d-Matrix® Corsair™ Redefines Performance and Efficiency for AI Inference at Scale

November 2024

Introducing d-Matrix® Corsair™

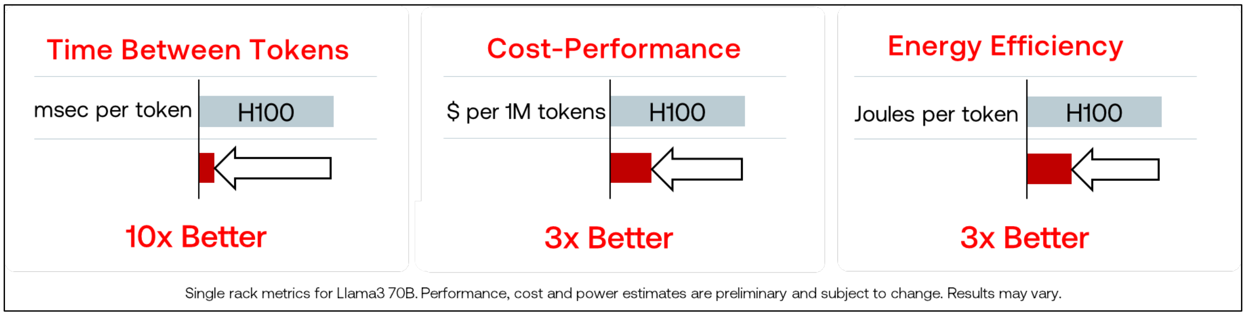
d-Matrix® was founded with the mission to transform AI from unsustainable to attainable, specifically targeting the datacenter inference market. d-Matrix took an unconventional first principles approach to optimize across all fronts — performance, power, cost, and ease of use. Since its founding in 2019, d-Matrix built multiple world-first innovations in silicon, software, chiplet packaging, and interconnects, and created Corsair™, a first-of-its-kind inference compute platform that excels at blazing fast token generation, performance-TCO, and energy efficiency.

Figure 1. d-Matrix Corsair dual card

d-Matrix Corsair is the world's densest integrated compute and memory solution packing 3200 mm^2 of silicon in a single PCIe card, including 2 GB on-chip integrated Performance Memory at 150 TB/s memory bandwidth, and up to 256 GB off-chip Capacity Memory, with peak dense compute of 2400 TFLOPs for MXINT8 and 9600 TFLOPs for MXINT4 equivalent formats.

The dual-card Corsair comprises of 6400 mm^2 silicon with 250B+ transistors, 16 chiplets with all-to-all connectivity, 19.2 PFLOPs of MXINT4 and 4 GB Performance Memory at 300 TB/s bandwidth, and up to 512 GB Capacity Memory.

The industry standard FHFL PCIe form factor integrates easily into AI servers and racks. d-Matrix is partnering with OEMs and System Integrators to qualify the Corsair PCIe cards in their servers and simplify deployments across datacenters. Corsair makes Generative AI (GenAI) commercially viable 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]. Corsair unlocks the full potential of GenAI, especially for Enterprise Large Language Models (LLMs), Reasoning, Agents, and Video generation use cases.



Single rack metrics for Llama3 70B. Performance, cost and power estimates are preliminary and subject to change. Results may vary.

Figure 2. Corsair offers industry-leading ultra-low latency, performance-TCO and energy efficiency

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

Unique Approach to GenAI Inference Acceleration

Generative workloads are based on Transformer architecture and have unique requirements. They need high peak-compute capacity to process user prompts, high memory bandwidth to generate tokens, and high memory capacity to store large models and user context. This leads to a significant increase in compute cost and power consumption, making the future of GenAI unsustainable. d-Matrix has taken a unique approach to address these challenges.

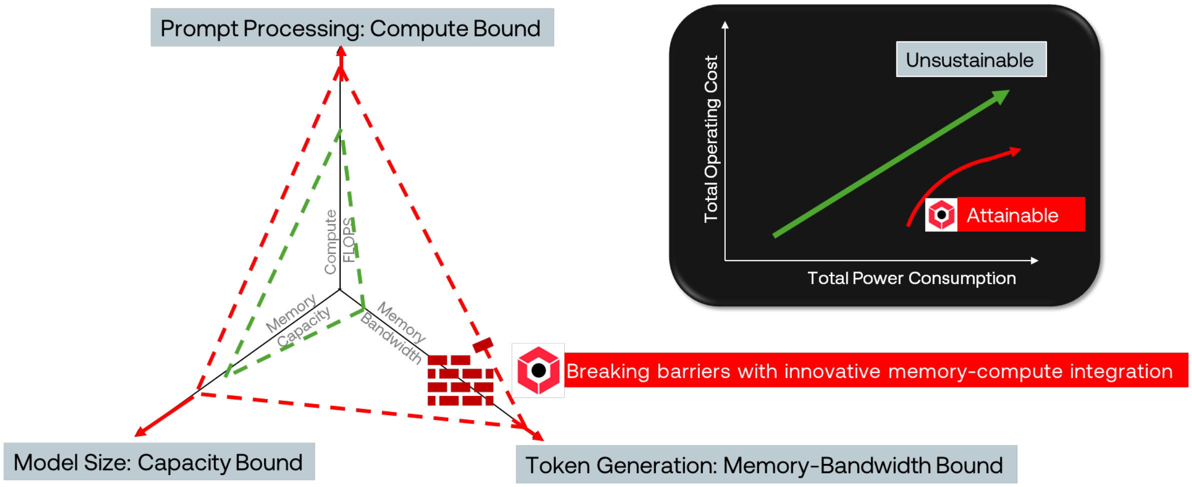

Prompt Processing: Compute Bound
Compute FLOPS
Memory Capacity
Memory Bandwidth
Model Size: Capacity Bound
Token Generation: Memory-Bandwidth Bound
Breaking barriers with innovative memory-compute integration
Total Operating Cost
Total Power Consumption
Unsustainable
Attainable

Figure 3. Generative inference workloads present unique challenges. d-Matrix's innovative solution is changing the trajectory from unsustainable to attainable

Digital In-Memory Compute Architecture (DIMC™) Breaks Memory Barrier

Traditional accelerators often use costly and power-hungry High Bandwidth Memory (HBM) to speed up token generation and are unable to scale or compute more effectively because of their Von Neumann architecture in which memory is physically separate from compute. d-Matrix breaks this memory barrier by integrating a multiplier directly into memory bit cell using a logic process, enabling both lower energy consumption and ultra-low latency. Digital IMC addresses the challenges of analog IMC, providing noise-free computation and greater flexibility to handle future AI needs. The integrated Performance Memory of the on-chip memory-compute complex enables fast token generation with its ultra-high bandwidth, an order of magnitude higher than the 4-8 TB/s of HBM available today.

