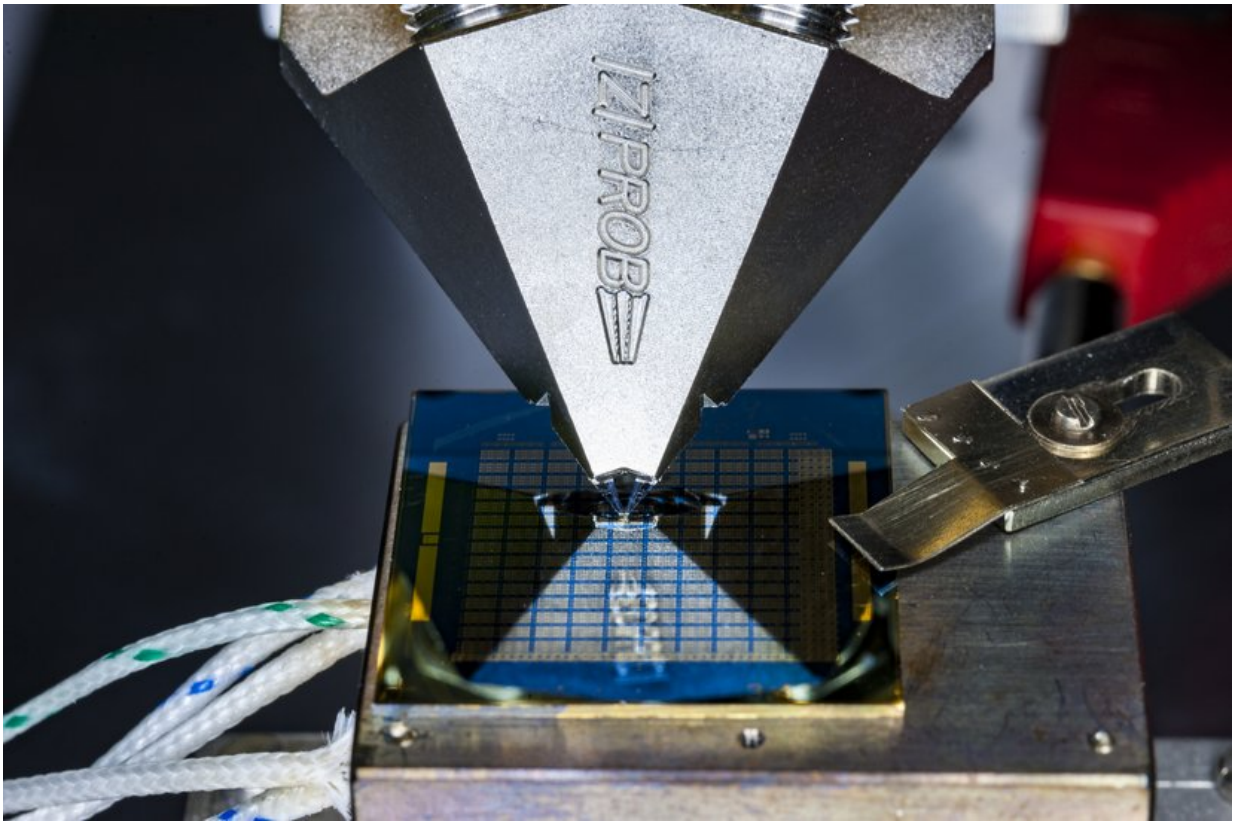
Dual 8-bit breakthroughs bring AI to the edge

December 3 2018, by Jeff Welser



A chip comprising several PCM devices. The electrical probes coming into contact with it is used to send signals to individual devices to perform the in-memory multiplication. Credit: IBM

This week, at the International Electron Devices Meeting (IEDM) and the Conference on Neural Information Processing Systems (NeurIPS),

IBM researchers will showcase new hardware that will take AI further than it's been before: right to the edge. Our novel approaches for digital and analog AI chips boost speed and slash energy demand for deep learning, without sacrificing accuracy. On the digital side, we're setting the stage for a new industry standard in AI training with an approach that achieves full accuracy with eight-bit precision, accelerating training time by two to four times over today's systems. On the analog side, we report eight-bit precision—the highest yet—for an analog chip, roughly doubling accuracy compared with previous analog chips while consuming 33x less energy than a digital architecture of similar precision. These achievements herald a new era of computing hardware designed to unleash the full potential of AI.

Into the post-GPU era

Innovations in software and AI hardware have largely powered a 2.5x per year improvement in computing performance for AI since 2009, when GPUs were first adopted to accelerate deep learning. But we are reaching the limits of what GPUs and software can do. To solve our toughest problems, hardware needs to scale up. The coming generation of AI applications will need faster response times, bigger AI workloads, and multimodal data from numerous streams. To unleash the full potential of AI, we are redesigning hardware with AI in mind: from accelerators to purpose-built hardware for AI workloads, like our new chips, and eventually quantum computing for AI. Scaling AI with new hardware solutions is part of a wider effort at IBM Research to move from narrow AI, often used to solve specific, well-defined tasks, to broad AI, which reaches across disciplines to help humans solve our most pressing problems.

