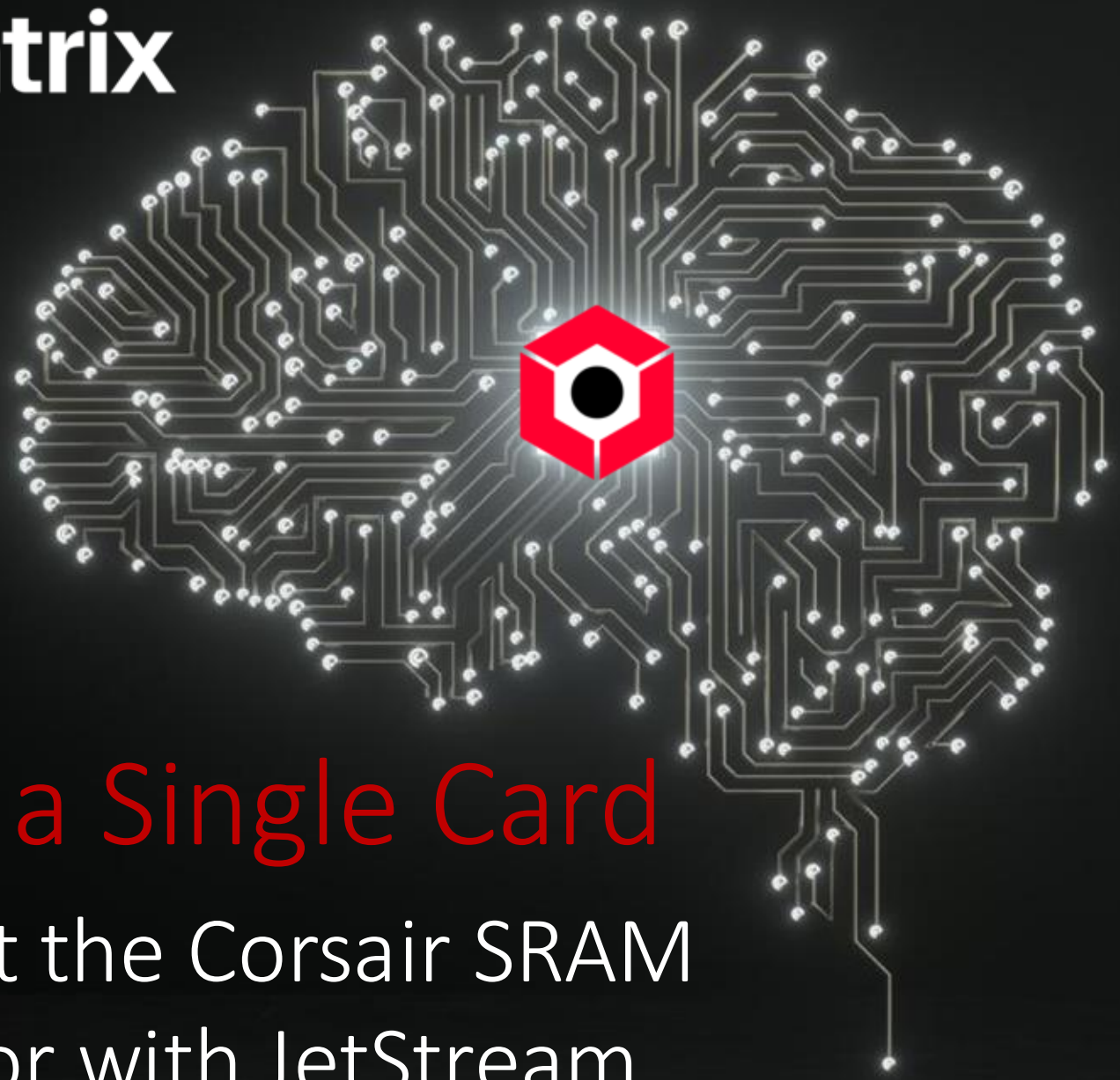
d-Matrix



Beyond a Single Card

Scaling Out the Corsair SRAM Accelerator with JetStream

Agenda

- **Corsair and JetStream** – Sai Rahul Chalamalasetti
- **JetStream Software** – David Karlov

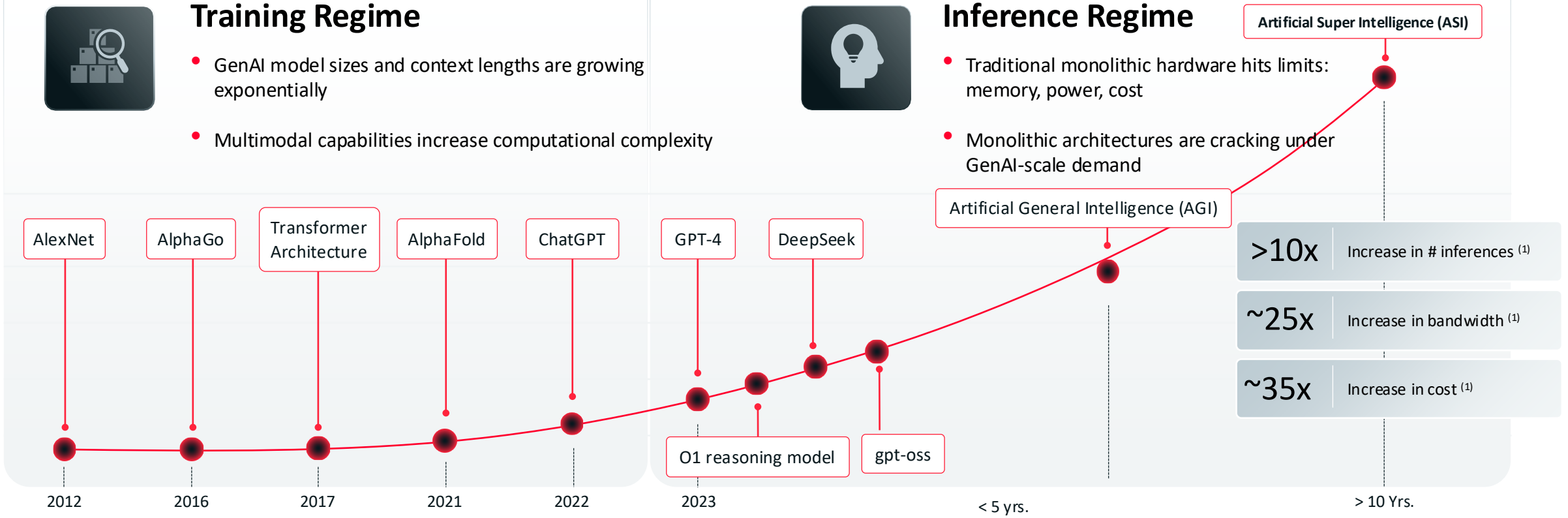
Rise of the Inference Regime

Training Regime

- GenAI model sizes and context lengths are growing exponentially
- Multimodal capabilities increase computational complexity

Inference Regime

- Traditional monolithic hardware hits limits: memory, power, cost
- Monolithic architectures are cracking under GenAI-scale demand

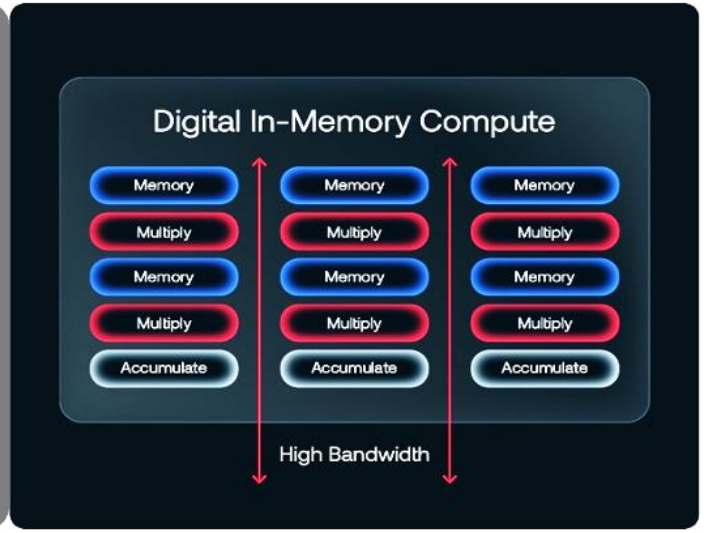


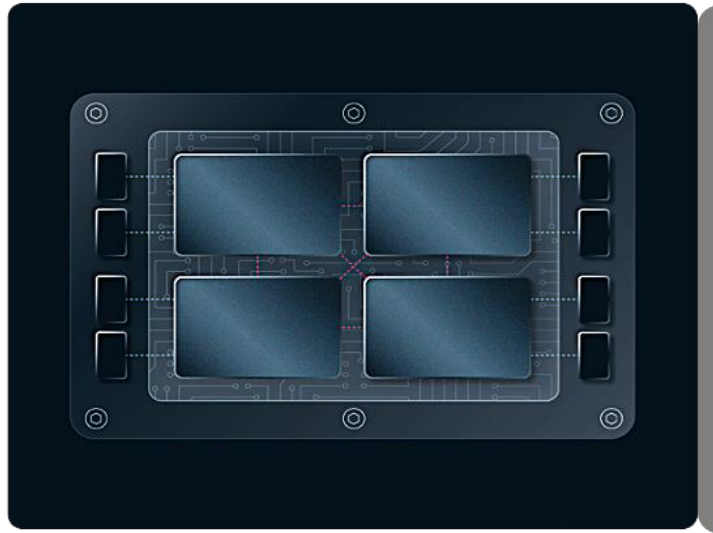
One shot → Multi-turn → Reasoning → Agentic …

