New AI Hardware Products and Trends

April 3, 2019

Alex Norton, HPC Research Analyst

In This Presentation

- This presentation will highlight some of the trends in emerging technologies associated with the AI ecosystem.
  - Much of the information in this presentation is broad trends of the overall market in terms of the emerging technology trends.
  - There will also be a few highlights of interesting technologies being produced and incorporated into the ecosystem.

An Exploding Ecosystem

- The AI ecosystem has been growing very quickly
- Targeted problems are more complex
- One result: new, AI-specific hardware
  - From companies large and small
  - China is very active
- The categories of processors and technologies continue to grow as well, and now include:
  - GPUs
  - TPUs
  - FPGAs
  - ASICs
  - Neuromorphic Chips
  - IPUs
  - Inference Chips
  - Training Chips
  - Dataflow processors
  - Vector processors
  - 3D stacking
  - Optical interconnects

Co-Design

- AI chips will be centered on co-design, with specific tasks in mind. Examples:
  - Low-power ASICs at the edge
  - Custom AI chip in hyperscale data centers or the cloud
- GPUs will remain important but not for all AI workloads.
- Software and model-designed hardware is the direction forward.

Power

- Power consumption is critical to chip design for AI workloads.
  - Low power chips can be placed closer to the edge.
  - Latency for near real time training and inference require the compute to be next to the stored data.
    - ADS use cases are a perfect example of this. Automotive chips for automated driving need to be able to process right where the data is being collected, with the lowest possible latency.
- Processing and memory also need to be closer together.
  - With faster interconnect/fabric speeds

Cloud Companies Joining the Processor Development Party

- Google uses tensor cores to accelerate machine learning workloads.
  - Only available on Google cloud for now
  - Google announced the third generation TPU last year.
- Amazon, at their re:Invent conference in November of 2018, announced their inference chip, Inferentia.
  - Designed to accelerate machine learning, especially inferencing.

Startups

- AI startups are finding strong investment interest from VC companies.
  - The first large wave of processors target image processing, especially live stream video processing and large batch image processing, for various tasks across many verticals.

Examples of Startups

- Cerebras Systems
  - A small company started by Andrew Feldman that designs efficient AI systems as a whole, working to eliminate bottlenecks instead of moving them around.
- Ayar labs
  - A company emerging from a DARPA funded lab that designs optical interconnect solutions to allow for a higher data throughput.
- DeePhi (Now a part of Xilinx)
  - Develops deep learning hardware and software aimed at markets from edge devices to the data center.

HYPERION RESEARCH

Questions?

Thank you!