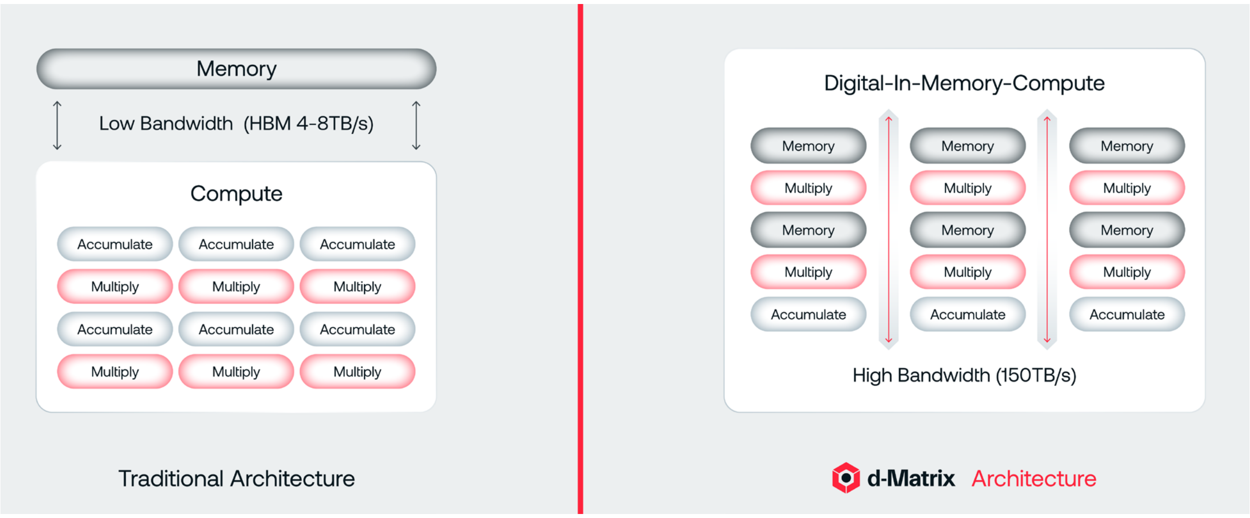


Figure 4. d-Matrix DIMC architecture breaks the memory barrier in traditional architectures, achieving extremely high memory bandwidth

Chiplet-based Architecture for Flexibility and Scalability

d-Matrix is among the first in the industry to embrace chiplet-based architecture. d-Matrix scales memory and compute by using smaller chiplet dies that improve yields, lower costs, as well as overcome reticle limits.

Figure 5 shows the high-level architecture block diagram of a single chiplet. It consists of four Quads, each containing four Slices, a RISC-V control core, and a Dispatch Engine. Each Slice contains DIMC cores, SIMD cores, and a Data Reshape Engine. The architecture achieves extremely high memory and compute density with each chiplet.

d-Matrix has designed a unique communication topology for efficient data movement within the chiplet (via proprietary network-on-chip) and between four chiplets in a package. DMX Link™ is a custom die-to-die interconnect based on OCP Open Domain-Specific Architecture (ODSA), enabling error-free, low-latency, and low energy transmission. Chiplets communicate with each other via the DMX Link. As shown in Figures 5 and 6, the four Slices in a quad, four Quads in a chiplet, and four chiplets in a package are connected using an all-to-all topology, which is critical for low latency communication and enables faster token generation speeds.

Each chiplet has a standard PCIe5 x 16 interface for scale-out communication and two LPDDR interfaces, in addition to the DMX Link interfaces to connect multiple chiplets. The off-chip LPDDR5 Capacity Memory enables GenAI workloads in offline batched inference use cases.

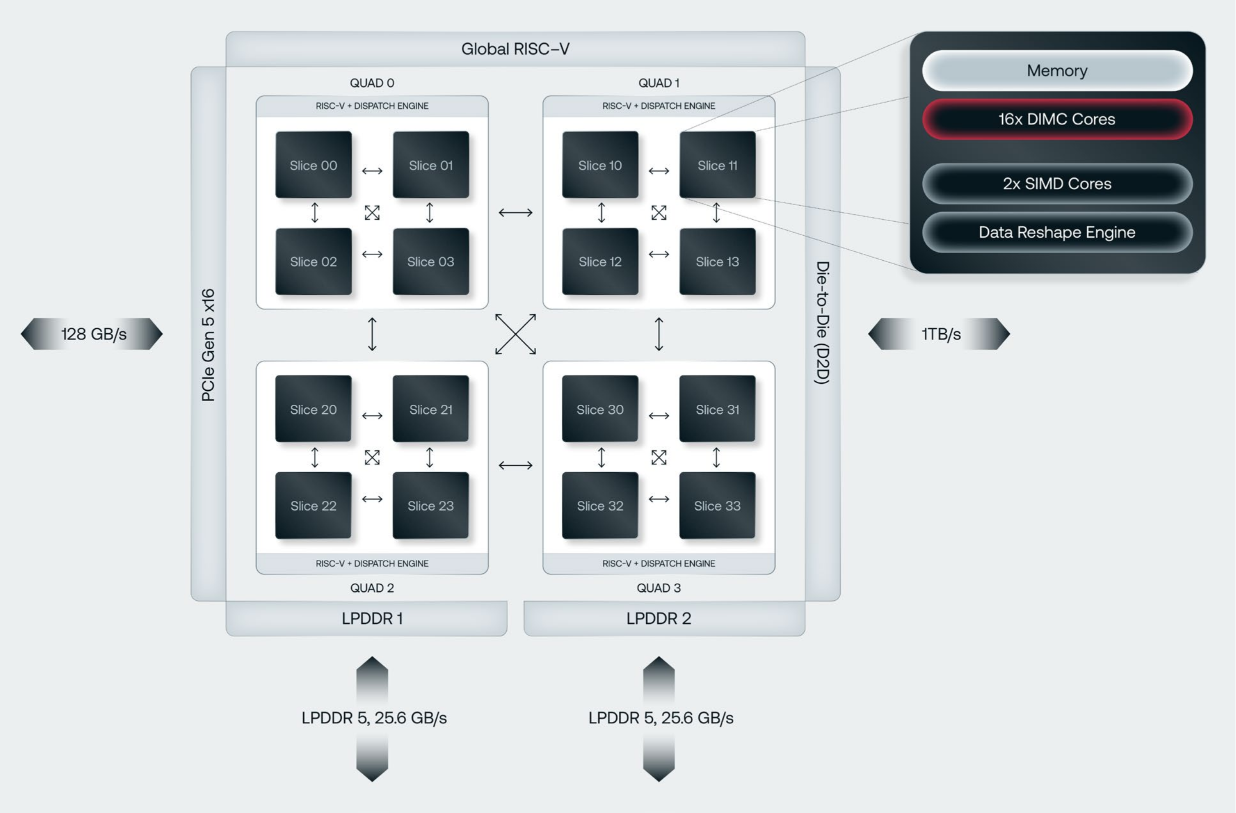


Figure 5. Corsair chiplet architecture consists of DIMC, SIMD, and RISC-V cores, along with PCIe5 x 16, LPDDR5, and die-to-die DMX Link interfaces