Building a System for Efficient AI Computing

Enormous memory bandwidth

High capacity



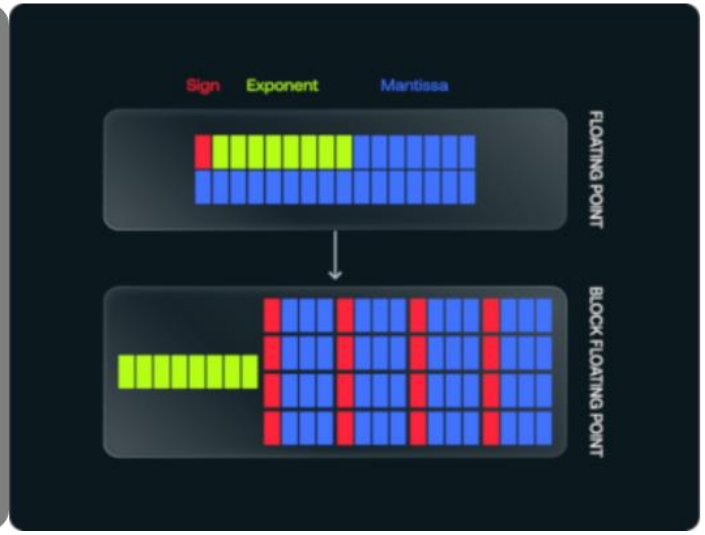


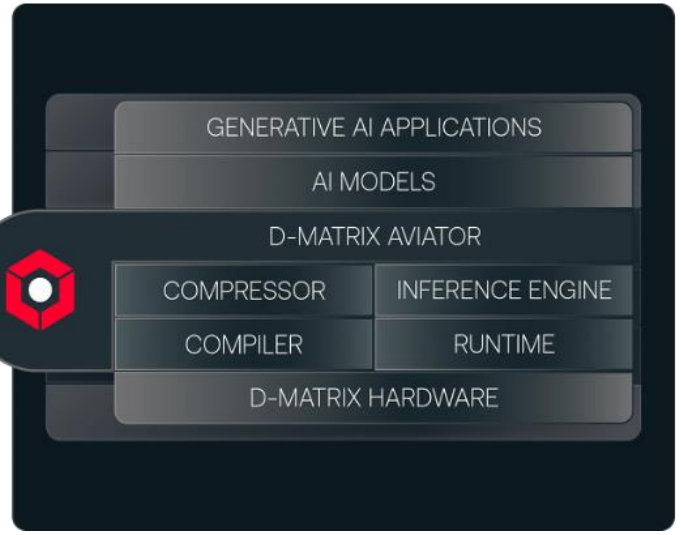
Scaling and yields

Efficient interconnections

High compute TOPS

Stable low precisions





Perf and flexible SW

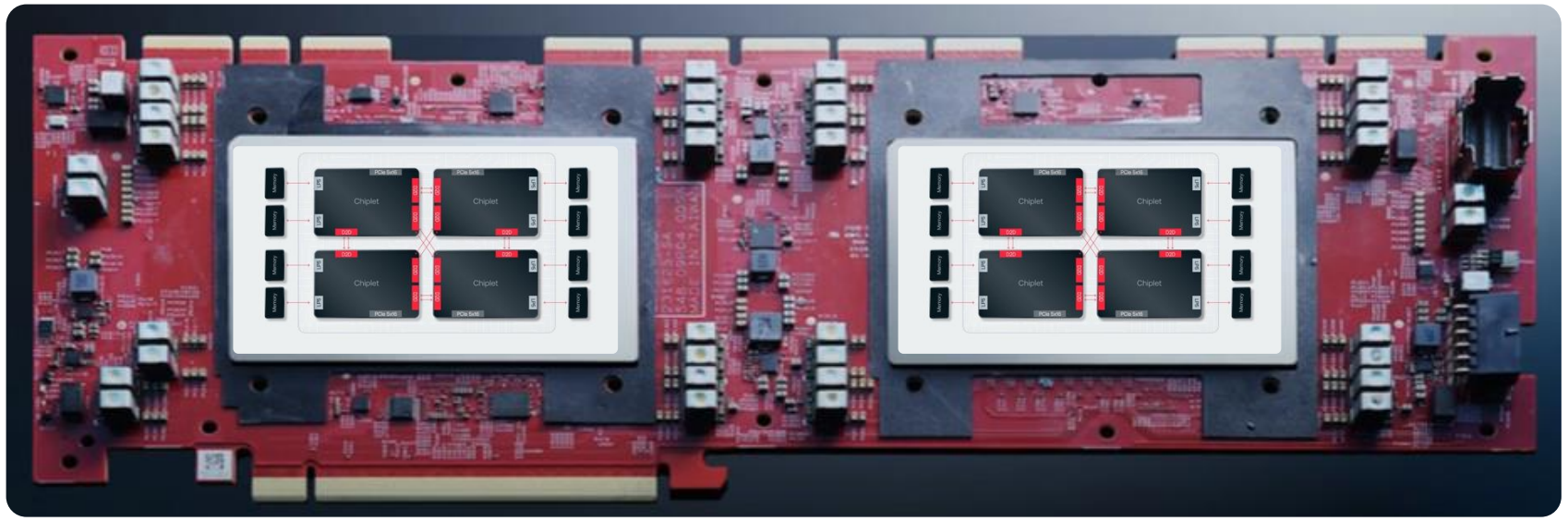
Custom features for AI

Corsair

8 Chiplets
TSMC 6nm

2400 TFLOPS MXINT8
9600 TFOPS MXINT4

PCIe: 128 GB/s
DMX Bridge: 512 GB/s



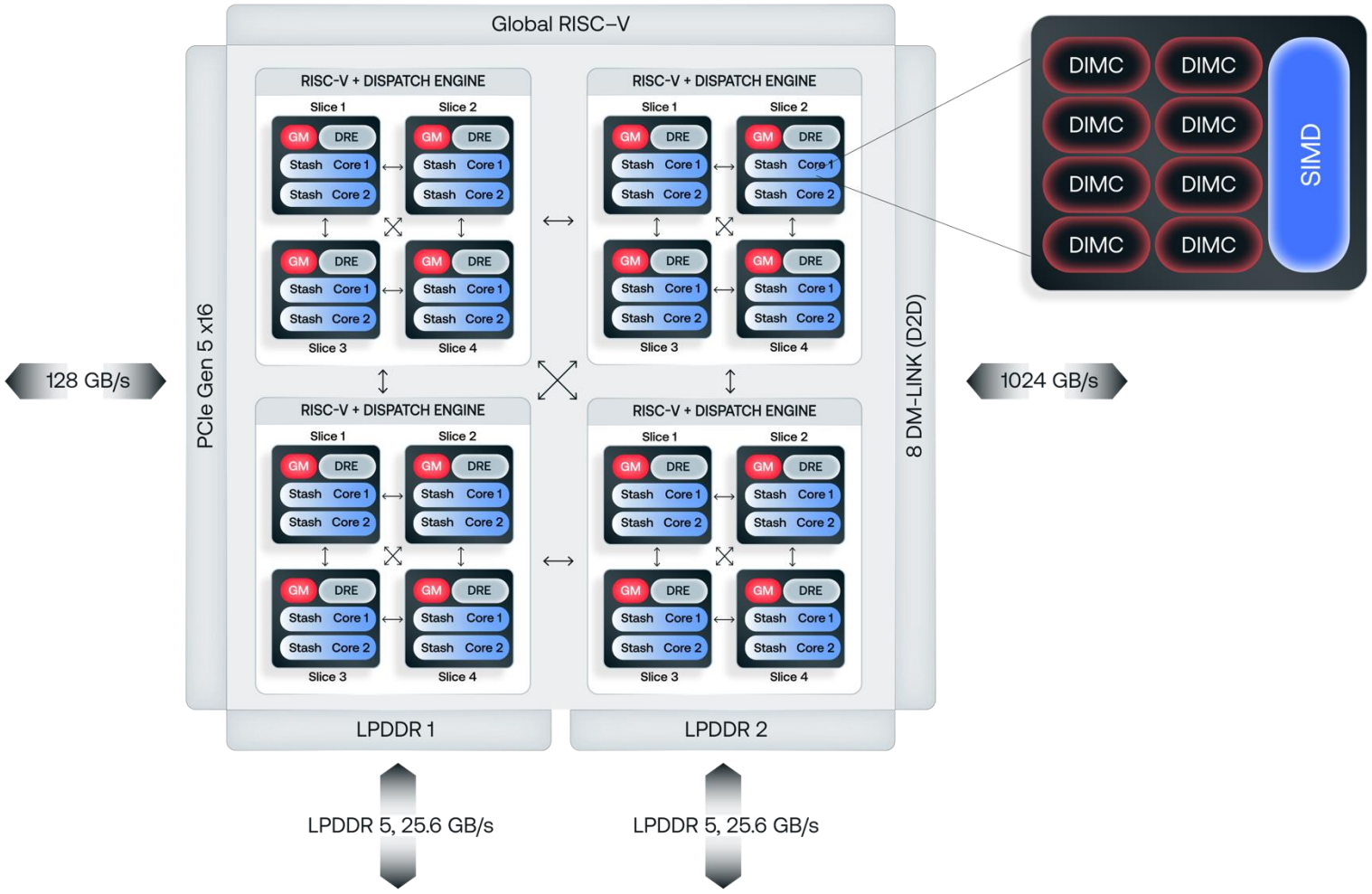
Deliver Inference with sub-millisecond latency

PCIe Gen5
600W TDP

2GB SRAM
Memory 150 TB/s

256GB LPDDR5
400 GB/s

Corsair Chiplet: Built Using Modular Hardware Blocks

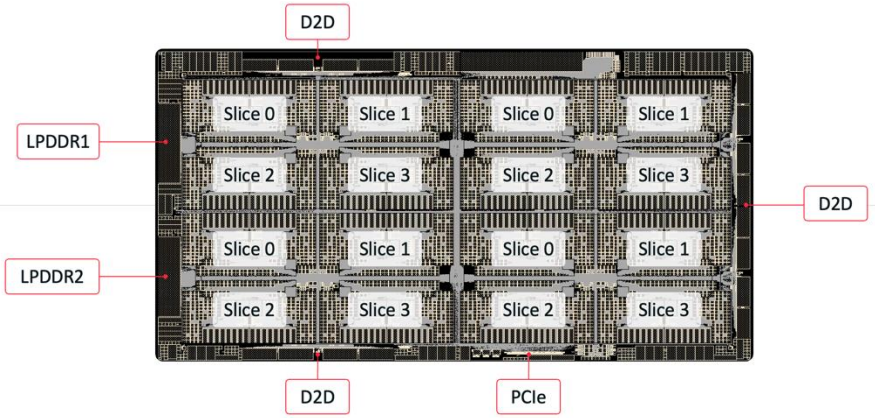


- Corsair SOC Chiplet
  - 4 Quads
  - Global RISC V Control Core
  - LPDDR & PCIe Subsystems
  - D2D
- Quad
  - Locus of control, RISC-V + DE
  - Ease of Programmability
  - 4 Slices clustered with a RISC V Control Core
  - Instruction Dispatcher (DE)
- Slice
  - 2x Cores sharing a Global Memory (GM)
  - Data Reshaper, Low-Latency NOC

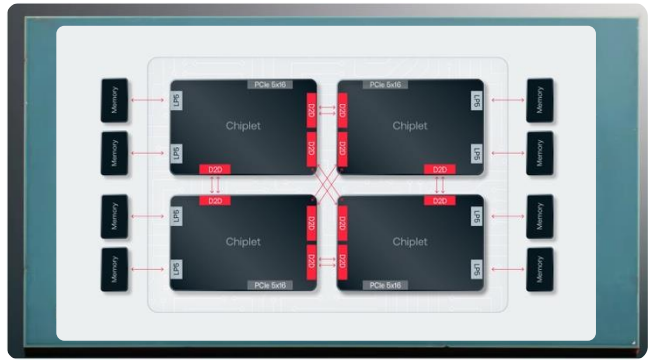
d-Matrix Connectivity Quadrant

(Different Transports and Push Mode Communication)

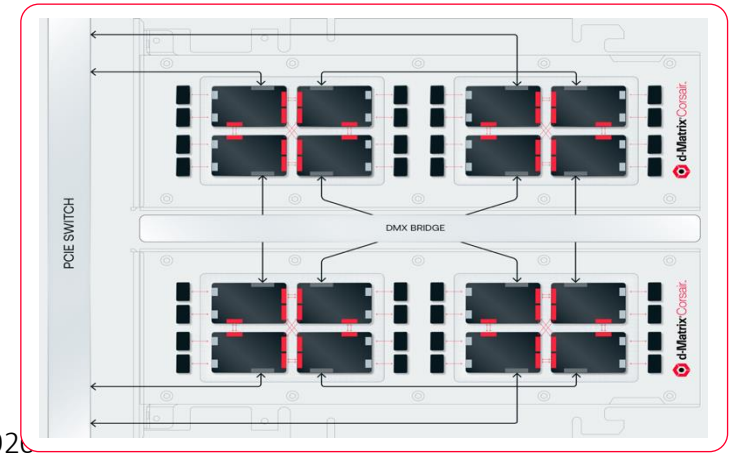
Chiplet (NOC)



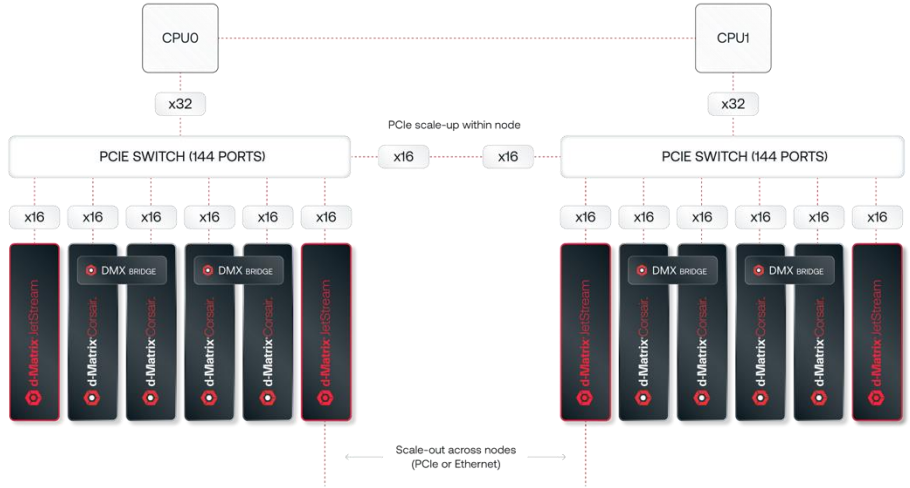
Package (D2D)



Card (TP-1 and TP-2) (PCIE)



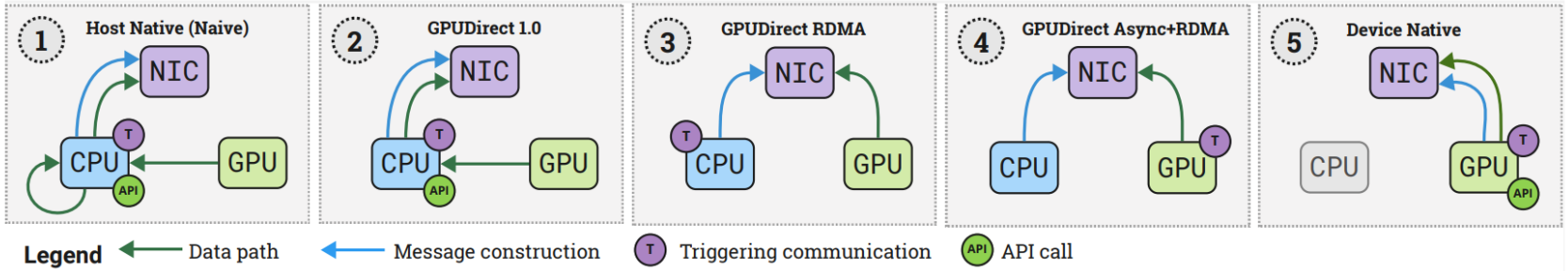
Node (PCIE)



Why does communication have to be different across Nodes ?

