NEWS RELEASE

Tachyum Democratizes AI for All with $5000 Prodigy ATX Platform

2/8/2024

LAS VEGAS--(BUSINESS WIRE)-- **Tachyum** ®, creator of Prodigy®, the world's first Universal Processor, today released a white paper that details how its Prodigy ATX Platform will democratize AI for those who may not normally have access to sophisticated AI models. The Prodigy ATX Platform allows everyone to run cutting-edge AI models for as low as $5,000.

Built from the ground up to provide leading-edge AI features that address the emerging demand for AI across a wide range of applications and workloads, Prodigy's AI subsystem incorporates innovative features that deliver the high performance and efficiency required of AI environments. The white paper shows how a single Prodigy system with 1 Terabyte (TB) of memory can run a ChatGPT4 model with 1.7 trillion parameters, whereas it requires 52 NVIDIA H100 GPUs to run the same model at significantly higher cost and power consumption.

Since LLMs (Large Language Models) are so memory capacity intensive, determining the memory footprint for an LLM is critical. Just as critical is the use of the latest technology that optimizes the memory footprint for state-of-the-art LLMs, which can have trillions of parameters. Prodigy benefits from its advanced AI subsystem that supports leading-edge data types such as 4-bit TAI and effective 2-bit weights with FP8 per activation that greatly reduces the memory footprint required for LLMs.

Since the Prodigy ATX Platform is intended to leverage pre-trained models and focus on inference, Tachyum reviews the assumptions for the memory footprint required for inference. Assuming an LLM with 1 trillion parameters running with FP8, the memory required for weights is 1 TB. There is additional overhead for inference memory that is typically 0.2x the model size, or 200 GB, that is added for runtime calculations for activations. For FP8, the total memory required for a 1 trillion parameter model is approximately 1.2 TB.

Considering Tachyum's 4-bit TAI with 4-bit weights, the memory needed for weights is reduced to 500 GB and the runtime inference memory is fixed at 200 GB for a total requirement of 700 GB. Now, considering running TAI sparse with 2-bit weights the memory required for weights is further reduced to 250 GB. With the 200 GB runtime inference memory, the total requirement is 450 GB.

If we repeat these steps for the 1.7 trillion parameter ChatGPT4 LLM: the inference memory is 0.2 x 1.7 trillion = 340 GB, and the total memory needed for FP8 is 1.7 TB for weights + 340 GB = 2.04 TB. Going from FP8 to 4-bit TAI the weights require 2x less memory, so the memory consumed by the weights is 1.7/2 = 850 GB, and the total memory requirement is 850 GB + 340 GB = 1.19 TB.

If we go to TAI sparse with 2-bit weights, the memory requirement for weights is reduced to 425 GB. If we now add the 425 GB to the memory required for inference, 425 GB + 340 GB, we come up with a total memory footprint requirement of 765 GB, which fits well below the 1 TB of commodity system memory that the Prodigy ATX Platform offers, and there is plenty of headroom, so even larger LLMs can be supported.

Key architectural components of the Prodigy ATX Platform outlined in the white paper include:

- Single-socket 96-core Prodigy Universal Processor running up to 5.7 GHz with 8 DDR5 memory controllers
- 16 64 GB commodity DIMMs (2 DIMMs/channel) running up to DDR5-6400 with a total memory capacity of 1 TB
- 3 PCIe 5.0 slots that support up to full height and full-length form factors:
    - 1 slot x16 with 16 lanes
    - 2 slots x16 with 8 lanes
- 3 M.2 NVMe slots supporting 22x80mm form factor
- 1200W power supply

The platform benefits from Prodigy's unique "half die" solution that allows a full 192-core device to function as two separate 96-core devices. This architecture provides Tachyum with increased yield for 96-core devices, lowering platform costs and helping make the Prodigy ATX Platform even more affordable.

The Prodigy ATX Platform addresses an extensive range of use cases, such as language generation, language translation, code generation, virtual tutoring, content summarization, sentiment analysis, fraud or cyber-attack detection, and content filtering. The platform benefits the many pre-trained LLMs available today with support for both proprietary and open-source models.

"Generative AI will be widely used far faster than anyone originally anticipated," said Dr. Radoslav Danilak, founder

and CEO of Tachyum. “In a year or two, AI will be a required component on websites, chatbots and other critical productivity components to ensure a good user experience. Prodigy’s powerful AI capabilities enable LLMs to run much easier and more cost-effectively than existing CPU+GPGPU-based systems, empowering organizations of all sizes to compete in AI initiatives that otherwise would be dominated by the largest players in their industry.”

Prodigy provides both the high performance required for cloud and HPC/AI workloads within a single architecture. As a Universal Processor offering utility for all workloads, Prodigy-powered data center servers can seamlessly and dynamically switch between computational domains. By eliminating the need for expensive dedicated AI hardware and dramatically increasing server utilization, Prodigy reduces CAPEX and OPEX significantly while delivering unprecedented data center performance, power, and economics. Prodigy integrates 192 high-performance custom-designed 64-bit compute cores, to deliver up to 4x the performance of the highest-performing x86 processors for cloud workloads, up to 3x that of the highest performing GPU for HPC, and 6x for AI applications.

Those interested in reading the “Tachyum’s Prodigy ATX Platform Democratizing AI for Everyone” white paper can download a copy at **https://www.tachyum.com/resources/whitepapers/2024/2/8/tachyums-prodigy-atx-platform-democratizing-ai-for-everyone/**.

Follow Tachyum

https://twitter.com/tachyum
https://www.linkedin.com/company/tachyum
https://www.facebook.com/Tachyum/

About Tachyum

Tachyum is transforming the economics of AI, HPC, public and private cloud workloads with Prodigy, the world’s first Universal Processor. Prodigy unifies the functionality of a CPU, a GPU, and a TPU in a single processor to deliver industry-leading performance, cost and power efficiency for both specialty and general-purpose computing. As global data center emissions continue to contribute to a changing climate, with projections of their consuming 10 percent of the world’s electricity by 2030, the ultra-low power Prodigy is positioned to help balance the world’s appetite for computing at a lower environmental cost. Tachyum recently received a major purchase order from a US company to build a large-scale system that can deliver more than 50 exaflops performance, which will exponentially exceed the computational capabilities of the fastest inference or generative AI supercomputers available anywhere in the world today. When complete in 2025, the Prodigy-powered system will deliver a 25x multiplier vs. the world’s fastest conventional supercomputer – built just this year – and will achieve AI capabilities 25,000x larger than models for ChatGPT4. Tachyum has offices in the United States and Slovakia. For more information, visit **https://www.tachyum.com/**.

Mark Smith
JPR Communications
818-398-1424
marks@jprcom.com

Source: Tachyum