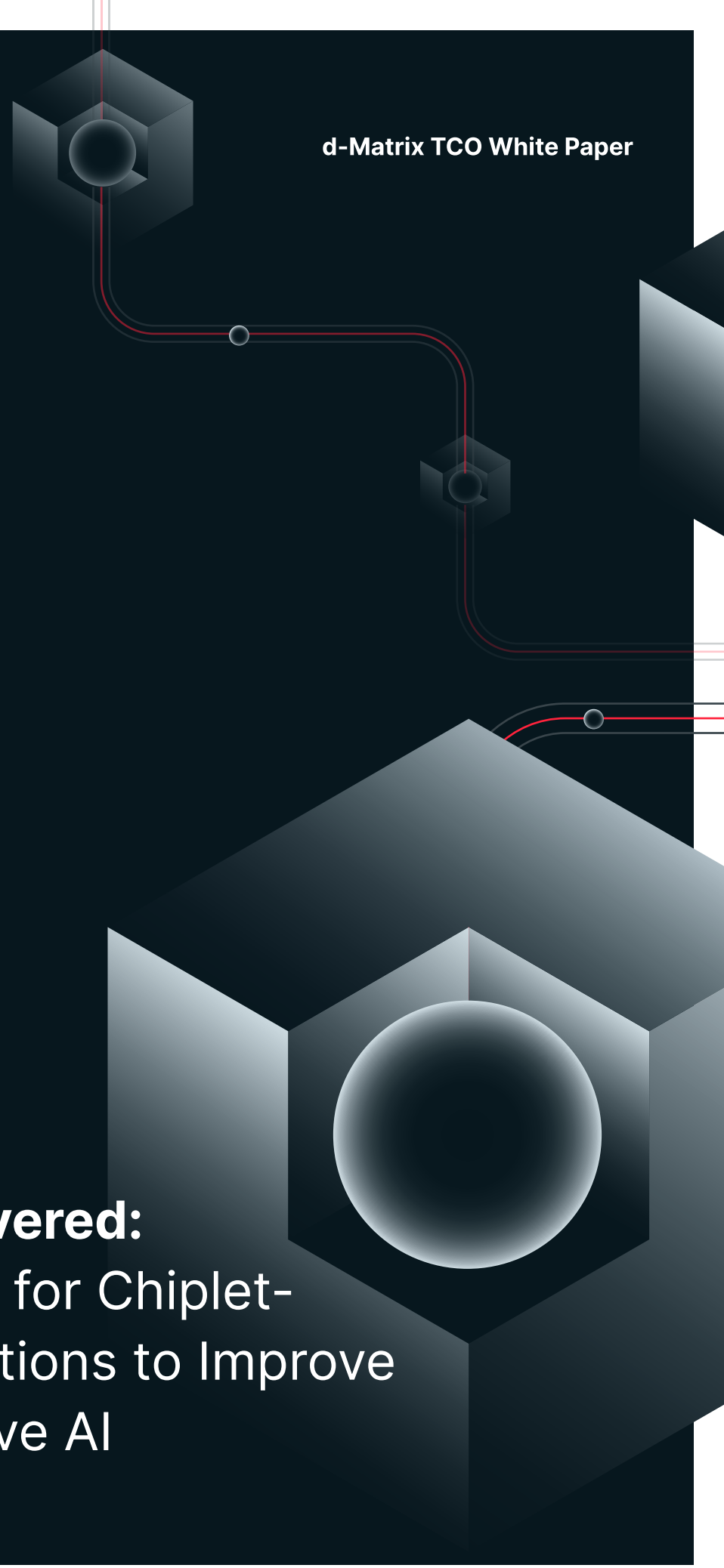
d-Matrix

d-Matrix TCO White Paper



Intelligence Delivered: The Urgent Need for Chiplet-based DIMC Solutions to Improve TCO for Generative AI

d-Matrix Total Cost of Ownership White Paper

Changing AI Economics: d-Matrix Corsair Generates the Entire Contents of Wikipedia in Less Than a Day

Table of Contents

V.1 - Sep 2023

Introduction

Generative AI is poised to reinvent how companies create new information and utilize the information they already have. Generative AI will dramatically improve the efficiency of knowledge workers across the globe and will change how business operates forever. This realization was crystallized in a "big-bang" moment with the release of ChatGPT. ChatGPT is one of OpenAI's Large Language Models (LLMs) and has produced an explosion of interest in generative AI.