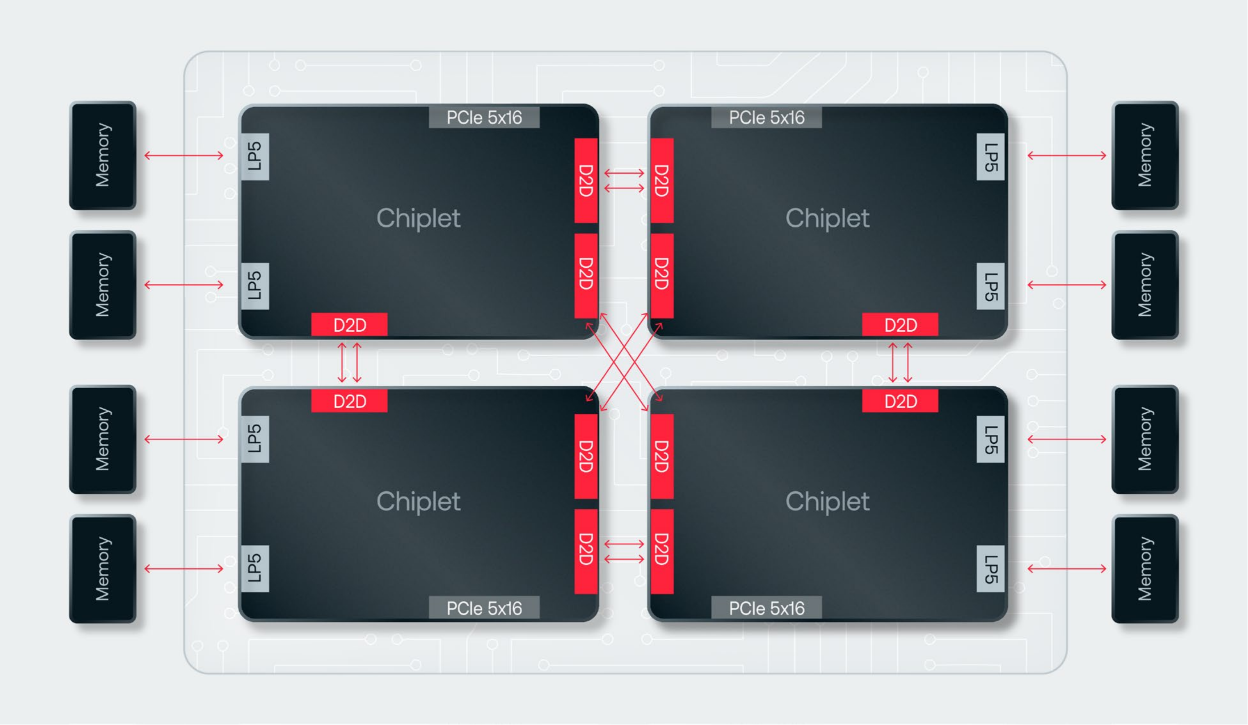


Figure 6 . FCBGA package with 4 chiplets connected in an all-to-all topology using DMX Link

Block Floating Point Numerics Lowers Memory Footprint

d-Matrix hardware is among the first in the industry to natively implement Block Floating Point numerical formats, now an OCP standard called Microscaling formats (MX). These numerical formats offer the best of both worlds—the energy efficiency of integer arithmetic with the high dynamic range of floating-point necessary for model accuracy. Corsair supports MXINT16, MXINT8, and MXINT4 numerical formats, all using native d-Matrix Block Floating Point format.

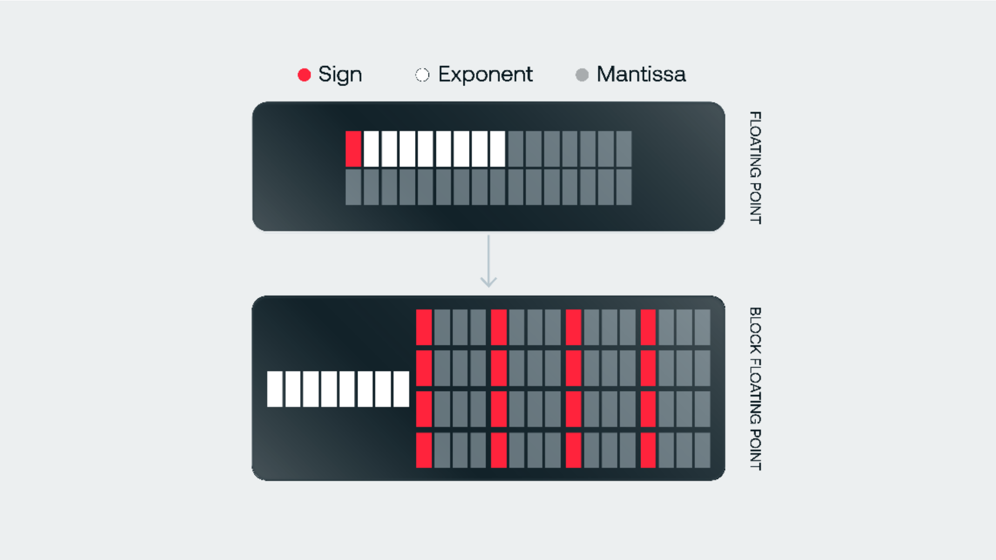


Figure 7 .Block floating point numerics are both energy efficient and provide high dynamic range

Table 1 shows the numerical formats and corresponding TFLOPS. Corsair also supports advanced features such as sparsity, on-the-fly quantization, and inline decompression for efficient storage and processing.

Numerical format	Corsair TFLOPS*
MXINT4	9600
MXINT8	2400
MXINT16	600

Table 1. Numerical formats and Compute FLOPS

*Both weights and activations in same format. TOPS can be higher with mixed precision. For example, MXINT4 weights and MXINT8 activations can provide 4800 TOPS.

Scalable Architecture to Serve a Variety of GenAI Models

GenAI models vary in size from a few billion to tens and hundreds of billion parameters, and do not fit on one card. Inference involves distributing such models across several cards in a server (Scale-up) or across servers in a rack environment (Scale-out). At the model level, this means distributing different model layers across different cards (pipeline parallelism) or splitting a tensor across multiple cards (tensor parallelism). Corsair hardware and software natively support distributed inference and have been co-designed to scale from chiplets to racks over a standard fabric.

As described earlier, d-Matrix has designed a unique communication topology for all-to-all data movement within the chiplet and between four chiplets in a package. This all-to-all topology is further extended between four packages across two cards via DMX Bridge. This allows efficient data transfers and low-latency collective communication needed for tensor parallel operation.

Figure 8 shows the Corsair PCIe card with two packages, where chiplets in each package are connected in an all-to-all topology via DMX Link.

