ANALYST BRIEF
Sponsored by: d-Matrix

Businesses benefit from customizing smaller, open generative AI models for inference at scale using new high-performance, low-latency, high-throughput, and cost-effective purpose-built inference accelerators, especially those available in broadly supported PCIe or OAM industry form factors.

Inference at Scale for Generative AI Models Can Be Faster, Cost Efficient, and Power Efficient

October 2024

Written by: Peter Rutten, Research Vice President, Performance-Intensive Computing, Worldwide Infrastructure Research

AT A GLANCE

KEY TAKEAWAYS

- Production inference deployments do not necessarily need the same high-end, expensive, and hard-to-purchase GPUs that are used for model training.
- Purpose-built PCIe and OAM accelerator cards are especially suitable for enterprises managing high-volume, real-time inference requests on GenAI models in high-performing, rack-scale environments.
- Built with disruptive technologies such as in-memory computing, they can deliver significantly lower latency and better throughput than GPUs.
- The cards can be added to existing hardware or bought bundled with newly purchased hardware. Disaggregation empowers enterprises and the ecosystem to build and deploy efficient GenAI inference solutions.
- Cards come with an inference-focused software stack that is integrated with popular open source software tools and frameworks, making it easy to use.

Introduction

How enthused — or concerned — should today's enterprises be about generative AI (GenAI)? The amount of attention given to GenAI in mainstream media suggests something revolutionary is happening. The investment in GenAI, especially in hardware, is remarkable. Tech giants are investing tens of billions of dollars in massive compute clusters that are equipped with thousands — sometimes tens of thousands — of expensive, high-end GPUs. The level of investments and the narrative around it indicate the coming of a new "AI race." AI-generated art, music, texts, and images are sparking both excitement and concern. Technology, culture, and even politics seem to be converging in this transformative moment. For any organization, the question is this: To what extent does this affect us, and how should we respond?

Let's take a step back. Cloud service providers and hyperscalers are making the majority of these significant investments because they have the means to do so. Their primary purpose is to train increasingly larger and more accurate GenAI models, which are then made available for commercial use by enterprises and other organizations. Businesses can leverage these models for a variety of use cases such as chatbots, content creation, and code generation. GenAI models hold the potential to transform employee efficiency and customer experience. However, there are several caveats when it comes to deploying them in production — often referred to as "running inference on the model." These commercial models are often priced by the token (the smallest unit of "intelligence" — text, image, or video — used by a model) and can get expensive when

deployed at scale. In addition, their capabilities tend to be generalist, not specifically fine-tuned to an industry or a use case. Often, inferencing at scale also requires performance-intensive computing (PIC) infrastructure focused on inference acceleration.

For small-scale AI initiatives that do not require a certain amount of customization to an organization's distinct business process, use case, or an industry, generalist models can be a workable approach. But once any of these characteristics come into play, alternative approaches are required, such as:

- **Customizing or fine-tuning:** When a generalist model is insufficiently tailored, organizations have no choice but to develop (train) their own models or use pretrained models that they can revise through additional training (fine-tuning) with their own data. Training a model from scratch can be prohibitively complex and expensive. Instead, organizations can use more focused, high-quality open models that are much more manageable (smaller) and easier to fine-tune with less costly infrastructure. With the rapidly growing open model ecosystem, there are several high-quality open models available for many specific use cases.
- **Retrieval-augmented generation (RAG):** With RAG, the model output can be grounded in an organization's own proprietary data. This technique involves organizing the company's data in a vector database and retrieving it in real time to improve the model's outputs. This is an increasingly popular strategy, as it helps improve the model's outputs, keeps sensitive or proprietary data out of the model itself, and makes its use more secure or compliant.
- **Inferencing at scale:** Once a model is ready to go into production, its intended scale of use becomes a critical factor. Will a few hundred sales associates use the model for sales support? Or will it serve as a real-time AI agent for thousands — or even hundreds of thousands — of end users? The latter scenario is becoming increasingly common.

But what about the compute infrastructure? Given the compute cost and power requirements for deploying GenAI models, as well as the memory bandwidth requirements for faster inference times, businesses need PIC infrastructure that enables efficient large-scale inferencing.