Digital AI accelerators with reduced precision

IBM Research launched the reduced-precision approach to AI model training and inference with a landmark paper describing a novel dataflow approach for conventional CMOS technologies to rev up hardware platforms by dramatically reducing the bit precision of data and computations. Models trained with 16-bit precision were shown, for the very first time, to exhibit no loss of accuracy in comparison to models trained with 32-bit precision. In the ensuing years, the reduced-precision approach was quickly adopted as the industry standard, with 16-bit training and eight-bit inferencing now commonplace, and spurred an explosion of startups and venture capital for reduced precision-based digital AI chips.

The next industry standard for AI training

The next major landmark in reduced-precision training will be presented at NeurIPS in a paper titled "Training Deep Neural Networks with eight-bit Floating Point Numbers" (authors: Naigang Wang, Jungwook Choi, Daniel Brand, Chia-Yu Chen, Kailash Gopalakrishnan). In this paper, a number of new ideas have been proposed to overcome previous challenges (and orthodoxies) associated with reducing training precision below 16 bits. Using these newly proposed approaches, we've demonstrated, for the first time, the ability to train deep learning models with eight-bit precision while fully preserving model accuracy across all major AI dataset categories: image, speech, and text. The techniques accelerate training time for deep neural networks (DNNs) by two to four times over today's 16-bit systems. Although it was previously considered impossible to further reduce precision for training, we expect this eight-bit training platform to become a widely adopted industry standard in the coming years.

Reducing bit precision is a strategy that's expected to contribute towards more efficient large-scale machine learning platforms, and these results mark a significant step forward in scaling AI. Combining this approach

with a customized dataflow architecture, a single chip architecture can be used to efficiently execute training and inferencing across a range of workloads and networks large and small. This approach can also accommodate "mini-batches" of data, required for critical broad AI capabilities without compromising performance. Realizing all of these capabilities with eight-bit precision for training also opens the realm of energy-efficient broad AI at the edge.

Analog chips for in-memory computing

Thanks to its low power requirements, high energy efficiency, and high reliability, analog technology is a natural fit for AI at the edge. Analog accelerators will fuel a roadmap of AI hardware acceleration beyond the limits of conventional digital approaches. However, whereas digital AI hardware is in a race to reduce precision, analog has thus far been limited by its relatively low intrinsic precision, impacting model accuracy. We developed a new technique to compensate for this, achieving the highest precision yet for an analog chip. Our paper at IEDM, "8-bit Precision In-Memory Multiplication with Projected Phase-Change Memory" (authors: Iason Giannopoulos, Abu Sebastian, Manuel Le Gallo, V. P. Jonnalagadda, M. Sousa, M. N. Boon, Evangelos Eleftheriou), shows this technique achieved eight-bit precision in a scalar multiplication operation, roughly doubling the accuracy of previous analog chips, and consumed 33x less energy than a digital architecture of similar precision.

The key to reducing energy consumption is changing the architecture of computing. With today's computing hardware, data must be moved from memory to processors to be used in calculations, which takes a lot of time and energy. An alternative is in-memory computing, in which memory units moonlight as processors, effectively doing double duty of both storage and computation. This avoids the need to shuttle data between memory and processor, saving time and reducing energy

demand by 90 percent or more.

Phase-change memory

Our device uses phase-change memory (PCM) for in-memory computing. PCM records synaptic weights in its physical state along a gradient between amorphous and crystalline. The conductance of the material changes along with its physical state and can be modified using electrical pulses. This is how PCM is able to perform calculations. Because the state can be anywhere along the continuum between zero and one, it is considered an analog value, as opposed to a digital value, which is either a zero or a one, nothing in between.

We have enhanced the precision and stability of the PCM-stored weights with a novel approach, called projected PCM (Proj-PCM), in which we insert a non-insulating projection segment in parallel to the phase-change segment. During the write process, the projection segment has minimal impact on the operation of the device. However, during read, conductance values of programmed states are mostly determined by the projection segment, which is remarkably immune to conductance variations. This allows Proj-PCM devices to achieve much higher precision than previous PCM devices.

The improved precision achieved by our research team indicates in-memory computing may be able to achieve high-performance deep learning in low-power environments, such as IoT and edge applications. As with our digital accelerators, our analog chips are designed to scale for AI training and inferencing across visual, speech, and text datasets and extending to emerging broad AI. We'll be demonstrating a previously published PCM chip all week at NeurIPS, using it to classify hand-written digits in real time via the cloud.

This story is republished courtesy of IBM Research. Read the original story

here.

Provided by IBM