The race is now on to develop and operate generative AI applications. Today, applications such as ChatGPT are run on Graphics Processing Units (GPUs). The problem with GPUs is they are not optimized for inference. Too many GPUs are needed and they use eye watering quantities of electricity. The result is a big upfront purchase, a huge electric bill and a devastating carbon footprint for the companies running AI applications. Currently, the promise of generative AI is unattainable. In order to make generative AI widely accessible, it needs to be delivered at an affordable cost and in an earth-sustainable way. This requires a purpose-built solution that is performance and power optimized for AI inference.

This white paper discusses the reason why it is so costly for companies to run generative AI inference on GPUs and introduces a solution that will reduce the cost to run LLMs by 10x to 20x and even up to 60x in some use cases.

Today, if someone wanted to use LLaMA 2 to generate as much content as all of Wikipedia, she would need to generate 5.7 billion tokens to produce the 4.3 billion words in the online encyclopedia. With a single inference node using the d-Matrix solution, an AI model could reproduce the entire Wikipedia repository in just 18 hours - completing the job in less than a day thanks to a 9x improvement in throughput compared to state of the art GPUs.

While GPUs are incredibly powerful for gaming or mining cryptocurrency, their performance is suboptimal for running generative AI. The unique memory bandwidth demands of running AI inference results in GPUs spending most of the time idle, waiting for more data to transfer in from Dynamic Random Access Memory (DRAM). Along with reduced throughput and added latency, moving data in and out of DRAM also requires energy that drives up power and cooling costs. But GPUs have been the best available solution until now.

The only way we can deploy AI in a cost effective and sustainable manner is by using compute solutions that are purpose-built for generative AI. By ensuring sufficient memory bandwidth for high throughput — and ultimately, efficiency — the inference cost can be reduced by orders of

magnitude. D-Matrix has created a Digital In Memory Compute (DIMC) architecture that significantly enhances key metrics for large transformer-based inference, improving performance by orders of magnitude.

- Total cost of ownership (TCO) improvement of 13-27x compared to GPUs when running LLaMA2-13B models with 4K context
- 20x better power efficiency
- 20x lower latency
- 40x higher memory bandwidth

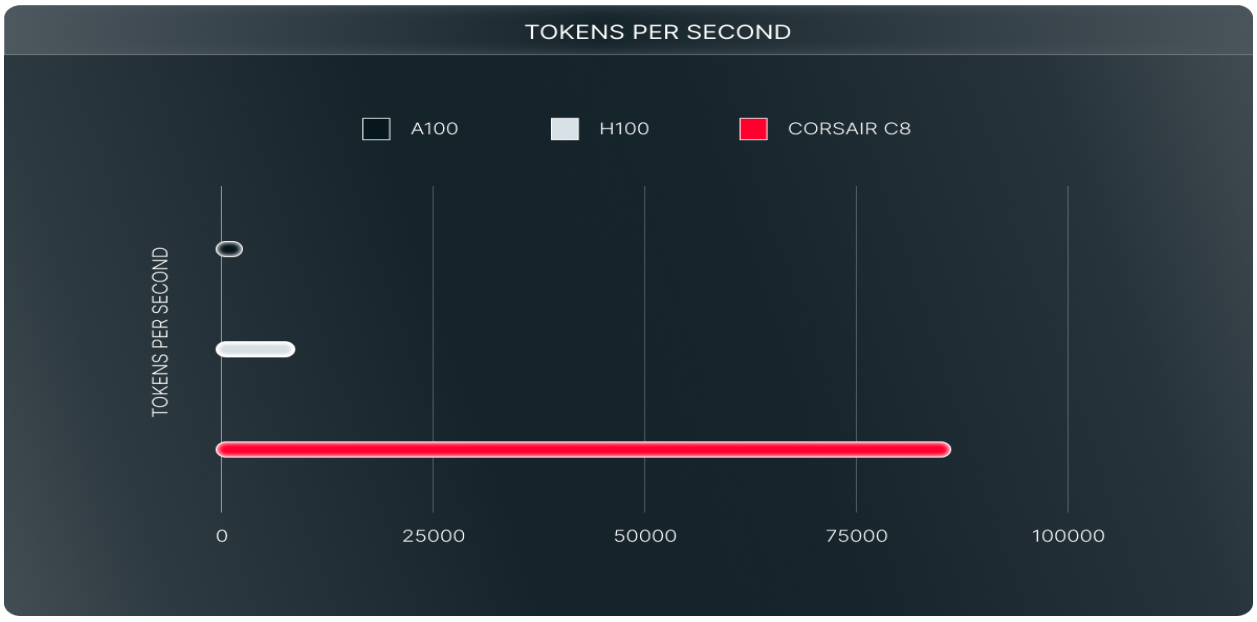
This white paper outlines a solution with:

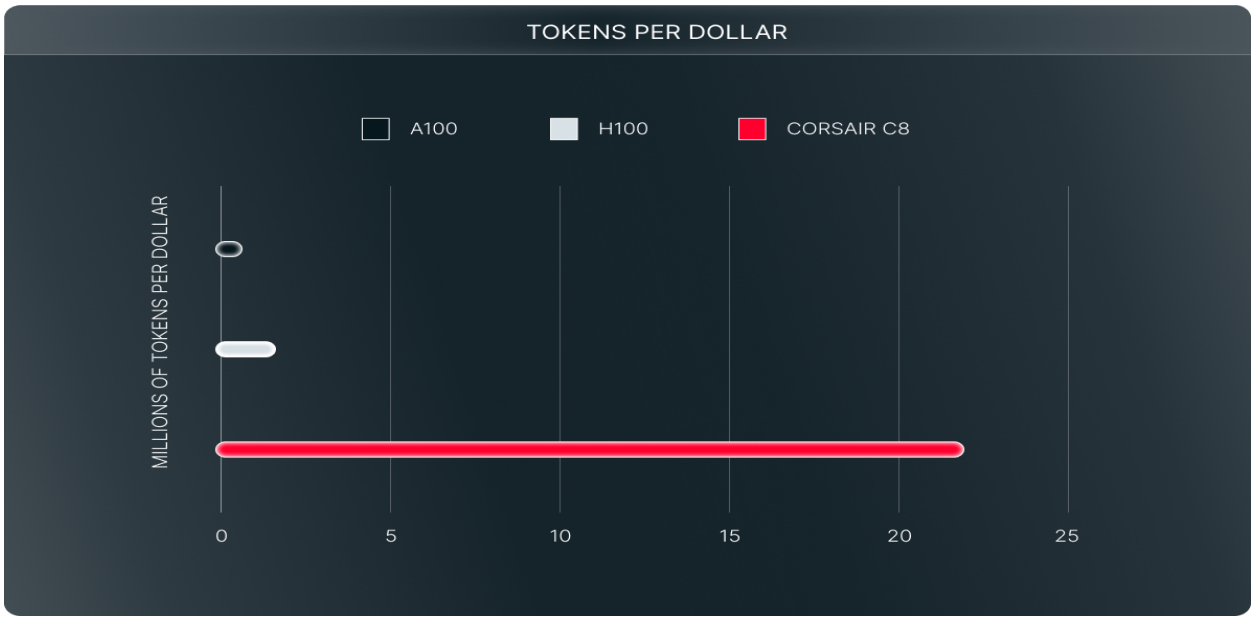
- Extremely high memory bandwidth using DIMC
- A super-fast chiplet interconnect for low latency communications
- A large number of FLOPS/W achieved by aggregating DIMC chiplets

d-Matrix Corsair TCO & Performance Comparison

	Corsair C8	A100	H100	Improvement C8 vs. A100	Improvement C8 vs. H100
Tokens per second	86,770	3,184	9,552	27.3x	9.1x
Millions of tokens per dollar	21.9	0.8	1.6	27.4x	13.7x