Type	API	Register	Trigger	Data path	Examples
① Host native	Host	Host	Host	Through host (2 copies)	GPU-Aware MPI before ②
② GPUDirect 1.0	Host	Host	Host	**Through host (1 copy)**	GPU-Aware MPI, NCCL
③ GPUDirect RDMA, ROCnRDMA	D/H	Host	Host	**Direct**	GPU-Aware MPI, NCCL, NVSHMEM
④ GPUDirect Async	D/H	Host	**Device**	Depends on ③	GPU-Aware MPI, NVSH-MEM
⑤ Device native	Device	**Device**	Device	Depends on ③	NVSHMEM with IBGDA, GPUrdma

Table 2. Types of inter-node communication methods. Cells in bold refer to where a change or optimization has been made. *D/H* means that both device-side and host-side API calls may belong to this type.



The Landscape of GPU-Centric Communication, https://arxiv.org/pdf/2409.09874v2

d-Matrix JetStream™

Purpose-built I/O for AI Inference Scaling

PCIe Gen5
150W TDP

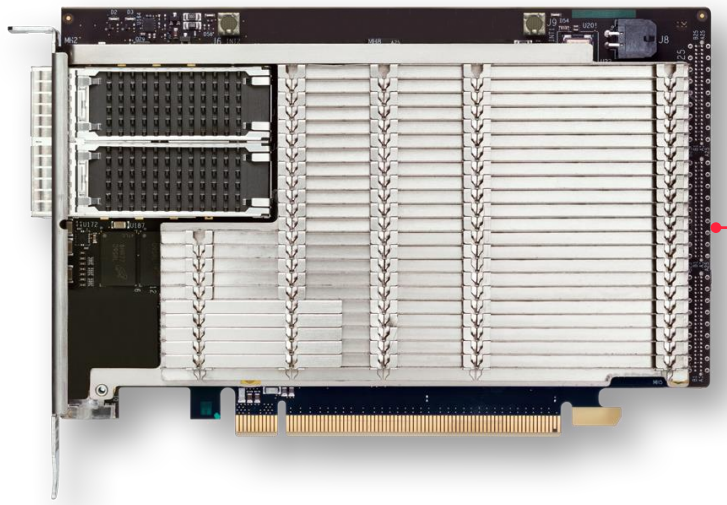
IEEE 802.3
Ethernet

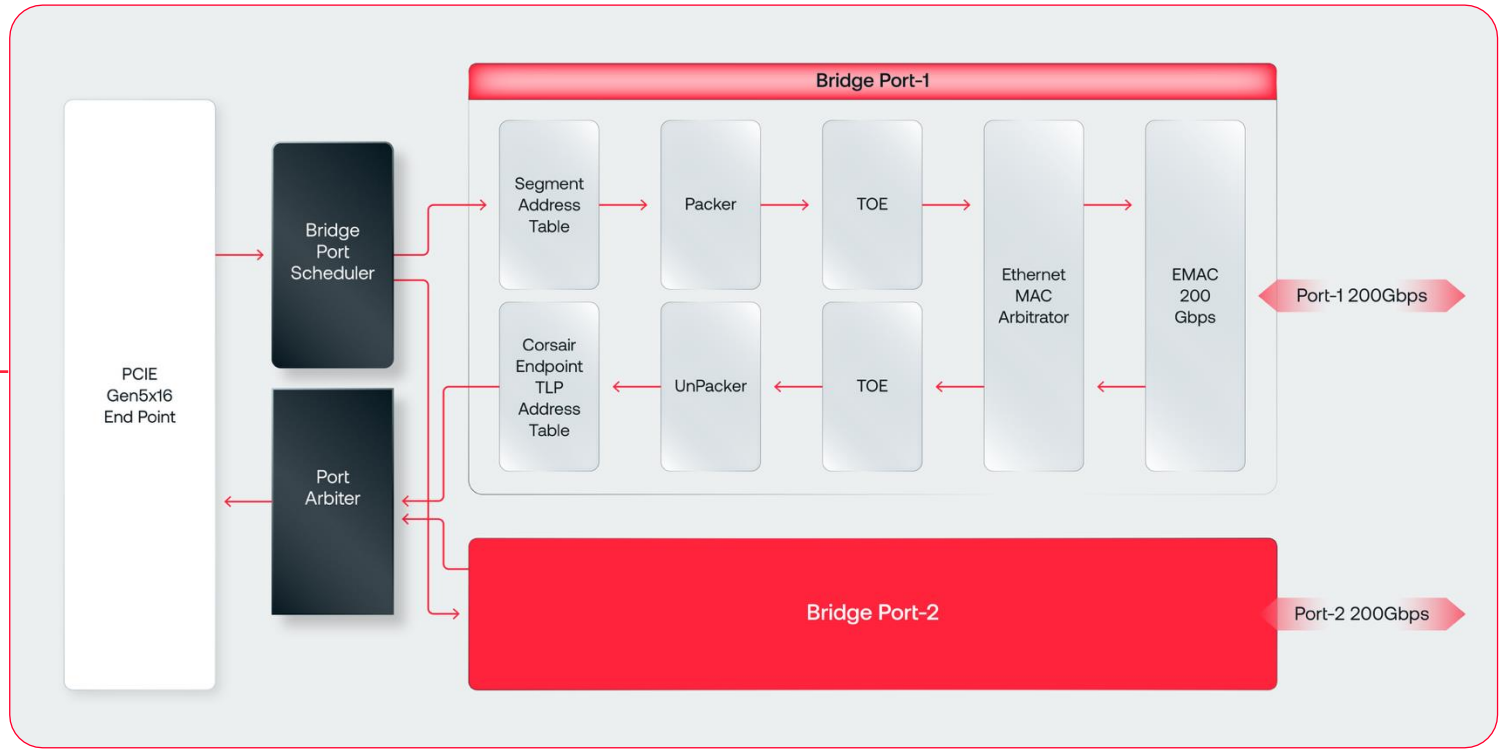
Latency 2us

400 Gbps

2x200Gbps

JetStream NIC

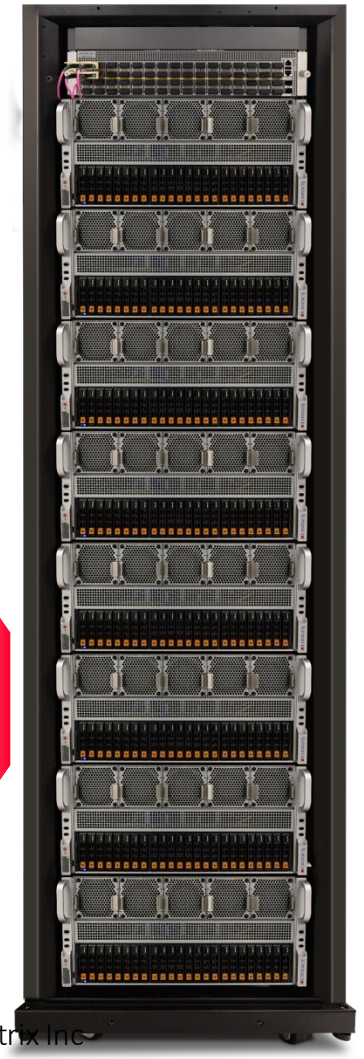




- Transport PCIE semantics over Ethernet
- Scale communication semantics from intra-chiplet to inter-node
- Lower latency vs RDMA

SquadRack™

Open Rack-scale Reference Solution Purpose-Built To Speed Up Time to Inference

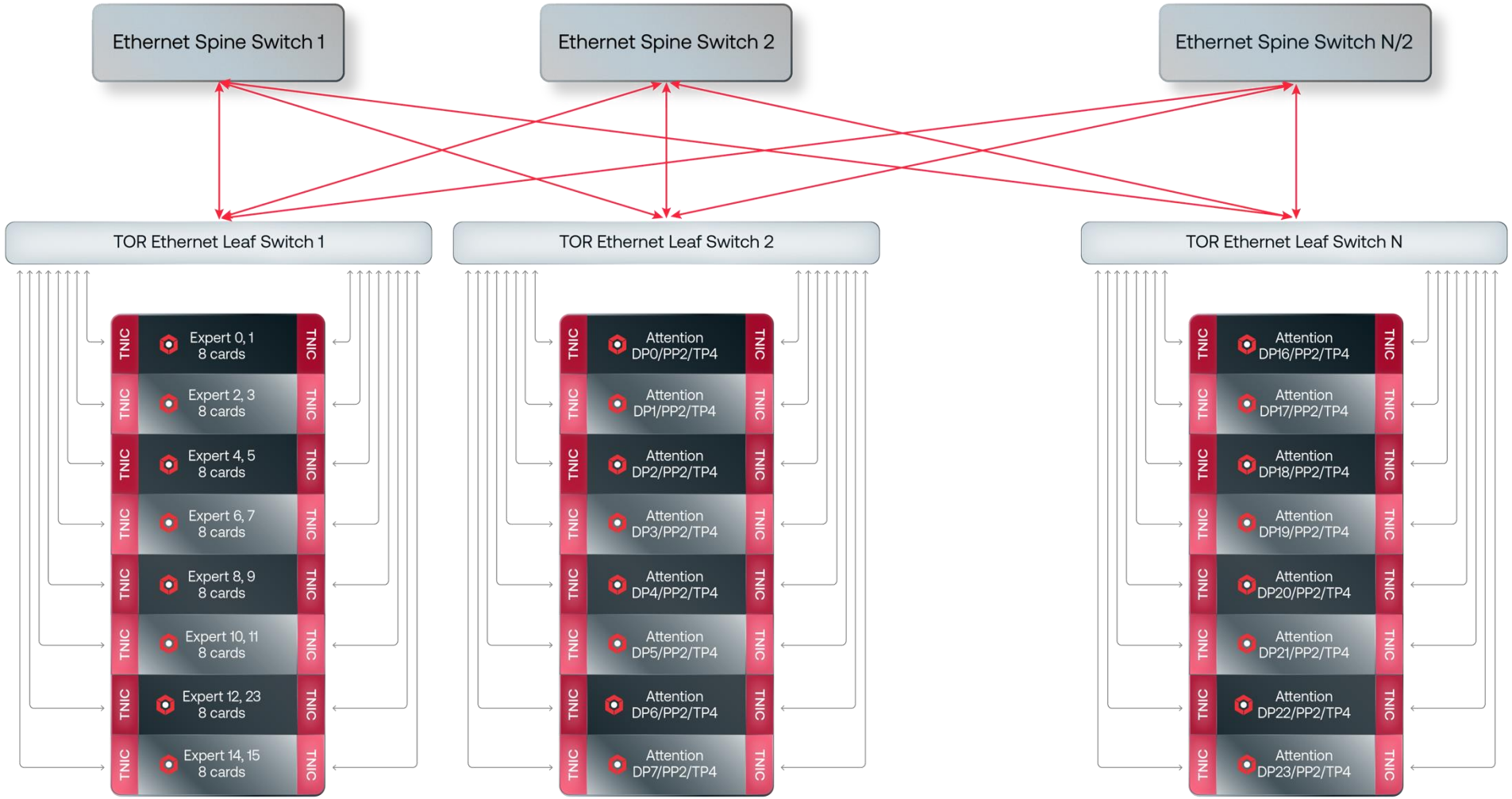


Reference Solution
Built in Collaboration With

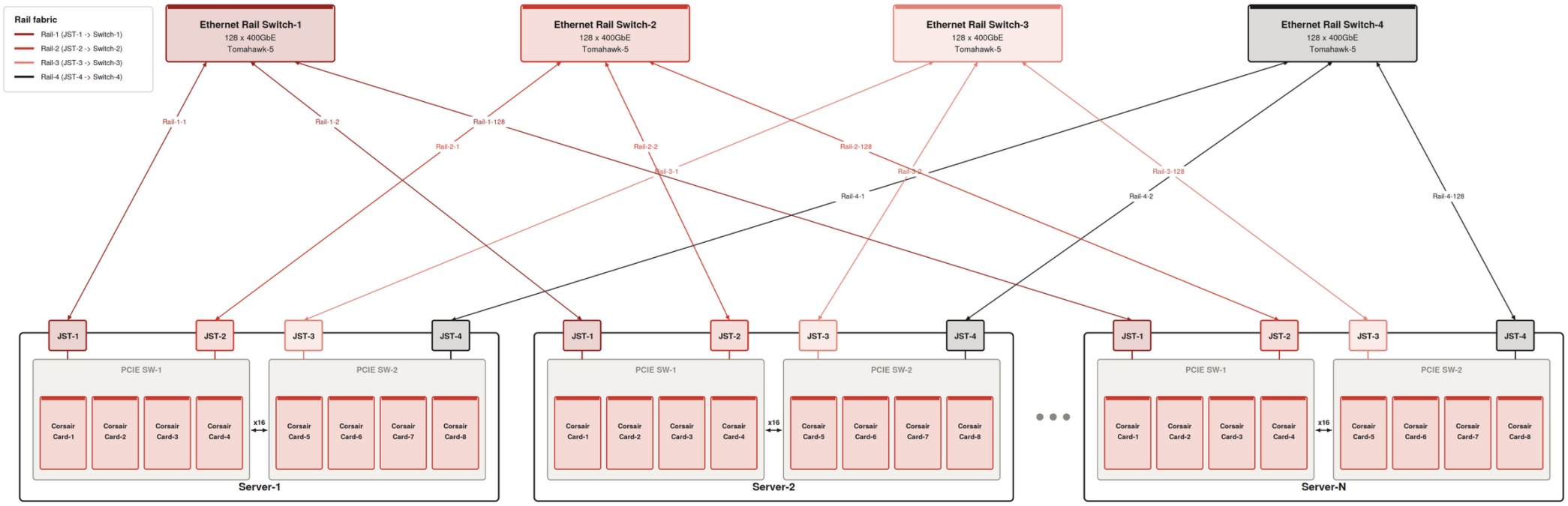
SUPERMICRO ARISTA BROADCOM

Air-cooled
Open Standards Based
No vendor lock-in
Customizable to fit inference demand
In Existing Datacenters

