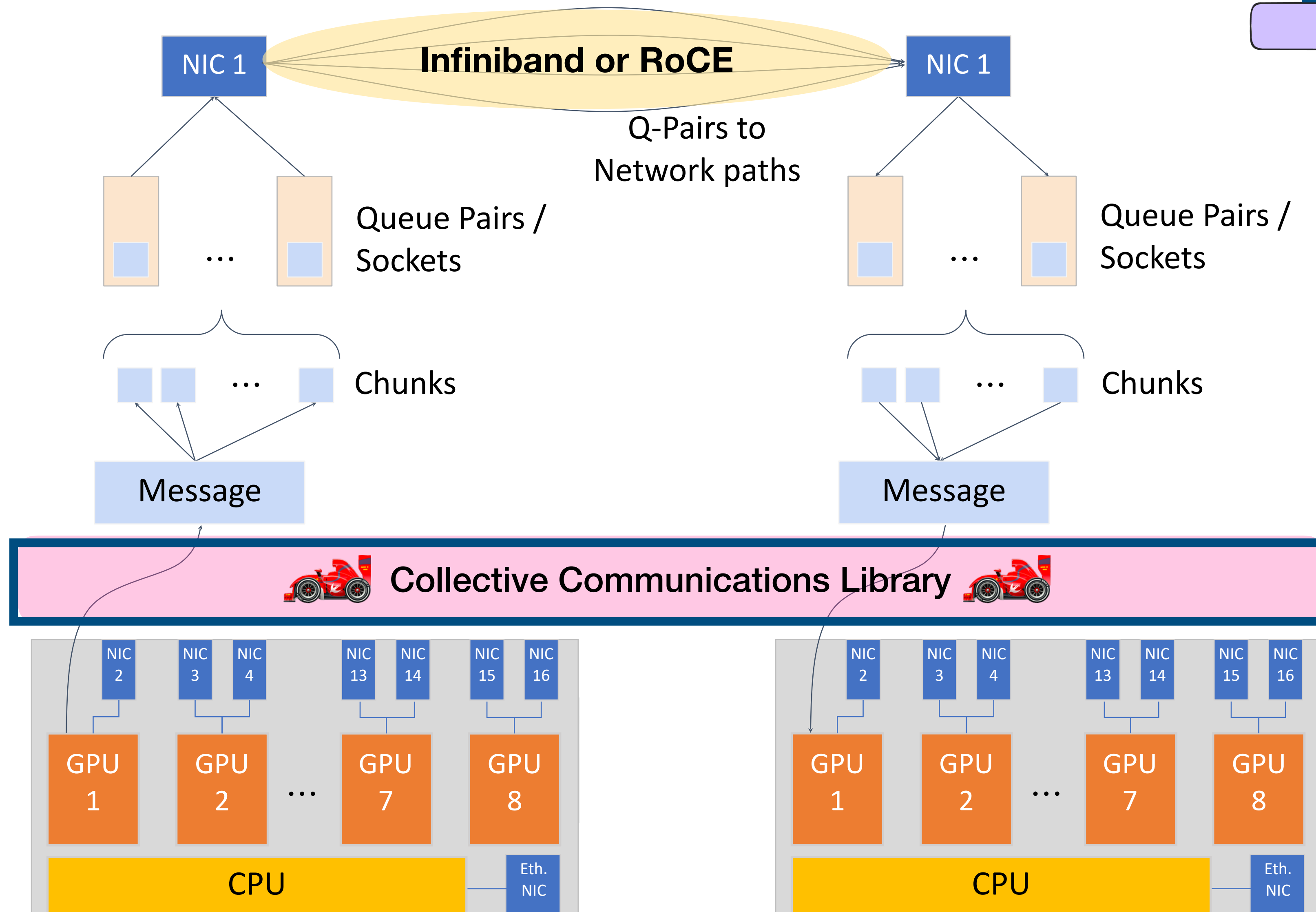
Collective communications



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



AI workload layers

Demands from AI networks

Challenges in AI networks

Key Takeaways

Time Force,

Department of Temporal Affairs

Tempus Mundi Servamus





Traditional networks



AI networks



Traditional networks



AI networks



Traditional networks

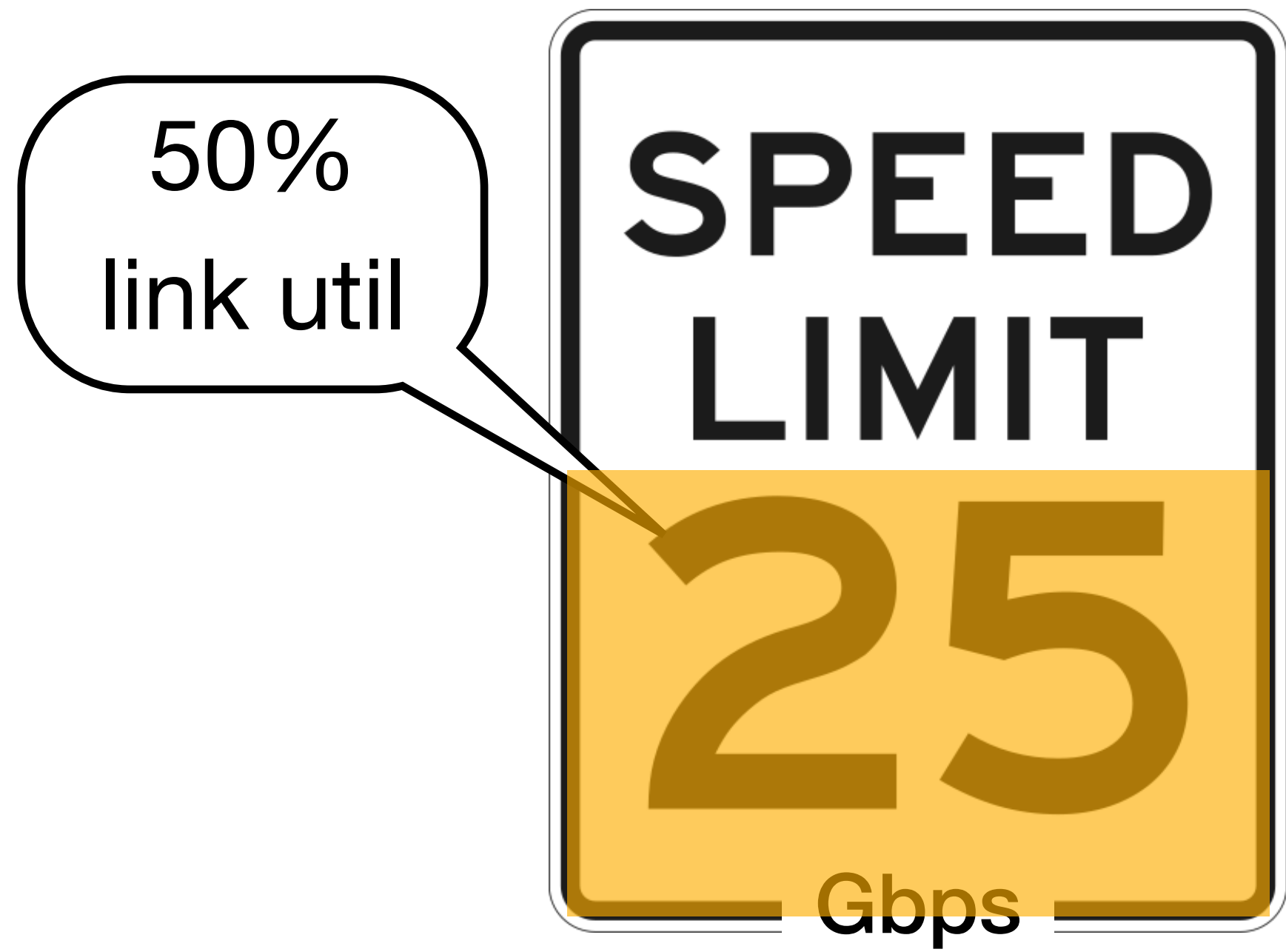


AI networks



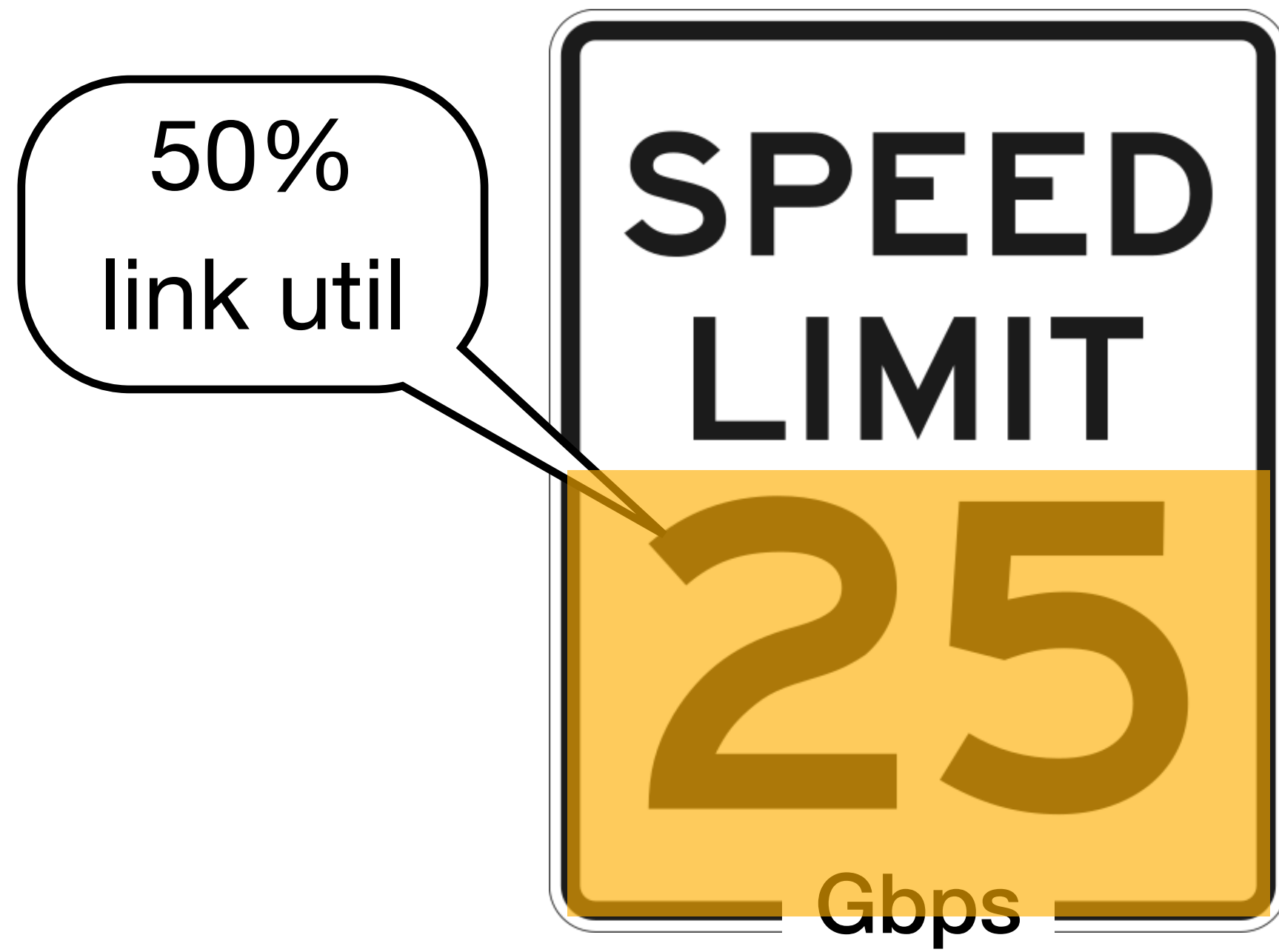
Traditional networks

AI networks



Traditional networks

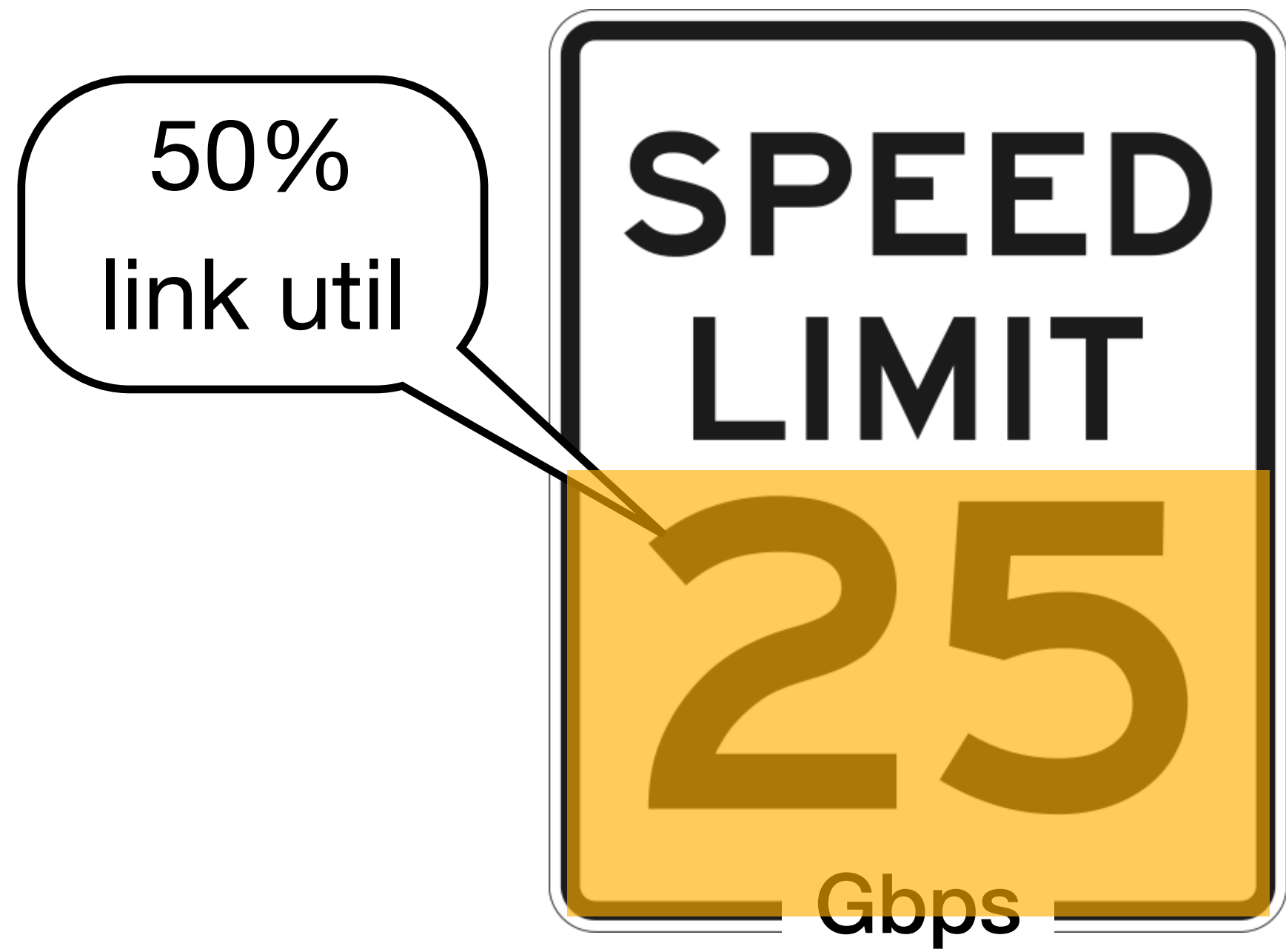