LLaMA2-13B, 4K context; results are preliminary and subject to change





Demand for Data Center Compute

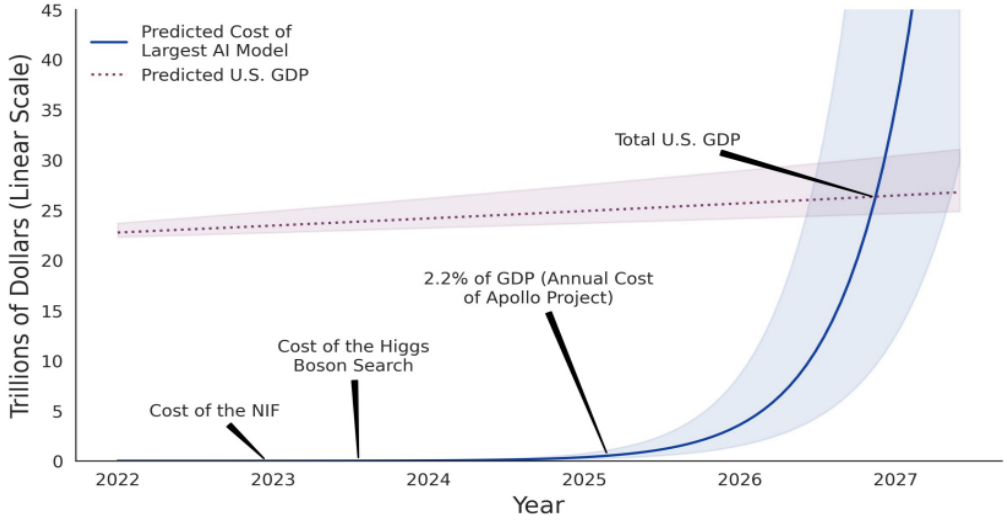
> *"The training phase for GPT-4 alone cost over $100 million," OpenAI CEO Sam Altman in Wired.*

Generative AI creates substantial challenges when it comes to environmental sustainability. Current hyperscale systems are not just prohibitively expensive to operate, but demand exceedingly high power requirements to operate and cool large arrays of GPU-based AI processors. To better grasp the energy-intensive character of AI, consider the figures associated with running models like ChatGPT. For instance, a SemiAnalysis study reveals that ChatGPT-3 relies on a staggering number of nearly 29,000 NVIDIA GPUs to serve up answers, and has a daily operational cost exceeding $694,000. [Source: SemiAnalysis]

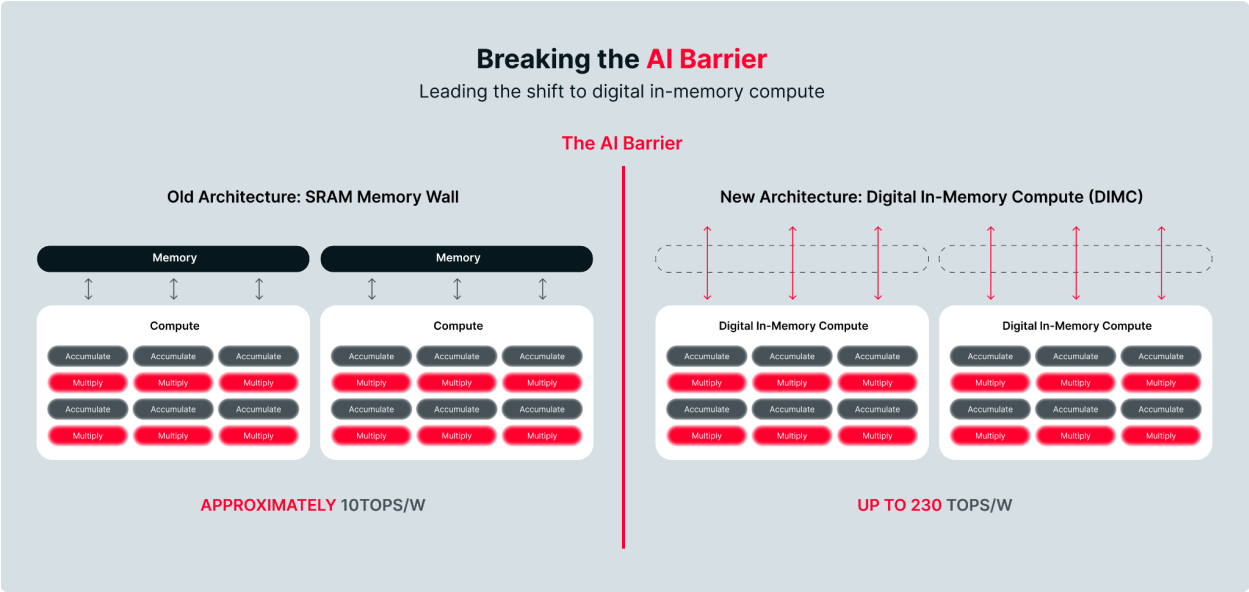
These numbers do more than just illustrate financial expenditure, they also underscore a deep-seated environmental issue. The power consumption—and corresponding CO2 emissions—of LLMs and other generative AI models accumulates rapidly. Consider that merely augmenting major search engines with LLMs could require up to five times more computational power. [Source: Wired] A 5x increase in computing power would require a tremendous investment in server farms. If data centers are going to meet burgeoning AI demands while mitigating their carbon footprints, they urgently need innovative solutions. The Center for Security and Emerging Technology graph below outlines how compute costs will soon overtake the total U.S. GDP.