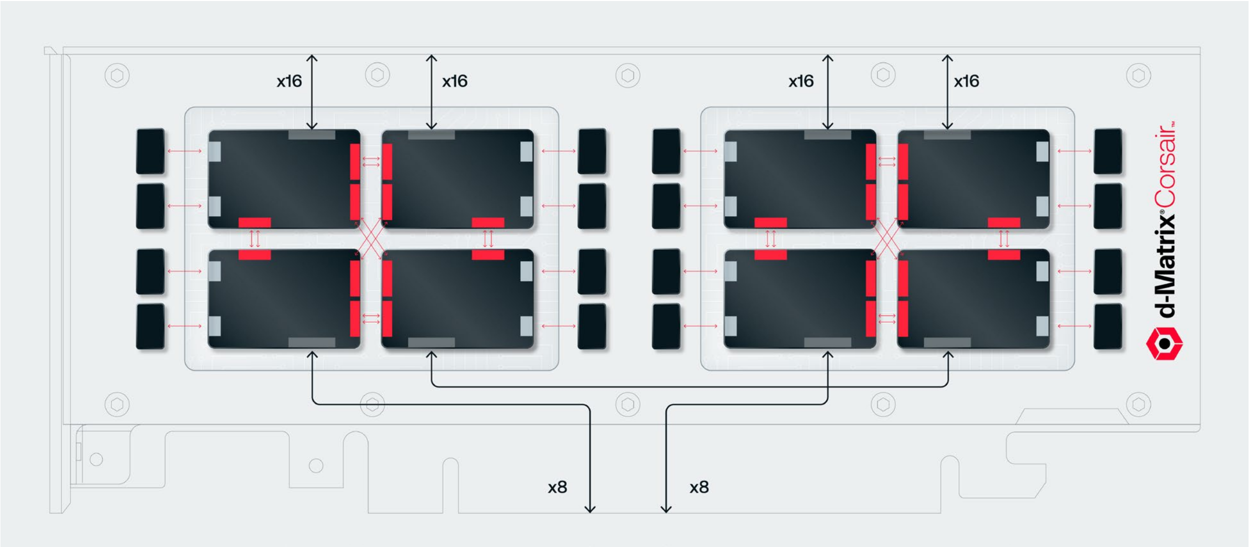


Figure 8. Corsair PCIe card with two packages connected using PCIe5 x 16 with integrated LPDDR5

Each PCIe card has four additional PCIe Gen5 x16 links, which enables a pair of cards to be connected via DMX Bridge in an all-to-all topology across four packages (i.e., 16 chiplets). This is shown in Figure 9.