AI networks



Traditional networks



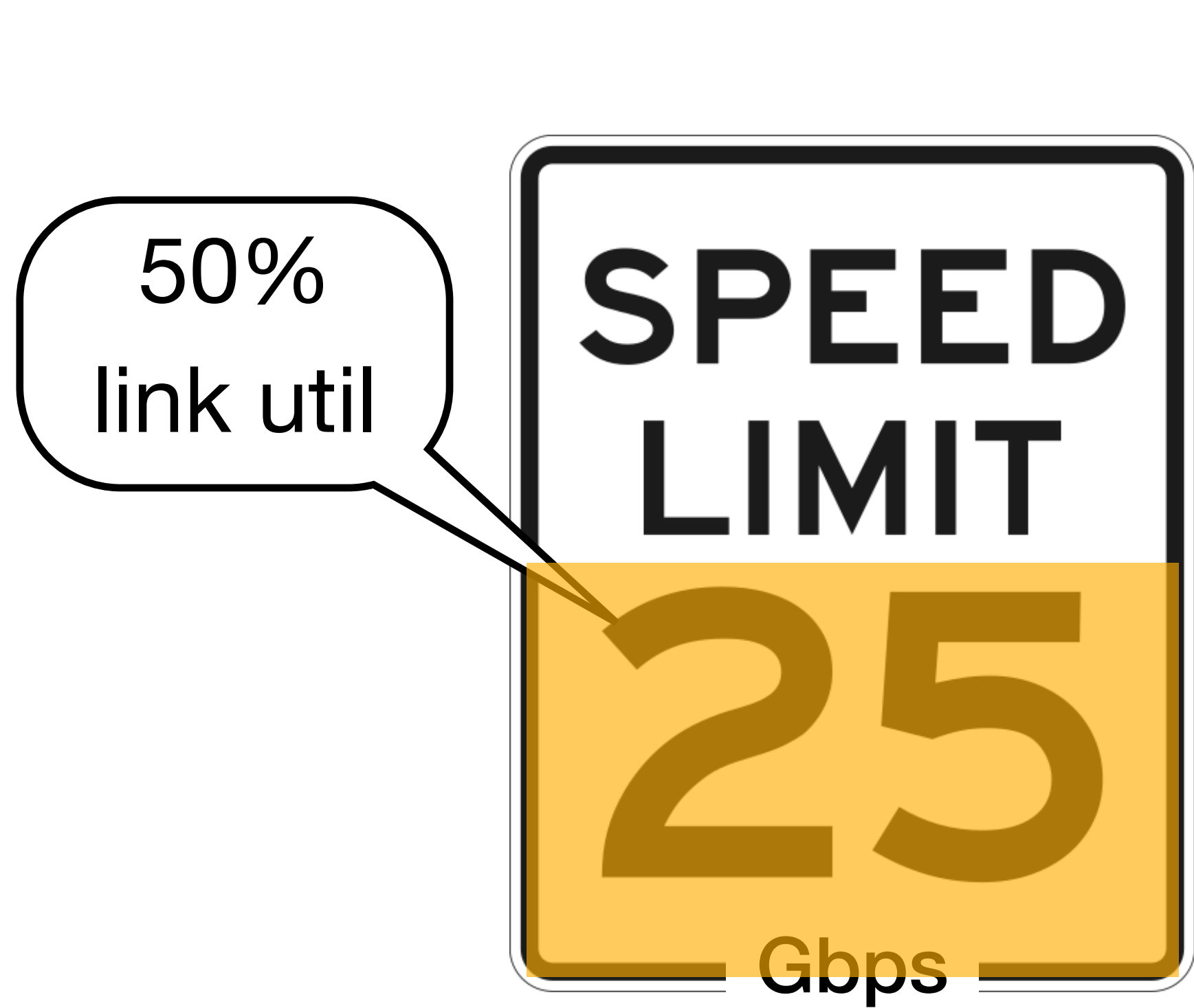
AI networks



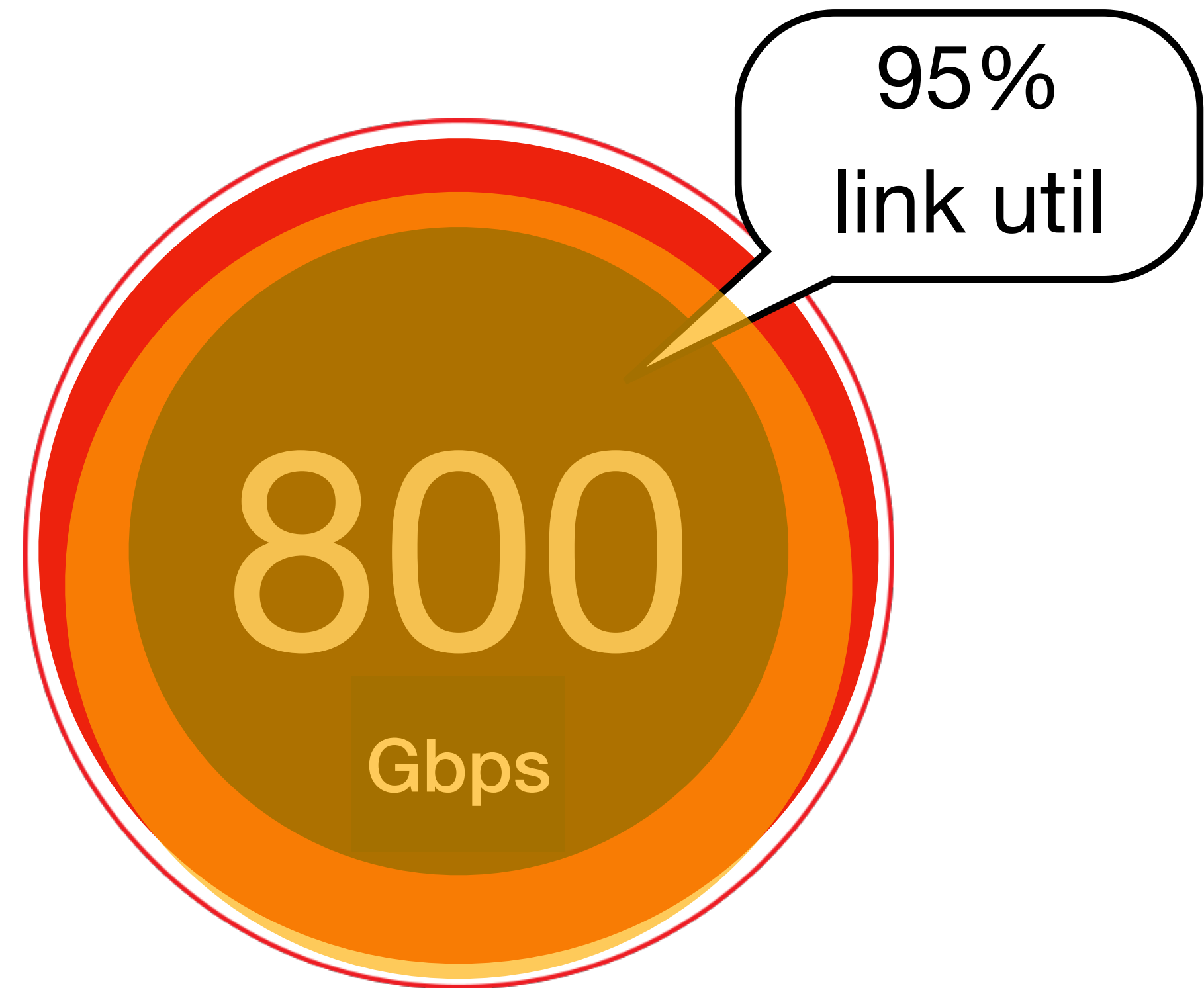
Traditional networks



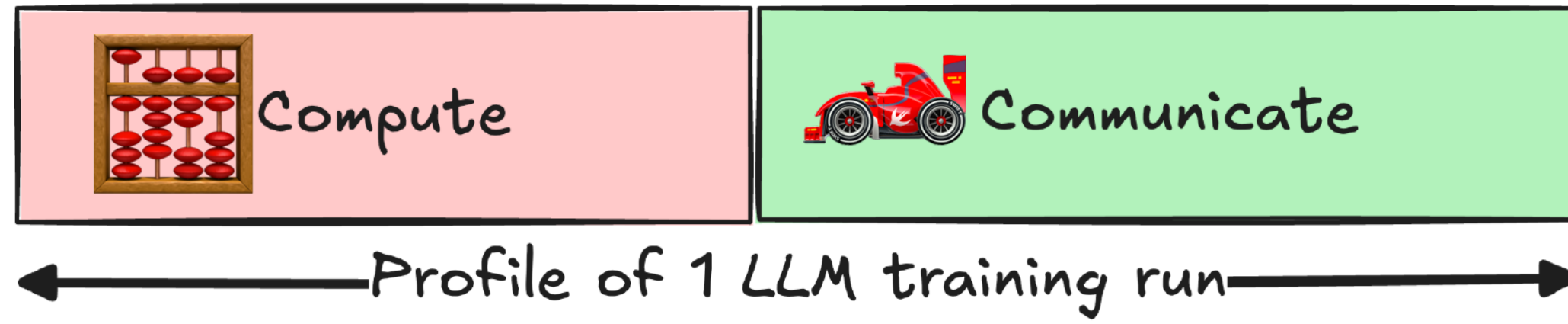
AI networks



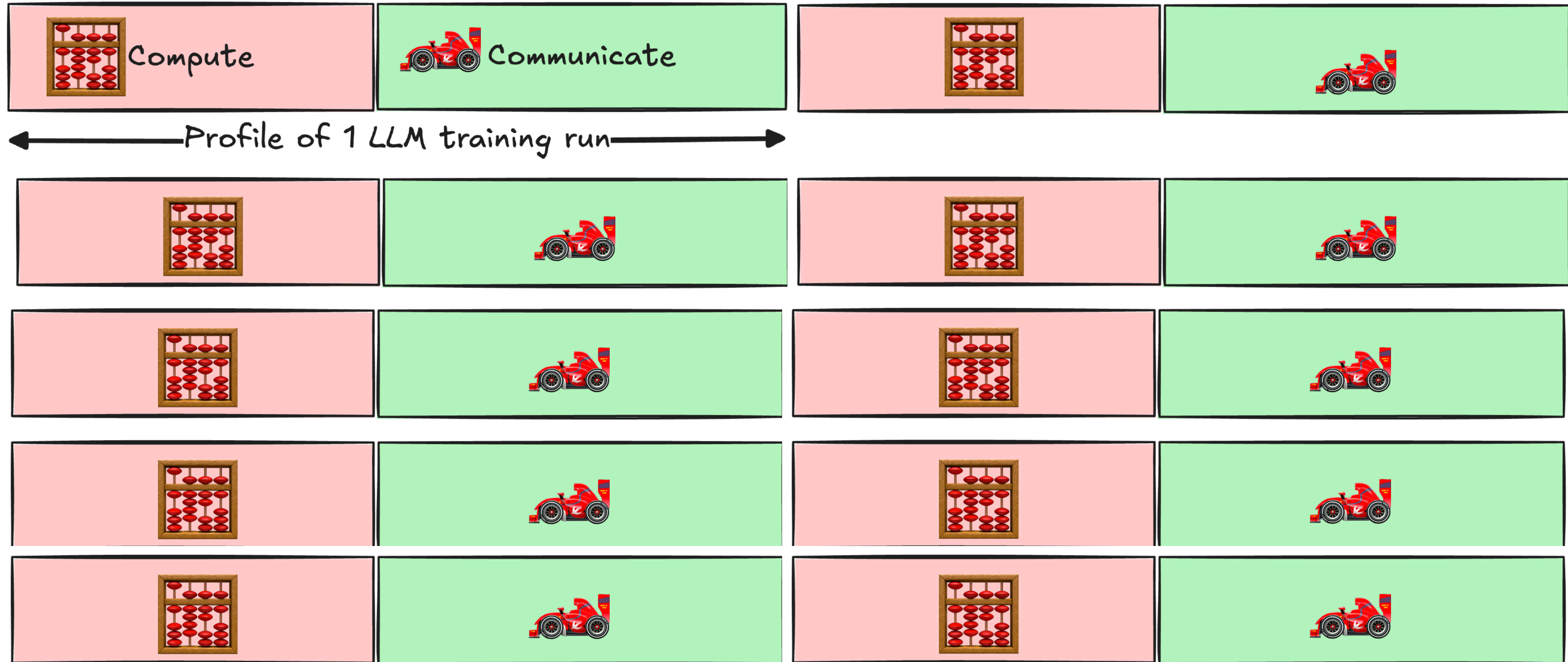
Traditional networks



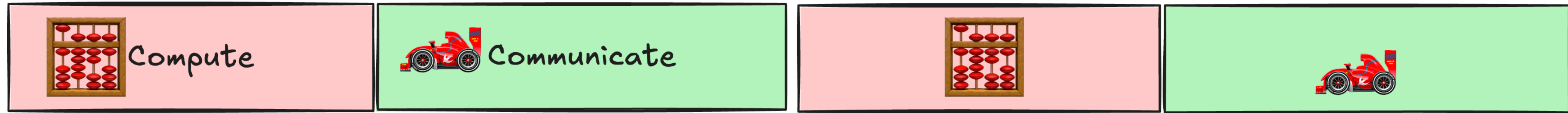
AI networks



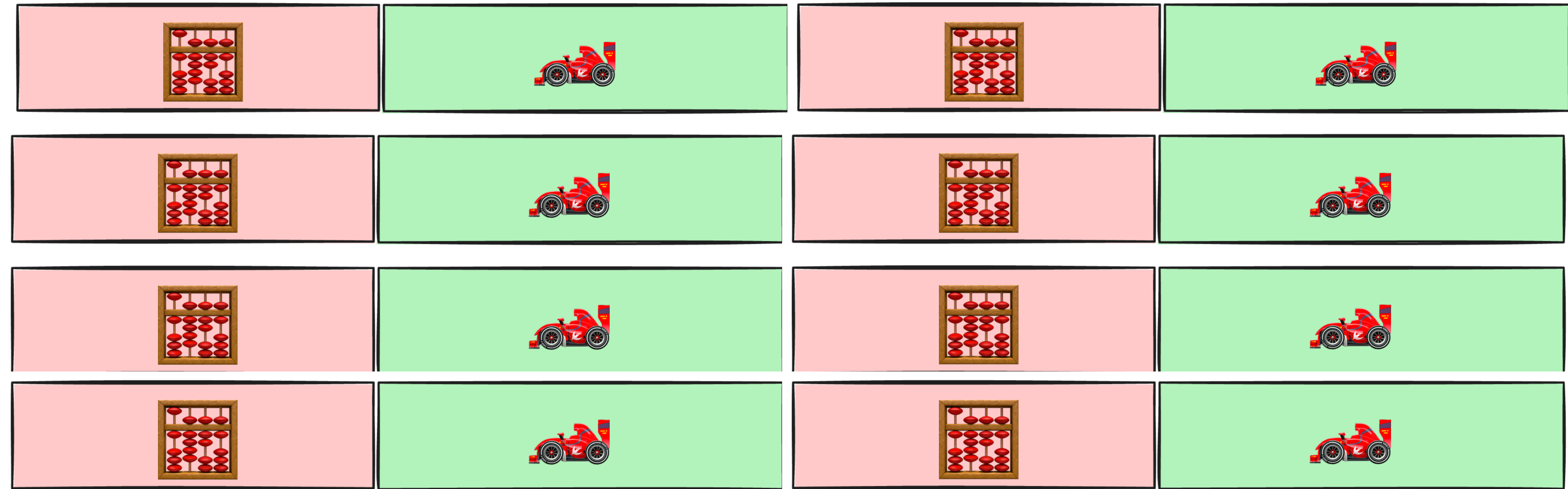
Source: Nvidia



Source: Nvidia



← Profile of 1 LLM training run →

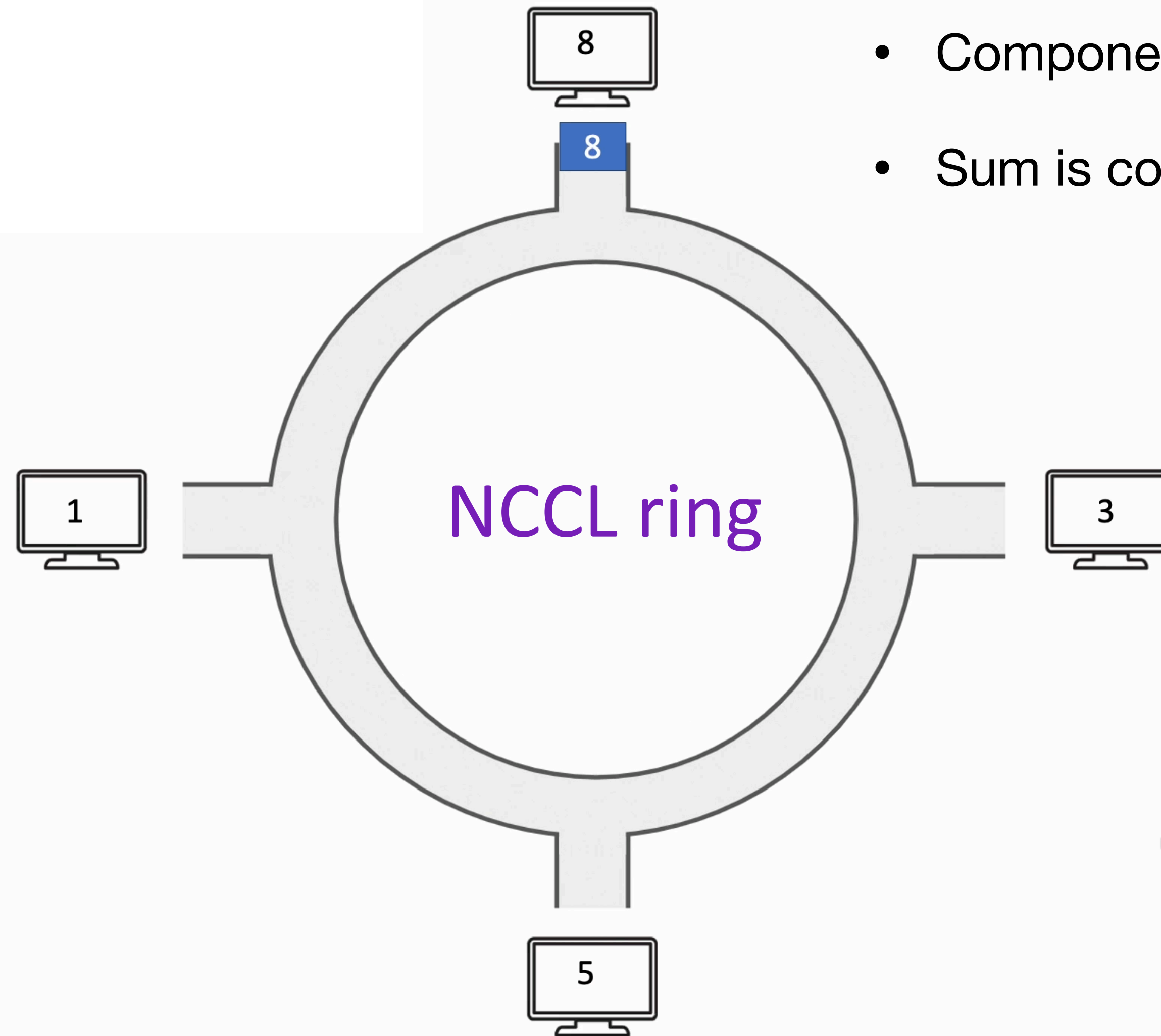


> 30% of total job completion time is spent on networking!



