d-Matrix®

Product Brief

d-Matrix® Corsair™

Performance and Efficiency for AI Inference in Datacenters

Transformative experiences with Generative AI (Gen AI) are emerging, driven by advances like chain-of-thought reasoning, interactive video generation, and agentic workflows. However, deploying Gen AI applications at scale is challenging due to the unique computational and memory demands, as well as high energy consumption.

To make AI inference commercially viable, d-Matrix built a new computing platform from the ground up: Corsair™, the world's most efficient compute solution for AI inference at datacenter scale. Corsair delivers 10x better interactive performance, 3x energy efficiency, and 3x cost-performance versus GPU alternatives[1].

Unique Architecture Purpose-Built for AI Inference

Corsair blends cutting-edge innovations from d-Matrix into a novel Digital-In Memory Compute Architecture (DIMC™). These include custom digital in-memory compute cores, block floating point numerical formats, chiplet-based scaling, fast die-to-die connectivity across chiplets via DMX Link™ and package-to-package connectivity across two cards via DMX Bridge™. Corsair comes in an industry-standard full-height, full-length PCIe card that is easy to integrate in AI datacenters.

Blazing fast
10x
interactive speed

Commercially Viable
3x
cost-performance

Sustainable
3x
energy efficiency

[1] *Performance, cost and power estimates are preliminary and subject to change. Results may vary.*

Performance Memory for Blazing Fast Interactivity

d-Matrix's novel DIMC architecture breaks the memory barrier by tightly integrating compute and memory. The integrated **Performance Memory** of the on-chip memory-compute complex enables fast token generation with its ultra-high bandwidth of 150 TB/s, an order of magnitude higher than HBM available today. In **Performance Mode**, Gen AI models fit in Performance Memory and can achieve up to 10x faster interactive latency compared to alternatives using HBM.

Capacity Memory for Offline Batched Inference

Corsair also comes with up to 256 GB of off-chip **Capacity Memory** that enables Gen AI workloads in offline use cases. In **Capacity Mode**, Corsair supports large models, large context lengths, and large batch sizes. For example, a server with 8 cards can fit models of 1T+ parameters.

Numerical Formats for Inference Acceleration

Corsair is among the first in the industry to natively adopt **Block Floating Point** numerical formats, now an OCP standard called Microscaling formats (MX). These are ideal for inference, since they have the energy efficiency of integer arithmetic with the high dynamic range of floating-point. Corsair also supports advanced features such as sparsity, on-the-fly quantization, and inline decompression for efficient storage and processing.

Built for Datacenter Scale

Each Corsair PCIe card contains two ASIC packages connected back-to-back using **PCIe Gen5**, where each package has four chiplets connected in an all-to-all topology using DMX Link™. Four packages across a pair of cards are connected via DMX Bridge™ in an all-to-all topology across 16 chiplets. This unique topology is critical for fast token generation speeds. Four such pairs of Corsair cards (i.e., 8 cards) can be connected with PCIe switches and scaled up to build an inference server that integrates easily into AI rack infrastructure, as shown in the figure below.