Rack Level Scale-Out: Multi-node and Multi-Rack



Rail Only Topology for Low Latency



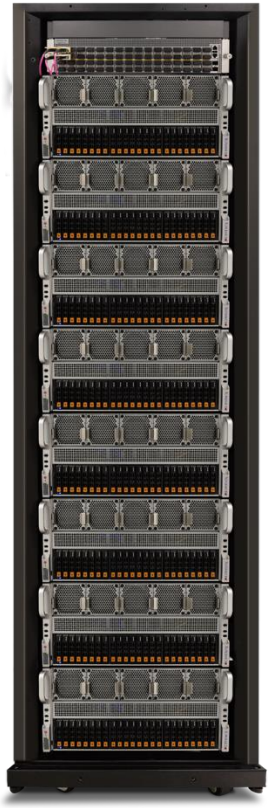
Use TP/PP/EP inside a Node
Use PP/EP between Nodes through Ethernet Switches
Use Push Mode to communicate across rails

SquadRack

Clip, Whisper
GPT-OSS-20B

Half-rack

GPT-OSS-120B



Full-rack

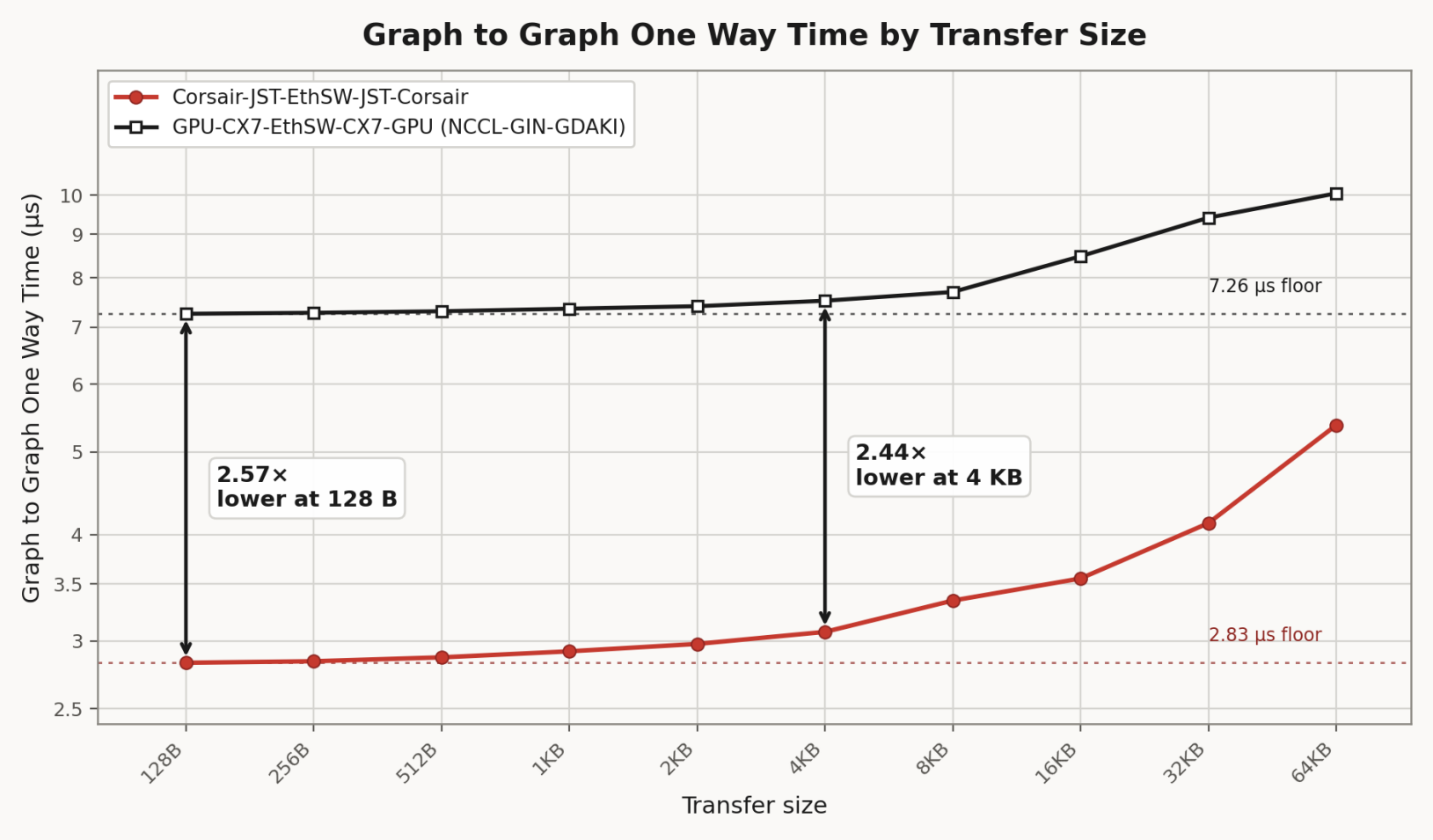
Inference at scale w/ Kubernetes pods

Leaf-spine cluster

2T parameter MOE

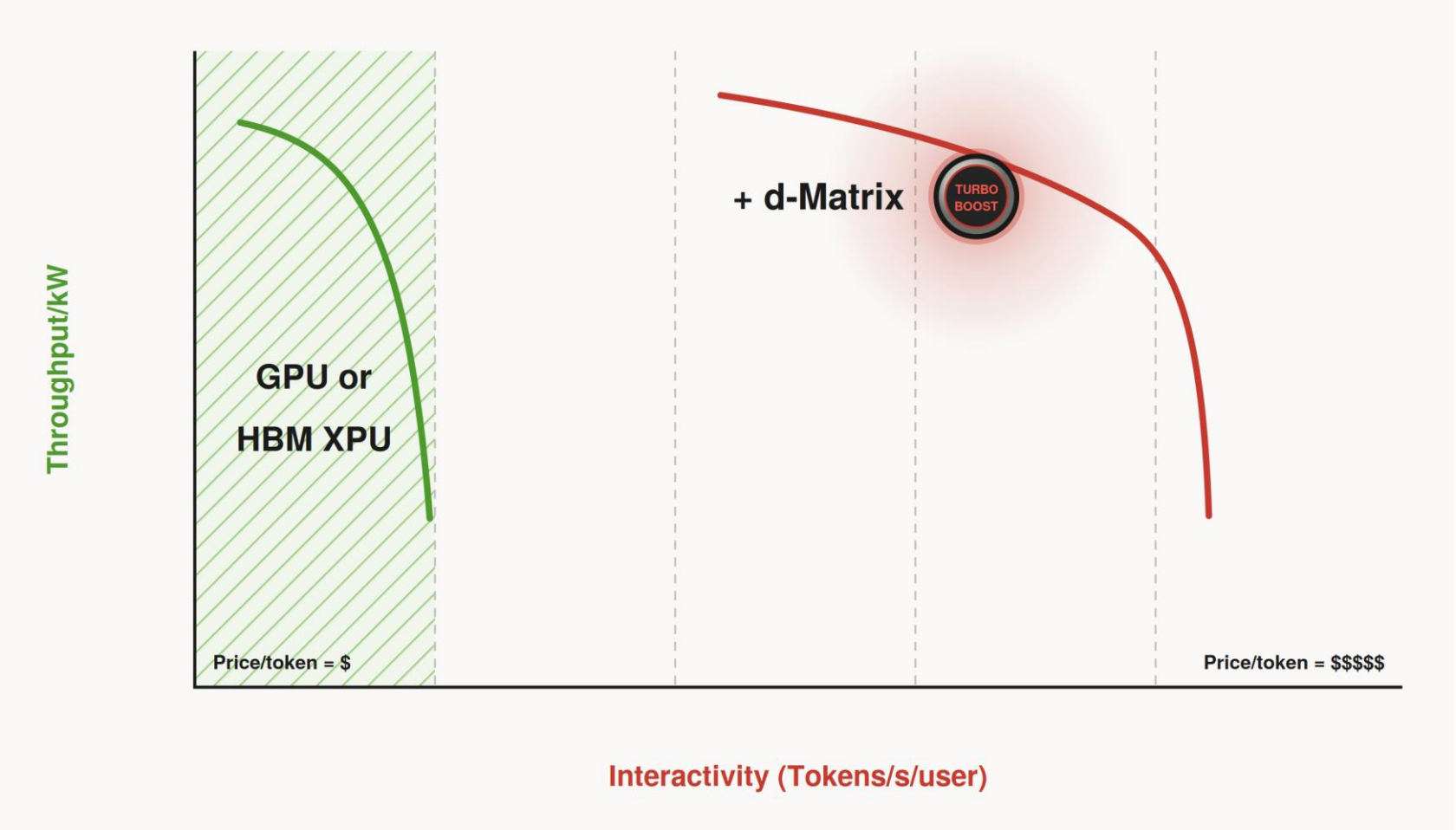
Rail-optimized topology

Graph-to-Graph Transfer and Synchronization

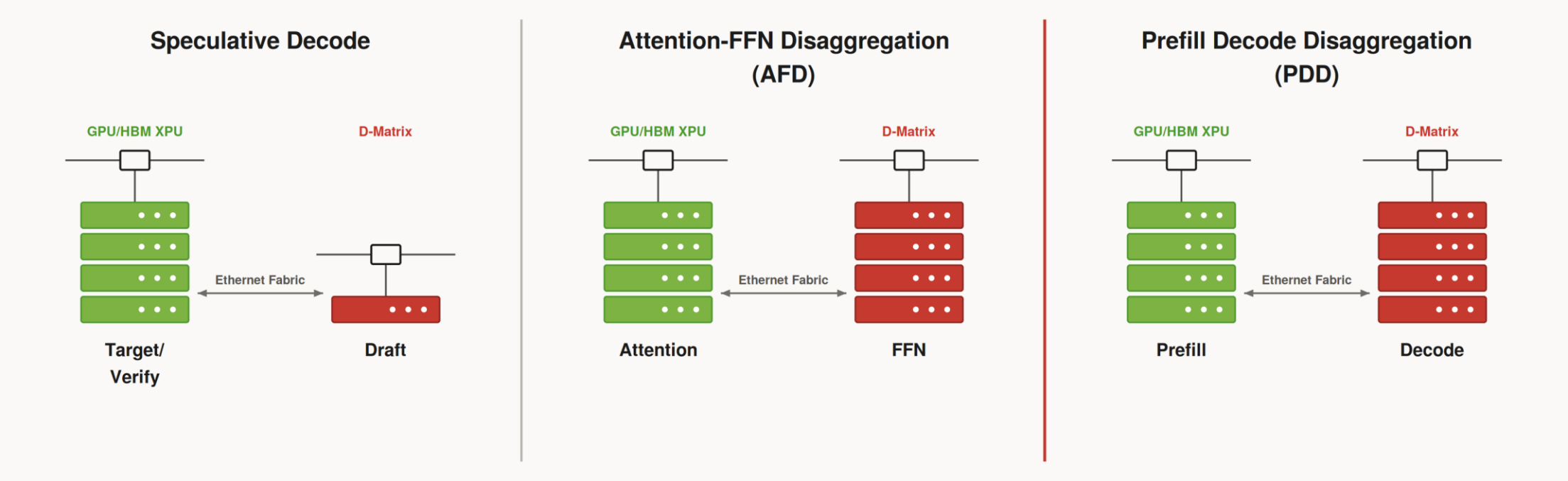


Enabling the Premium (Fast) Token Economy

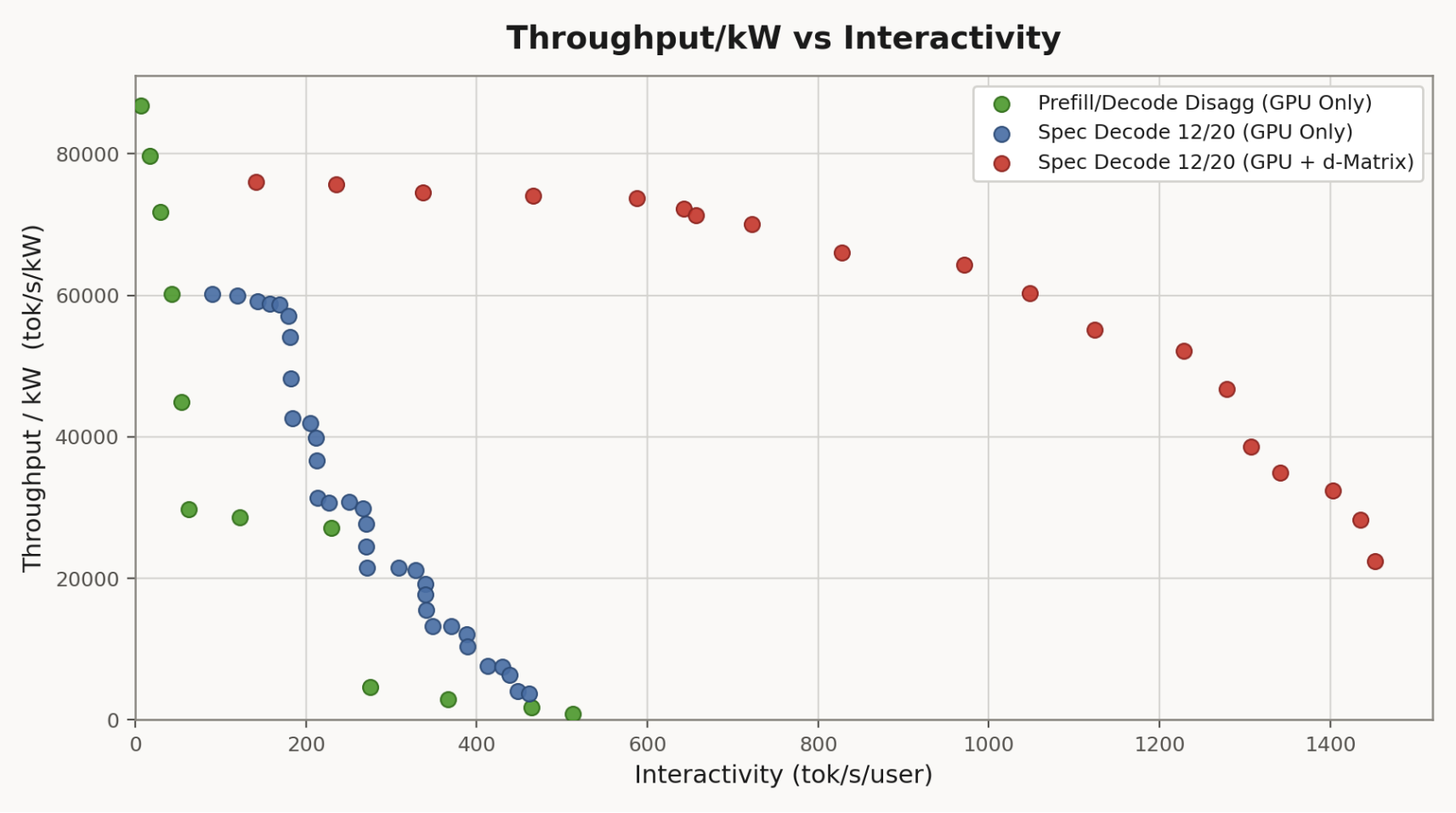
d-Matrix Gives Boost Needed for Higher Interactivity



Heterogeneous Compute: Disaggregated Fast LLM Decode

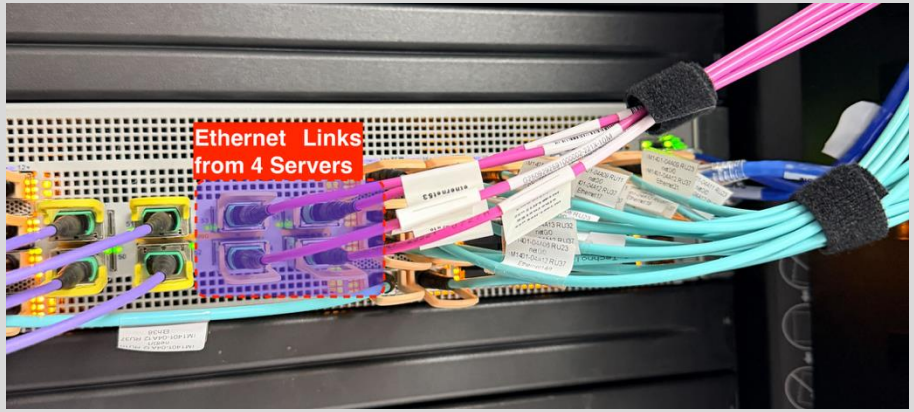


Heterogenous deployment: B200 + Corsair



GPT-OSS-120B

Datacenter Setup (2-B200s + 2-Corsair Servers)



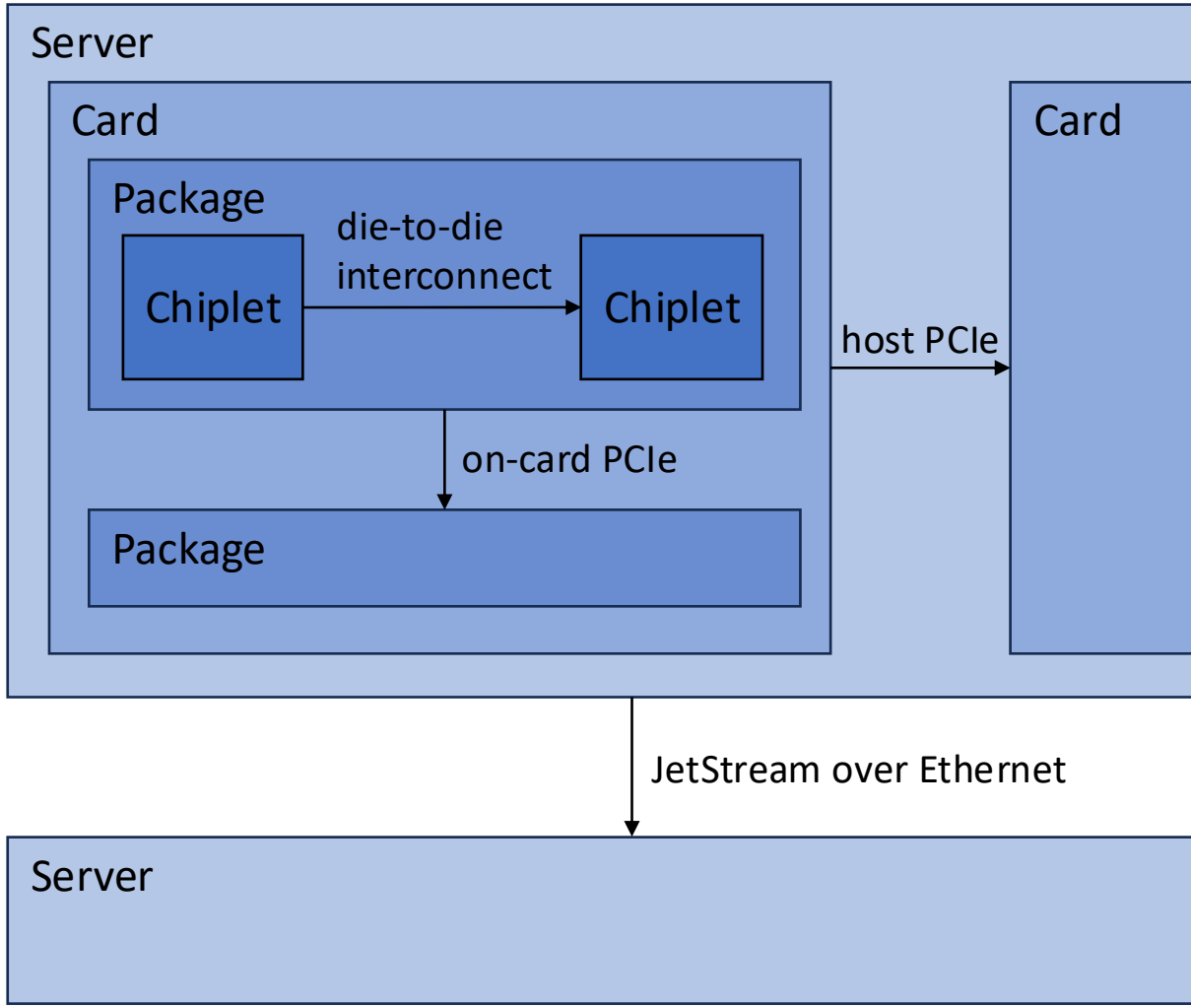
https://www.linkedin.com/posts/check-out-our-gpu-d-matrix-corsair-server-share-7445515001798438913-Dhdy/?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAGaTfsBzMZiq5lpOUCtoA5jjZiEPy3yjbc