> *"AI will consume more energy than the human workforce," by 2025, unless significant strides are made in efficiency, according to Gartner.*

Source:CSET, Georgetown University, 2022



Digital In Memory Compute Creates New Efficiency



Artificial intelligence applications require substantial data manipulation and computation, which present unique memory-related challenges. Traditional von Neumann architectures that separate compute from memory rely on continuous data and weight value transfers between memory and the processing unit. This process, especially when intermediate computation results need to be saved back to memory, wastes energy in data transfer and increases latency due to data movement. These inefficiencies result in substantial heat generation, driving additional cooling costs. DIMC is a new approach that changes how data is stored, moved, and processed, which dramatically improves efficiency and latency.

To manage data transfer challenges, conventional accelerators use a grid-organized method. This technique allows for simultaneous operation (or parallelization) and efficient data reuse (spatial multicasting of the operands) as the data is fetched from memory. These accelerators consist of a two-dimensional grid of processing elements (PEs). Each PE contains a multiply-accumulate (MAC) unit for basic math operations and small storage files to hold the required data (operands). This data is retrieved from larger external memory storage and distributed across the PEs in a 'spatial unrolling' strategy. However, this transfer process from external storage to PEs continues to be a performance bottleneck in modern accelerators. [Source: Arxiv]

To mitigate data transfer inefficiency, technology vendors are developing innovative approaches, such as In-Memory Computing solutions. The core concept of IMC is to bring compute and memory together, and there are both analog and digital approaches to IMC. While Analog IMC

offers significant energy efficiency and supports massive parallelization, its spatial mapping capabilities are limited, and factors like circuit noise can compromise results. Digital IMC, on the other hand, is less energy efficient but offers noise-free operation and more flexibility in spatial mapping. [Source: Ibid]

D-Matrix has pioneered a way to move compute calculations into Static Random Access Memory (SRAM) in a way that fundamentally alters the physics of memory-compute integration. D-Matrix leverages a unique DIMC engine composed of custom digital circuits that directly integrate compute into programmable memory. D-Matrix created a multiplying element by using a 6T SRAM bit cell with four additional transistors. However, unlike traditional SRAM solutions, d-Matrix activates all word lines simultaneously. This technique reduces the memory footprint and relieves memory bandwidth pressure, helping to save power while enhancing the performance of transformer models.

Chiplets Provide Flexibility

As semiconductor device fabrication has improved, we've seen the smallest possible gap between transistors shrink from 22 nm in 2012 down to 3 nm today, with 2 nm chips available soon. These advancements have come from improved clean room techniques, more precise manufacturing and improvements in materials science. But similar to the slowdown in Moore's Law, the basic laws of physics are making it harder to continue to shrink the distance between transistors on an individual die. At the same time, the large number of relatively simple AI calculations combined with the need for high-bandwidth interconnects has pushed AI processor dies to ever larger sizes, which results in more manufacturing errors and chips with defects that can't be used by customers.