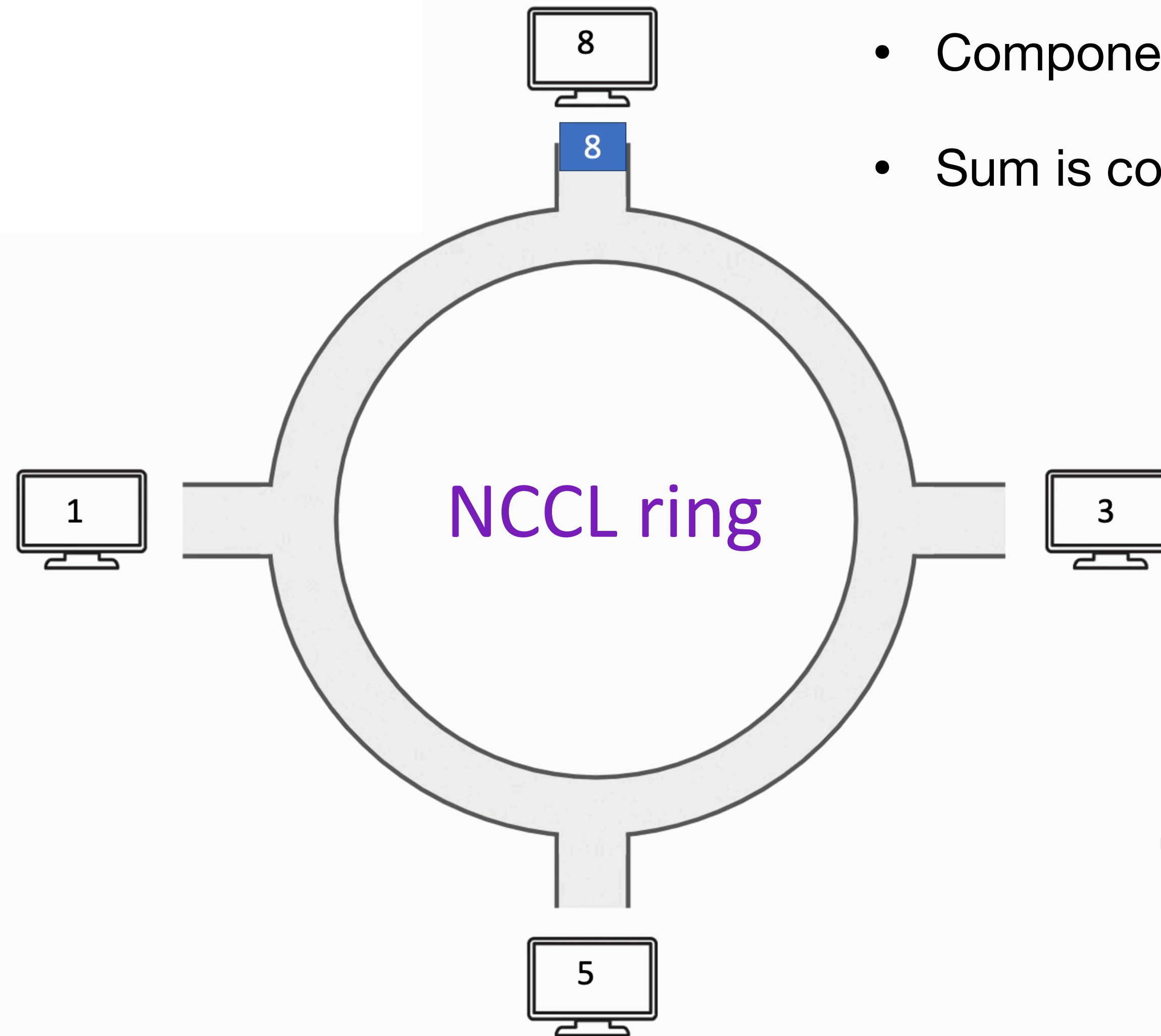
Source: AMD

All-reduce



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

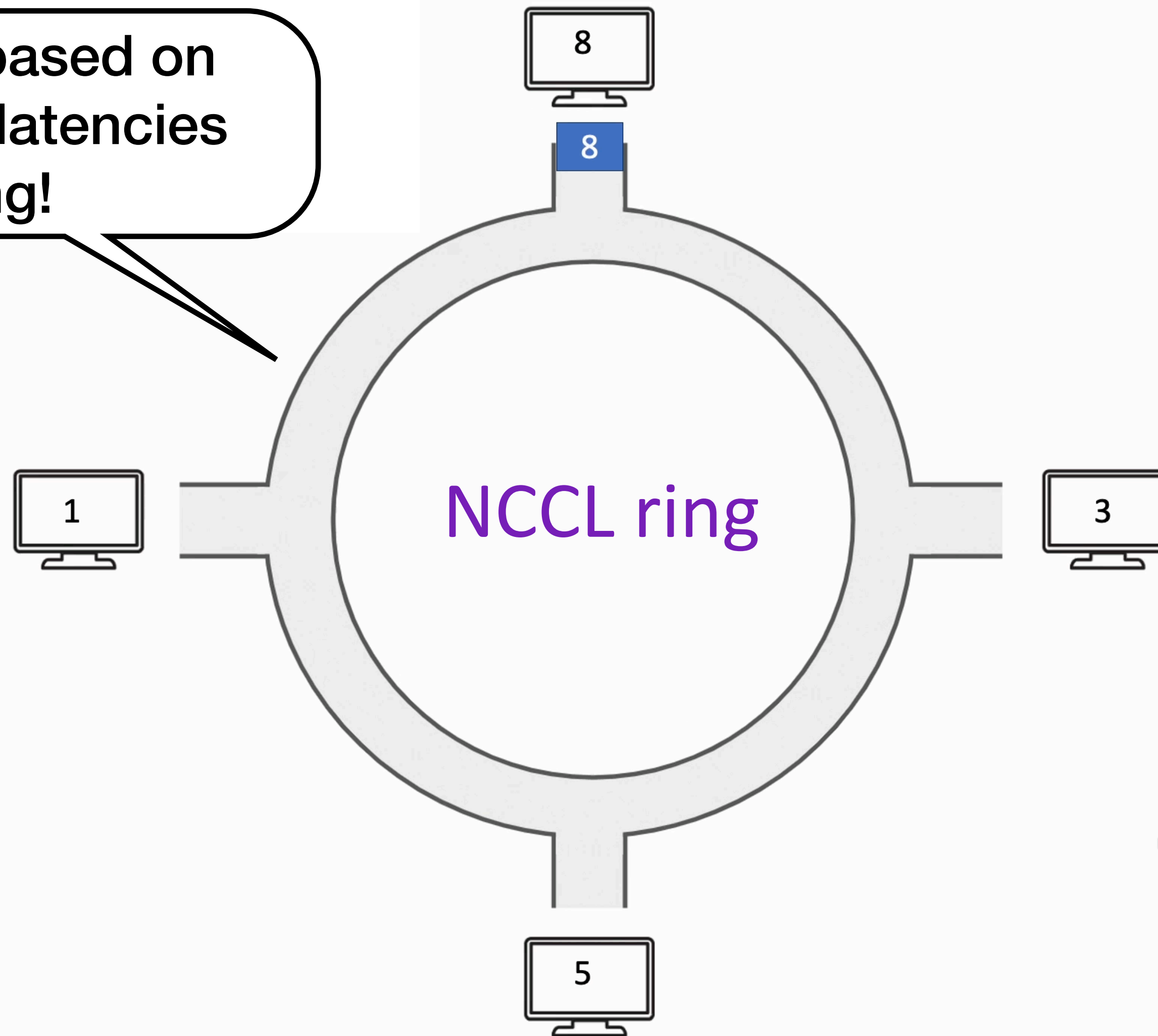
All-reduce



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

All-reduce

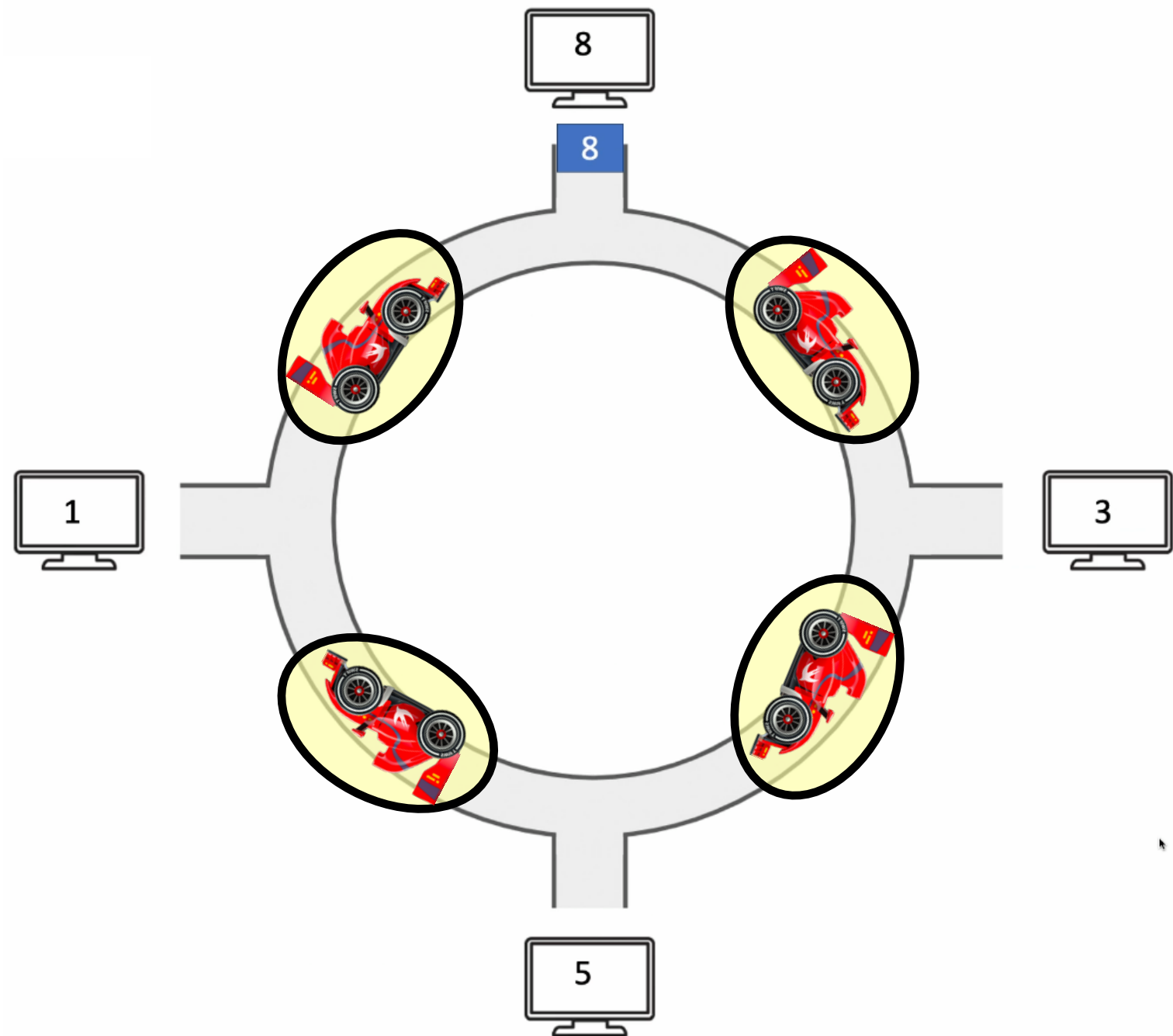
Finish time is based on the sum of the latencies in the ring!





Communicate

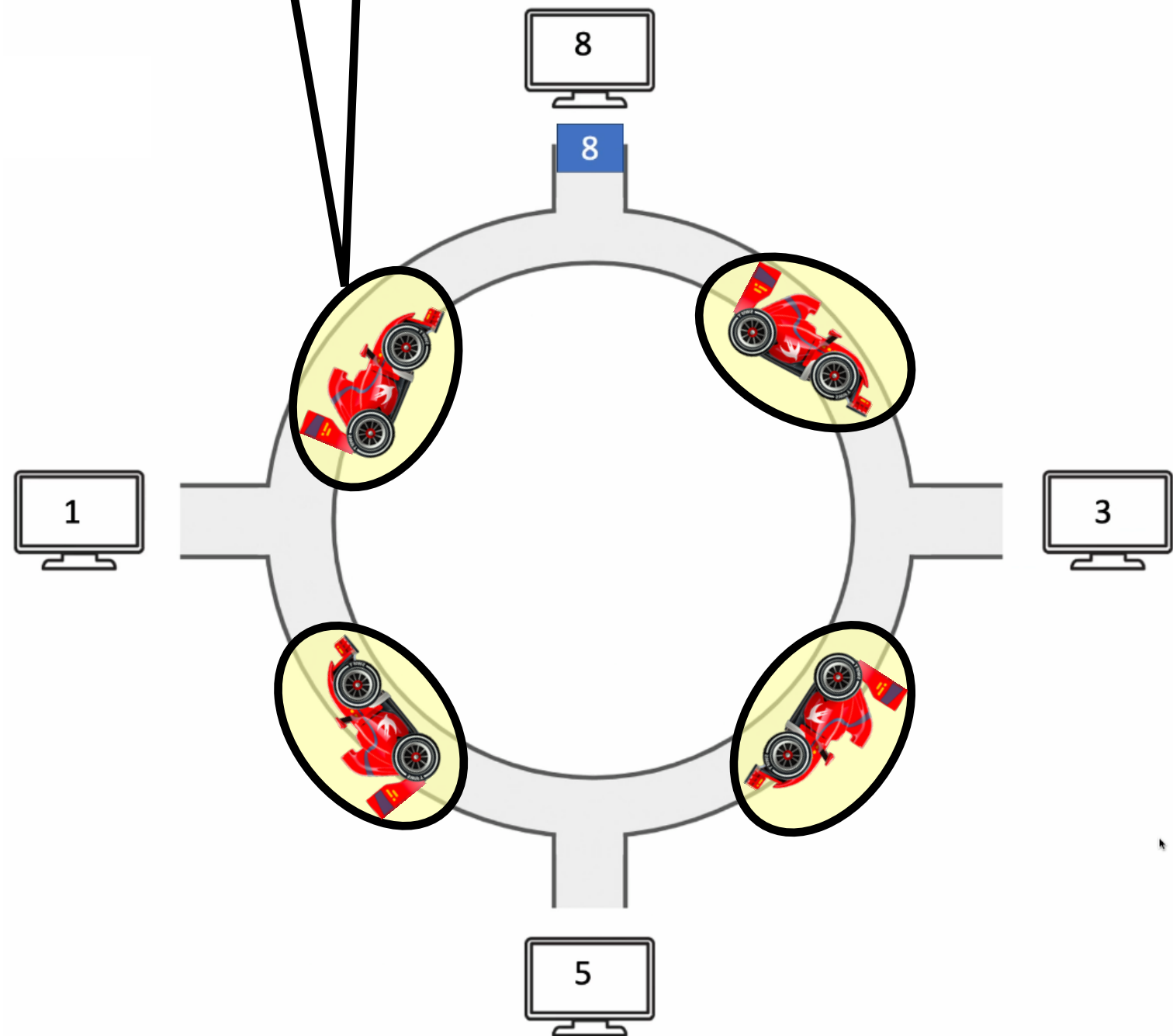
Pit crew is not done until the slowest member is done.





