d-Matrix® JetStream™

Ultra-low Latency Scaling of AI Inference in Datacenters

The growth of Agentic AI with reasoning, multi-modal interactive content and inference-time scaling is pushing the boundaries of AI infrastructure. Deploying AI inference services at scale in the datacenter requires building and scaling multi-node inference clusters. This necessitates optimizing both compute and networking aspects for latency, cost-performance and energy efficiency at scale.

Both Ethernet and PCIe-based architectures impose limits on scalability, latency, and reliability. For example, existing PCIe fabrics are limited in terms of the maximum numbers of accelerators per host, introduce single-point-of-failure risks, and struggle to keep pace with the massive data flows required for next-generation AI workloads. And the challenge with ethernet solutions such as RDMA (RoCE, IB) is the latency overhead in communication.

d-Matrix JetStream™ is a purpose-built network interface card (NIC) enabling efficient scaling of AI workloads and delivering ultra-low latency inference with Corsair clusters. As a "Transparent NIC" and streaming solution, JetStream unlocks seamless accelerator-to-accelerator communication across nodes, leveraging PCIe peer-to-peer memory writes to bypass the host-initiated communications of existing PCIe and Ethernet based approaches.

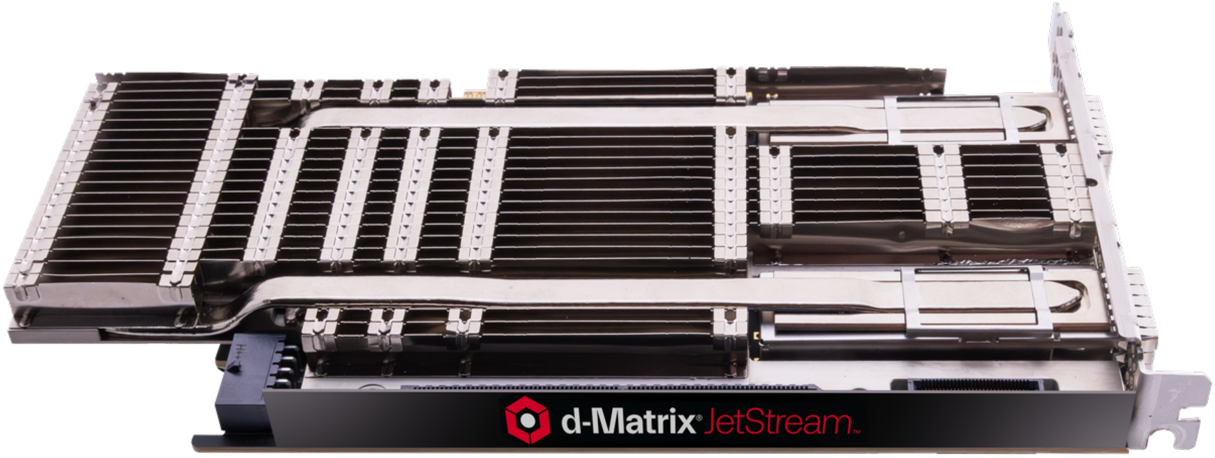

Figure 1. d-Matrix JetStream: purpose-built ultra-low latency Transparent NIC

With d-Matrix Corsair's blazing fast inference performance, it is imperative to have device-initiated communications to keep up with the computing speed. d-Matrix JetStream decouples the data-plane from the control-plane, minimizing host-device communication overheads and thereby avoiding an IO bottleneck. By extending Corsair's PCIe-based communication semantics for multi-node communication, JetStream enables ultra-low latency scale-up communication with traditional scale-out infrastructure.

JetStream comes in an industry-standard PCIe form factor and connects to standard off-the-shelf Top-of-Rack ethernet switches. This makes it convenient and easy to deploy with existing datacenter infrastructure.