In response, the semiconductor industry has been undergoing a significant paradigm shift over the past decade from using monolithic processor dies to using a "chiplet" based approach with smaller chips, connected with high-bandwidth interconnect. This shift is a response to the increasing challenges associated with traditional Moore's Law scaling, where the continuous miniaturization of transistor components has become exceedingly difficult and expensive. [Source: Futurum Research]

In chiplet architectures, manufacturers create different functions separately rather than fabricating all components on a single, large chip. The chiplet-based approach offers greater scalability, improved manufacturing efficiency and increased design flexibility. By stitching these nodes together with high-speed interconnects, d-Matrix can ensure quick and efficient communication between chiplets and make them function as a cohesive unit.

Optimizing DIMC Chiplets for Generative Inference

The d-Matrix solution includes a blend of advancements that tackle difficult challenges related to memory bandwidth, memory capacity and computational needs during AI inference. The d-Matrix Jayhawk 2 solution delivers up to 150 TB/s throughput in memory bandwidth, a significant jump from the 3 to 4 TB/s that High Bandwidth Memory (HBM) is able to stream on modern GPUs.

d-Matrix leverages chiplets to pack the compute, memory, and networking needed for large language models into a single PCIe card, which would be impossible otherwise. D-Matrix employs 8 chiplets to achieve a large 2GB memory capacity. This, coupled with the previously mentioned 150 TB/s memory bandwidth, offers considerable advantages over GPUs for specialized models.

To ensure that the chiplets appear and function as monolithic chips, d-Matrix uses an 8 TB/s die-to-die interconnect.

Write Once, Read Many

LLMs are trained using techniques like Next Sentence Prediction and masking to determine what words are missing in a sample sentence (e.g. “I like to XXX YYYY cream” would return the words “eat” and “ice” to complete the sentence). During training, many unique calculations determine the average weight and likelihood that one token of information will follow another specific token of data.

With AI inference, each query is hitting the same database of pre-trained weights to predict the next token in the sequence. While data changes frequently during training, inference calls for each token to be calculated against a relatively static set of weights. New calculations are conducted on the same set of weights for each token generated.

One advantage of the d-Matrix architecture is that instead of transferring data in and out of the GPU as you would in a traditional architecture, d-Matrix keeps weights loaded in SRAM to rapidly compute multiple inferences without transferring data over the system bus.

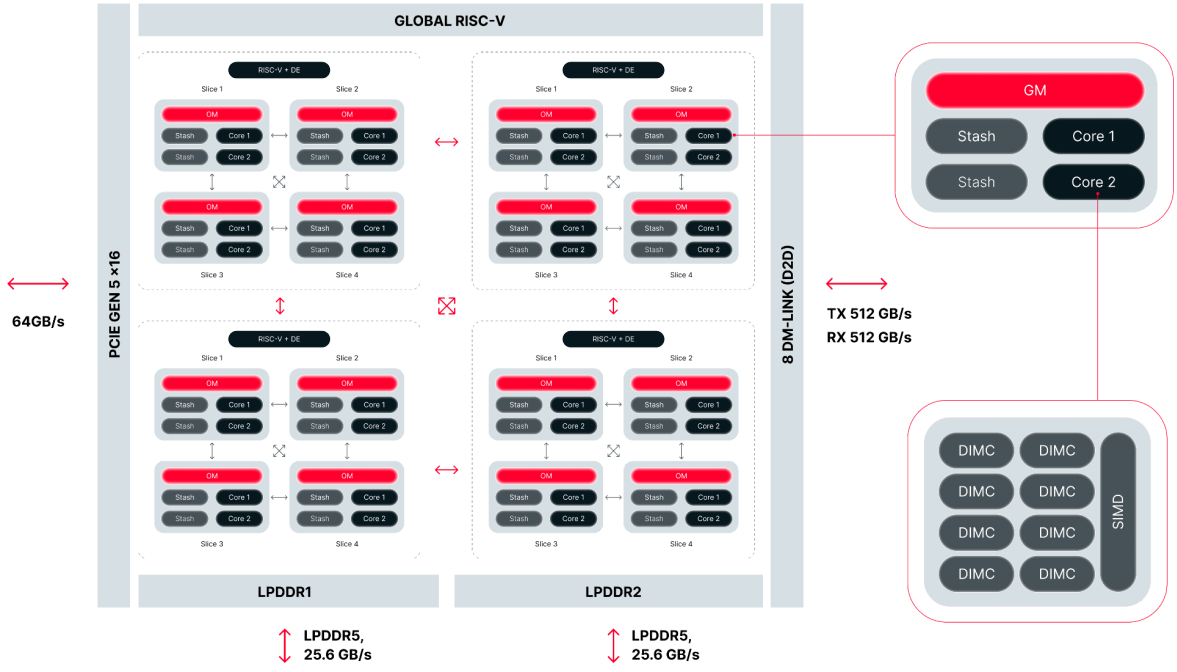
Corsair - a Unique Approach to Latency and Efficiency