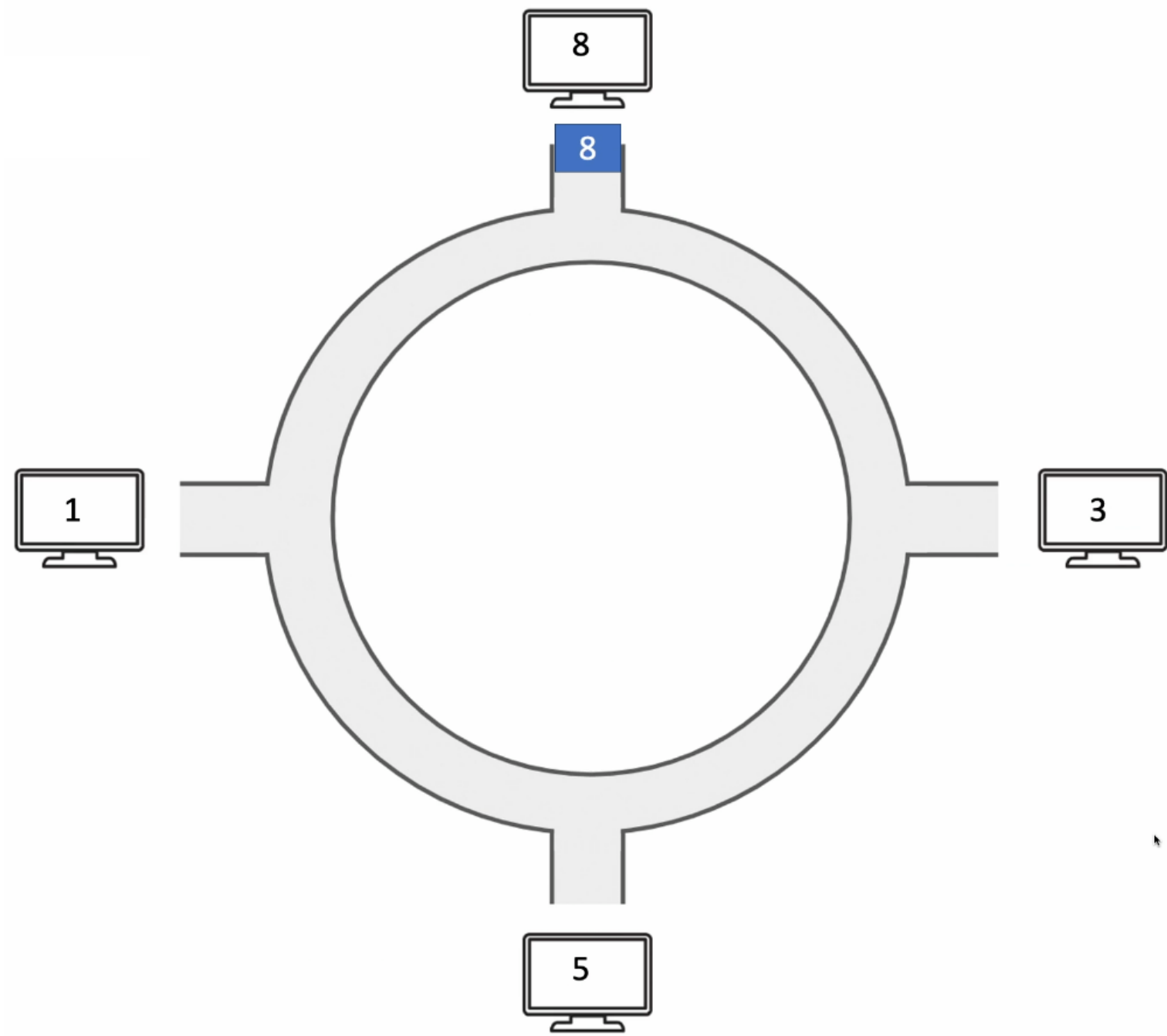
Communicate

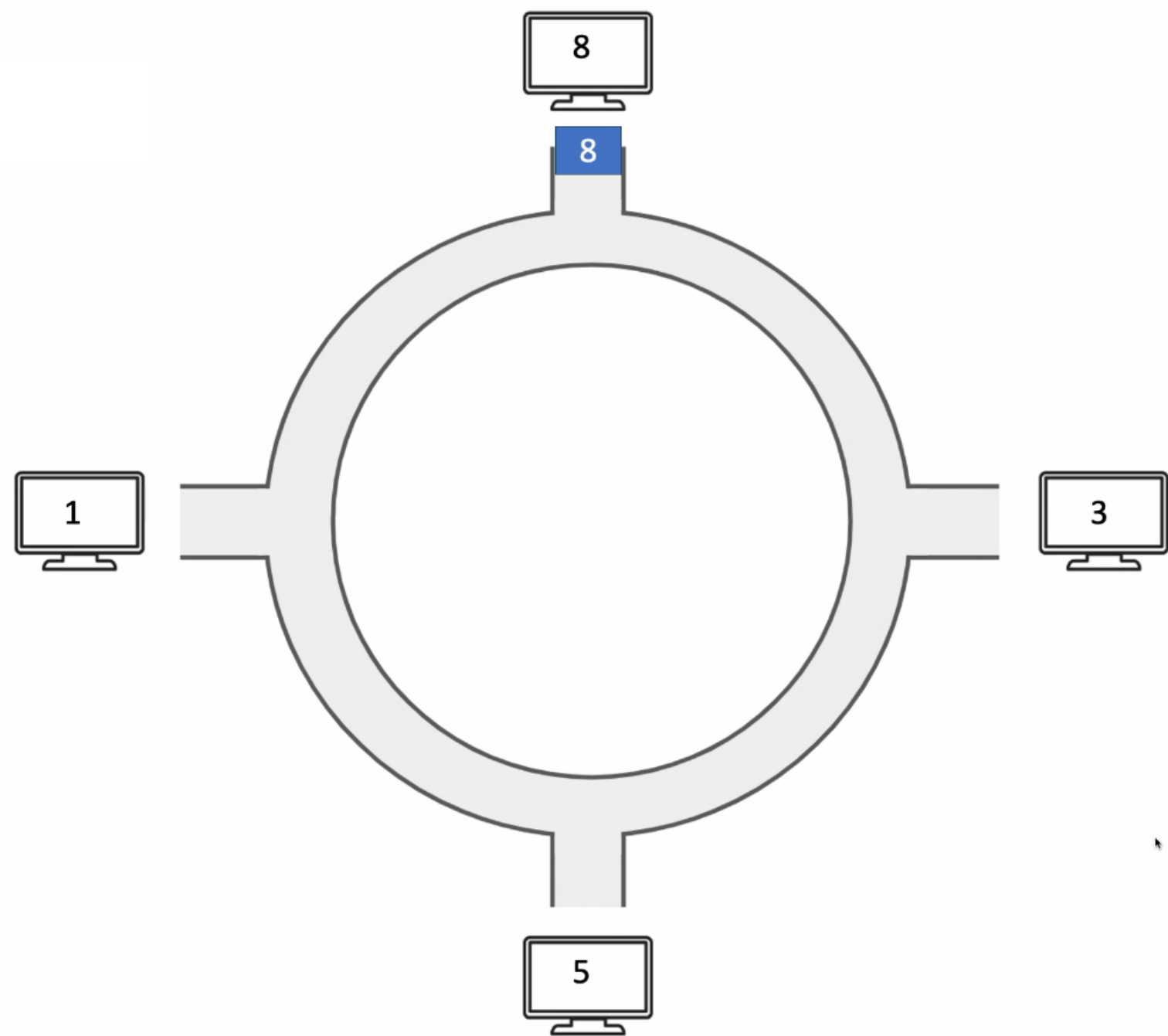
GPU's won't proceed to next step until all synchronize partial results.



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

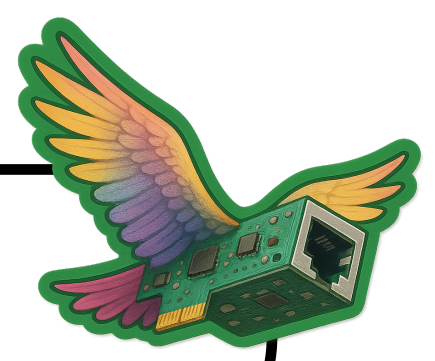




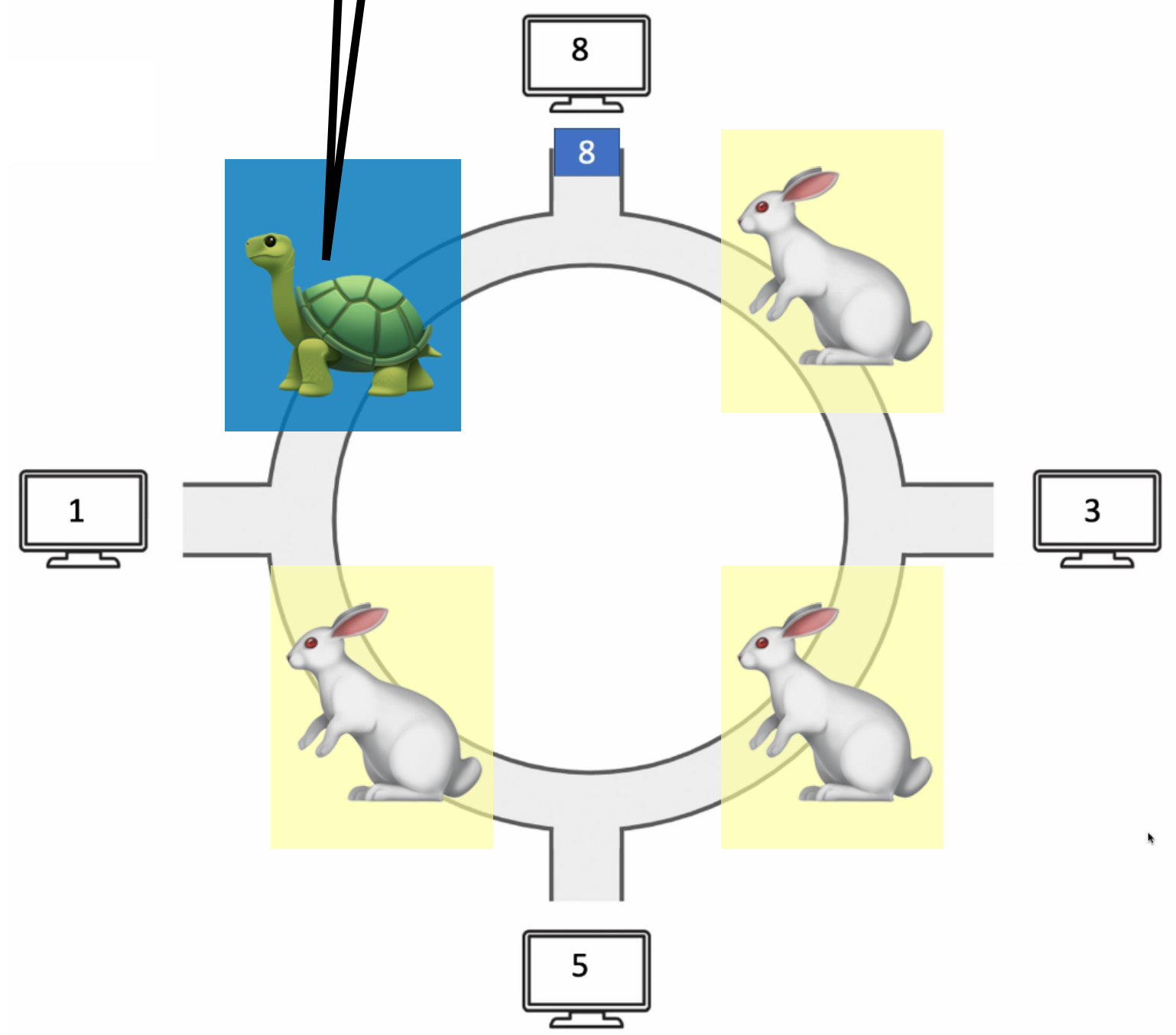


Slowest member determines speed.





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

Tribble on
Deep Space Explorer



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



Challenges



Visibility

Challenges



Visibility

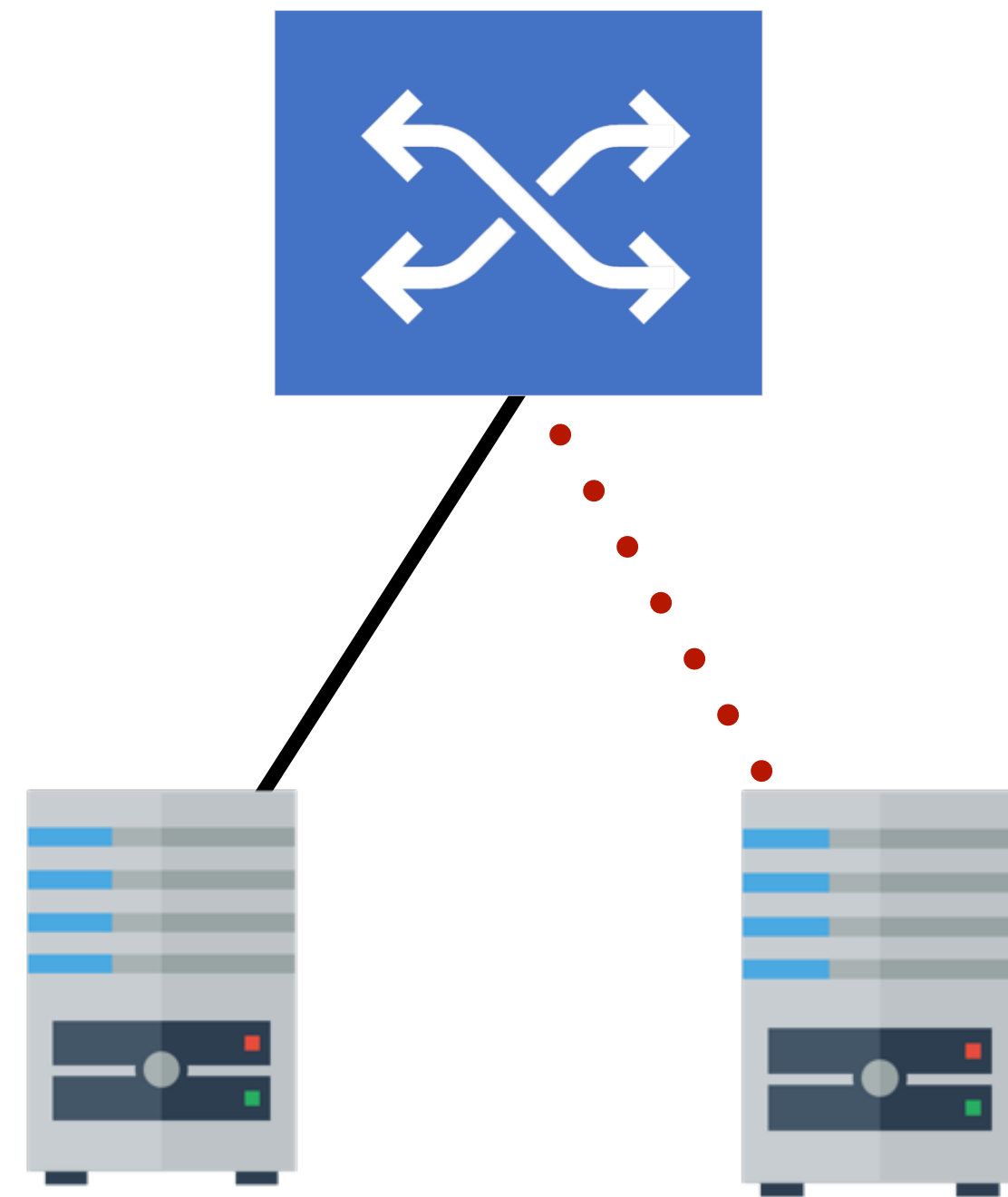


Reliability

Lack of visibility

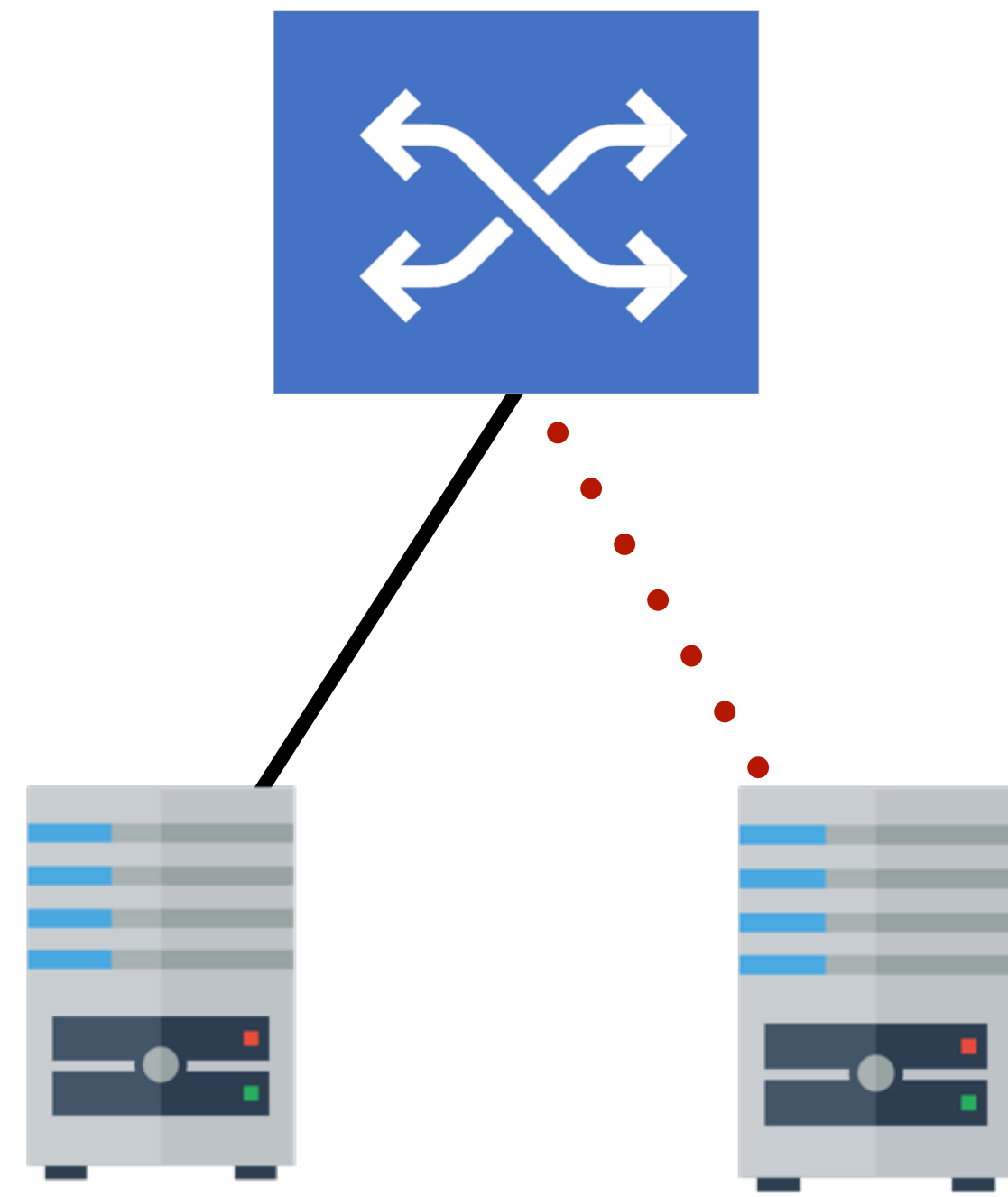


Lack of visibility



6 months ago...

