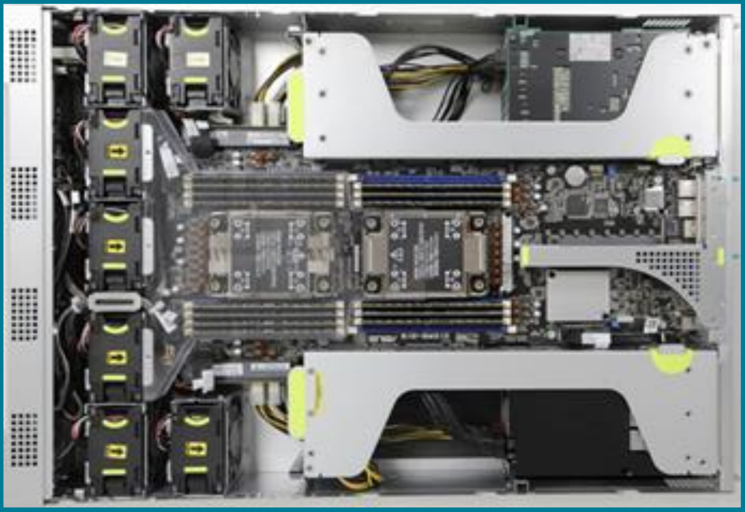


HyperAccel Taps AMD Accelerator Card and FPGAs for New AI Inference Server

AMD Alveo™ U55C Accelerator Card and AMD Virtex™ UltraScale+™ FPGAs Help HyperAccel's Orion Servers Accelerate Transformer-Based Large Language Models, Including Llama 3

PARTNER

HYPER ACCEL

INDUSTRY

Consumer

CHALLENGES

As the use of large language models (LLMs) expands, there is growing demand for efficient, fast, and cost-effective inference solutions. Conventional hardware, with its limited performance and frequently high cost, has slowed the adoption of Generative AI in many applications.

SOLUTION

HyperAccel's Orion server implements a specialized processor tailored for the inference of LLMs, based on multiple, high-performance AMD FPGAs.

RESULTS

An AMD Alveo card and AMD Virtex UltraScale+ FPGAs are helping HyperAccel deliver the performance and scalability that it needed for various workloads.

AMD TECHNOLOGY AT A GLANCE

AMD Alveo™ U55C accelerator card
AMD Virtex™ UltraScale+™ FPGA

HyperAccel is a South Korea-based startup that was founded in January 2023. The company specializes in developing AI-specific semiconductors and hardware for inference systems that maximize memory bandwidth usage and increase cost efficiency by applying this solution to a large language model. The company provides hyper-accelerated silicon IP/solutions for emerging Generative AI applications. It has created a fast, efficient, and affordable inference system that accelerates transformer-based large language models (LLMs) with multi-billion parameters, such as OpenAI's ChatGPT and Meta's Llama LLM–including Llama 3. Its AI chip, named Latency Processing Unit (LPU), is a hardware accelerator dedicated for end-to-end inference of LLMs.

CHALLENGE

As the applications of LLMs expand, there is an escalating demand for efficient, fast, and cost-effective inference solutions. For cloud service providers, fast and cost-effective inference hardware is essential for hosting Generative AI applications with high performance and reduced total cost-of-ownership (TCO). For AI companies, a platform with an intuitive software stack is required to seamlessly deploy their applications or models. For service businesses, the provision of a full end-to-end solution is necessary to integrate state-of-the-art AI technology for more effective and advanced service.

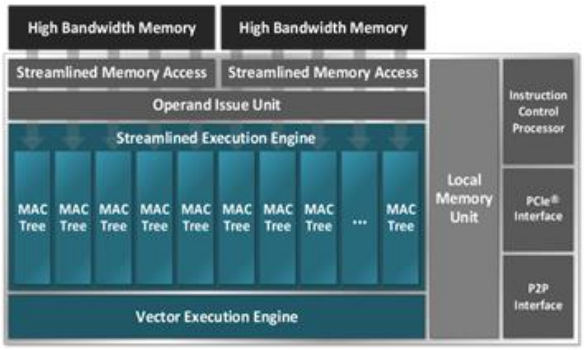
SOLUTION

HyperAccel proposed to address the cost and performance issue by developing "Orion," which is a server that implements a specialized processor tailored for the inference of LLMs, on multiple high-performance AMD FPGAs. Orion fully leverages the memory bandwidth usage and hardware resources of each FPGA for maximum performance. The scalable architecture supports the latest LLMs, which consist of multi-billion parameters.