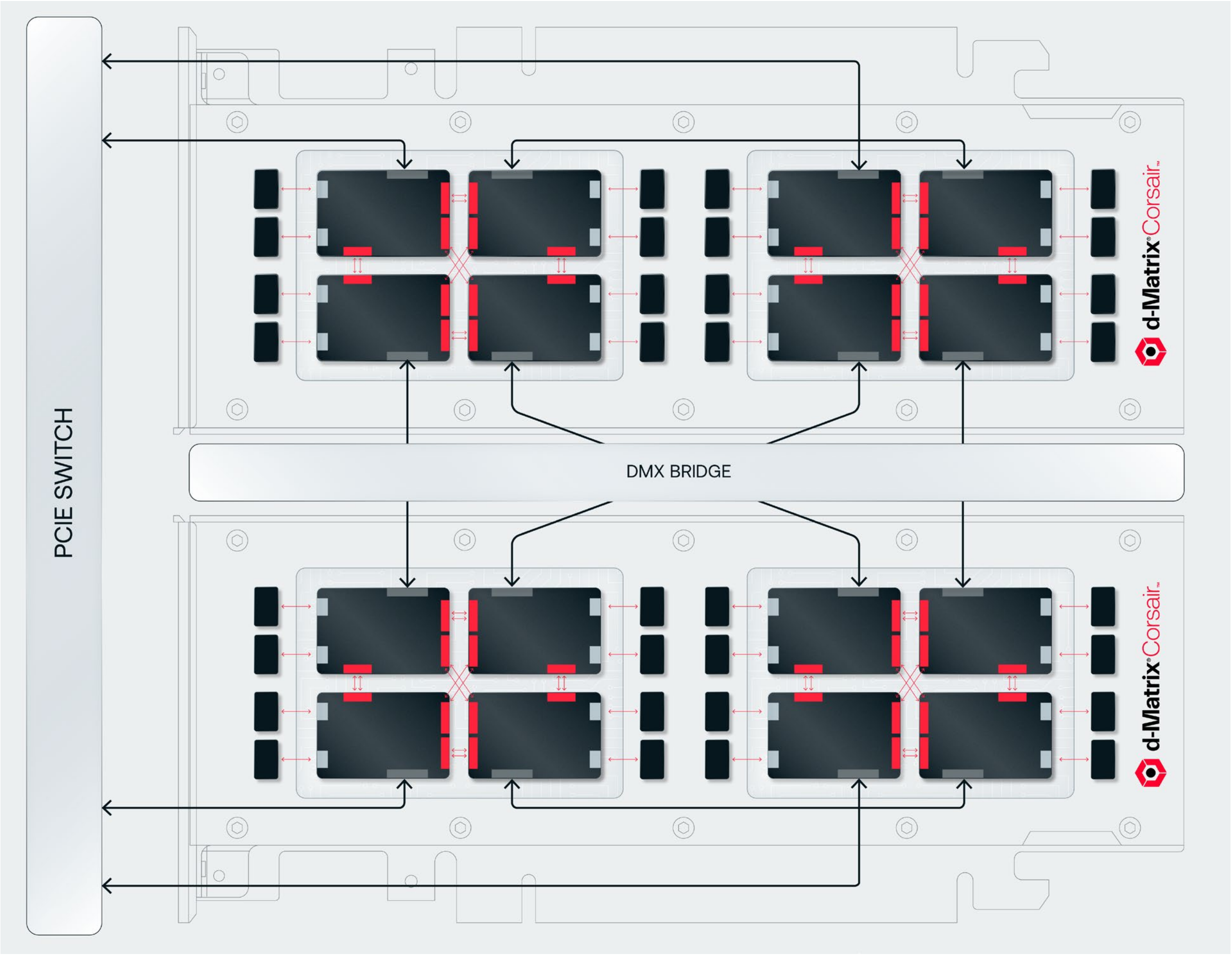


Figure 9. Rich connectivity between two Corsair cards using DMX Bridge™

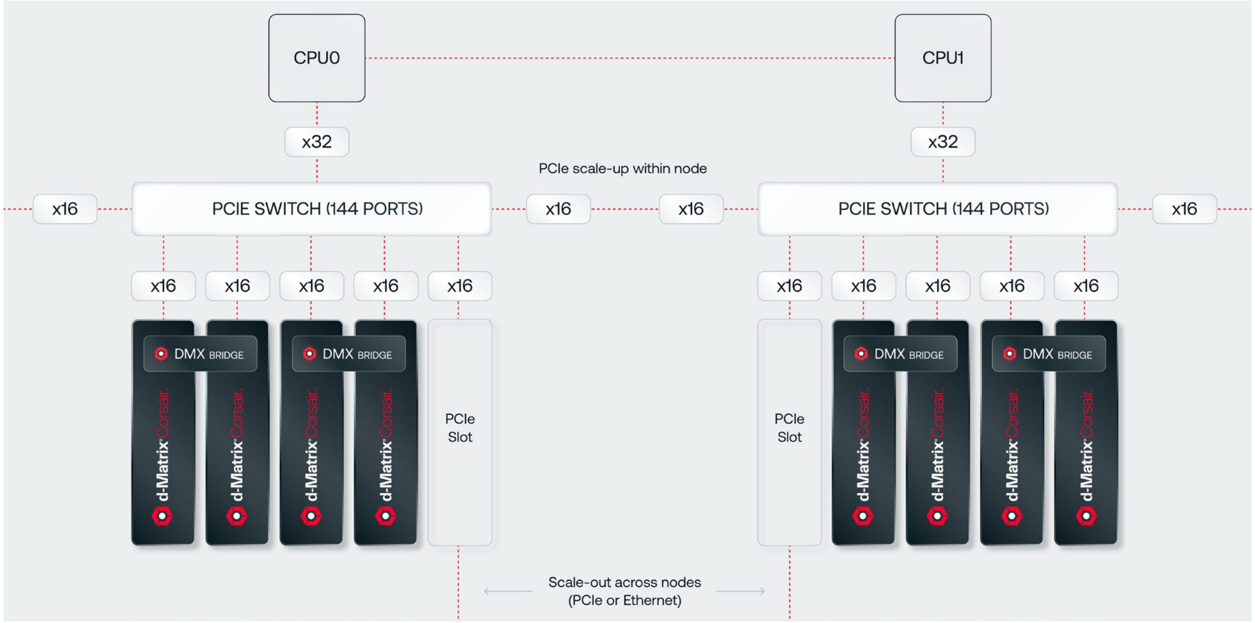


Figure 10. Reference system block diagram of an 8-card Corsair server

Four such pairs of Corsair cards (i.e., 8 cards) are connected with PCIe switches and scaled up in an 8-card inference server as shown in the reference configuration in Figure 10. Overall, a server with 8 Corsair cards has the following specifications, shown in Table 2.

Features	8X Corsair Inference Server System
MXINT4 (4-bit, dense)	76.8 PFLOPs
MXINT8 (8-bit, dense)	19.2 PFLOPs
Performance Memory	16 GB, 1200 TB/s
Capacity Memory	Up to 2 TB, 3.2 TB/s
DMX Bridge bandwidth	2 TB/s

Table 2. Specifications for an Inference server with 8x Corsair cards

Scaling out across multiple servers in a rack can be enabled over industry standard PCIe or ethernet. d-Matrix is collaborating with OEMs and System Integrators to qualify Corsair cards in AI inference servers.

Taking an open-standards approach ensures that Corsair is compatible with a wide range of datacenters and AI servers, ensuring our customers can easily integrate d-Matrix as part of their existing infrastructure. By prioritizing compatibility with the open ecosystem, d-Matrix makes it easy for enterprises to adopt Corsair without overhauling or interfering with existing AI solutions and ensuring the broadest access to advanced generative inference compute.

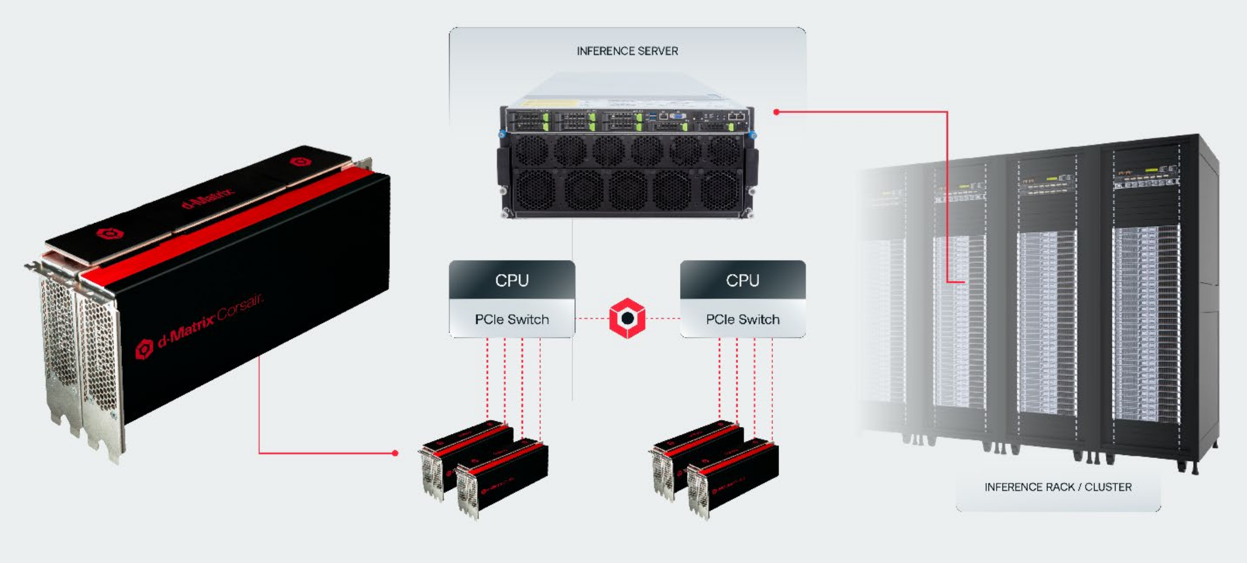


Figure 11. Corsair comes in an industry standard PCIe form factor and is easy to scale from a card to server to racks and integrate in existing AI rack infrastructure.

Aviator: State-of-the-Art Software Co-designed with Hardware

d-Matrix Aviator™ is an enterprise-grade software stack co-designed with d-Matrix hardware to maximize performance while being easy to use. Aviator meets developers where they are by enabling integration with the developer ecosystem and abstracting proprietary d-Matrix software components. Aviator enables easy model conversion and users can bring trained models from GPUs or other systems onto d-Matrix hardware. Built with open-source software — such as OpenBMC, MLIR, PyTorch and the Triton DSL for custom kernel creation — Aviator includes native support for distributed inference across multiple Corsair cards, servers, and racks for handling large-scale, memory-intensive GenAI models.

The Aviator user workflow consists of two phases: (1) the Build Flow, which converts models from popular frameworks (such as PyTorch) into compiled binaries, and (2) the Execution Flow, in which a distributed runtime framework executes the binaries on one or more cards.