The team at d-Matrix has a history of working on leading technologies at a variety of companies including Inphi, Broadcom, Intel and Lucent. Corsair is the first generative AI computing platform developed by d-Matrix. The d-Matrix Corsair computing platform is focused on addressing inefficiencies and the unique needs of AI inference compute.

Introduction to Corsair

With the exploding popularity of Chat GPT, GPT4, Bard, Falcon, LLaMA and many other LLMs, efficient generative inference has become one of the biggest challenges in datacenters. The key reason that this is challenging is because tokens are generated one by one, and causes the application to have a very low arithmetic intensity and be severely memory bandwidth limited. Since no other solutions are available, GPUs like Nvidia's A100 and H100 are deployed to manage the demand. These and other GPUs have been designed for ML training applications, which are very compute intensive – but not efficient for memory-bound workloads like token generation.

Corsair is a chiplet optimized for generative inference. A high-level block diagram follows.

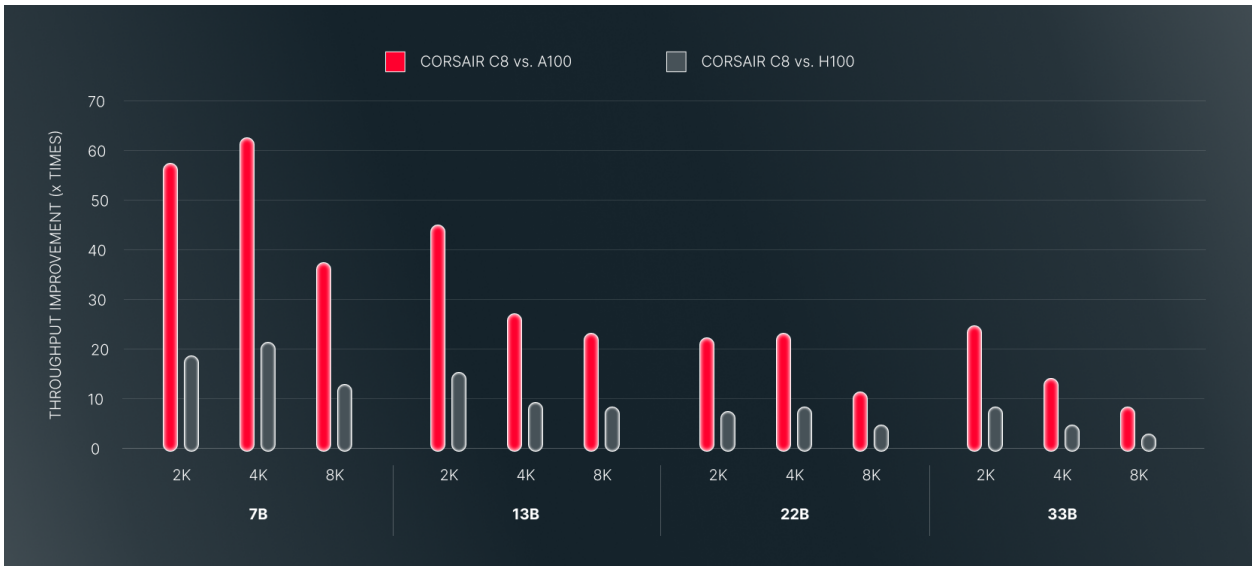


Corsair Throughput And TCO Advantage

In this section, we discuss the token throughput and TCO performance achieved by the d-Matrix Corsair C8. Corsair C8 is compared to projected performance of Nvidia A100 and H100 systems.

