Quick Take

The AI Hardware Summit: A Recap

Alex Norton, Bob Sorensen, Steve Conway, and Earl Joseph
November 2018

HYPERION RESEARCH OPINION

This inaugural, two-day AI Hardware summit held in Mountain View, California, at the Computer History Museum, brought together researchers, vendors, and users to explore the development of the AI ecosystem from a hardware perspective. Large companies, startups, and analysts joined to hear over 30 speakers and roundtables. Although the overall theme was all AI hardware, many of the presentations focused on AI processors and the work that is being done to design hardware to further the development of the evolving and growing field of AI, machine learning, and deep learning.

There were consistent observations about the present and future status of the AI hardware ecosystem in the presentations, the most common being that Moore's Law is nearing its end, if it is not already there, and the community needs to start exploring new technologies, architectures, and materials for chip design. Most of the presenters agreed that the concept of AI-specific hardware is essential to further the progress of AI.

HIGHLIGHTS OF THE CONFERENCE

There were three common threads to the presentations in the conference: precision and efficiency, power, and co-design.

On precision, many of the speakers noted that the precision needed to train an AI model does not need to be as high as for other traditional HPC computations, such as modeling and simulation. They indicated that to train AI models of the size needed in the future, developers will need to reduce precision—while maintaining accuracy--and this will drive down training time to workable levels.

- Many discussed the transition from floating point 32 to floating point 16, and the move to INT8 as effective, but there was also positive interest in bfp16, which alters half precision, fp16, slightly with the mantissa and exponent to enable more efficient computation.

On power, an issue not unique to AI processors, presenters noted that many accelerators today consume large amounts of power. In addition, it requires power to transfer data, and many current AI models require significant data transfer. In keeping with the co-design aspect involving software engineers and processor designers, presenters considered attacking the power issue from both sides: chip designers are working to reduce the power consumption while software engineers are attempting to make their models more computationally efficient to reduce the power on their side.

Many engineers discussed the idea of co-design, highlighting that without co-design, chips could fall short of reaching the proper functionality necessary for the growing complexity of the AI models. Presenters stressed that software engineers need to play an integral part in the design of these

processors because the most effective way to tackle the hard problems of AI moving forward is through processors designed with a specific AI task in mind.

- These chips will not be one size fits all for AI models but rather will be targeted at specific types problems that will execute the model efficiently and effectively.
- There were some engineers who said that training and inference are two separate processes, so they must run on processors with separate architectures, each tailored to either training models or inference models.

Exciting technology that emerged from the conference was abundant. A few of many examples:

- Habana Labs unveiled their extremely impressive inference chip, which is designed to process 15,000 images per second.
- Ayar Labs showed their optical I/O technology, attempting to break through the bottleneck of copper pin I/O.
- RAIN Neuromorphics displayed their attempt to mimic a neuromorphic chip after the brain to capture the power of the brain for artificial intelligence problems.

These are just a few of the many interesting presentations given.

FUTURE OUTLOOK

This was a stimulating display of what the AI hardware ecosystem has already been doing, and more importantly, what the plans are moving forward.

Co-design in hardware specific for AI is going to be crucial to effectively train and execute the models for the AI problems of today and of the future. There was a great quote during the presentation by Andrew Feldman from Cerebras Systems: *"The largest problems of today are the smallest we should consider for the future.*"

As the space of AI and the problems it will tackle expand, hardware is going to have to move at the same pace, if not faster. This conference showed there is a widespread effort to push ahead into the unknown space of post-Moore's law technologies, and exciting breakthroughs are happening around the community.

For more details about information gathered at the conference, please contact Alex Norton at anorton@hyperionres.com.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