JetStream software stack - Overview

- JetStream software:
  - Enables transparent addressing of any destination across any transport
  - Provides one addressing model, consistent from chiplet to rack

The addressing problem

- A transfer's destination may be another package on the same card, or a card across a rack
- Device addresses and physical paths differ completely at each scale
- Workload must not be rewritten or regenerated for each physical placement
- Goal: one addressing abstraction; transport resolved transparently underneath



Inference workloads on Corsair

- Corsair writes across PCIe are declarative instructions in a workload (e.g., MOVE_DDR2PCI)
- PCIe destination is expressed as a symbolic handle, not a hardcoded address
- Handles resolve to physical addresses when workload is executed
- Same instruction stream regardless of where the destination physically sits

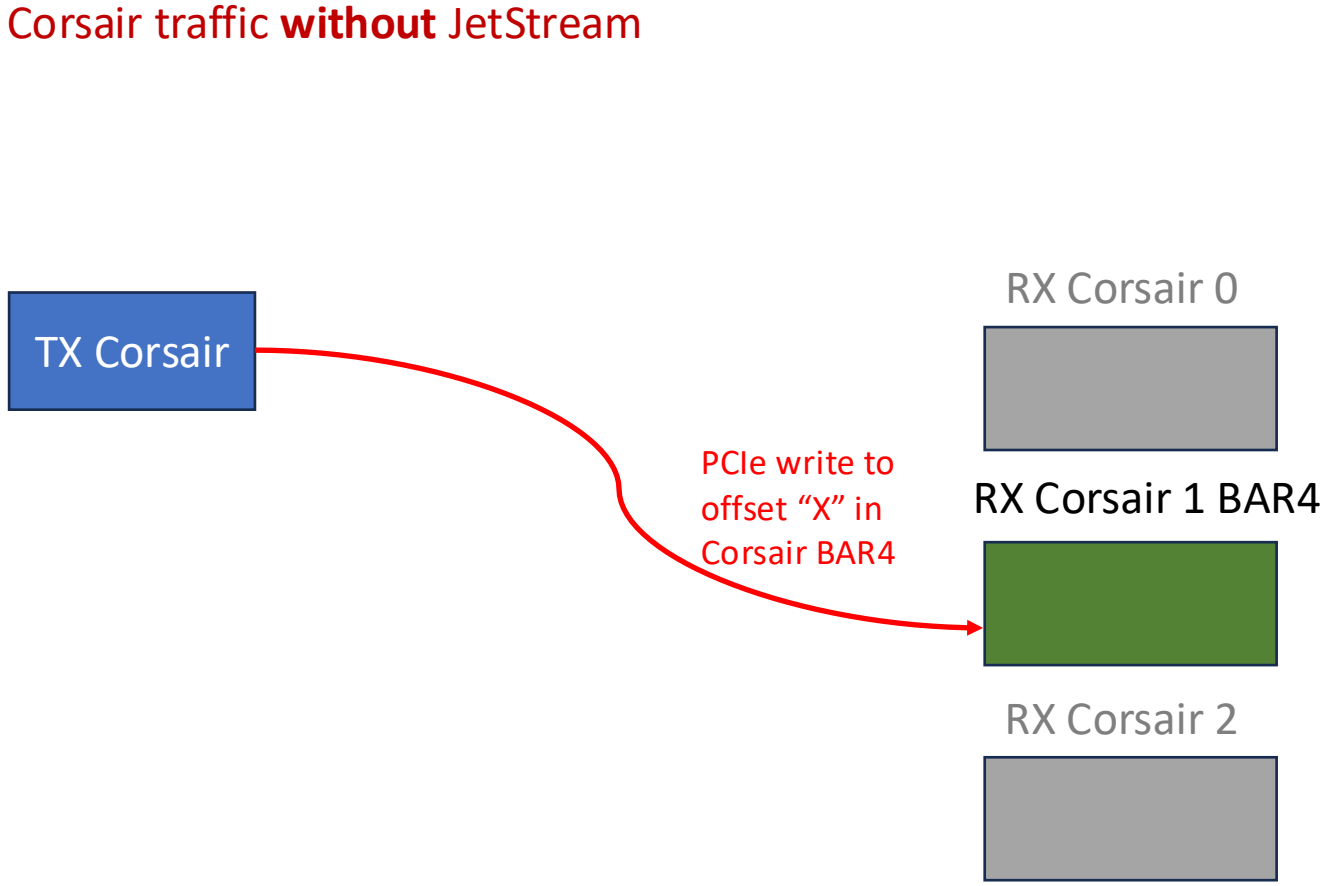
```
 {
  "ID": 4,
  "OP": "MOVE_DDR2PCI",
  "REM_PCIE_ADDR": "pci_r1_m1_c0",
  "DST_CHIPLET": 0,
  "TRF_SIZE": 2048,
  ...
},
```

Handle	Address
pci_r1_m0_c0	?
pci_r1_m1_c0	?
pci_r2_m0_c0	?
...	?

