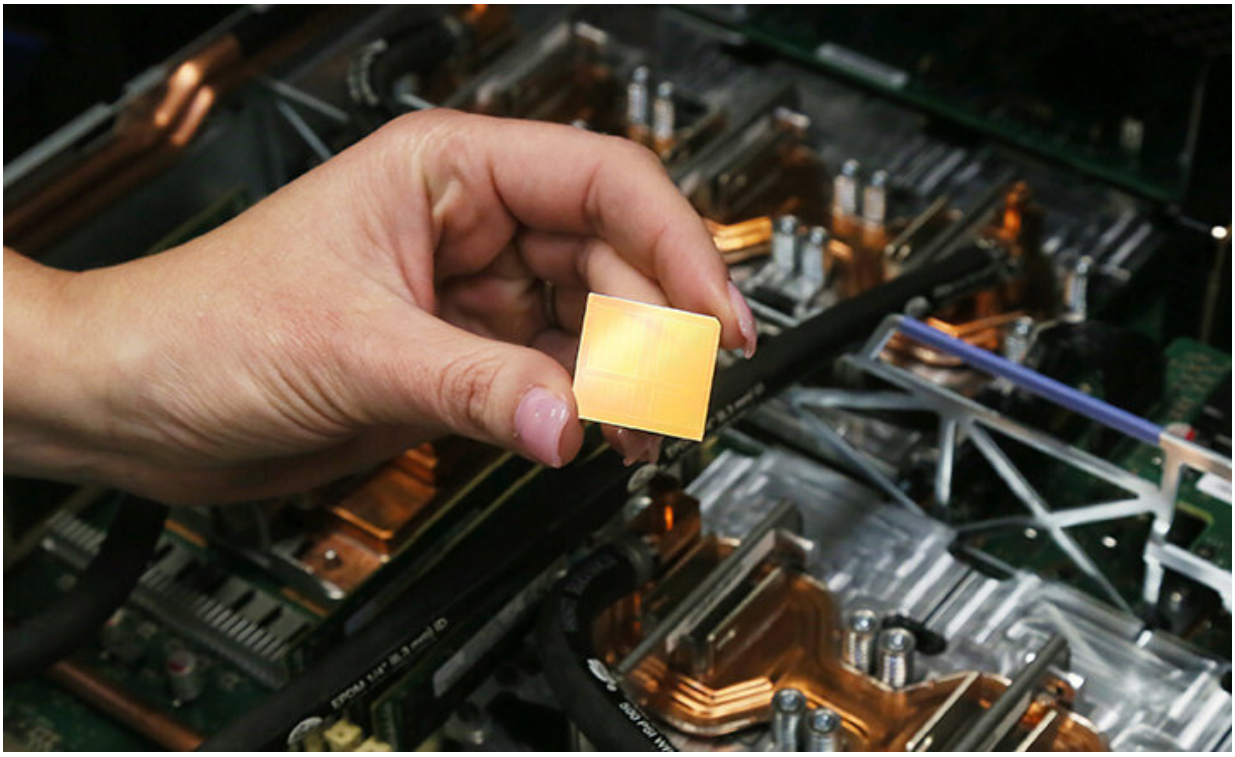
The future of hardware is AI

December 7 2017



IBM’s new POWER9 processor. Credit: IBM

AI workloads are different from the calculations most of our current computers are built to perform. AI implies prediction, inference, intuition. But the most creative machine learning algorithms are hamstrung by machines that can't harness their power. Hence, if we're to make great strides in AI, our hardware must change, too. Starting with GPUs, and then evolving to analog devices, and then fault-tolerant

quantum computers.

Let's start in the present, with applying massively distributed deep learning algorithms to Graphics processing units (GPU) for high speed data movement, to ultimately understand images and sound. The DDL algorithms "train" on visual and audio data, and the more GPUs should mean faster learning. To date, IBM's record-setting 95 percent scaling efficiency (meaning improved training as more GPUs are added) can recognize 33.8 percent of 7.5 million images, using 256 GPUs on 64 "Minsky" Power systems.

