NAND RESEARCH

RESEARCH NOTE

MLPERF 4.0 INFERENCE RESULTS

STEVE MCDOWELL, CHIEF ANALYST
APRIL 3, 2024

CONTEXT

MLCommons released the results of its MLPerf Inference v4.0 benchmarks, which introduced two new workloads, Llama 2 and Stable Diffusion XL.

Since its inception in 2018, MLPerf has established itself as a crucial benchmark in the accelerator market. The benchmarks offer detailed comparisons across a variety of system configurations for specific use cases. They are instrumental in guiding buyers and industry players in understanding system performance for relevant workloads, with the latest round incorporating cutting-edge generative AI tasks.

MLPERF 4.0 INFERENCE RESULTS

The MLPerf Inference v4.0 benchmarks results included submissions from multiple tech companies, each highlighting unique strengths in the rapidly evolving AI and machine learning landscape.

Here's a summary of the results for each vendor:

Nvidia

- Dominated the benchmark submissions, being the only vendor to tackle all workloads.
- Showcased the performance of H100 and the newly introduced H200 GPUs, emphasizing advancements in TensorRT-LLM compiler for enhanced performance on Llama 2 workload.

Qualcomm

- Introduced the Cloud AI 100 Ultra in the "Closed Preview" category, showcasing significant performance improvements at lower power consumption compared to its predecessor, Cloud AI 100 Pro.

- Demonstrated exceptional performance across ML benchmarks, with notable improvements in performance efficiency for both NLP and Computer Vision networks.

Intel/Habana

- Gaudi 2 accelerator demonstrated solid performance results for state-of-the-art models on MLPerf Inference, with strong results in both Stable Diffusion XL and Llama v2-70B benchmarks.
- 5th Gen Xeon Scalable processors made their first appearance, and due to hardware and software enhancements, they showed a notable improvement in performance for general-purpose AI workloads.

Juniper Networks

- A first-time participant that highlighted the importance of networking in ML performance, submitting tests for Llama 2 running over a spine-leaf network topology.

Red Hat & Supermicro

- Collaborated on a submission leveraging Red Hat OpenShift AI, demonstrating the flexibility of OpenShift AI's model serving stack to support open-source LLM runtimes such as vLLM.

Wiwynn

- New submitter that benchmarked its ES200G2 server in both edge and data center categories, demonstrating the platform's capability to handle popular AI frameworks efficiently.

Others

- Among others, Dell, Fujitsu, Google, Hewlett Packard Enterprise, Lenovo, and Oracle submitted results across various models, showcasing their systems' capabilities in handling AI workloads efficiently.
- The submissions covered a range of platforms, from edge devices to data center solutions, and included proprietary and open-source software stacks.

ANALYSIS

The MLPerf Inference v4.0 results paint a comprehensive picture of the current state and competitive dynamics within the AI accelerator market.

The introduction of Qualcomm's Cloud AI 100 Ultra in the "Closed Preview" category shows Qualcomm's increasing presence in the AI inference space with a solution that emphasizes performance efficiency at a lower power envelope. This strategy positions

Qualcomm as a key player in the market and caters to a growing demand for energy-efficient AI processing in edge and data center applications.

Intel/Habana's results with the Gaudi 2 accelerator are particularly noteworthy, showing competitive diversity. Its strong showing in demanding workloads like Stable Diffusion XL and Llama v2-70B show that specialized architectures can provide compelling alternatives to NVIDIA’s offerings, especially in scenarios where workload-specific optimizations are critical.

Juniper Networks' participation shows the often-underappreciated role of networking in AI system performance, emphasizing the need for a holistic approach to AI infrastructure where data movement and system interconnects directly impact overall system performance.

The MLPerf Inference v4.0 benchmarks show a dynamic and evolving AI accelerator market. While Nvidia continues to dominate, the emergence of strong contenders like Qualcomm and Intel/Habana, along with the increasing importance of holistic system considerations and open-source ecosystems, show that we're heading toward a vibrant and competitive landscape.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.