Addresses unknown in workload

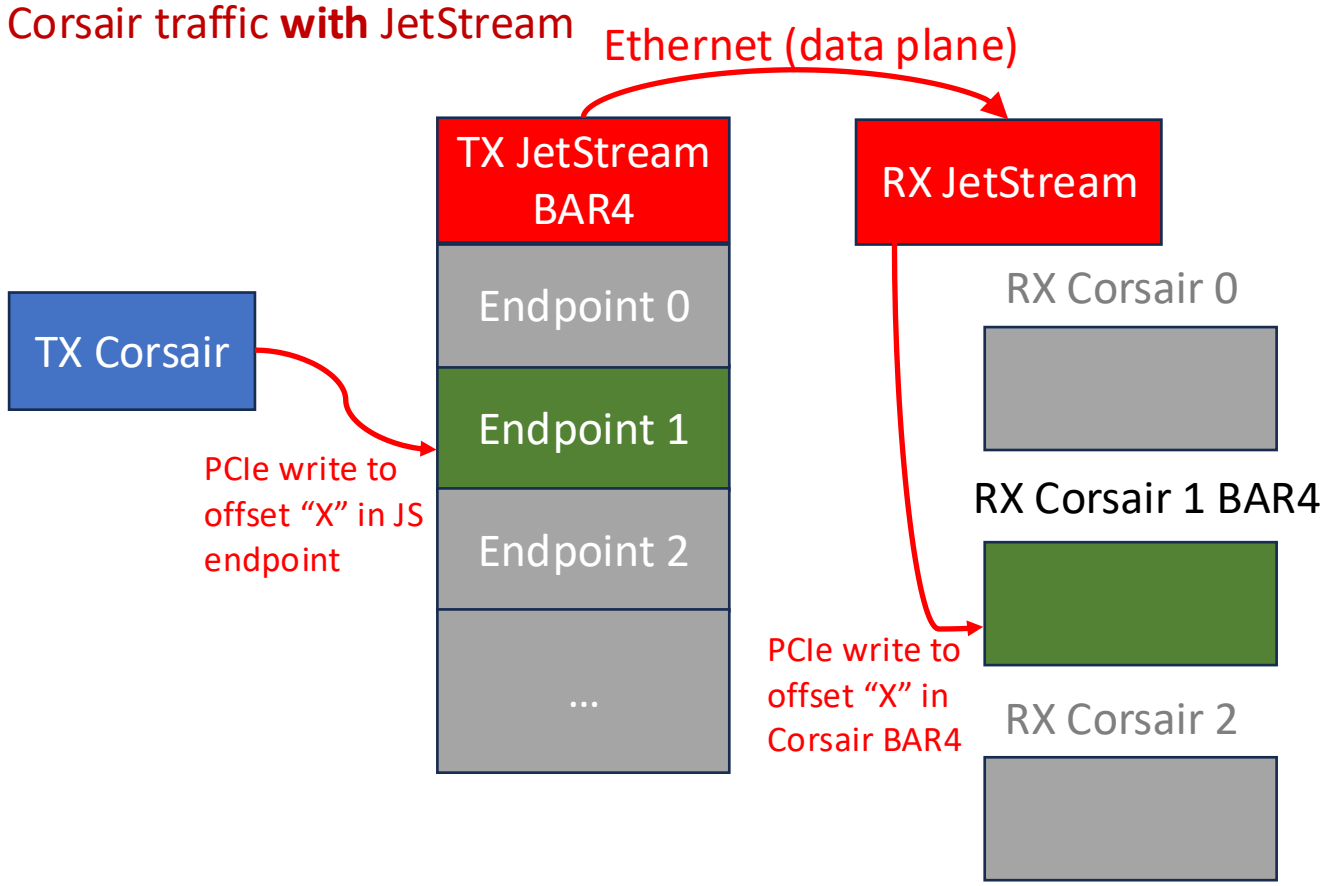
Mirrored BAR addressing

- Full address space of multiple remote Corsairs are mirrored in a local JetStream's PCIe BAR
- Transfers to another Corsair across Ethernet look like local PCIe writes



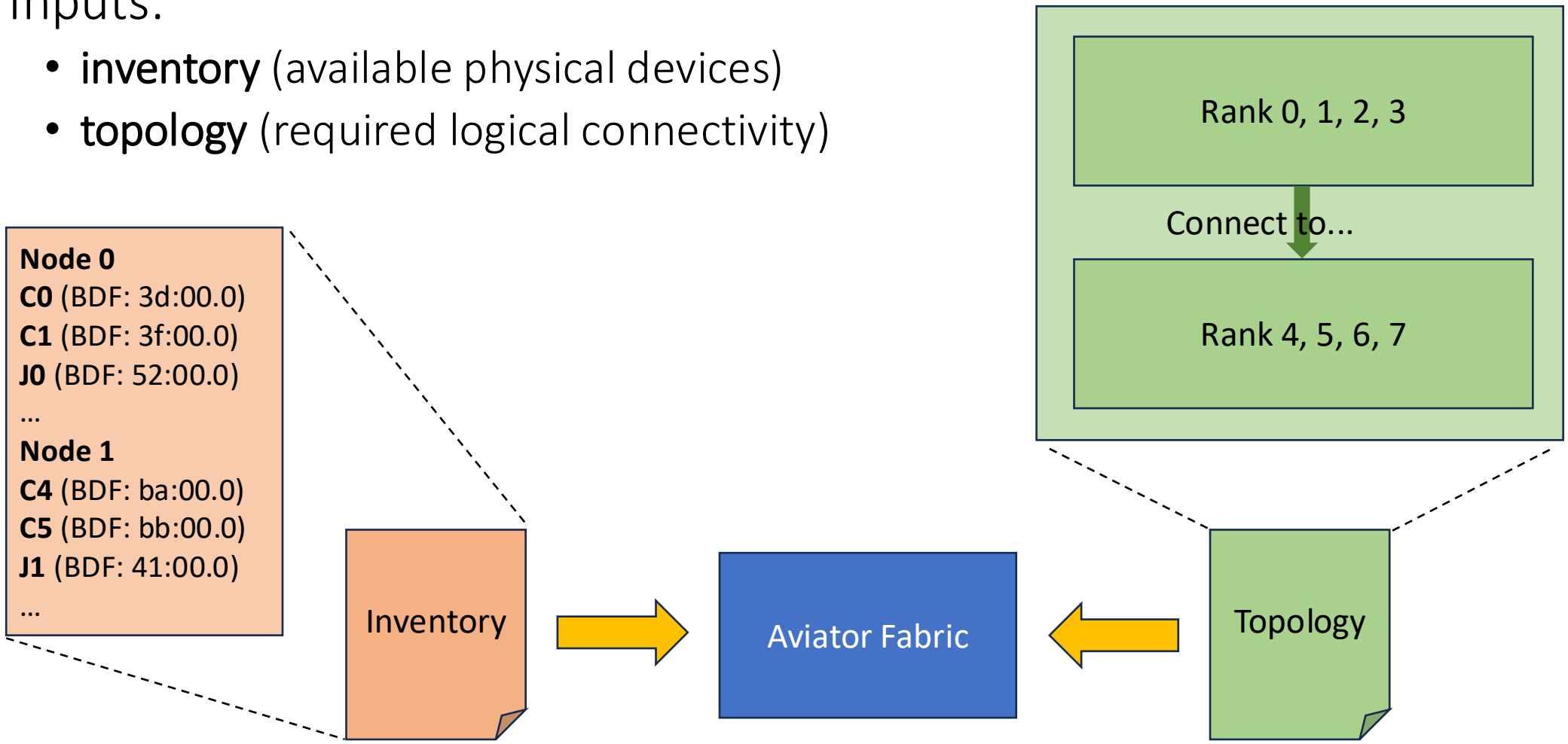
Mirrored BAR addressing

- Full address space of multiple remote Corsairs are mirrored in a local JetStream's PCIe BAR
- Transfers to another Corsair across Ethernet look like local PCIe writes



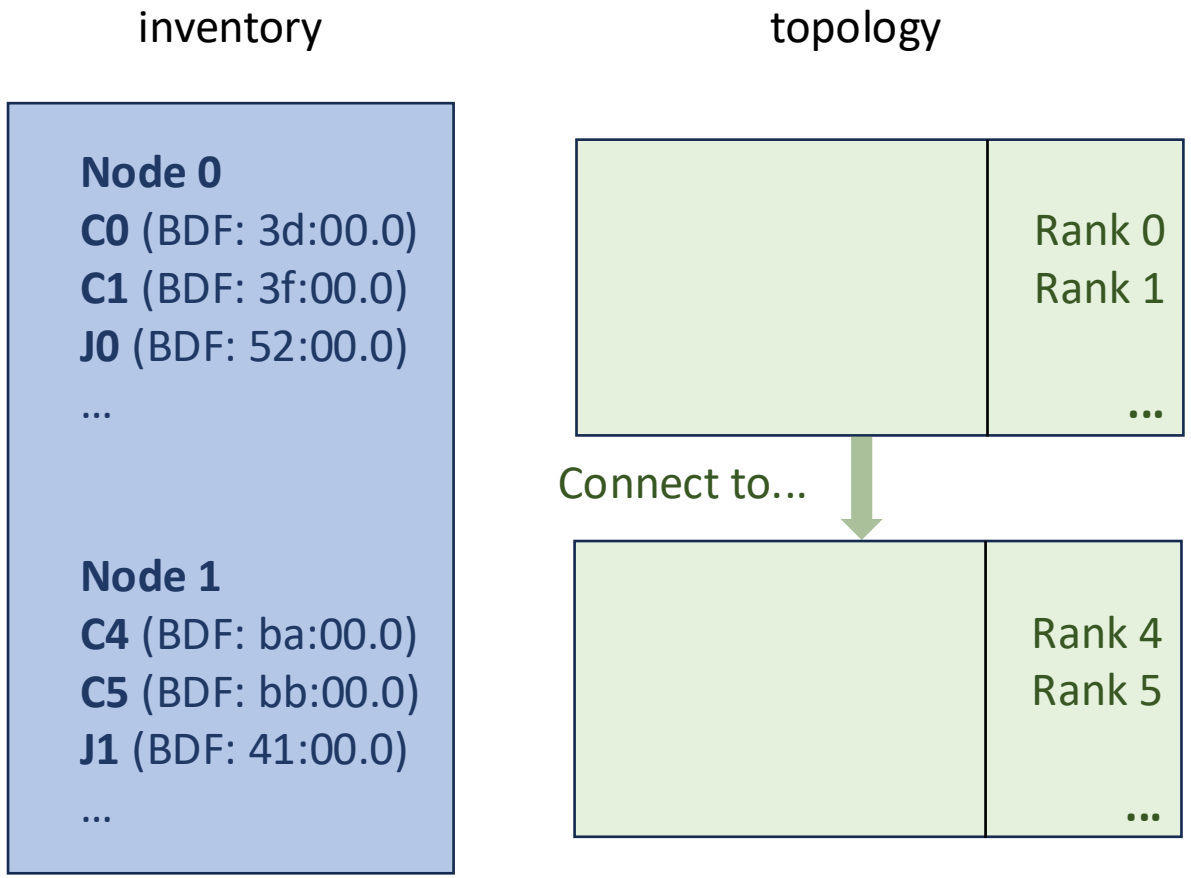
Control plane: Aviator Fabric

- Inputs:
  - **inventory** (available physical devices)
  - **topology** (required logical connectivity)