Lack of visibility



6 months ago...

The ToR has down links
but I don't know why!



Today

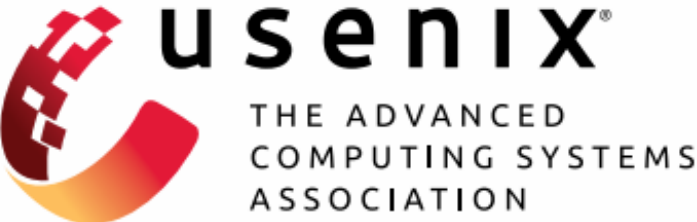
Lack of visibility

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



Huygens clock synchronization

- ▶ Software, probe mesh based
- ▶ High precision at scale



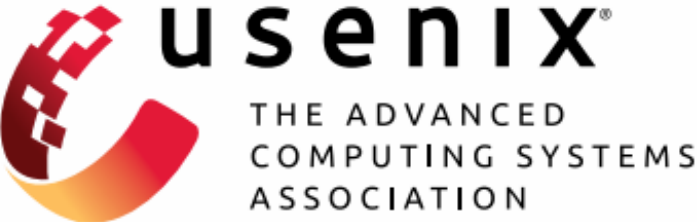
Exploiting a Natural Network Effect for Scalable, Fine-grained Clock Synchronization

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 clock synchronization

- ▶ Software, probe mesh based
- ▶ High precision at scale



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>



⚠ Demo Disclaimer ⚠

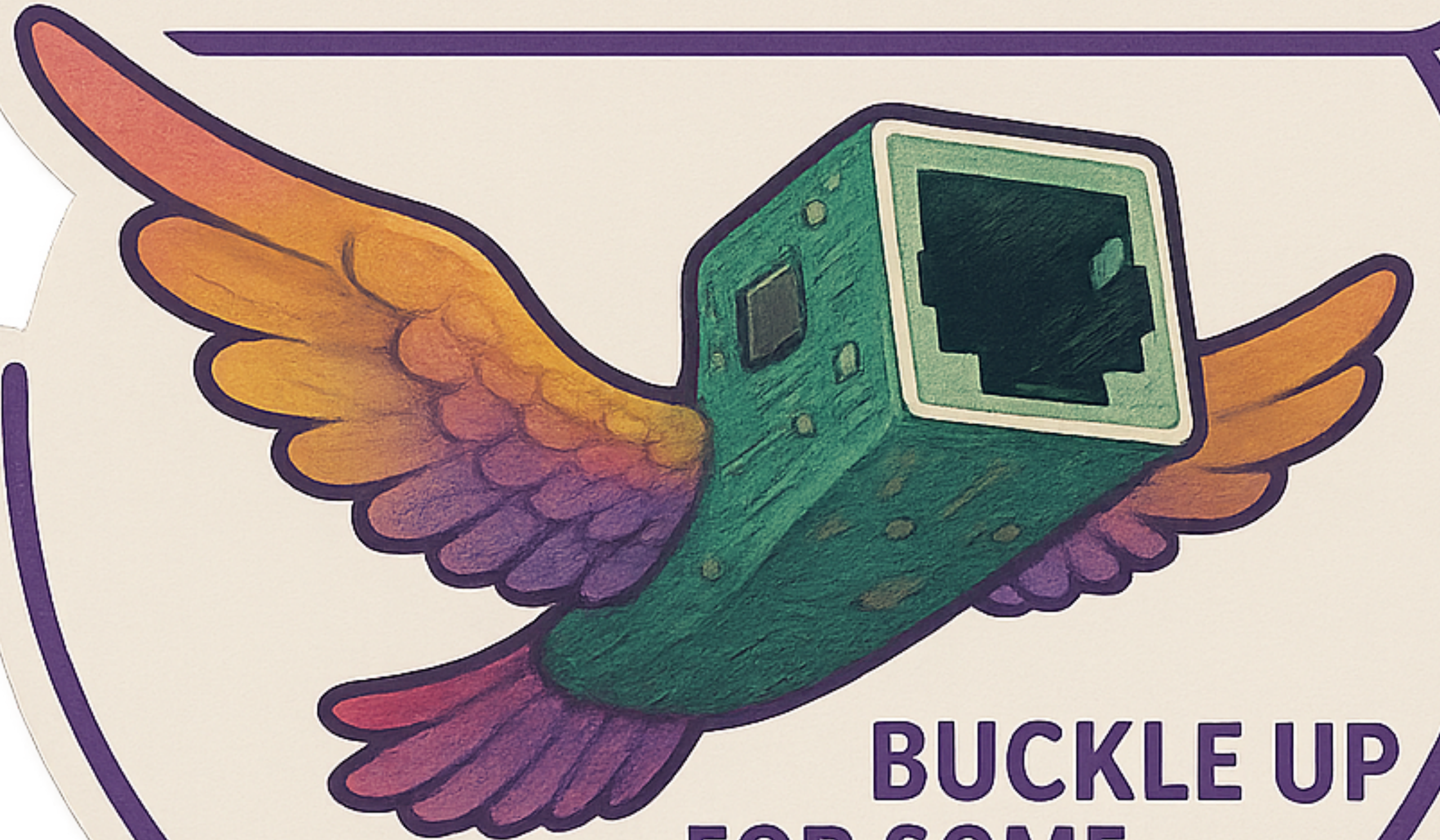
The upcoming demo
may not be
warp-speed thrilling...

unless you really, *really* love networking.

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



DEMO DISCLAIMER



**BUCKLE UP
FOR SOME
NETWORKING**

Visibility demo

The screenshot displays the 'IB Mesh' monitoring interface. At the top, there is a search bar for filtering agents by Name, Address, or Tag, and a time filter set to 'Last 2 minutes'. Below this, the 'Monitoring Mode' is set to 'Fleet', and 'NIC Alerts' are configured for 'Inbound' and 'Outbound'. Key performance indicators are shown: 'Healthy/Total NICs' at 64/64, 'Average NIC Throughput' at 2.96Mbps, 'Control' is 'Disabled', and 'Running Jobs' is 0. A legend indicates 'NIC' (green), 'Unhealthy NIC' (red), and 'NCCL Job' (orange), with a 'Probe Delay' range from ≤ 5μs to > 5μs. The main area shows a grid of 16 numbered nodes (1-16) arranged in a 4x4 layout, grouped under four ECMP categories: ecmp4, ecmp5, ecmp6, and ecmp7. A white callout box with the text 'In the same rack' is positioned above the grid. The interface includes a sidebar with navigation icons and a bottom status bar with a notification bell, help icon, and user profile.

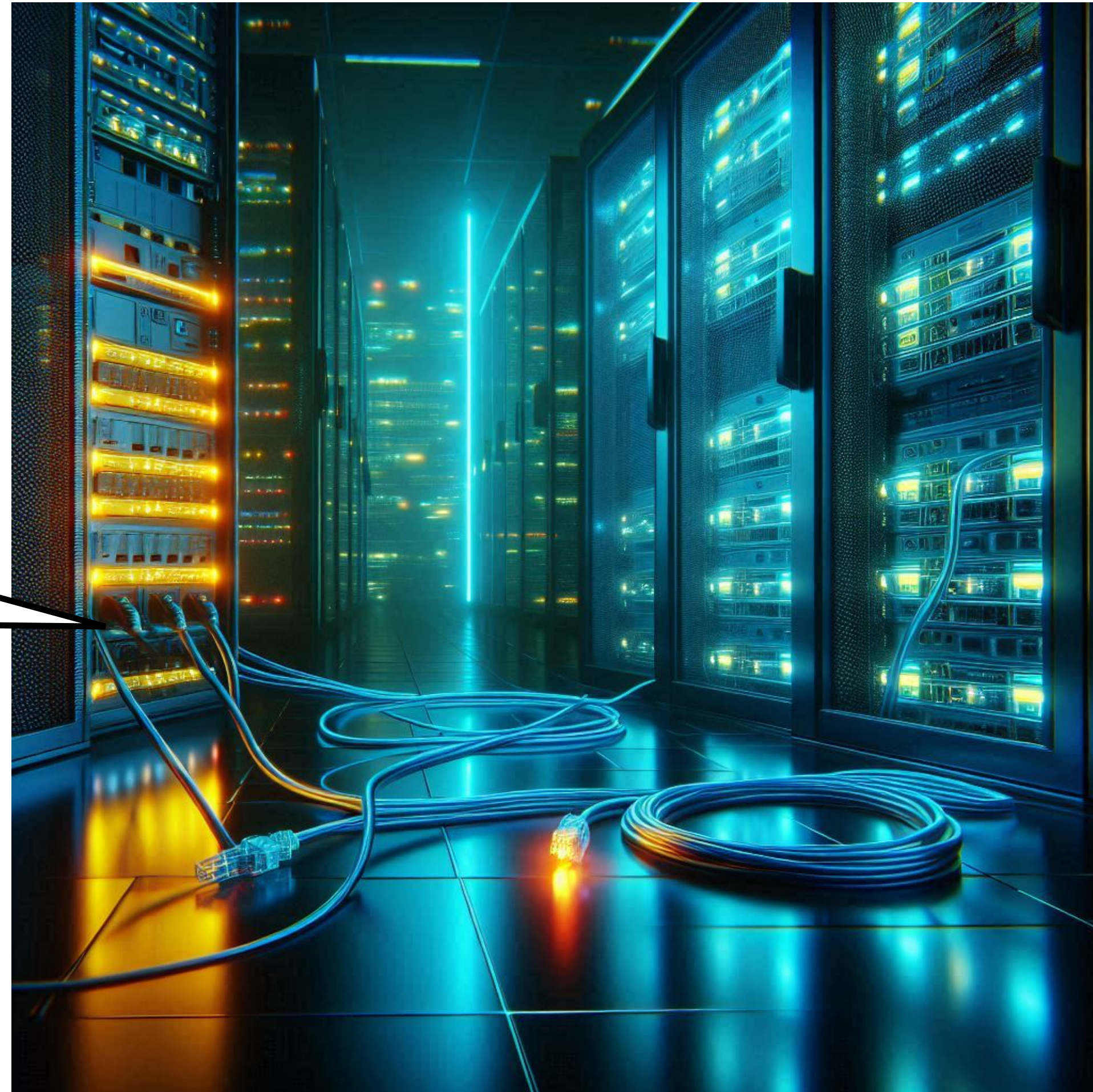
Visibility demo

The screenshot displays the 'IB Mesh' monitoring interface. At the top, there is a search bar for filtering agents by Name, Address, or Tag, and a time range selector set to 'Last 2 minutes'. Below this, the 'Monitoring Mode' is set to 'Fleet', and 'NIC Alerts' are configured for 'Inbound' and 'Outbound'. Key metrics are shown in a summary row: 'Healthy/Total NICs' at 64/64, 'Average NIC Throughput' at 2.96Mbps, 'Control' status as 'Disabled', and 'Running Jobs' at 0. A legend indicates 'NIC' (green), 'Unhealthy NIC' (red), and 'NCCL Job' (orange), with a 'Probe Delay' range from $\leq 5\mu\text{s}$ (green) to $> 5\mu\text{s}$ (red). The main area shows a grid of agents, with a white callout box pointing to a specific agent labeled 'In the same rack'. Below this, four ECMP groups (ecmp4, ecmp5, ecmp6, ecmp7) are visible, each containing a 4x4 grid of 16 numbered agent tiles (1-16).

Lack of reliability



**Link flapped 10 times!
Send technicians
to swap cables!**



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?

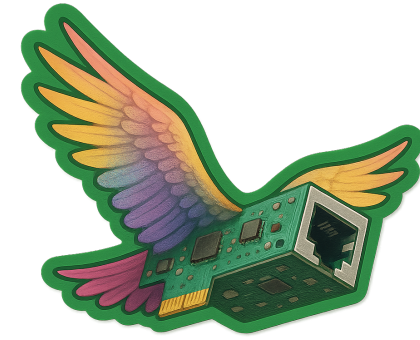


- Network link temporarily drops. (up => down)
- It can last in the order of seconds.
 - A few seconds => may self heal and recover from it if < timeout.
 - > a few seconds => collective communication times out => job fails.

Lack of reliability



Lack of reliability



- ▶ Fall back to previous checkpoint
- ▶ Lost 🕒 💰

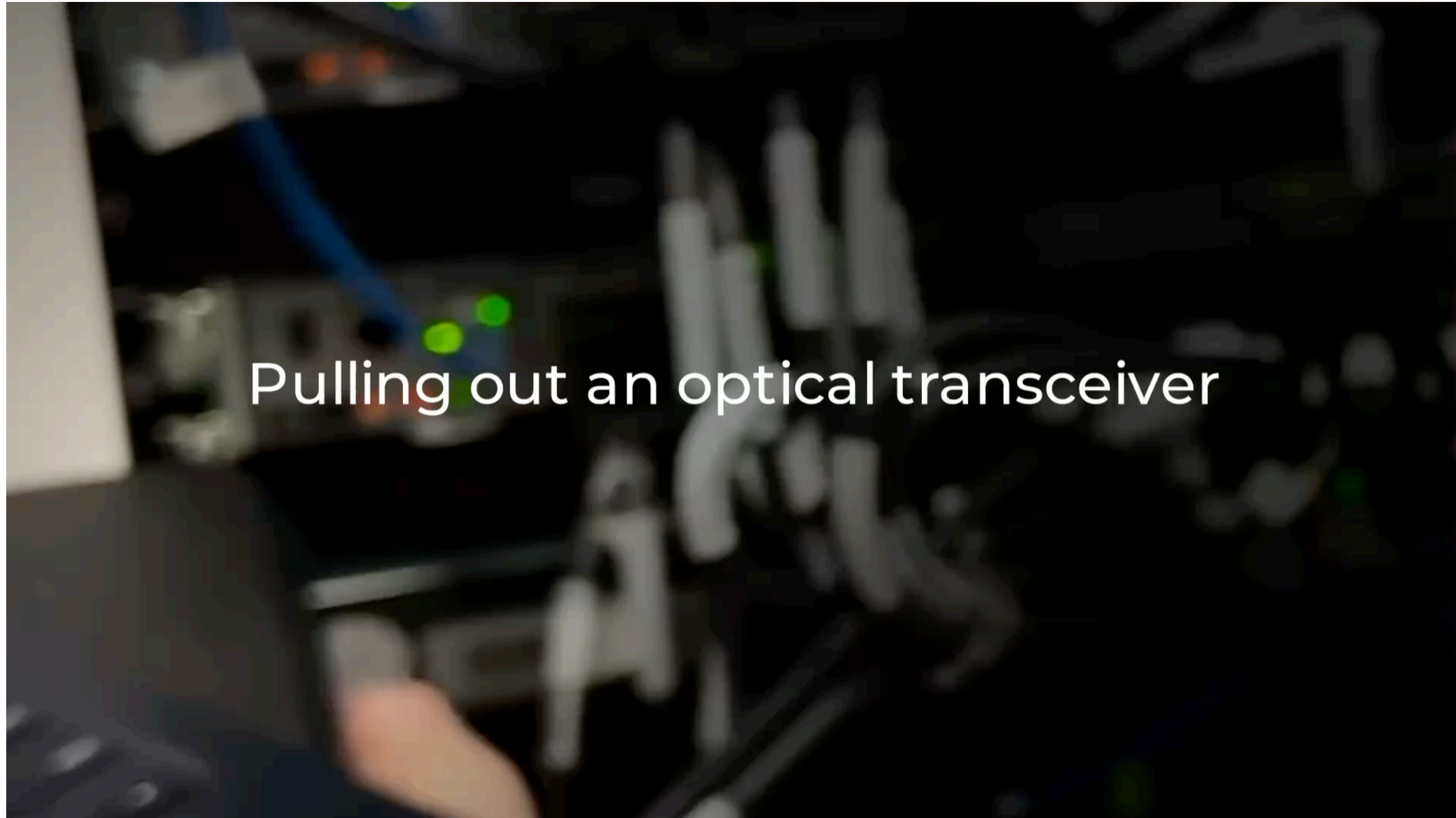


Poll

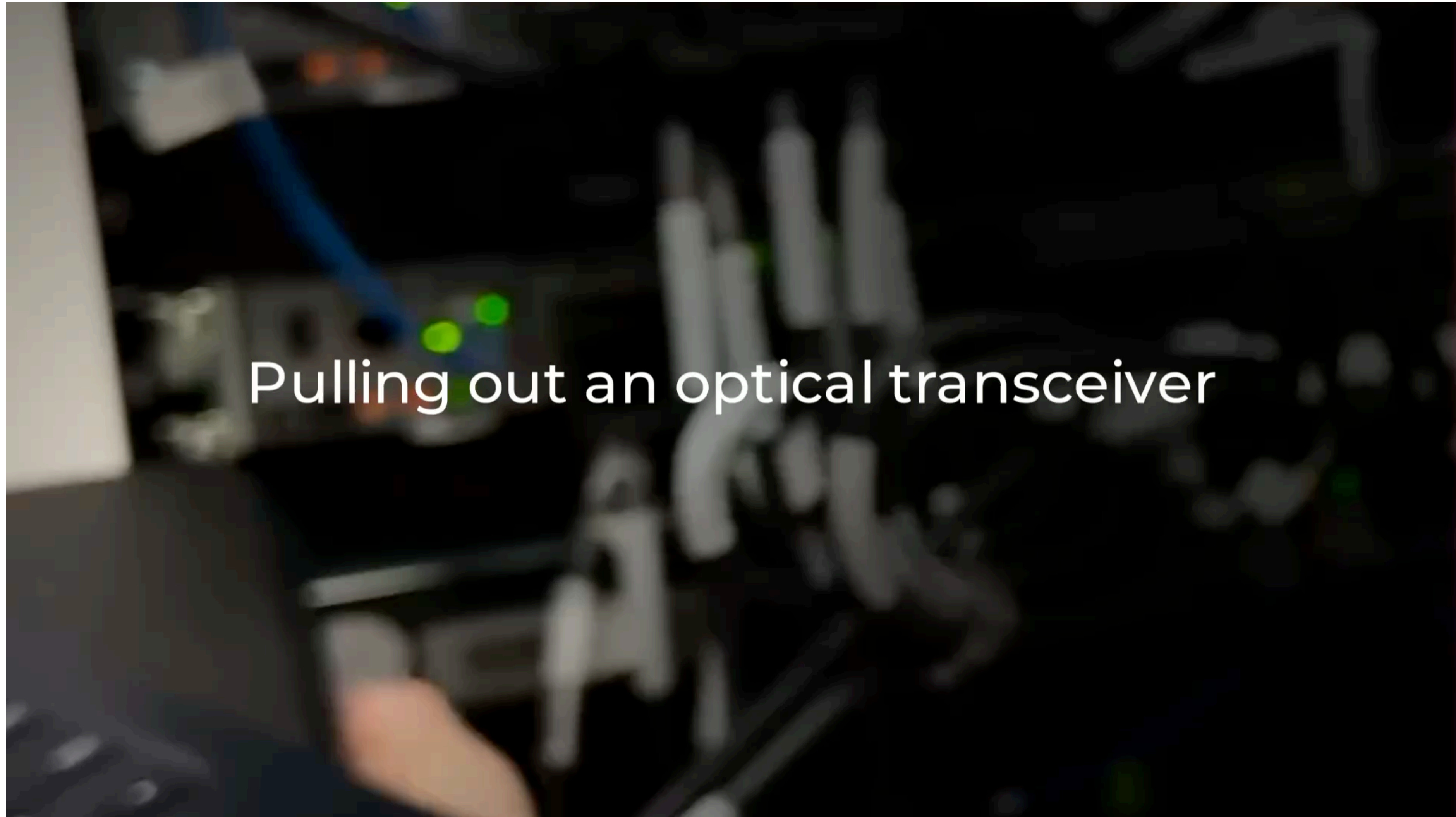
👉 *If your cluster had a personality, which would it be?*