Benefits of Specialized PCIe Cards for Inference at Scale

Production inference deployments do not necessarily need the same high-end, expensive, and hard-to-purchase GPUs that are used for model training. Cost, power, and high memory bandwidth requirements for faster inference times mean businesses need specialized compute (i.e., PIC infrastructure that enables large-scale inferencing efficiently). There are alternative accelerated AI compute solutions in the market from large incumbents as well as small start-ups. These come in the form of processors, PCIe cards, and even entire systems, while some are available only on the cloud.

An industry standard form factor (PCIe or OAM) is especially suitable for deploying in a high-performing, rack-scale environment that can manage high-volume, real-time inference requests on GenAI models:

- PCIe cards are simply added to existing hardware or bundled with newly purchased hardware by a systems integrator, without introducing either new architectures or entirely new platforms into the datacenter. This enables the datacenter operators to optimize inference compute efficiency with their existing rack infrastructure, space, and power constraints.

- These accelerator cards come with an inference-focused software stack that is integrated with popular open source software tools and frameworks, making it seamless to run inference on models that are trained on GPUs. Even better, sometimes, the vendors that develop these cards have independent software vendor (ISV) relationships that allow them to optimize their product for specific ISV software.
- These custom-built PCIe cards can perform better than some incumbent solutions on GenAI inference. They are purpose built to address memory bandwidth–bound GenAI workloads. Built with disruptive technologies such as in-memory computing, they deliver significantly lower latency and better throughput. As a result, these cards enable a better user experience for GenAI applications. They are especially promising for emerging agentic and reasoning applications.
- With both lower cost and lower power consumption than incumbents, these new PCIe accelerator cards offer a more sustainable solution to address the increasing power demands of GenAI inference deployments. They are typically less costly to purchase and operate and can help deliver an attractive ROI on TCO for large-scale deployments.

Considerations

Organizations should take a pragmatic approach to their GenAI development and deployment. They should not let the hype that surrounds cloud service provider and hyperscaler AI development — which boasts ever-larger commercial and generalist models, ever-larger clusters to train them on, and ever-larger infrastructure investments to build these clusters — blind them.

Instead, organizations can customize smaller open source GenAI models that are widely available and of very high quality to their specific needs. And they can run inference at scale using new disruptive acceleration technologies available in the PCIe form factor with better performance (latency and throughput), lower power, and lower cost than high-end GPUs.

Organizations should take a pragmatic approach to their GenAI development and deployment and not let the hype that surrounds cloud service provider and hyperscaler AI development blind them.

Conclusion

Once a GenAI model is ready to go into production, its intended scale of use becomes a critical factor. Given the compute cost and power requirements for deploying GenAI models, as well as the memory bandwidth requirements for faster inference times, businesses need PIC infrastructure that enables large-scale inferencing efficiently. Production inference deployments do not necessarily need the same high-end, expensive, and hard-to-purchase GPUs that are used for model training. There are alternative accelerated AI compute solutions in the market. Specialized PCIe cards are especially suitable for deployment in a high-performing, rack-scale environment that needs to manage high-volume, real-time inference requests on GenAI models. They are simple to add, come with an inference-focused software stack, and can deliver higher performance than GPUs at lower power consumption and cost.

About the Analyst

Peter Rutten, *Research Vice President, Performance-Intensive Computing, Worldwide Infrastructure Research*

Peter Rutten's runs IDC's Performance Intensive Computing research practice, which covers infrastructure stacks and deployments for AI, HPC, and quantum computing.

MESSAGE FROM THE SPONSOR

d-Matrix Is Transforming GenAI From Unsustainable to Attainable and Making GenAI Commercially Viable

As organizations embrace Generative AI, they must consider newer compute acceleration solutions to deploy GenAI applications efficiently at scale and maximize their ROI. d-Matrix is a visionary, five-year-old company focused on redefining the economics of Generative AI inference at scale, making it broadly viable for commercial applications. Built on a foundation of many world-first innovations across silicon, software, chiplet packaging, and interconnect technology, the seasoned and experienced d-Matrix team has created a GenAI inference compute platform that excels at blazing fast token generation (tokens/s), cost efficiency ($/token) and power efficiency (tokens/watt.) To learn more, visit www.d-matrix.ai.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com