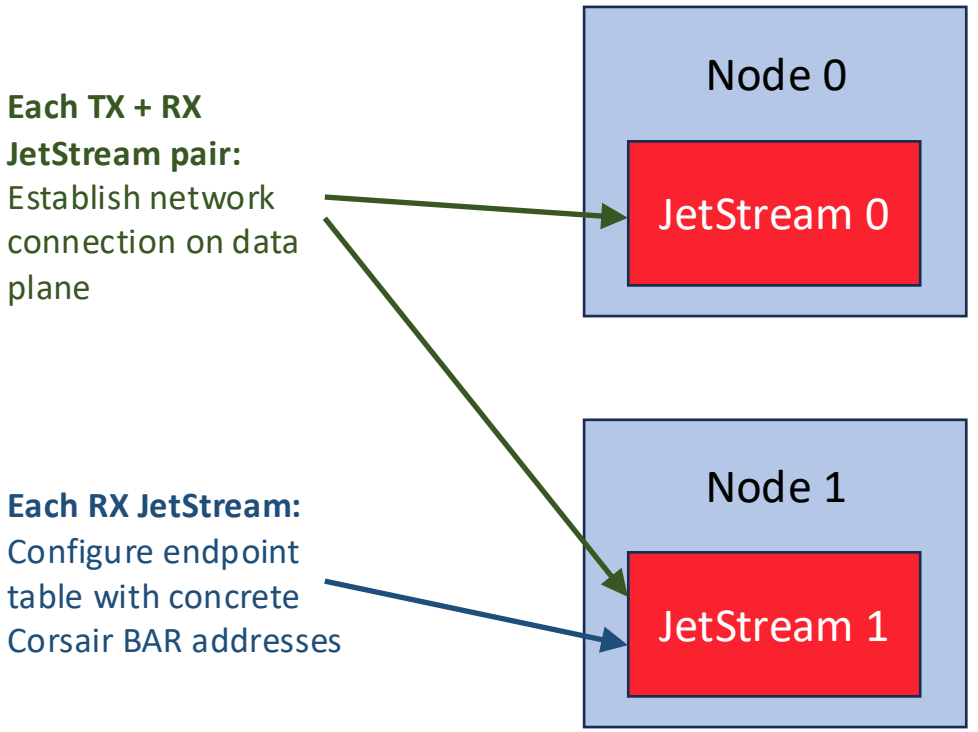
Control plane: Aviator Fabric

- Inputs:
  - **inventory** (available physical devices)
  - **topology** (required logical connectivity)
- Operation:
  1. Map topology requirements onto physical devices and connections



Control plane: Aviator Fabric

- Inputs:
  - **inventory** (available physical devices)
  - **topology** (required logical connectivity)
- Operation:
  1. Map topology requirements onto physical devices and connections
  2. Configure JetStream cards



Control plane: Aviator Fabric

- Inputs:
  - **inventory** (available physical devices)
  - **topology** (required logical connectivity)
- Operation:
  1. Map topology requirements onto physical devices and connections
  2. Configure JetStream cards
  3. Resolve each workload handle to a concrete address

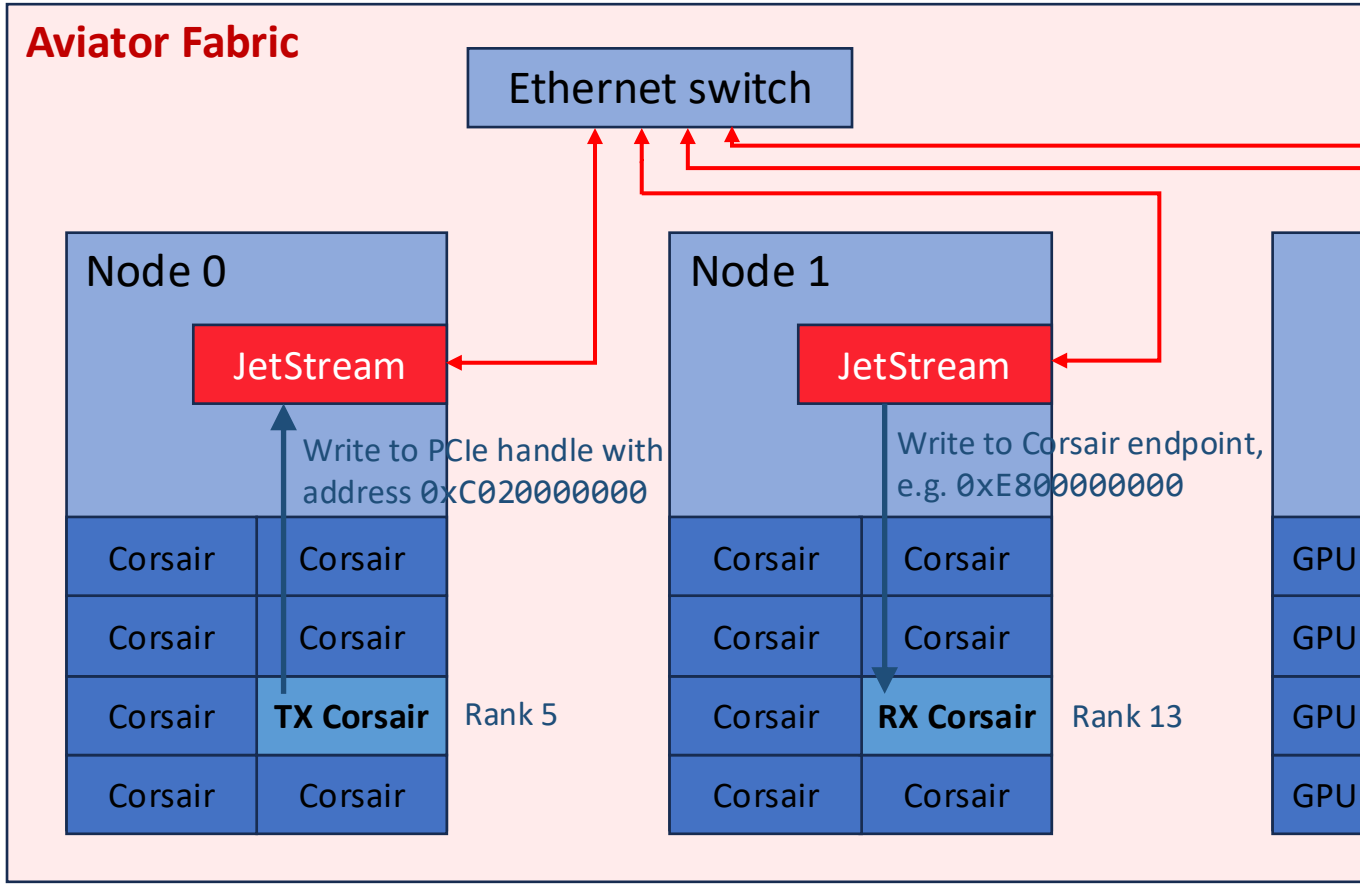
```
{
    "ID": 4,
    "OP": "MOVE_DDR2PCI",
    "REM_PCIE_ADDR" : "pci_r1_m1_c0",
    "DST_CHIPLET": 0,
    "TRF_SIZE": 2048,
    ...
},
```

Handle	Address
pci_r1_m0_c0	0xC000?000000
pci_r1_m1_c0	0xC020?000000
pci_r2_m0_c0	0xE800?000000
...	.?.

Addresses resolved by Aviator Fabric based on how the rank is reached in the instantiated fabric

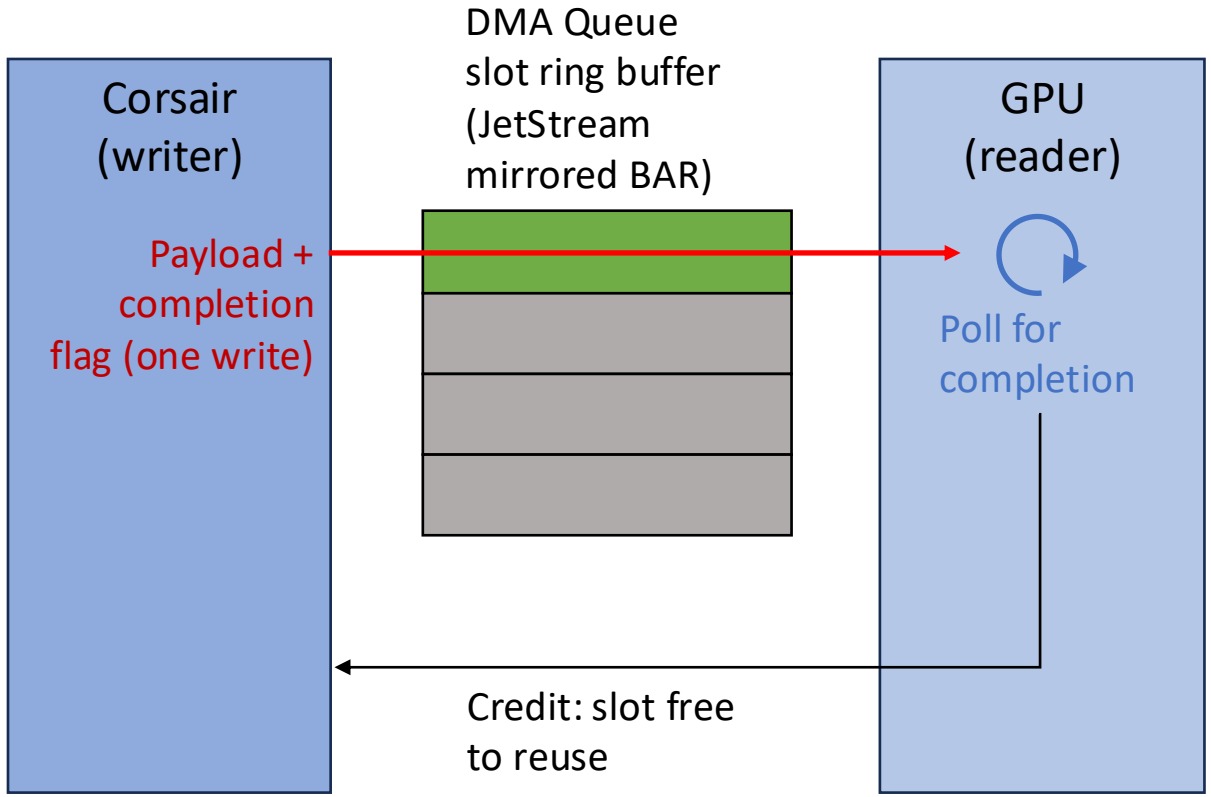
Extending addressing

- Aviator Fabric scales out the addressing model to Ethernet-connected nodes
- Handles in the workload are associated with a target rank
- Handles operate the same whether destination is direct or via JetStream
- Same scheme applies to Corsairs and other accelerator types (e.g. GPUs)



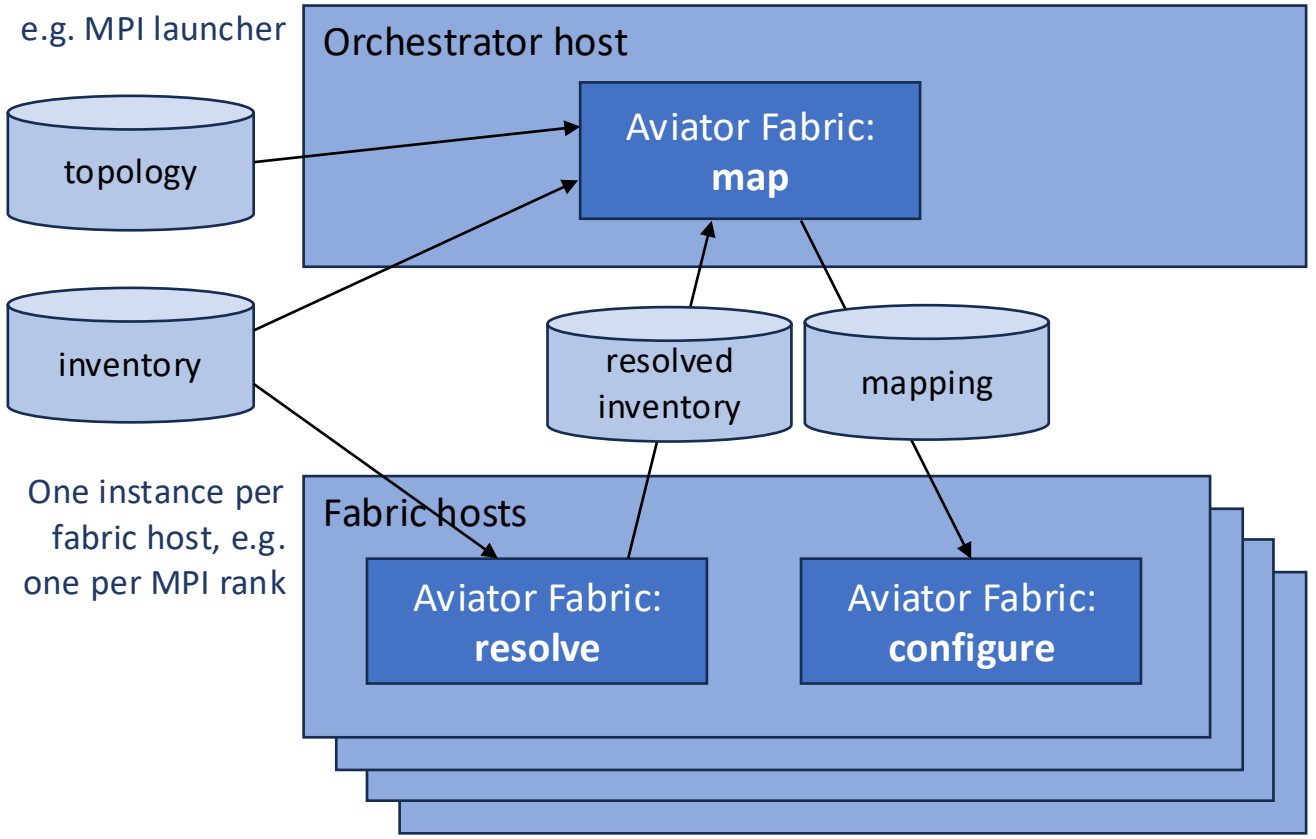
Completions for a GPU peer

- **Problem:** Corsair's completion mechanism doesn't reach workloads running on a GPU
- **Solution:** DMA Queue library extends the d-Matrix push-with-completion model to a GPU peer
- Completion flag is sent with or after payload; receiver detects “ready” by polling its own memory
- Avoids overwriting unconsumed buffer via return credit