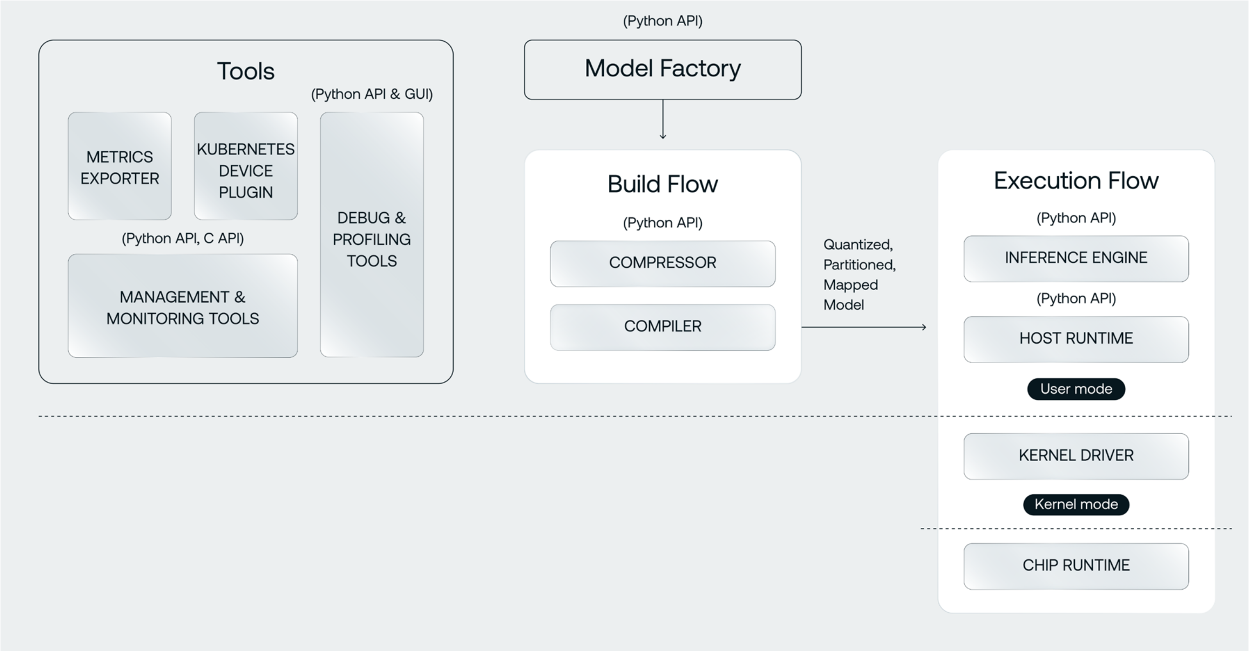


Figure 12. Aviator software

The Build Flow starts with the Model Factory, which is a template of PyTorch models that includes augmentations to run the models efficiently on d-Matrix hardware; this is similar to what users would do with other tools such as vLLM or TRT-LLM. Model Factory supports several popular open generative models, making it easy for users to use as-is, and provides tools to allow users to parallelize their models, make modifications, or bring their own models. Users may specify model weights from Hugging Face or provide their own weights.

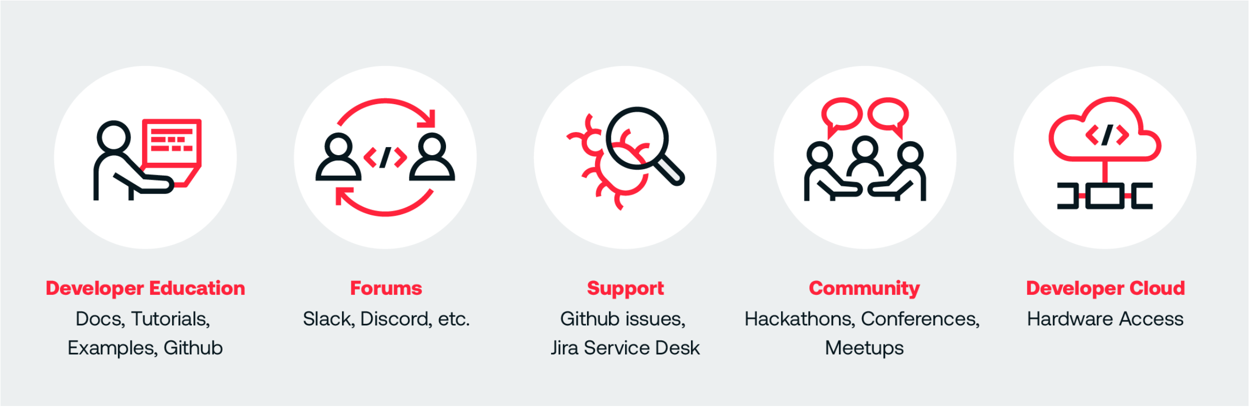
Compressor is a user-friendly model compression toolkit built on top of torch.fx. It simplifies Post-Training Quantization (PTQ) and Quantization Aware Training (QAT) while fully exploiting novel datatypes, such as MX formats, and weight sparsity available in Corsair. Compressor augments the model by configuring precision and sparsity for the different operators in the graph. This compressed model is provided as input to the Aviator Compiler for creating d-Matrix specific binaries.

The Aviator Compiler ingests the FX graph from Compressor and performs a series of graph optimizations targeted to the d-Matrix hardware. This includes tensor sharding, padding, integrating optimized kernels such as KV Cache, performing tiling and operator fusion, scheduling, and minimizing data movement across multi-level memory hierarchies. These optimizations are seamlessly applied to create compiled binaries without requiring explicit user intervention.

During the Execution phase, Aviator Inference Engine (IE) deploys compiled binaries efficiently on d-Matrix accelerator cards. It operates as a distributed system, orchestrating the requests over a cluster of 'worker' processes, one for each card. Workers manage cards using an efficient host runtime that interfaces with the on-card runtime (firmware). The Execution stack improves upon existing serving frameworks by decoupling host and on-card runtimes, allowing the host to enqueue work independently of the card's current execution state. This setup reduces launch latencies and enables Corsair to immediately take on new tasks as soon as it finishes a previous one.

For management and monitoring, Aviator provides both in-band and out-of-band tools. For orchestration, Aviator supports Kubernetes and comes with a Kubernetes Device Plugin and Metrics Exporter. Tools such as Aviator Debugger and Profiler are available for developers to dive deeper to profile and optimize their inference workloads.

d-Matrix is building a variety of resources for developers to easily deploy Corsair for their applications. This includes developer documentation, tutorials, videos, and reference models on GitHub. The d-Matrix developer community will be able to collaborate with each other and the d-Matrix team via online forums and hands-on developer focused events. Finally, d-Matrix will provide qualified developers with early access to the hardware through the Developer Cloud.



Developer Education
Docs, Tutorials,
Examples, Github

Forums
Slack, Discord, etc.

Support
Github issues,
Jira Service Desk

Community
Hackathons, Conferences,
Meetups

Developer Cloud
Hardware Access

Figure 13. Developer resources

The Corsair Advantage

With the unique combination of cutting-edge hardware and software innovation and codesign, Corsair delivers a significant advantage on interactive performance, power efficiency, and total cost of ownership.