Orion has 16 latency processing units (LPUs) on two 2U chassis to provide a total of 7.36 TB/s HBM bandwidth and 144K DSPs. LPU accelerates hyperscale Generative AI workloads that are both memory- and compute-intensive. Orion and its 256 GB HBM capacity support state-of-the-art LLMs with up to 100B parameters. The figure above shows one of the two 2U chassis with 8 LPUs.

The figure below shows the LPU architecture where the vector execution engine is enabled by the AMD Alveo™ U55C high-performance compute card. The Alveo U55C card packs high bandwidth memory (HBM2), addressing the most critical performance bottleneck to delivering low latency AI, which is memory bandwidth. In addition, they are capable of 200 Gbps of high-speed networking into a single slot, small form factor card, and are designed for deployment in any server.

In turn, each Alveo accelerator card is powered by FPGA fabric. The inherently low-latency nature of FPGAs is perfectly suited for real-time AI services as is needed in LLMs, given the FPGAs' massive hardware parallelism and adaptable memory hierarchy. The Alveo card features the powerful Virtex™ XCU55 UltraScale+™ FPGA, which delivers up to 38 TOPS of DSP compute performance optimized for fixed- and floating-point compute, including INT8 for AI inference. This FPGA enables hardware adaptability to change the architecture of their processor (LPU) based on customer feedback, e.g., requesting something non-standard in a Llama model, and thereby becoming a flexible solution adaptable to changing market and thus LLM parameter conditions.



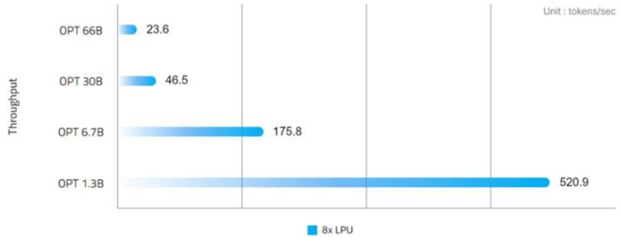
RESULT

Orion's high-performance and scalability is achieved by the LPU, enabled by the AMD Alveo accelerator card and the associated FPGAs, as well as HyperAccel's Expandable Synchronization Link (ESL) technology, which maximizes memory bandwidth usage for agile processing and eliminates the synchronization overhead in P2P computing. ESL is a communication link optimized for data transfer in LLM inference. Notably, Orion maintains impeccable accuracy with the hardware that supports standard FP16 data precision.

HyperAccel Orion Performance

Latency-optimized HyperAccel Orion seamlessly integrates with transformer-based LLMs like GPT, Llama, and OPT to generate more than 520 tokens/sec with 1.3B model and 175 tokens/sec with 7B model. In addition to outstanding performance, Orion showcases exceptional energy efficiency, generating a single token in 24 ms for a 66B model with a modest power consumption of only 600W.

Performance of HyperAccel LPU (source: https://www.hyperaccel.ai)



HyperAccel Orion – Workload Versatility

Orion delivers end-to-end solutions serviceable as a cloud service. For AI companies that have private LLMs or a professional sector that possesses internal data requiring data privacy and security, Orion can be installed as an on-premises solution. Orion is capable of the following workloads/applications:

- *Customer service:* Power virtual chatbots and virtual assistants to handle inquiries in real-time to free human agents to handle more complex issues.
- *Human-machine interface*: Enable language-related features in kiosks, robots, and other devices for better customer engagement.
- *Text generation*: Assist in producing, summarizing, and refining sophisticated textual content to provide convenience to users.
- *Language translation*: Translate customer queries and responses to break down language barriers and expand the reach of businesses globally.
- *Q&A:* Tailor responses to individual customers based on large pools of data along with previous interactions and preferences to increase customer satisfaction.

WANT TO LEARN MORE?

About AMD Virtex UltraScale+ FPGAs

About AMD Alveo U55C Accelerator Card

About HyperAccel

About HyperAccel
HyperAccel is a South Korea-based startup that was founded in January 2023. The company specializes in developing AI-specific semiconductors and hardware for Inference systems that maximize memory bandwidth usage and increase cost efficiency by applying this solution to a large language model. More information at: https://hyperaccel.ai/.

About AMD
For more than 50 years, AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work, and play. AMD employees are focused on building leadership, high-performance, and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and Twitter pages.