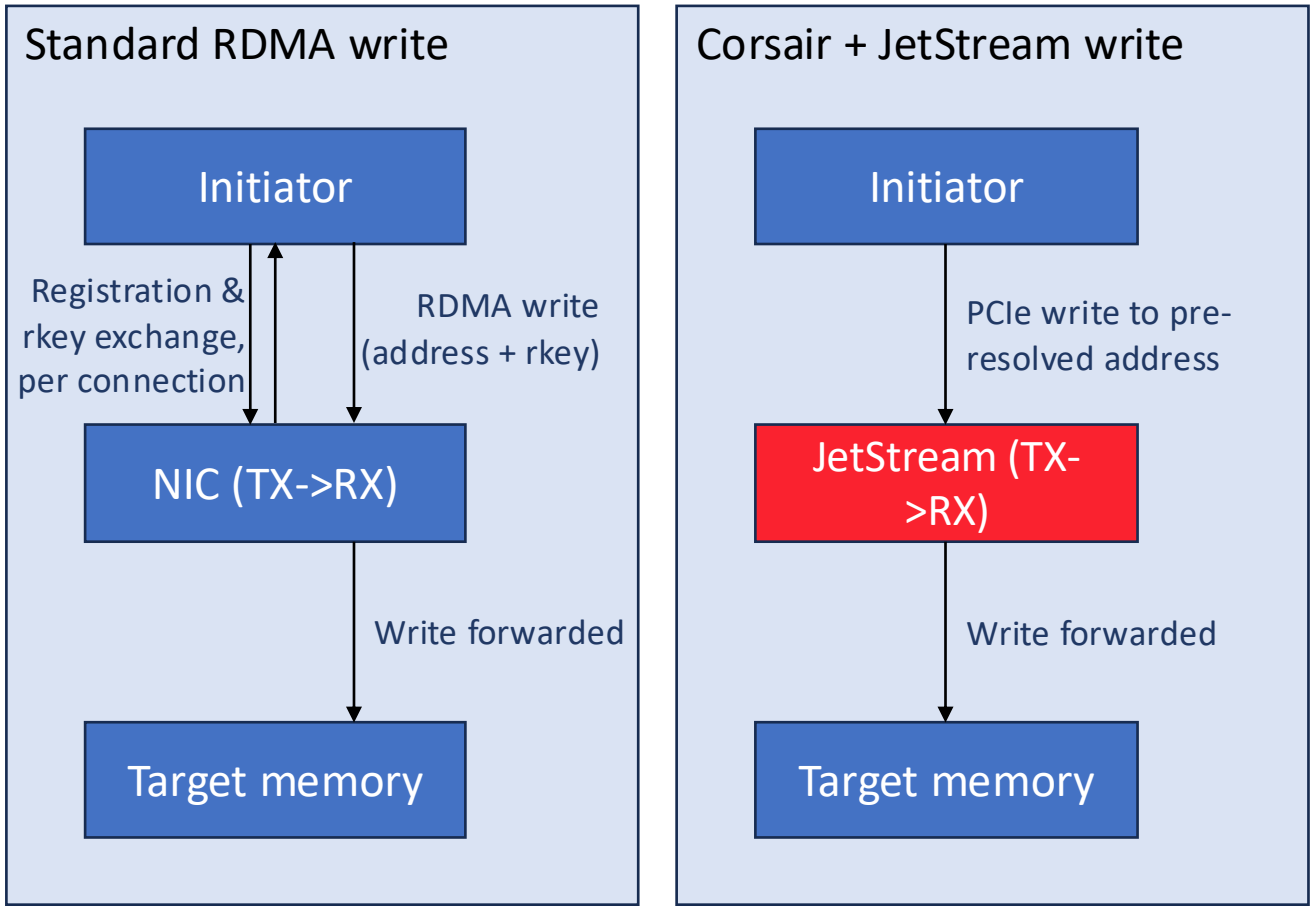
Fabric orchestration

- Aviator Fabric exposes modular building blocks for control-plane configuration
- Orchestrator invokes these blocks in parallel across the cluster hosts
- Orchestrator adds cluster-scale concerns: health monitoring, recovery, scheduling
- Annotated here using MPI as a concrete example; Kubernetes or other schemes work the same way
- Data-plane traffic itself remains fully device-driven throughout



RDMA Verbs comparison

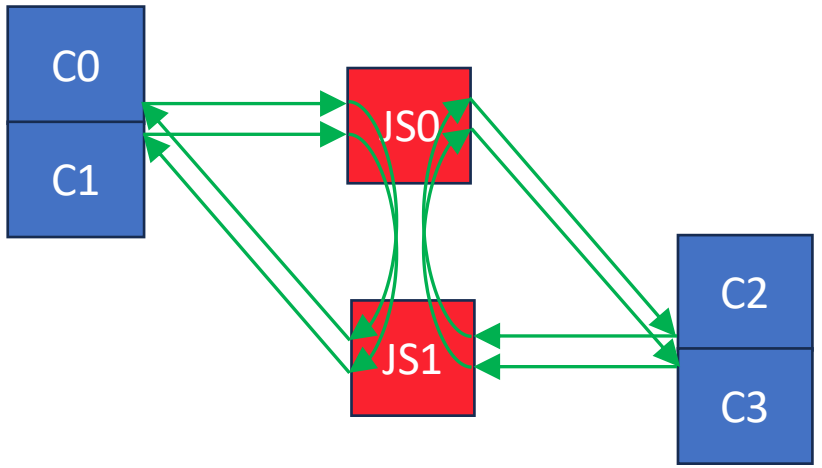
- RDMA's one-sided Write still requires per-connection round-trips to set up the write before any data can move
- Aviator Fabric resolves addressing once in advance for the whole fabric
- Enables end-to-end push with no round-trips



Performance - experiment

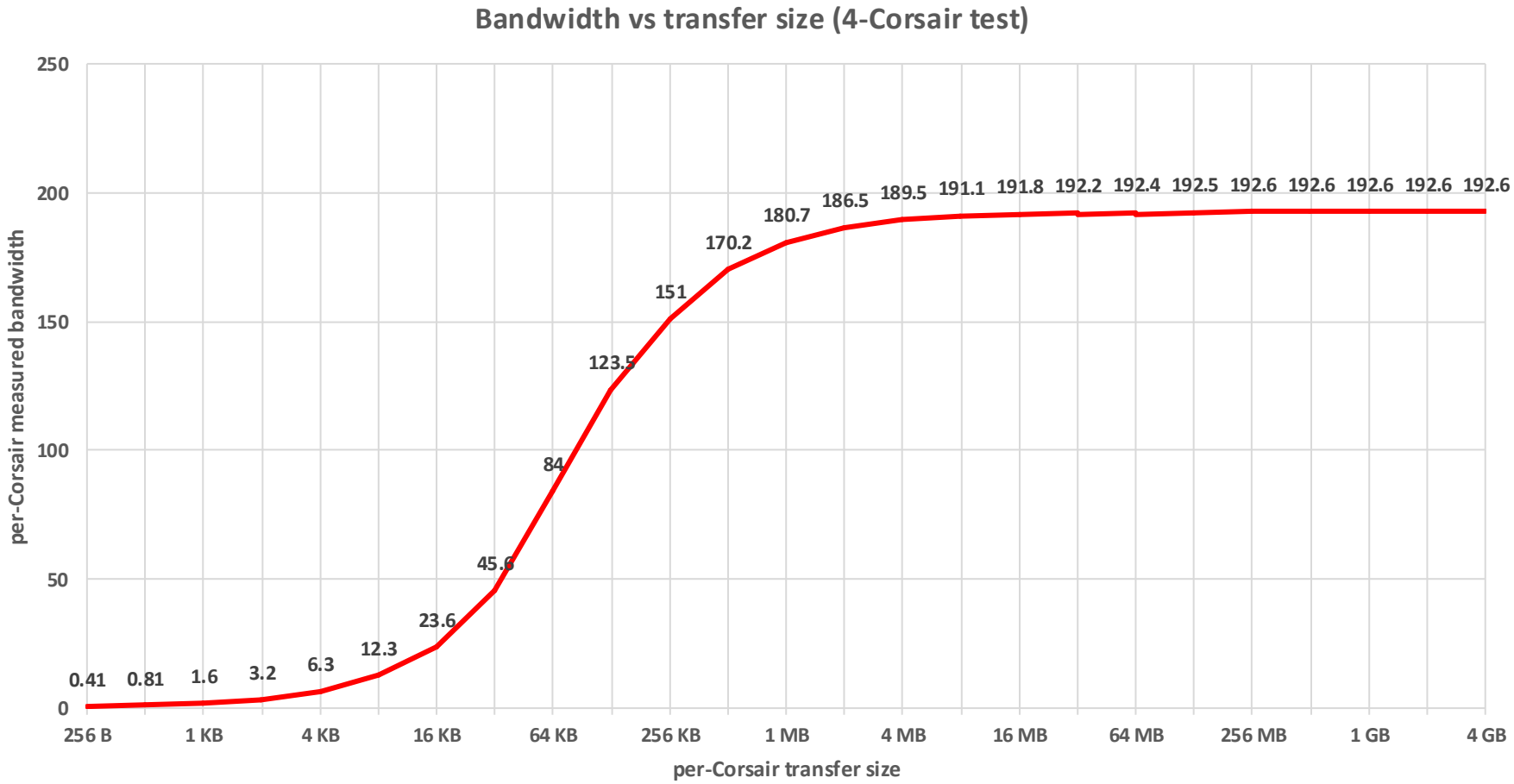
- Setup: 4 Corsairs, 2 JetStreams
- Simultaneous 4-way push traffic (2 directions)
- Time for all traffic to return measured using Corsair's internal CPU clock
- In this test, 400 Gb/s theoretical JetStream bandwidth is split between 2 Corsairs in each direction, so theoretical max bandwidth per Corsair is 200 Gb/s

Corsair 0 -> JetStream 0 -> JetStream 1 -> Corsair 0
Corsair 1 -> JetStream 0 -> JetStream 1 -> Corsair 1
Corsair 2 -> JetStream 1 -> JetStream 0 -> Corsair 2
Corsair 3 -> JetStream 1 -> JetStream 0 -> Corsair 3



Performance - results

- Bandwidth: 192.6 Gb/s per Corsair (385.2 Gb/s per direction per JetStream)
- 96% of theoretical limit



Summary

- Single addressing and data flow model, consistent from chiplet to multi-rack fabrics
- Key benefits:

Transparency: No change to workload regardless of fabric transport across any given link

Latency: Latency advantage of d-Matrix hardware is maintained at scale

- Also demonstrated:

Throughput: Near-theoretical bandwidth maintained when fabric is concurrently shared

d-Matrix

formidable to tyrants only. He has called together legislative bodies at places unusual, uncomfortable, and distant from the depository of their public Records, for the sole purpose of fatiguing them into compliance with his measures. He has dissolved Representative Houses repeatedly, for opposing with manly firmness his invasions on the rights of the people. He has refused for a long time, after such dissolutions, to cause others to be elected; whereby the Legislative powers, incapable of Annihilation, have returned to the People at large for their exercise; the State remaining in the mean time exposed to all the dangers of invasion from without, and convuls, and when so suspended, he has utterly neglected to attend to them. He has refused to pass other Laws for the accommodation of large districts of people, unless those people would relinquish the right of Representation in the Legislature, a right inestimable to them and formidable to tyrants only. He has called together legislative bodies at places unusual, uncomfortable, and distant from the depository of their public Records, for the sole purpose of fatiguing them into compliance with his measures. He has dissolved Representative Houses repeatedly, for opposing with manly firmness his invasions on the rights of the people. He has refused for a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security.--Such has been the patient sufferance of these Colonies; and such is now the necessity which constrains them to alter their former Systems of Government. The history of the present King of Great Britain is a history of repeated injuries and usurpations, all having in direct object the establishment of an absolute Tyranny over these States. To prove this, let Facts be submitted to a history of repeated injuries and usurpations, all having in direct object the establishment of an absolute Tyranny over these States. To prove this, let Facts be submitted to a candid world. He has refused his Assent to Laws, the most wholesome and necessary for the public good. He has forbidden his Governors to pass Laws of immediate and pressing importance, unless suspended in their operation till his Assent should be obtained; and when so suspended, he has utterly neglected to attend to them. He has refused to pass other Laws for the accommodation of large districts of people, unless those people would relinquish the right of Representation in the Legislature

Thanks...

Output from Llama 3 model prompted to complete the Declaration of Independence

Running on multi-node Corsair + Jetstream + Aviator system (BFP16, no finetuning)

d-Matrix