Distributed deep learning has progressed at a rate of about 2.5 times per year since 2009, when GPUs went from video game graphics accelerators to deep learning model trainers. So a question I addressed at Applied Materials' Semiconductor Futurescapes: New Technologies, New Solutions event during the 2017 IEEE International Electron Devices Meeting:

What technology do we need to develop in order to continue this rate of progress and go beyond the GPU?

Beyond GPUs

We at IBM Research believe that this transition from GPUs will happen in three stages. First, we'll utilize GPUs and build new accelerators with conventional CMOS in the near term to continue; second, we'll look for ways to exploit low precision and analog devices to further lower power and improve performance; and then as we enter the quantum computing era, it will potentially offer entirely new approaches.

Accelerators on CMOS still have much to achieve because machine learning models can tolerate imprecise computation. It's precisely because they "learn" that these models can work through errors (errors

we would never tolerate in a bank transaction). In 2015, Suyong Gupta, et al. demonstrated in their ICML paper Deep learning with limited numerical precision that in fact reduced-precision models have equivalent accuracy to today's standard 64 bit, but using as few as 14 bits of floating point precision. We see this reduced precision, faster computation trend contributing to the 2.5X-per-year improvement at least through the year 2022.

That gives us about five years to get beyond the von Neumann bottleneck, and to analog devices. Moving data to and from memory slows down deep learning network training. So finding analog devices that can combine memory and computation will be important for neuromorphic computing progress.

Neuromorphic computing, as it sounds, mimics brain cells. Its architecture of interconnected "neurons" replace von-Neumann's back-and-forth bottleneck with low-powered signals that go directly between neurons for more efficient computation. The US Air Force Research Lab is testing a 64-chip array of our IBM TrueNorth Neurosynaptic System designed for deep neural-network inferencing and information discovery. The system uses standard digital CMOS but only consumes 10 watts of energy to power its 64 million neurons and 16 billion synapses.

But phase change memory, a next-gen memory material, may be the first analog device optimized for deep learning networks. How does a memory – the very bottleneck of von-Neumann architecture – improve machine learning? Because we've figured out how to bring computation to the memory. Recently, IBM scientists demonstrated in-memory computing with 1 million devices for applications in AI, publishing their results, Temporal correlation detection using computational phase-change memory, in *Nature Communications*, and also presenting it at the IEDM session Compressed Sensing Recovery using Computational Memory.

Analog computing's maturity will extend the 2.5X-per-year machine learning improvement for a few more years, to 2026 or thereabout.

While currently utilizing just a few qubits, algorithms run on the free and open IBM Q experience systems are already showing the potential for efficient and effective use in chemistry, optimization, and even machine learning. A paper IBM researchers co-authored with scientists from Raytheon BBN, "Demonstration of quantum advantage in machine learning" in Nature Quantum Information demonstrates how, with only a five superconducting quantum bit processor, the quantum algorithm consistently identified the sequence in up to a 100-fold fewer computational steps and was more tolerant of noise than the classical (non-quantum) algorithm.

IBM Q's commercial systems now have 20 qubits, and a prototype 50 qubit device is operational. Its average coherence time of 90μs is also double that of previous systems. But a fault tolerant system that shows a distinct quantum advantage over today's machines is still a work in progress. In the meantime, experimenting with new materials (like the replacement of copper interconnects) is key – as are other crucial chip improvements IBM and its partners presented at IEDM in the name of advancing all computing platforms, from von Neumann, to neuromorphic, and quantum.

More information: Diego Ristè et al. Demonstration of quantum advantage in machine learning, *npj Quantum Information* (2017). DOI: 10.1038/s41534-017-0017-3

Provided by IBM

https://phys.org/news/2017-12-future-hardware-ai.html