Performance Mode and Capacity Mode

Corsair offers two modes of operation. Users choose their preferred mode based on latency and throughput requirements as well as cost-performance tradeoffs. In Performance Mode, the best inference latency and throughput performance is achieved when the model weights and KV Cache fit in the integrated Performance Memory of the Corsair cards. Depending on the model size, context length, and batch size, the user would distribute the workload across multiple Corsair cards. On the other hand, Capacity Mode is recommended when the workloads are not latency sensitive and can take advantage of large off-chip Capacity Memory (up to 256 GB). Users can run large models, longer context lengths or large batch sizes, and optimize for performance-TCO.

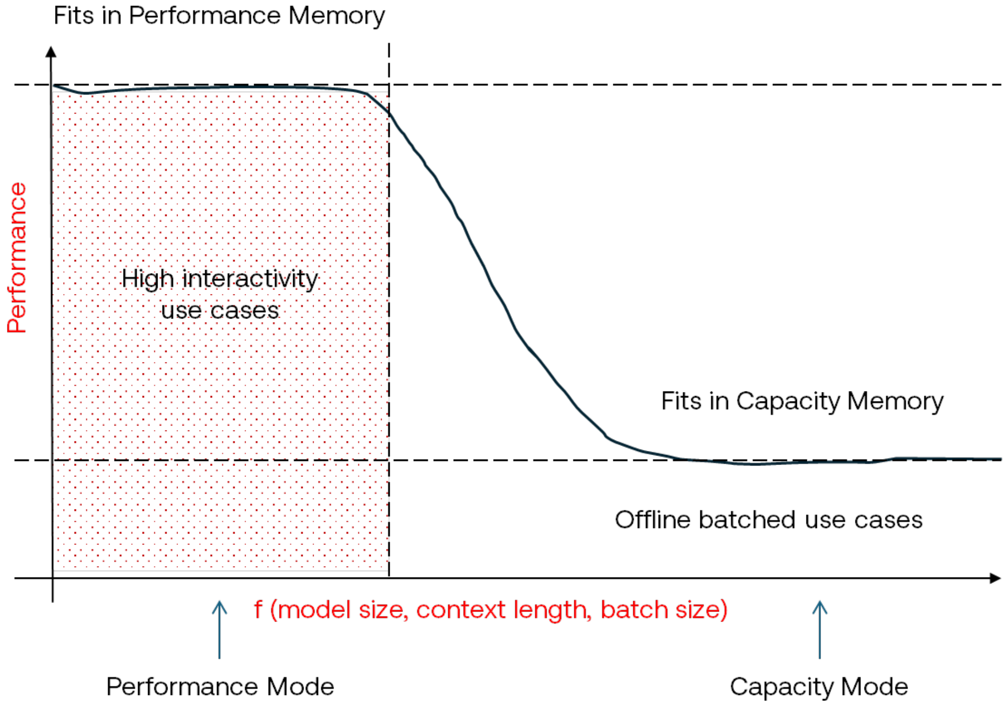


Figure 14. Corsair offers two modes of operation, covering a broad range of enterprise applications

Aviator enables users to make the choice of Performance Mode versus Capacity Mode, based on their latency, throughput, and cost constraints. By enabling both modes in the same product, Corsair unlocks value in both latency-sensitive real-time scenarios as well as offline batched scenarios.

Latency-bound Batched Inference

For inference, batching is a trade-off between throughput and latency. Higher batch sizes can enable better hardware utilization and thereby higher throughput, but this increases latency since batching requests implies performing more computation to get the first results. Let us consider the roofline performance model to understand why batching comes at the cost of latency. As shown in Figure 15 below, for lower batch sizes, the workload is memory bound. When batch size increases beyond the knee of the curve, it switches to the compute bound regime. Beyond this point, increasing batch size just increases the latency without increasing throughput. This is the optimal batch size that achieves best throughput with minimal impact to latency.

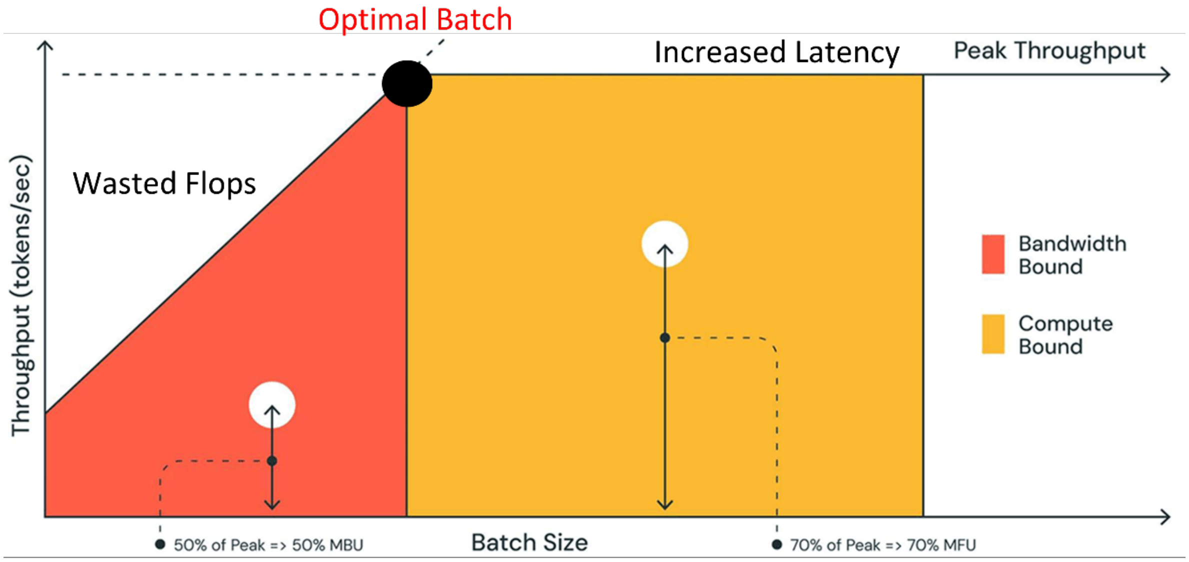


Figure 15. Source: https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices

Corsair versus GPU Performance Comparison[2]

This section describes the Corsair advantage compared to H100 GPU with various input and output token ratios for Llama2 7B model, which uses Multi-Head Attention (MHA), and Llama3 70B model, which uses Group-Query Attention (GQA). Corsair performance projections are for Performance Mode using MXINT8 numerical format. For GPU, the performance was measured using TensorRT-LLM on 8x H100 SXM cards with FP8 numerical format.

For this comparative analysis, both Corsair and GPU are operated with a latency constraint, and the batch size that enables the best latency-bound throughput is selected. Interactivity is measured using time per output token metric. This is representative of the token generation speed. Lower is better for interactivity and user experience. The efficiency advantage is measured using the throughput per card metric. In order to fit the Llama3 70B model weights

[2] Performance projections are preliminary and subject to change; results may vary.

and KV Cache in Performance Memory, 64 Corsair cards are used. For GPU, 8 cards are used. Hence, normalized throughput per card is used for the comparison. Batch sizes for Corsair versus H100 are shown on top of the bars.

