IBM Advances Research in Analog AI Computing

IBM IS TARGETING 2 ORDERS OF MAGNITUDE IMPROVEMENT IN POWER EFFICIENCY

In these early days of AI acceleration, speed reigns supreme. As the market and technologies mature, however, power efficiency and cost will become prime factors. NVIDIA GPU's currently rule the roost, however custom ASICs such as those offered by Intel Habana, Google, Graphcore, and others are just now coming to market to challenge the leader. Meanwhile, some researchers have begun to experiment with an entirely different approach: analog computation in memory. Analog minimize data movement from memory to compute registers, lowering costs and power. While startup Mythic has announced their first product using this innovative approach, IBM Research has announced their own effort, along with advances in digital AI cores and packaging technologies. These new technologies could lower power demands and increase performance for inference and even for training in the future. In this brief, we take a look at the IBM analog technology.

The IBM Thomas J. Watson Research Center in New York.

What is Analog Computing and what are the benefits?

Analog computing is a dramatic shift from typical digital calculations. Traditional digital computers are programmed with step-by-step logic, storing values in binary formats. Analog "computers" embody the desired calculations in their circuit topology, and the analog output of the circuit is the "answer" to the calculation.

Analog In-Memory Calculation

Training a large deep neural net (DNN), or executing the trained model (inference), requires trillions or quadrillions of calculations to determine and apply the synaptic weights in the network. Traditional von Neumann-based computers store their data in RAM or other memory, and they shuffle the data in and out of the CPU or GPU for calculation. These shuffle operations consume a lot of time and energy. DNN training passes can stretch on for weeks, and DNN model sizes are doubling roughly every 3.5 months.

In addition to developing digital AI innovations, IBM is developing an entirely different approach that enables massively parallel *analog* computation in-memory using continuously variable circuit values to represent numbers. As an analogy, think of your stereo. The musical output – the voltage that's applied to the speakers -- is an analog signal that's controlled by the circuitry and the sound input. The stereo circuits do analog "calculations" to determine the appropriate output.

Instead of using sound signals and volume knobs as the inputs, analog computation uses analog values and circuit topology to calculate the desired result. The output signal of the analog circuit represents the answer almost instantly, just like the stereo produces the desired audio signal.

IBM's analog calculations take place in memristive phase-change memory (PCM). Each memory cell contains memristive components that can store analog values (synaptic weights in AI models). Each cell can also be configured with the circuit topology required to perform analog calculations – locally, in the memory cell, without having to send the data to the CPU. The memory thus acts as a large numeric array capable of high-speed parallel in-memory matrix multiplications and other operations – which is exactly what is needed for DNN training and inference.

This architecture has major advantages for performance and for power consumption, which is an important consideration for low-power mobile and Internet-of-Things (IoT) applications. And in the data center, the US Department of Energy projects that Data Centers will consume 10% of the world's energy output by 2030. If analog can significantly reduce this demand, it will win.

Being analog, the PCM cells do not hold values as precisely as a digital memory cell. IBM has determined that the precision required to train and run AI can be relaxed, without damaging the end result. IBM researchers are developing techniques of Approximate Computing (called hardware aware training) to work with inexact values and derive a result of suitable accuracy.

INFERENCE AND TRAINING

IBM Research initial focus is inference processing but has longer-term plans for training in analog. Inference is less compute-intensive than training, but it has more stringent latency requirements, while training requires more precision. Inference and training both suffer today from the von Neumann bottleneck, shuffling data back and forth between memory and CPUs. Both analog inference and training avoid this limitation. Analog could consequently deliver very low latency for inference, and high performance for training, at lower power levels than today's solutions in the digital domain. The challenge for training will be developing novel algorithms for the stochastic gradient descent to ensure required accuracy in analog implementations. If IBM can deliver on this vision, it will create a step function in improving the affordability of DNN training.

Benchmarks of a mixed-precision analog chip on the common MNIST image datasets, show the analog solution actually matching digital accuracy trained off-chip with hardware-aware training. Scalability to large neural nets is evidenced by a ResNet-9 network, running at 85.6% classification accuracy on the CIFAR-10 dataset at 10.5 TOPs per Watt and 1.6 TOPs per square millimeter.

THE FUTURE POTENTIAL IS LARGE

IBM is demonstrating some impressive results with these new technologies, but they are of course not yet product-ready. From a hardware standpoint, we could envision IBM building multi-chip modules that attach one or more Analog AI accelerators to systems, possibly using IBM's future DBHi, or Direct Bonded Heterogeneous Integration, to interconnect the accelerator to the CPU. Also note that IBM recently announced on-die digital AI accelerators as part of the next generation Z system's Telum processor. The reduced-precision arithmetic core was derived from the technology developed by the IBM Research AI Hardware Center.

Many other potential applications are possible, within IBM and with their partners. However, the technology must demonstrate excellent performance, low power, and low cost to be widely adopted. If analog can demonstrate a durable 10X performance per watt advantage over, say, GPUs or edge ASICs, then it has a good shot at success. And since IBM is targeting a 100X advantage they should achieve that goal.

CONCLUSIONS

Companies around the world are developing AI acceleration processors. Some are building GPUs, some are designing custom ASICs, and others are focused on FPGA's. But nearly all are using digital cores to try to compete with GPUs for the data center AI market. It is possible that the future could belong to those who take an entirely different path. It could be quantum computing, or photonic circuits, or analog computing in memory. IBM is uniquely positioned to apply their broad technology portfolio to any of these areas, either as proprietary products or as licensed IP. All of these technologies can claim significant potential advantages over the GPU or ASIC. And all are nascent technologies that bear watching.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTORS

Gary Fritz, Senior Analyst at Cambrian-AI Research
Karl Freund, Founder and Principal Analyst at Cambrian-AI Research

PUBLISHER

Karl Freund, Cambrian-AI Research, LLC.

INQUIRIES

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Cambrian-AI Research". Non-press and non-analysts must receive prior written permission by Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by IBM Inc. Cambrian-AI Research provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements considering new information or future events.