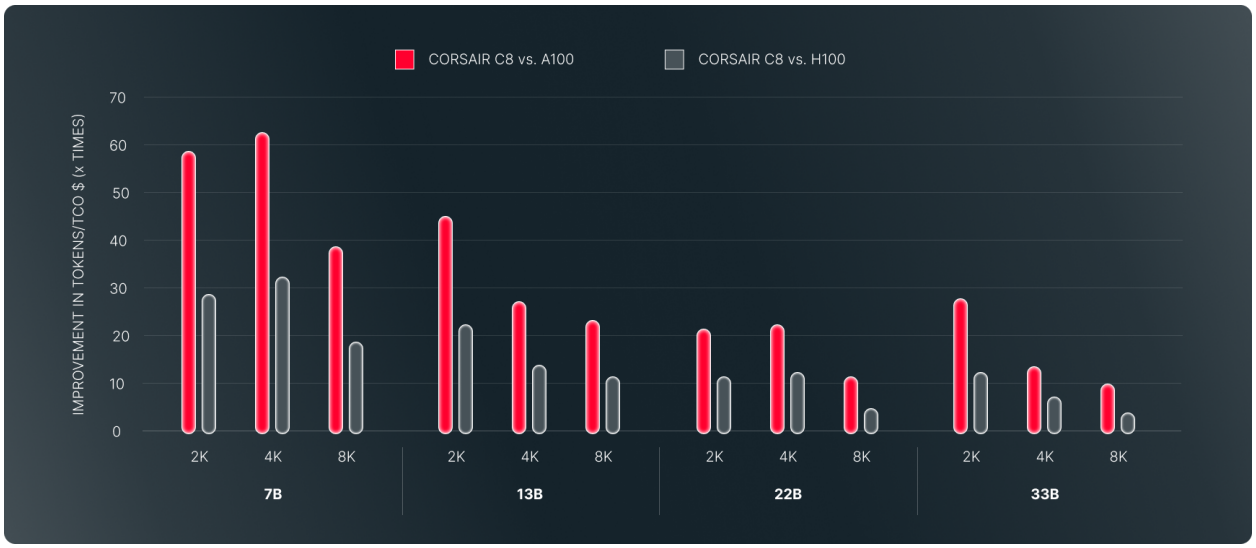
While we restrict competitive comparisons to only these options, we expect similar trends to hold for other accelerators whose main storage for model weights and KV cache is HBM-memory (e.g. AMD MI300X and Google TPU).

Throughput Improvement



Corsair throughput (generated tokens/s) improvement for LLMs of different sizes and different context lengths for d-Matrix Corsair (modeled) vs Nvidia A100 and H100 (projected). Beam search decoding with beam=4 and 75:25 input:output tokens is assumed. Results are preliminary and subject to change.

TCO Improvement



TCO improvement with Corsair for LLMs of different sizes and different context lengths. Beam search decoding with beam=4 and 75:25 input:output tokens is assumed. Results are preliminary and subject to change.

Conclusion

Generative AI stands at the forefront of transforming industries and revolutionizing how businesses harness and create information. The advent of LLMs, exemplified by ChatGPT, has ignited a paradigm shift in how we approach AI-driven innovation. However, this transformative potential comes at a significant cost, both economically and environmentally, due to the energy-hungry nature of current GPU-based inference solutions.

This white paper has delved into the inherent challenges posed by GPU-based inference, shedding light on the inefficiencies arising from memory bandwidth limitations. To meet the growing demand for AI-powered applications, it is crucial to address these limitations with innovative solutions that offer improved efficiency and sustainability.

The d-Matrix Digital In Memory Compute architecture, presented in this paper, emerges as a pioneering breakthrough in the field of generative AI. By leveraging chiplet-based design principles, d-Matrix will offer an AI inference solution that offers a 20x better performance-power profile, up to 40x more memory throughput, and a TCO improvement of 13-27x compared to traditional GPUs. By moving compute calculations into SRAM, d-Matrix is able to fundamentally alter the physics of memory-compute integration.

With a combination of extremely high memory bandwidth using DIMC, a super-fast chiplet interconnect for low latency communications, and a large number of FLOPS/W achieved by aggregating DIMC chiplets, d-Matrix is poised to pave the way for cost-effective generative AI innovation and a new era of AI advancements with the Corsair processor.