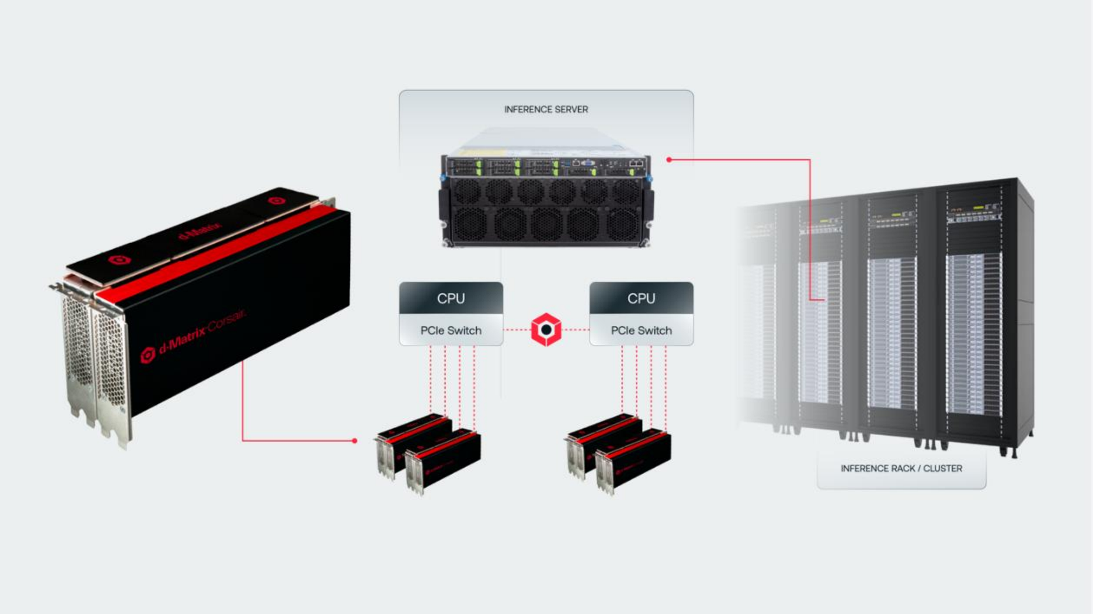

Figure 1. Easy to scale from a card to server to racks and integrate in existing AI infrastructure

Corsair Product Specification

Corsair comes in an industry-standard PCIe form factor with both single card and dual card configuration.

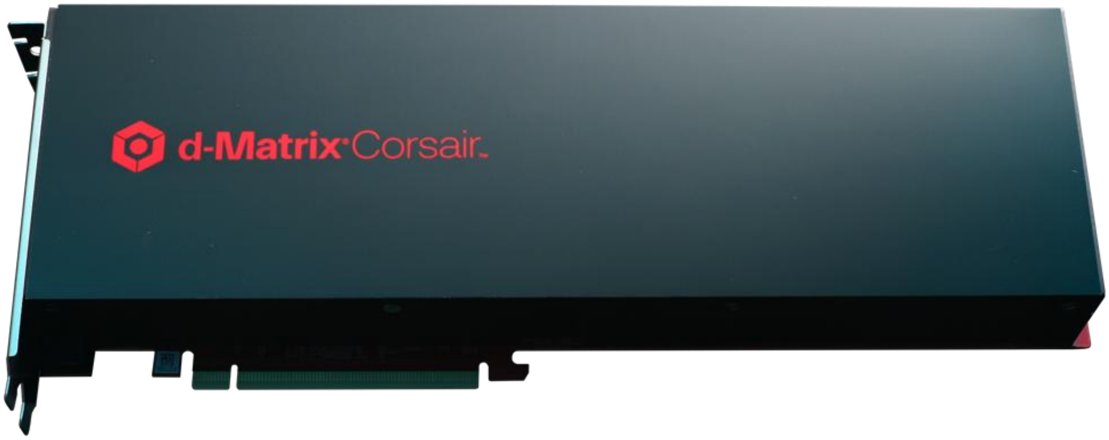

Table 1 lists the technical specifications of the Corsair PCIe card.

Specification	Value
DIMC compute cores	2048
Compute TFLOPS (8 bit dense)	2400
Compute TFLOPS (4 bit dense)	9600
Performance Memory	2 GB
Performance Memory bandwidth	150 TB/s
Capacity Memory	Up to 256 GB
Capacity Memory bandwidth	400 GB/s
Interconnect fabric	PCIe Gen5 x16: 128 GB/s bi-directional
Power supply inputs	12V with Aux connector (600 W) 12V (75 W) from PCIe edge fingers
Security	Secure boot
Management	Redfish, PLDM, SPDM in development
TDP	600 W
Thermals	Air cooling (dual slot)

Table 1. Corsair technical specifications

Corsair Dual-Card Specification

Table 2 lists the technical specifications of the Corsair PCIe dual card.

Specification	Value
DIMC compute cores	4096
Compute TFLOPS (8 bit dense)	4800
Compute TFLOPS (4 bit dense)	19200
Performance Memory	4 GB
Performance Memory bandwidth	300 TB/s
Capacity Memory	Up to 512 GB
Capacity Memory bandwidth	800 GB/s
Interconnect fabric	PCIe Gen5 x16: 128 GB/s bi-directional DMX Bridge: 512 GB/s bi-directional
Power supply inputs	12V with Aux connector (600 W) 12V (75 W) from PCIe edge fingers
Security	Secure boot
Management	Redfish, PLDM, SPDM in development
TDP	1200 W
Thermals	Air cooling (dual slot)

Table 2. Corsair dual card technical specifications