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

Lack of reliability



Lack of reliability



Flapping demo

The dashboard displays the following metrics and controls:

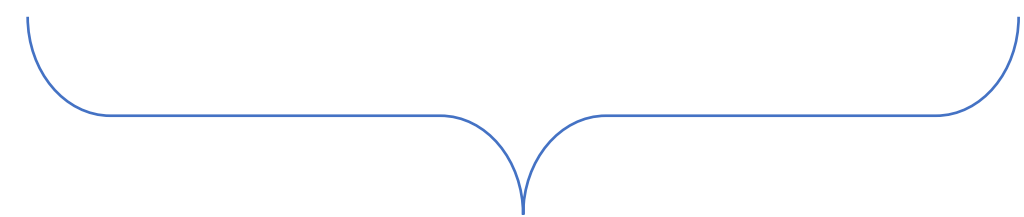
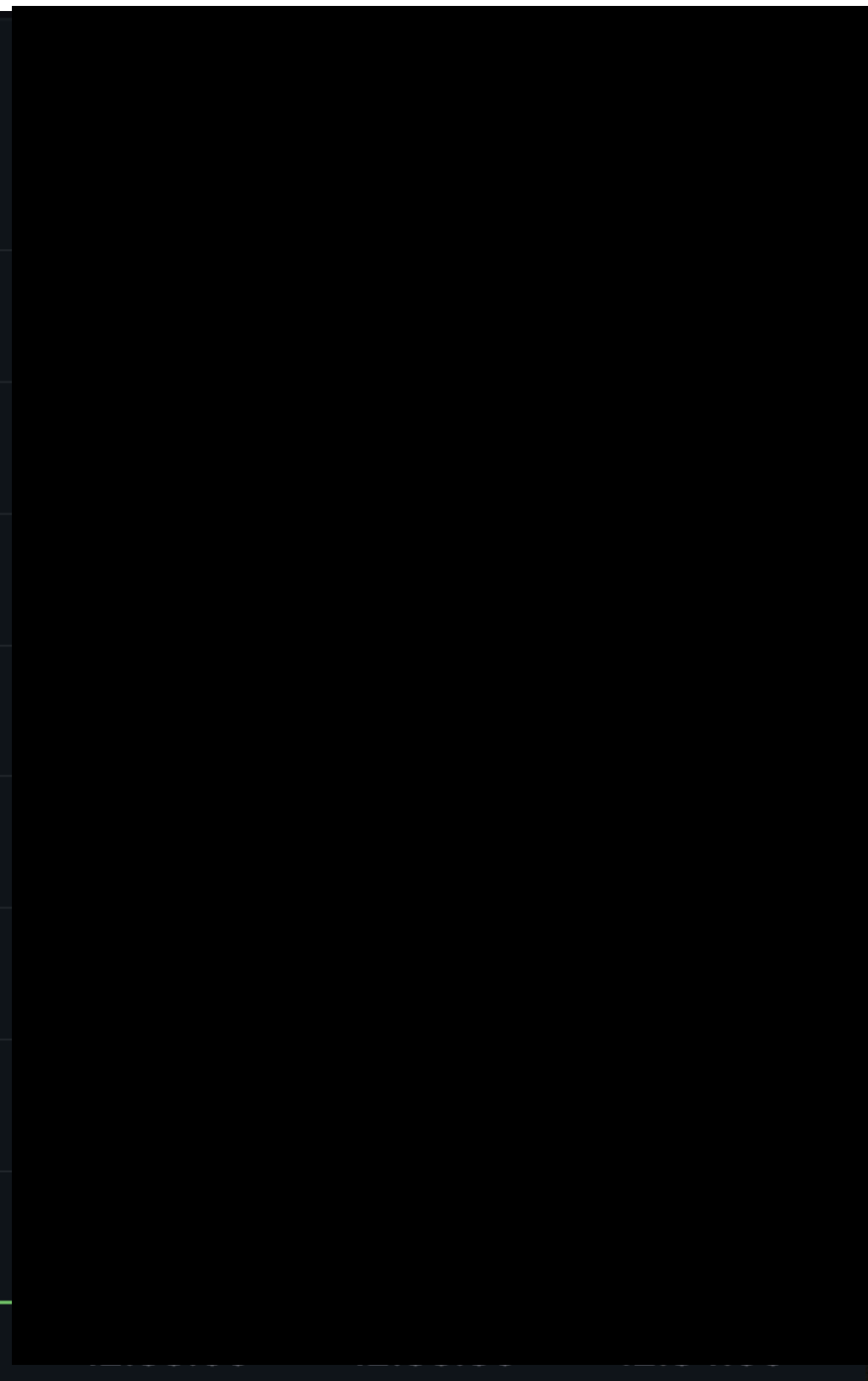
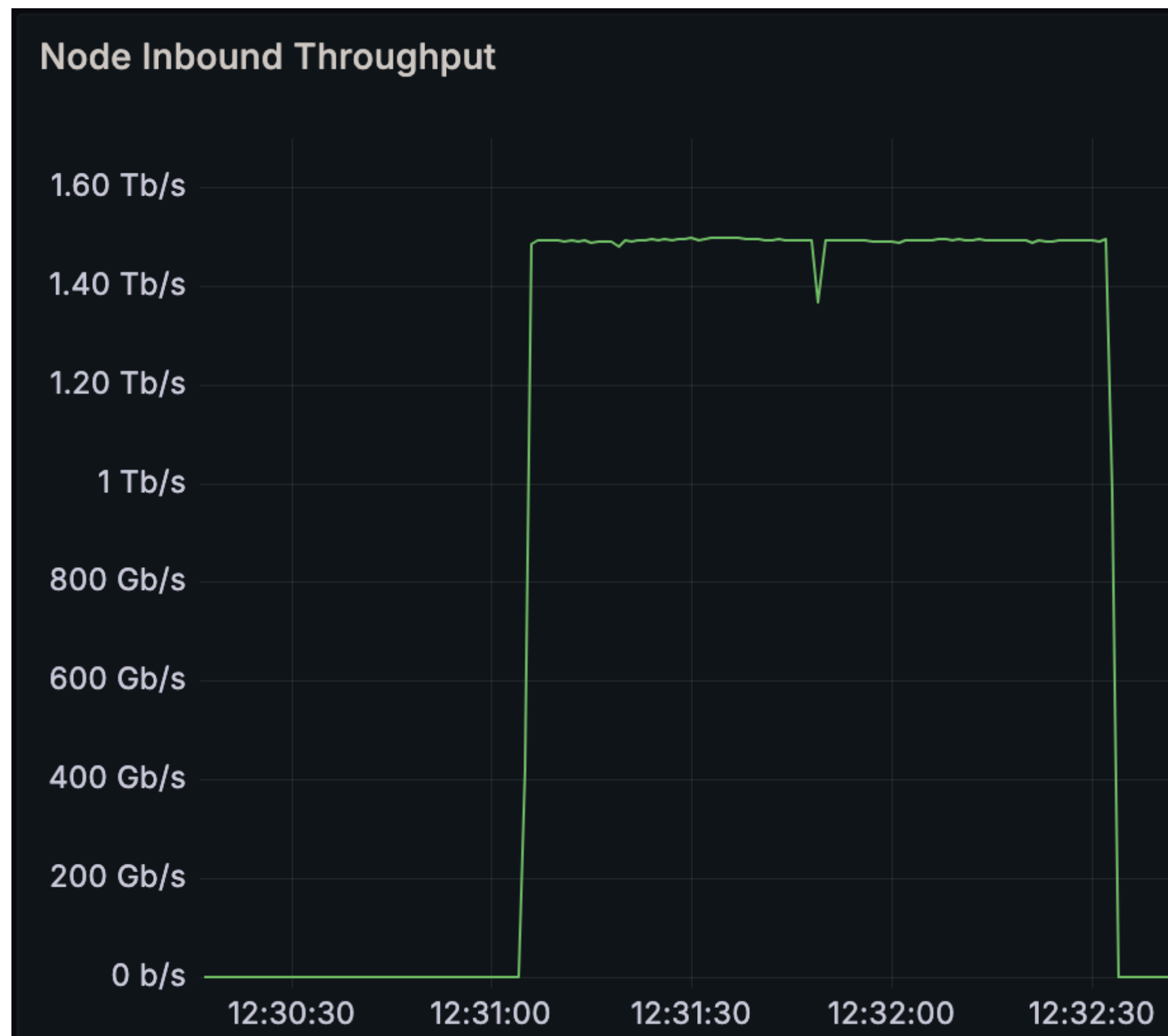
- IB Mesh** (Title)
- Filter agents by Name, Address, or Tag** (Search bar)
- Control** (Icon)
- Navigate** (Icon)
- Monitoring Mode**: Workload, **Fleet** (Selected)
- NIC Alerts**: Inbound, **Outbound** (Selected)
- Healthy/Total NICs**: 64/64
- Average NIC Throughput**: 2.96Mbps
- Control**: Disabled
- Running Jobs**: 0
- Legend**: 1 NIC (Green), 1 Unhealthy NIC (Red), NCCL Job (Yellow), **Probe Delay**: ≤ 5μs (Green), > 5μs (Red)
- ECMP Groups**: ecmp4, ecmp5, ecmp6, ecmp7. Each group contains a 4x4 grid of 16 numbered nodes (1-16).

Flapping demo

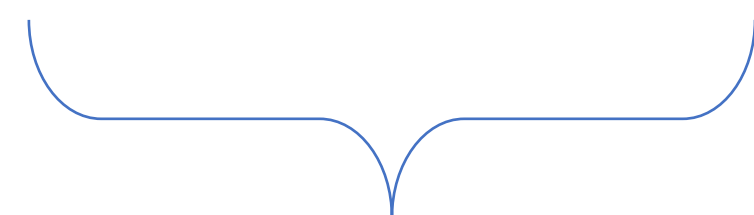
The dashboard displays the following metrics and controls:

- IB Mesh** (Title)
- Filter agents by Name, Address, or Tag** (Search bar)
- Control** (Menu)
- Navigate** (Menu)
- Monitoring Mode**: Workload, **Fleet** (Selected)
- NIC Alerts**: Inbound, **Outbound** (Selected)
- Healthy/Total NICs**: 64/64
- Average NIC Throughput**: 2.96Mbps
- Control**: Disabled
- Running Jobs**: 0
- Legend**: 1 NIC (Green), 1 Unhealthy NIC (Red), NCCL Job (Yellow), **Probe Delay**: ≤ 5μs (Green), > 5μs (Red)
- Grids**: Four ECMP groups (ecmp4, ecmp5, ecmp6, ecmp7) each containing a 4x4 grid of 16 numbered nodes (1-16).

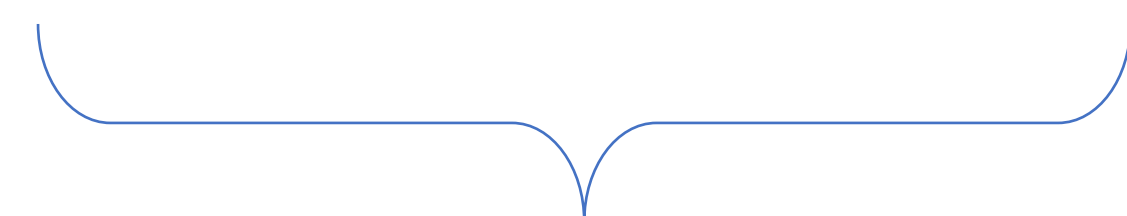
NIC Flaps in a cluster on Oracle Cloud



Run 1: No Failure

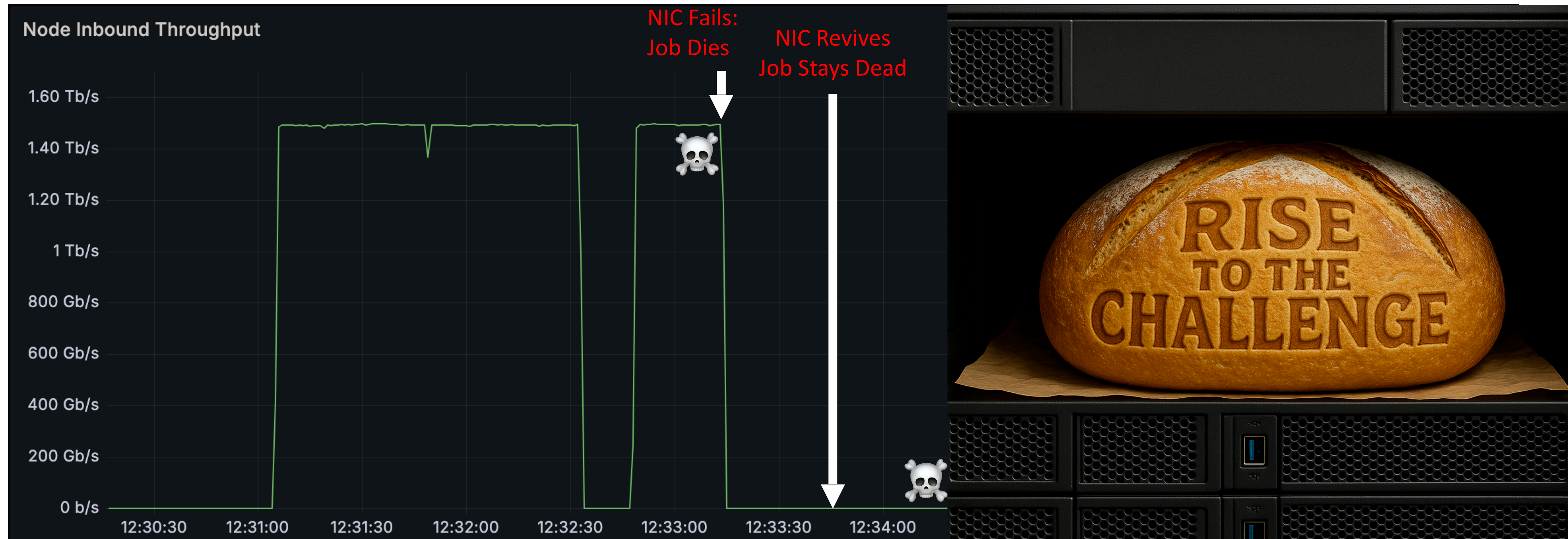


Run 2: NIC Failure
(What happens today)



Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

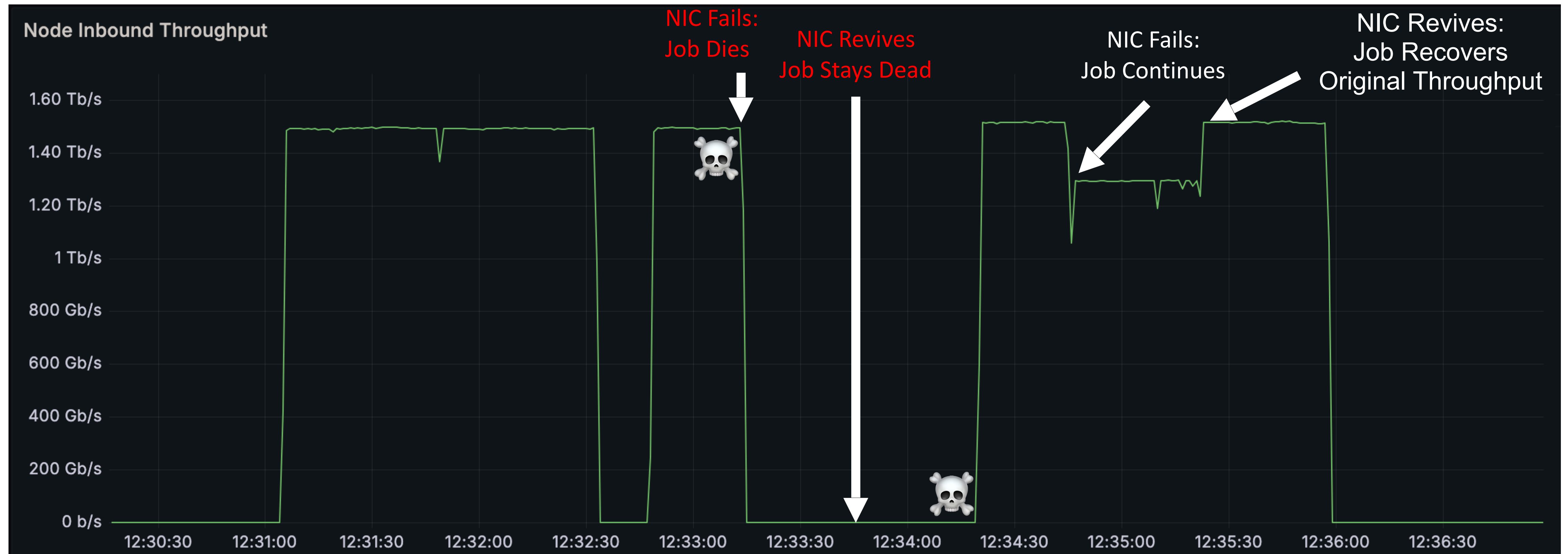


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

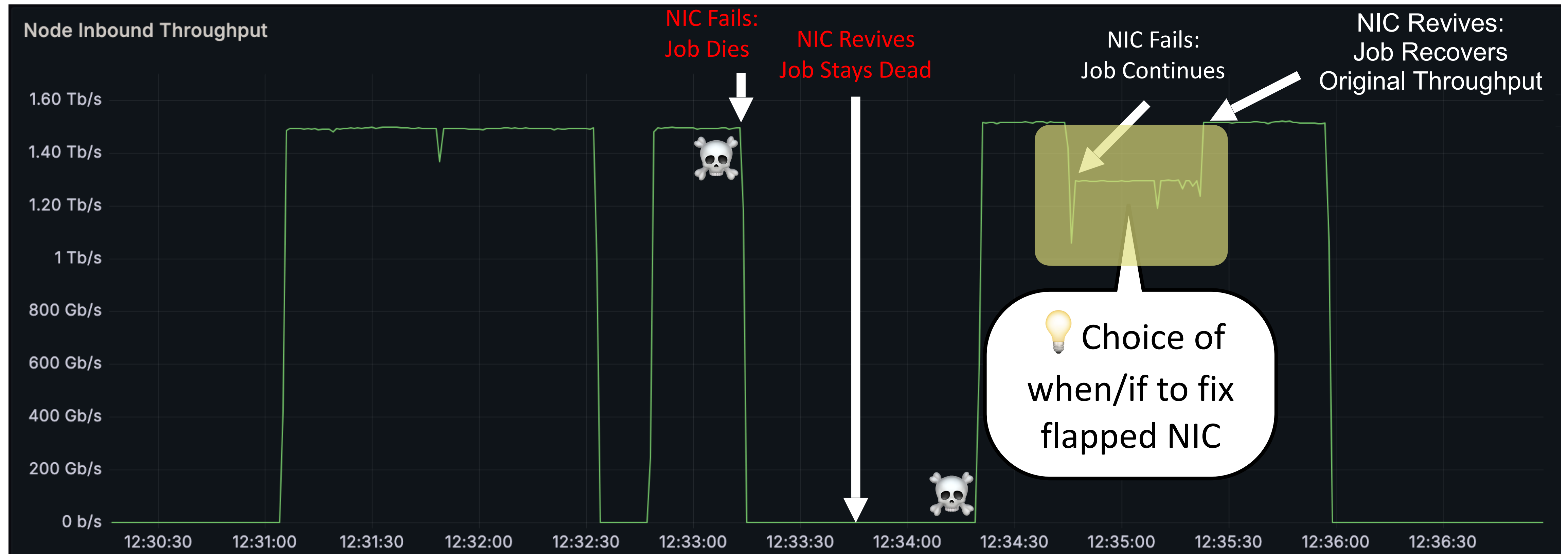


Run 1: No Failure

Run 2: NIC Failure
(What happens today)

Run 3: A solution
NIC Failure + Auto recovery

NIC Flaps in a cluster on Oracle Cloud

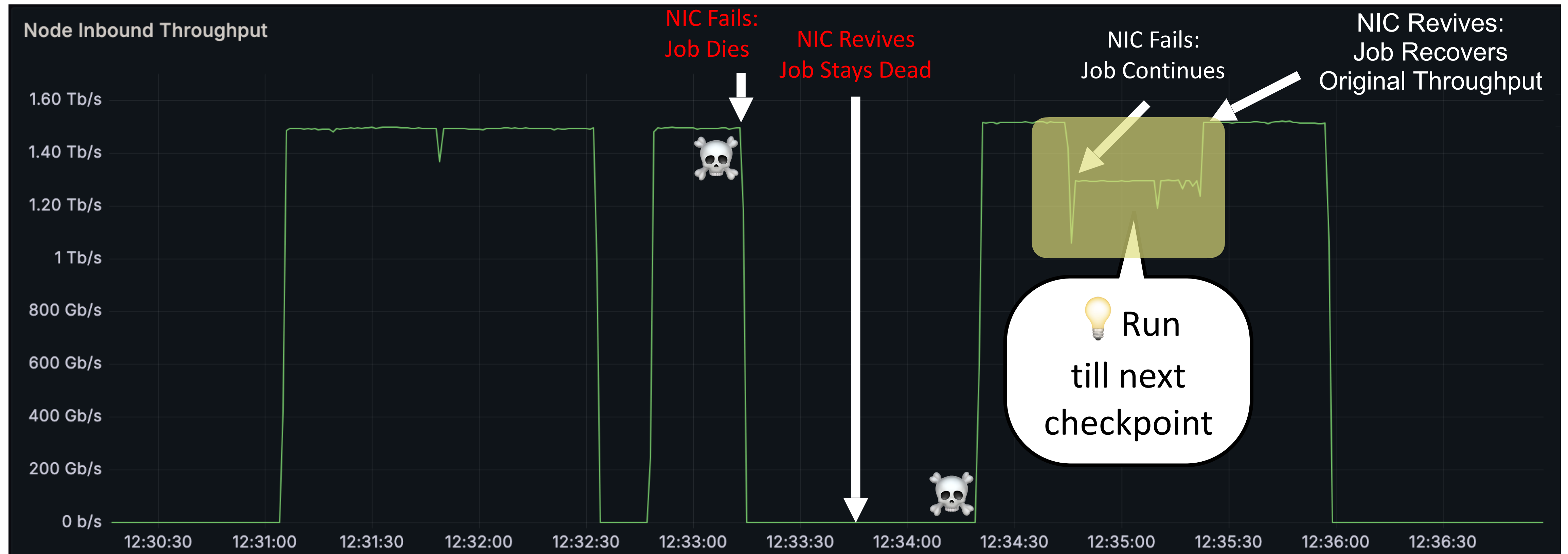


Run 1: No Failure

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

The dashboard displays the following information:

- IB Mesh** header with a search filter: "Filter agents by Name, Address, or Tag".
- Monitoring Mode:** Workload (selected) and Fleet.
- NIC Alerts:** Inbound and Outbound.
- Summary Metrics:**
 - Healthy/Total NICs: 64/64
 - Average NIC Throughput: 2.96Mbps
 - Control: Disabled
 - Running Jobs: 0
- Legend:** 1 NIC (purple), 1 Unhealthy NIC (grey), NCCL Job (orange), Probe Delay: ≤ 5μs (green), > 5μs (red).
- Agent Grids:** Four groups of agents labeled ecmp4, ecmp5, ecmp6, and ecmp7. Each group contains a 4x4 grid of 16 numbered agents (1-16), all of which are currently green.

Contention demo

IB Mesh Filter agents by Name, Address, or Tag Last 2 minutes Select Workspace

Control **Navigate** **Monitoring Mode** **NIC Alerts**

Workload **Fleet** Inbound Outbound

Healthy/Total NICs: **64/64** Average NIC Throughput: **2.96Mbps** Control: **Disabled** Running Jobs: **0**

1 NIC 1 Unhealthy NIC NCCL Job **Probe Delay:** ≤ 5μs > 5μs

ecmp4 ▾ ecmp5 ▾ ecmp6 ▾ ecmp7 ▾

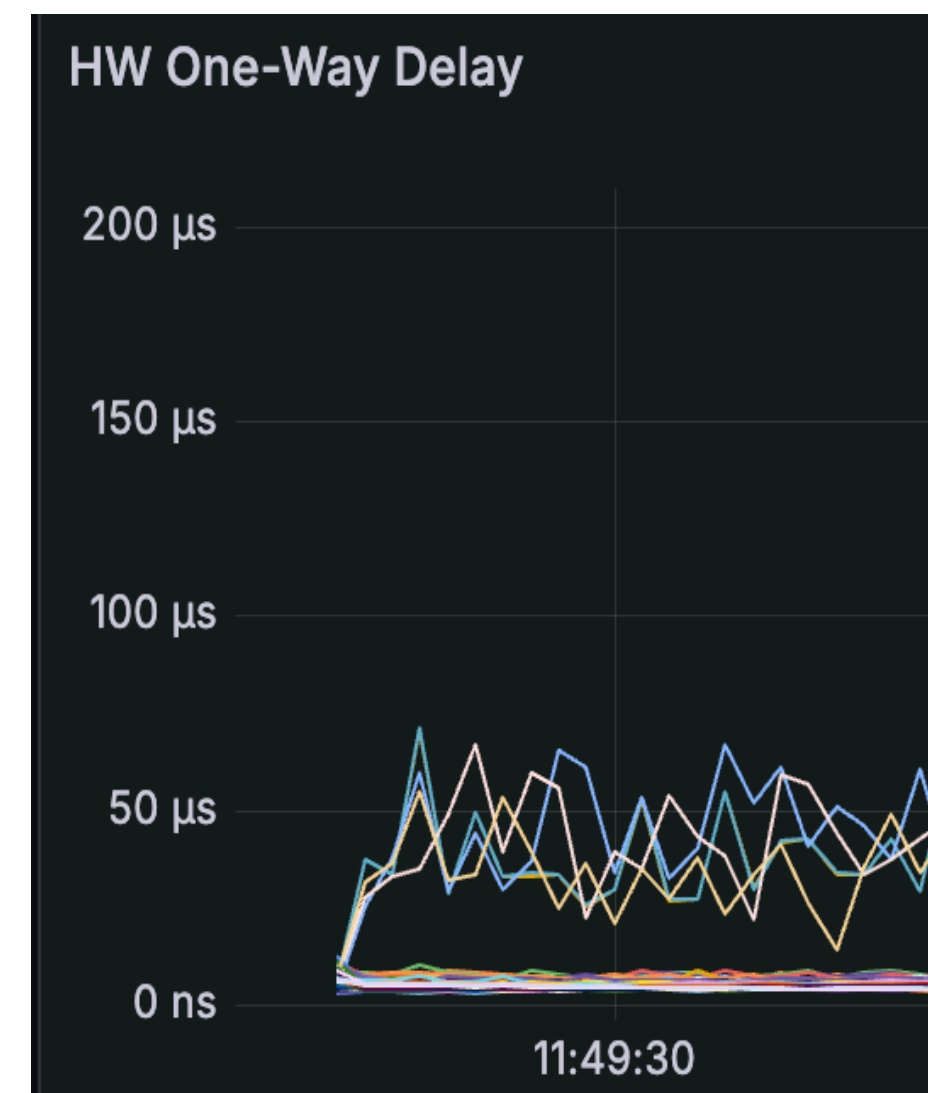
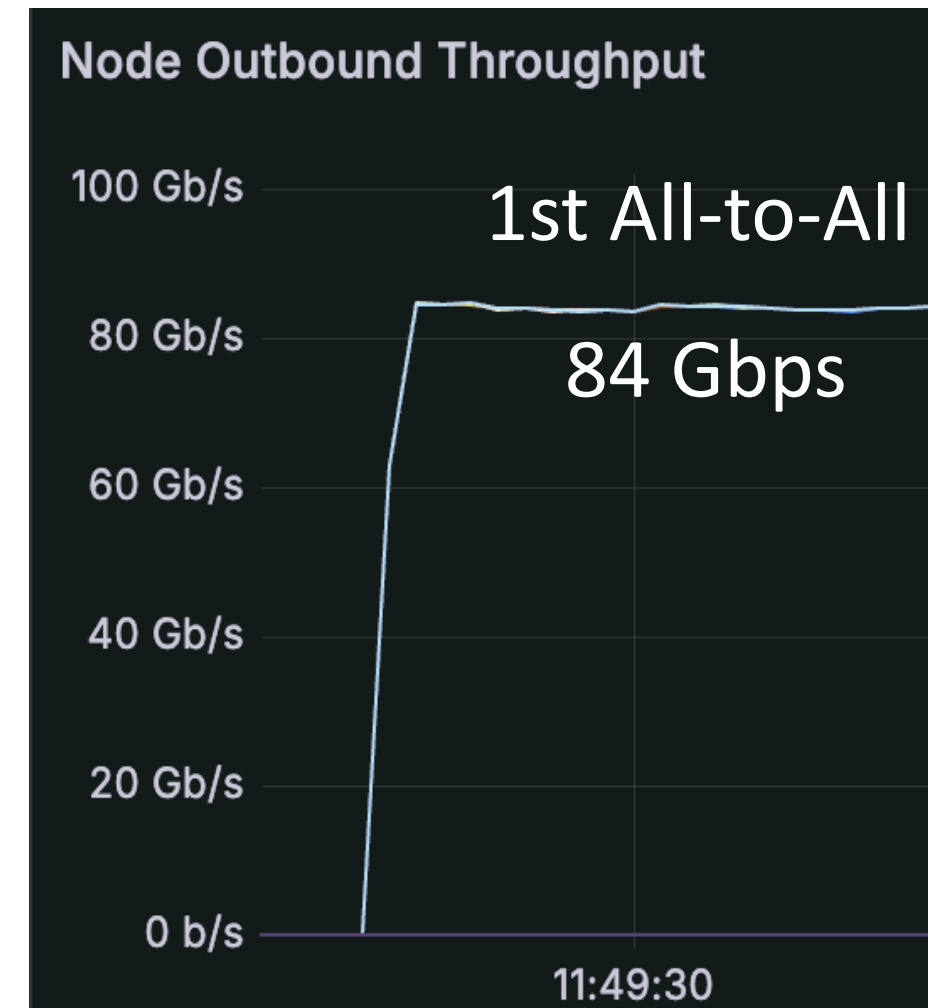
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

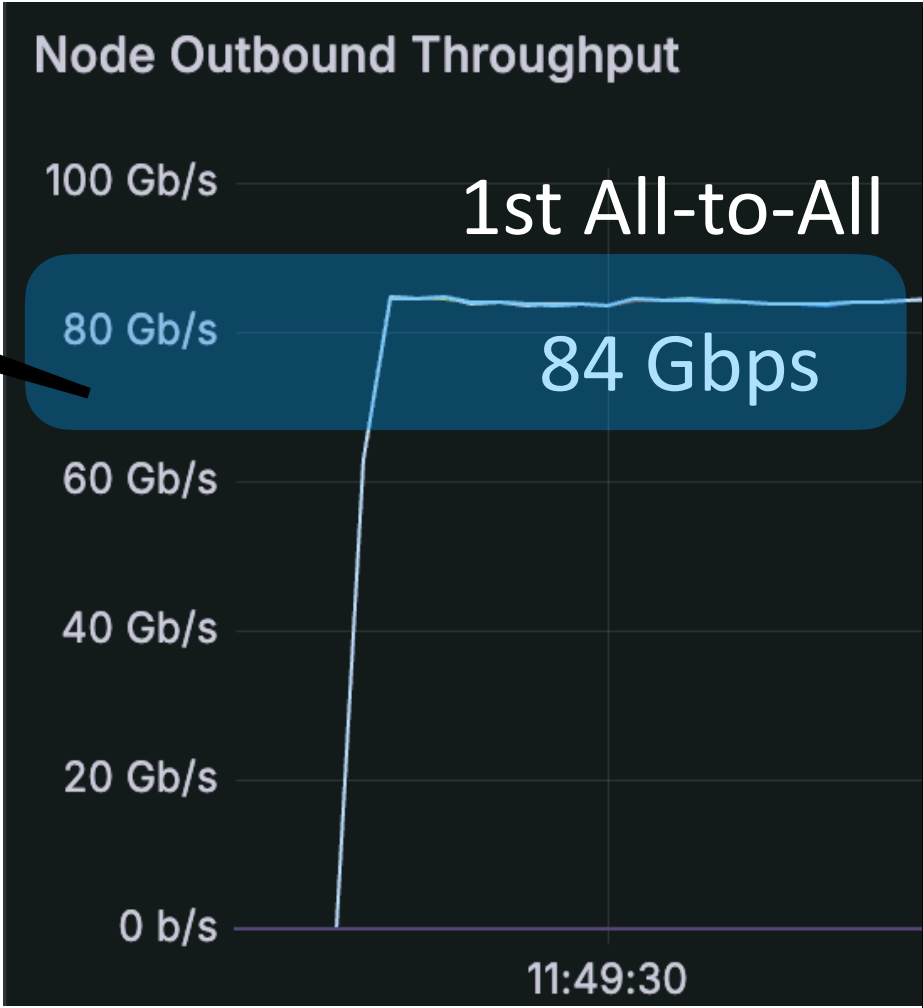
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Contention on ECMP Test Bed