Figures 16 and 17 shows the relative performance of Corsair versus H100 for the Llama2 7B model (MHA) and Llama3 70B model (GQA), respectively. As seen in the charts, the time per output token for Corsair is an order of magnitude faster than H100 for both models, while delivering higher throughput per card. Since each Corsair server consumes lower power and typically costs less than an H100 server, d-Matrix Corsair delivers significant overall cost-performance as well as an energy efficiency advantage over H100.

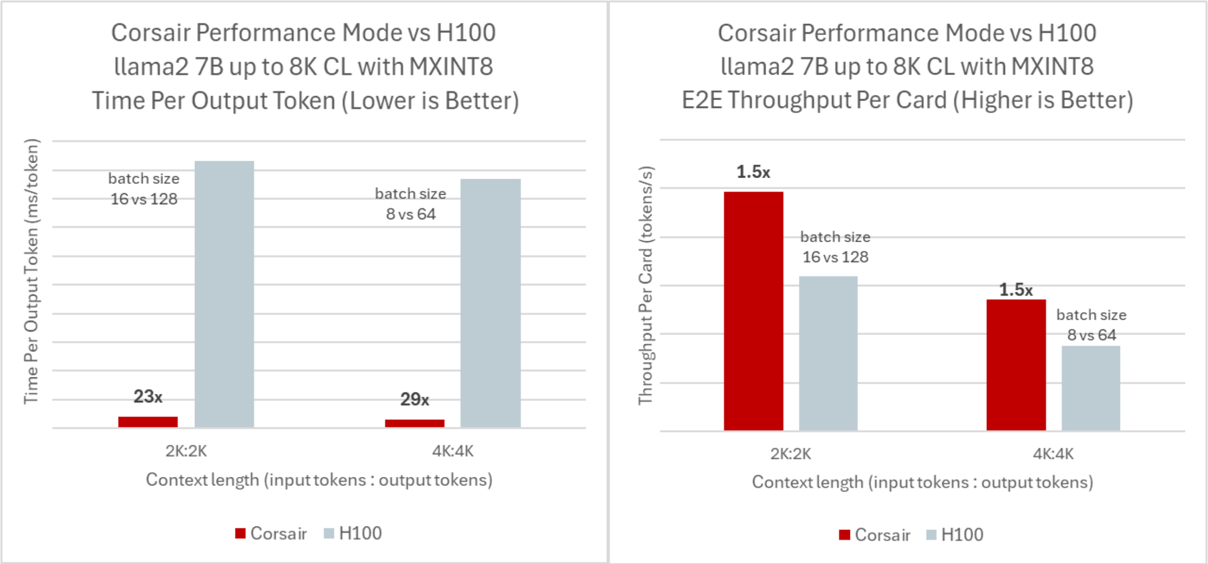


Figure 16. Corsair vs H100 'Time Per Output Token' and 'Throughput Per Card' advantage for Llama2 7B model

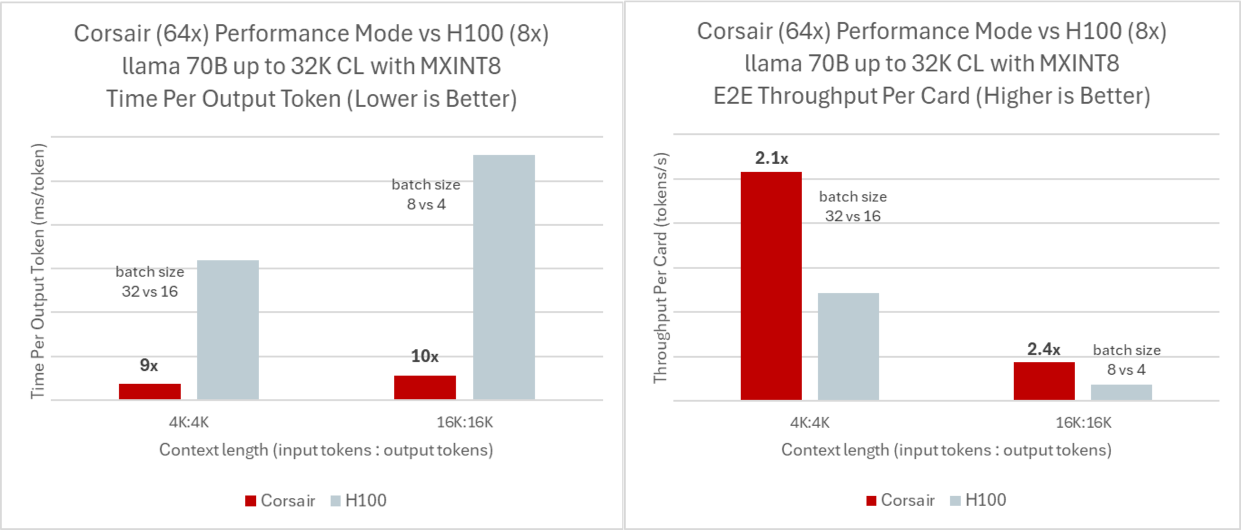


Figure 17. Corsair vs H100 'Time Per Output Token' and 'Throughput Per Card' advantage for Llama3 70B model

The table below shows Corsair performance projections for Llama3 8B in a single server with 8 cards and Llama3 70B in a single rack with 8 servers (64 cards) using MXINT8 and running in Performance Mode.

Llama3 Model	Precision	Input Length	Output Length	Batch Size	Time Per Output Token (ms/token)	Throughput (tokens/s)
8B	MXINT8	1024	1024	48	1	60,000
70B	MXINT8	1024	3072	64	2	30,000

Table 3. Corsair performance projections for Llama3 8B and 70B models

In summary, Corsair delivers significant advantages over alternatives for latency-bound throughput performance, and its ultra-high memory bandwidth enables ultra-low latency interactivity.

Conclusion

GenAI has the potential to transform industries and revolutionize how businesses harness and create information. However, this transformative potential comes at a significant cost, both economically and environmentally, due to the energy-hungry and cost prohibitive nature of currently available inference solutions. d-Matrix is ushering in a new era of AI computing with Corsair—a purpose-built inference solution that leverages novel Digital In-Memory Compute (DIMC) architecture, chiplets-based scaling, efficient block floating point numerics, and state-of-the-art Aviator software.

Corsair redefines single rack economics for AI inference, achieving 10x higher interactivity, 3x better cost-performance, and 3x better energy efficiency compared to GPU alternatives. With Corsair, generative inference deployments at scale can benefit from latency-bound batched inference, with high throughput and ultra-low latency at the same time.

To unleash the full potential of GenAI and make it widely accessible, it needs to be delivered in an affordable and sustainable way, without sacrificing performance. With Corsair, d-Matrix is transforming generative AI from unsustainable to attainable and commercially viable.