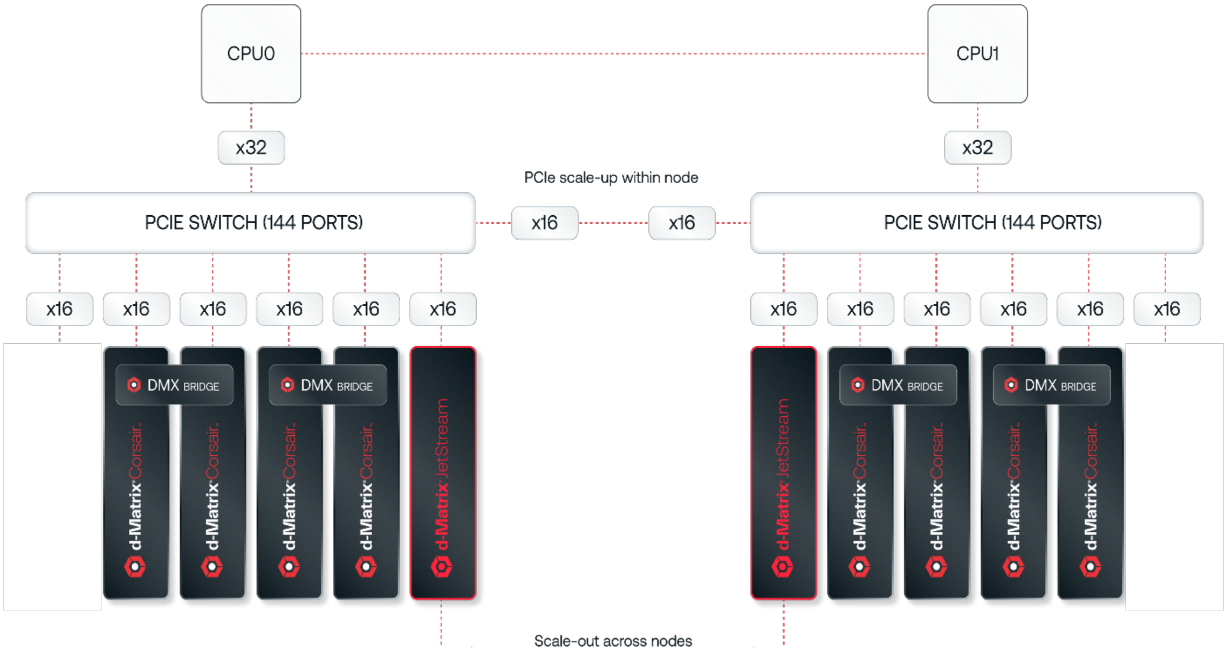

Figure 2. JetStream PCIe card in Corsair inference server, co-optimized for multi-node scaling

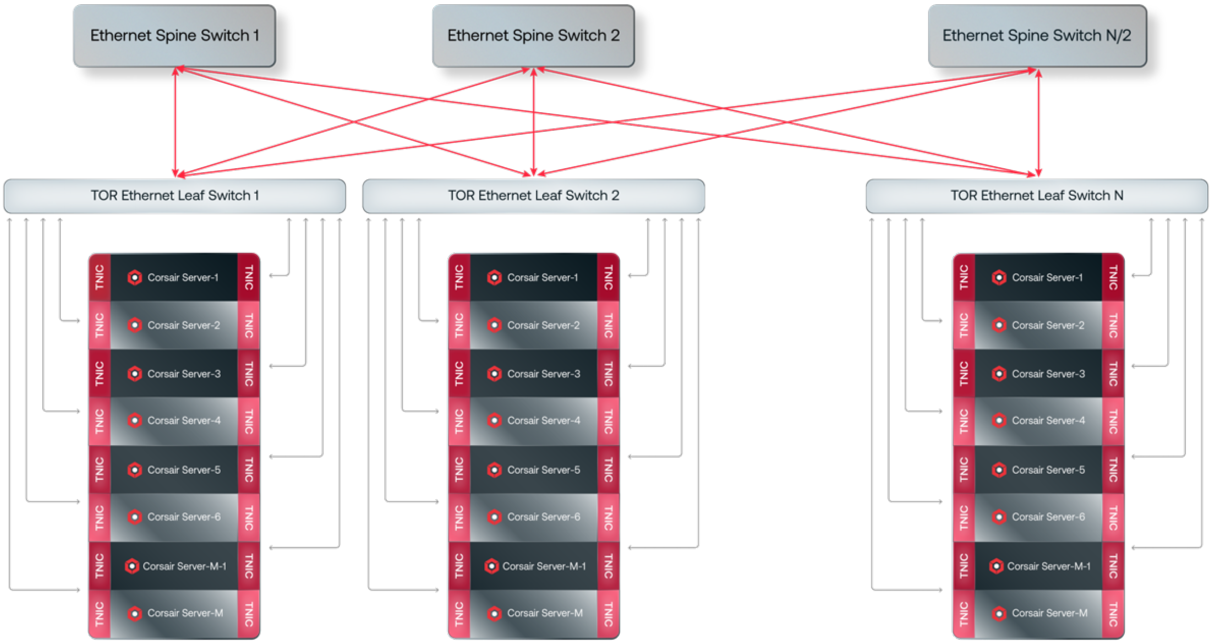

Figure 3. JetStream cards connected with standard off-the-shelf top-of-the-rack ethernet switches

Together d-Matrix Corsair and JetStream make AI inference commercially viable for hyperscale, public and private clouds by improving the cost-performance of AI inference by up to 3 times, improving energy efficiency by up to 3 times, and accelerating token generation speeds by up to 10 times[1].

Blazing fast	Commercially Viable	Sustainable
10x	3x	3x
interactive speed	cost-performance	energy efficiency

[1] For Llama70B model on 8 Corsair servers. Performance, cost and power estimates are preliminary and subject to change. Results may vary.

JetStream Product Specification

d-Matrix JetStream is a full-height, 3/4-length PCIe card that seamlessly integrates into AI servers and racks using Corsair.

The following table provides an overview of the card specifications.

Specification	Value
Network Protocol	IEEE 802.3 Ethernet
Maximum bandwidth	400 Gbps
Recommended interfaces	(copper) 400 Gbps-DAC, (optical) 400 Gbps-SR8, 400 Gpbs-VSR4
Transceiver form factor	QSFP-DD
Transceiver signaling formats (line, per lane)	56G-PAM4 / 28G-NRZ
Transceiver signaling formats (client, per lane)	56G-PAM4 / 28G-NRZ
Host bus interface	PCIe Gen5 x16, 32 GT/s
Max TDP (w/ transceivers)	150 W
Power supply inputs	12V with Aux connector (600W) 12V (75 W) from PCIe edge fingers
Secure boot	Supported

Key Benefits

- **Efficient Scaling** → Overcome PCIe and Ethernet limitations to connect more accelerators efficiently.
- **Ultra-Low Latency** → Direct accelerator-to-accelerator data transfers initiated by the device minimize communication overhead.

- **Optimized for modern AI workloads** → Designed for model-parallel communication, purpose-built for modern generative AI inference workloads.
- **Plug-and-play ecosystem** -> Integrates seamlessly with standard ethernet infrastructure in the datacenter