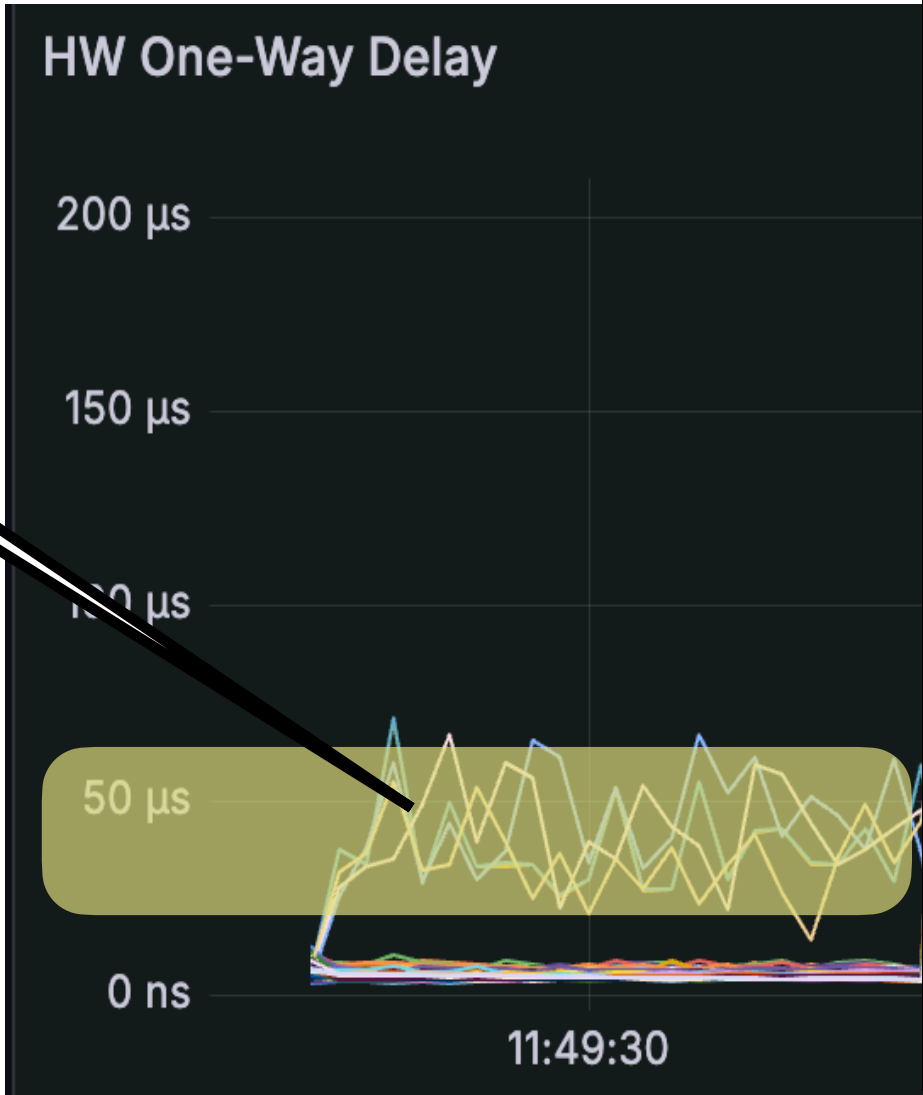


Contention on ECMP Test Bed

↓ Lower Throughput!

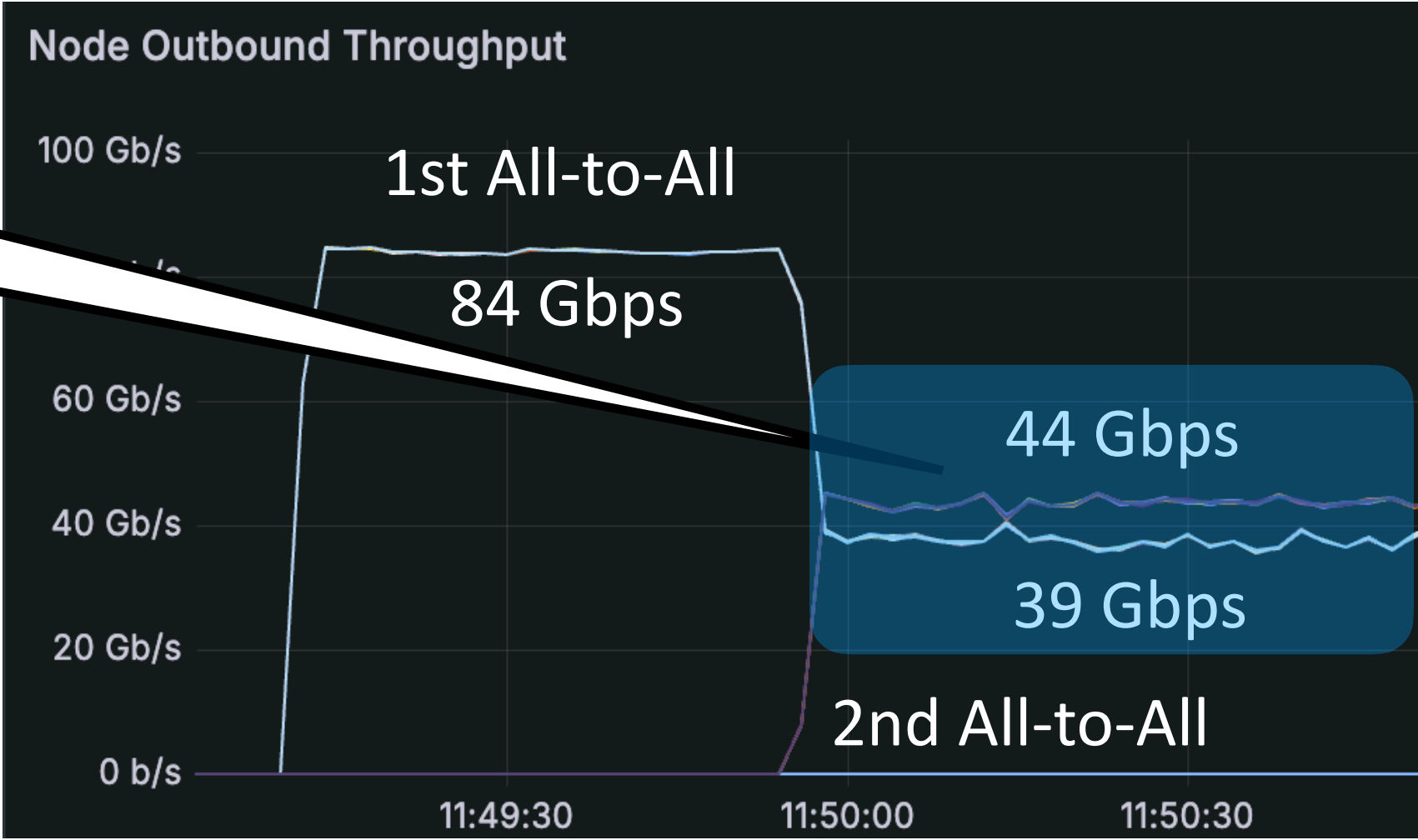


🐢 Queue pairs with high one way delays!

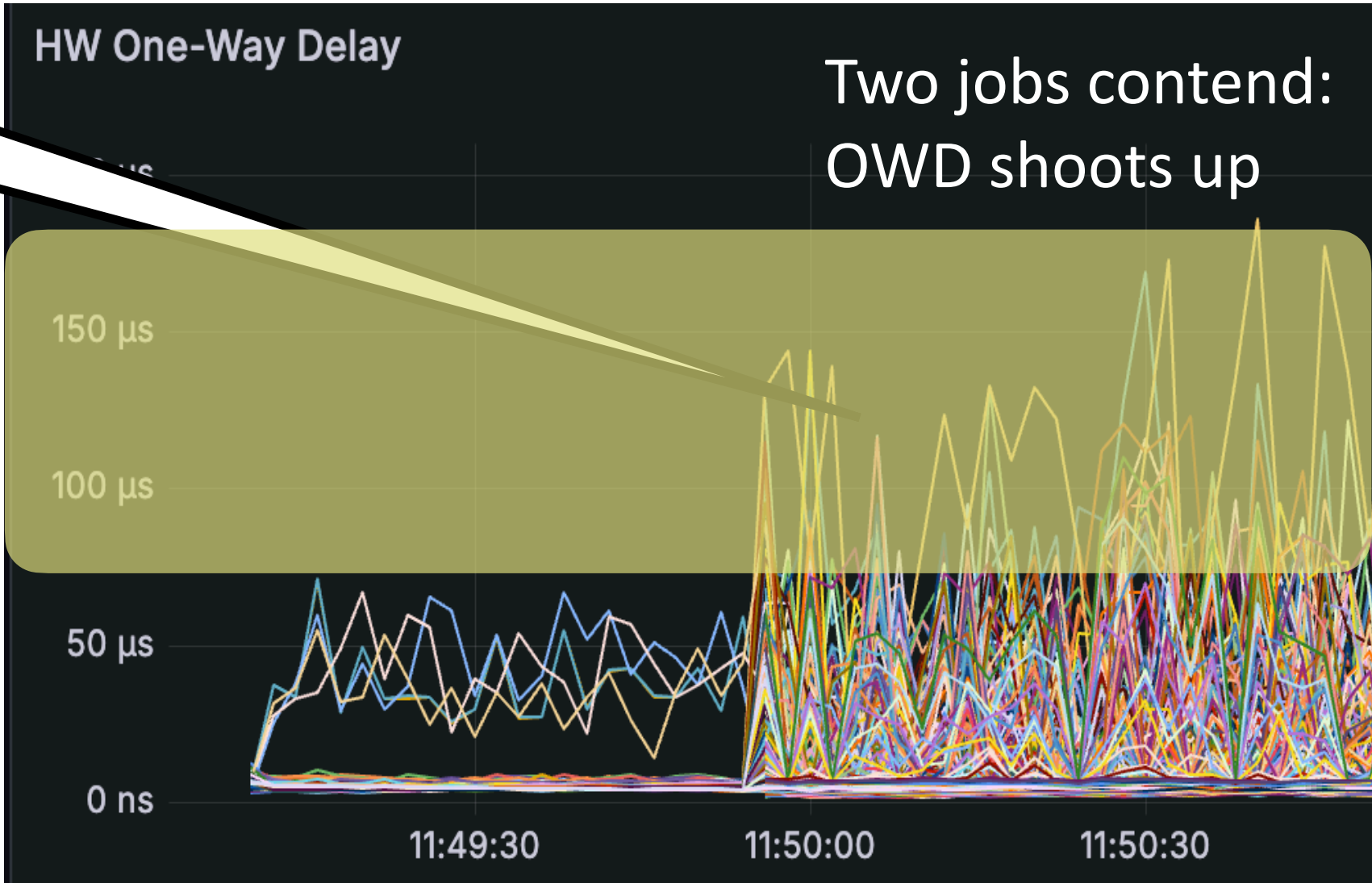


Contention on ECMP Test Bed

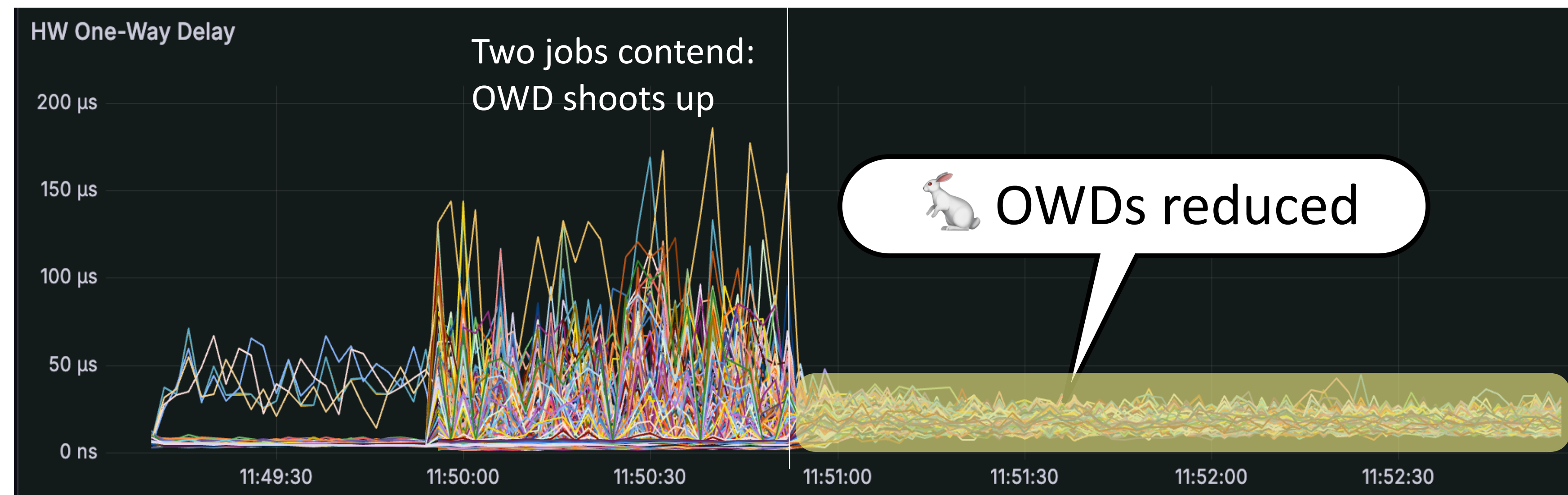
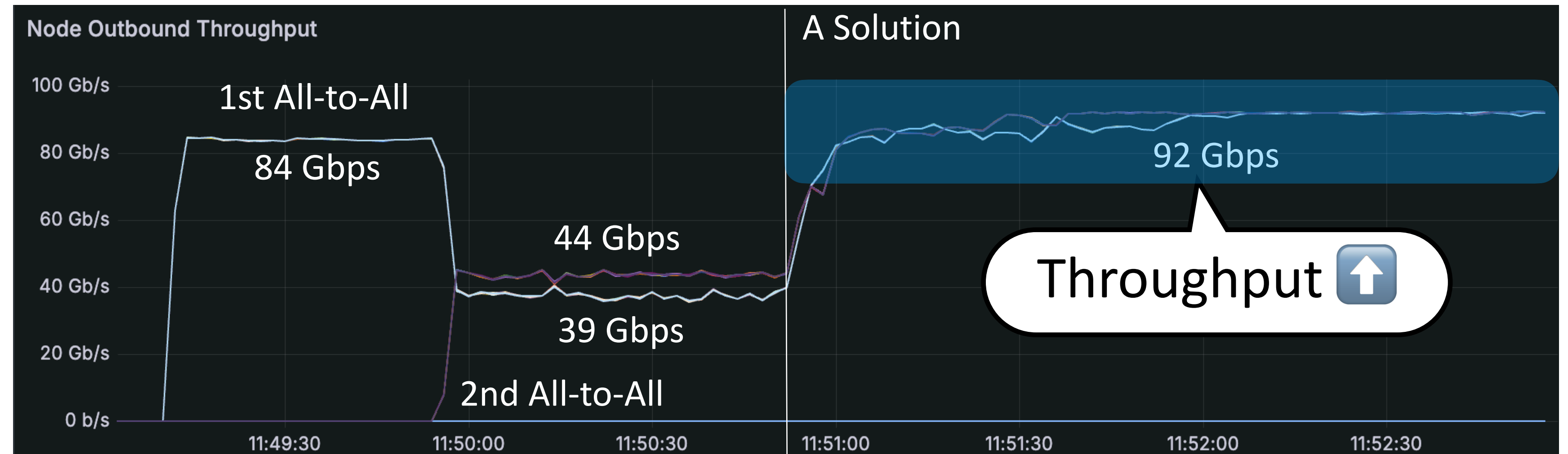
↓ Lower Throughput!



🐢 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.



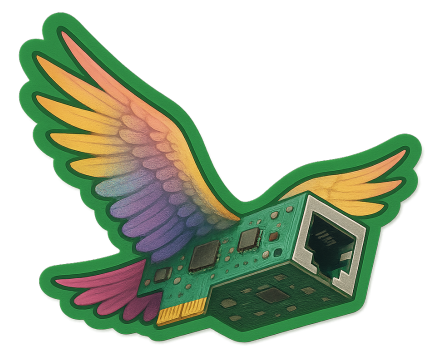
🙈 Visibility

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](https://www.linkedin.com/in/lerna)



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

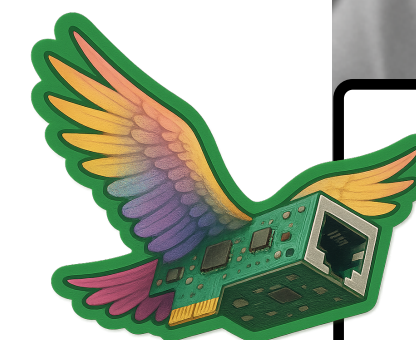


Thank You

[linkedin.com/in/lerna](https://www.linkedin.com/in/lerna)



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



Lucille Ball => Star Trek