Technology Insights... November 10, 2022

J.Gold Associates, LLC. Northborough, MA 01532 USA
www.jgoldassociates.com +1-508-393-5294
Research, Analysis, Strategy, Insight

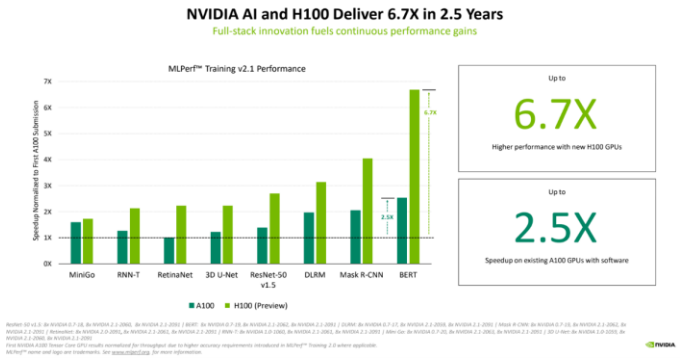
A PUBLICATION FOR CLIENTS OF J.GOLD ASSOCIATES

Nvidia's Impressive H100 MLPerf Benchmark

In the complex world of AI/ML processing, it can be hard to compare products from various vendors due to the wide range of models and workloads in use. MLPerf is a consortium of major industry players and research organizations that provides agreed-upon benchmark tests to try and standardize test results across various vendor offerings to give users a chance to evaluate competing performance claims.

Nvidia has previously provided MLPerf test results for its A100 product. It has just released its MLPerf benchmarks for its new high end device, the H100. It sports an impressive 6.7X performance gain over the older A100 devices in certain workloads, and is still being optimized with software that could eventually push the performance even higher. But as can be seen from the chart below provided by Nvidia, the performance improvements vary widely depending on the actual test. Still, even a 2X improvement in some of the other tests is impressive.

"...The H100 is an impressive device that Nvidia plans to continuously upgrade through software improvements. Indeed, this is a leading edge product that others will have trouble catching up to, at least in the short term. Nevertheless, while Nvidia is a powerhouse at the high end of the AI/ML market in cloud and on-prem instances, it does have a weaker position in the Edge market where power/performance selection criteria necessitates a reduced power footprint....."



Wide range of Workloads

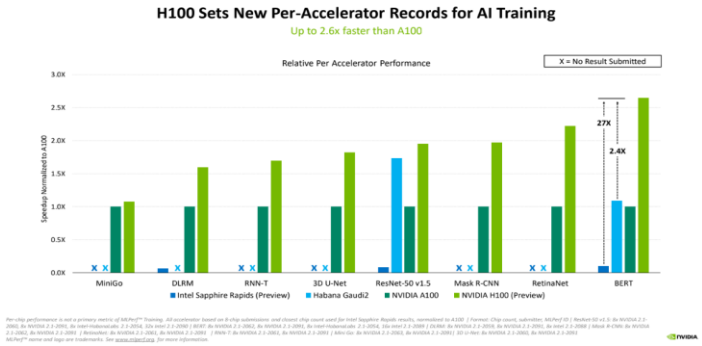
Nvidia has provided test results across a broad set of tests within MLPerf, which other companies typically fail to do. They generally select only a couple of tests to run that favors their parts (e.g., Resnet). By showing a broader range of test results, Nvidia is providing a much more nuanced view of how its products work under various workloads. Credit goes to Nvidia for doing this and we would like to see other vendors do the same for a more complete comparison of performance.

Nvidia vs Intel

Clearly Nvidia is not alone in expounding on its AI/ML performance. Intel also claims major capabilities. As a result, and in an attempt to show its advantage, Nvidia provides comparisons for the H100 to devices from Intel. This includes the Sapphire Rapids AI-accelerated CPU in its Xeon family, and the Habana Gaudi 2 AI-specific processor for select MLPerf tests for which Intel has published data. Of course, since the competing products did not test to the other benchmark tests that Nvidia includes in the chart below, there is no way to know how they would perform against the H100 (or A100 for that matter). And with Intel releasing its long awaited Pontevecchio GPU for HPC solutions, it would be good to get a benchmark comparison for it and not just for the Sapphire Rapids chip by itself, not only in training but also in inference tasks as well.

(Intel has provided some test results for comparison with the older A100, but the ones currently available are not MLPerf tests).

"…It has just released its MLPerf benchmarks for its new high end device, the H100. It sports an impressive 6.7X performance gain over the older A100 devices in certain workloads, and is still being optimized with software that could eventually push the performance even higher…."



Lots of Power Required

The H100 is a 700 Watt device, compared to the A100 which is "only" 400 watts TPD. The A100 won't be going away anytime soon, as it fills a niche for a lower cost, lower power device that can be used not only for training, but for increasingly complex inference tasks as well. Of course, there is an argument to be made that the higher performance of the H100 will still use less overall energy than the A100 as it processes the data and then goes into standby more quickly. This is a valid argument for a cloud implementation where the overall megawatt hour measurement is the governing expense metric for hyperscalers and large data centers. But in limited power applications, even a 400 watt part may be prohibitive from both a power and heat-produced standpoint.

Moving to the Edge

The A100 provides a potential Edge computing component for situations that the H100 just can't. Many edge deployments are power sensitive and the additional 300 watts could be a deal-breaker in some situations. Indeed, for many edge solutions, even the 400 Watt requirements of the A100 may be prohibitive. For such solutions, performance per watt becomes a critical benchmark. However this is not yet a key measurement criterion within the MLPerf benchmark suite (although there are discussions about making it part of the benchmarks in the future). Nvidia did not provide such a test result in its recent disclosures, but this would be a very interesting (and informative) test measurement for potential edge-related solutions.

Nvidia still needs to provide lower power training devices to compete with the likes of Qualcomm and even Intel that are upping their game to include AI training and not just inference in their product offerings. While many of the lower power devices are explicitly used for inference tasks, which generally are less compute intensive, the need to also provide edge-based model training is growing in popularity as new solutions take hold and proliferate out of the data center. Out of fairness, the H100 is clearly not targeted at that market, nor for that matter was the A100. These are squarely targeted at the hyperscaler and on-prem large system/HPC markets. Still, the Edge market is ever expanding towards the higher end and these devices may be adapted to that functionality over time.

Bottom Line: The H100 is an impressive device that Nvidia plans to continuously upgrade through software improvements. Indeed, this is a leading edge product that others will have trouble catching up to, at least in the short term. Nevertheless, while Nvidia is a powerhouse at the high end of the AI/ML market in cloud and on-prem instances, it does have a weaker position in the Edge market where power/performance selection criteria necessitates a reduced power footprint. It would be good for Nvidia to provide some sort of Edge-optimized AI/ML training solution for lower power needs.

J.Gold Associates, LLC.
6 Valentine Road
Northborough, MA 01532 USA

Phone:
+1-508-393-5294

Web:
www.jgoldassociates.com

Email:
info@jgoldassociates.com

Research, Analysis, Strategy, Insight

J.Gold Associates provides advisory services, syndicated research, strategic consulting and in-context analysis to help its clients make important technology choices and to enable improved product deployment decisions and go to market strategies.