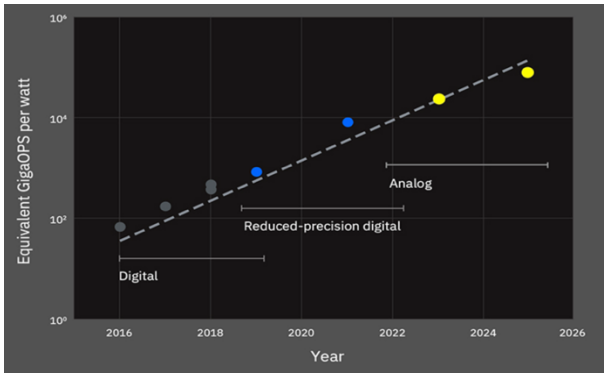
THE IBM RESEARCH AI HARDWARE CENTER: AN UPDATE

CELEBRATING ITS TWO-YEAR ANNIVERSARY, IBM ANNOUNCES INNOVATIVE AI ACCELERATION TECHNOLOGIES AND NEW INDUSTRY AND ACADEMIC PARTNERS

The IBM Research AI Hardware Center is the nexus of a group of academic and industry leaders contributing to the next wave of AI technologies. The Center's mission is to develop technologies that will deliver 2.5 times annual improvement in AI hardware compute efficiency, attaining a 1000-fold improvement in ten years and culminating in what IBM terms “Fluid Intelligence”. Now celebrating its second anniversary, IBM is tracking to that pace or better.

GOALS AND ACCOMPLISHMENTS

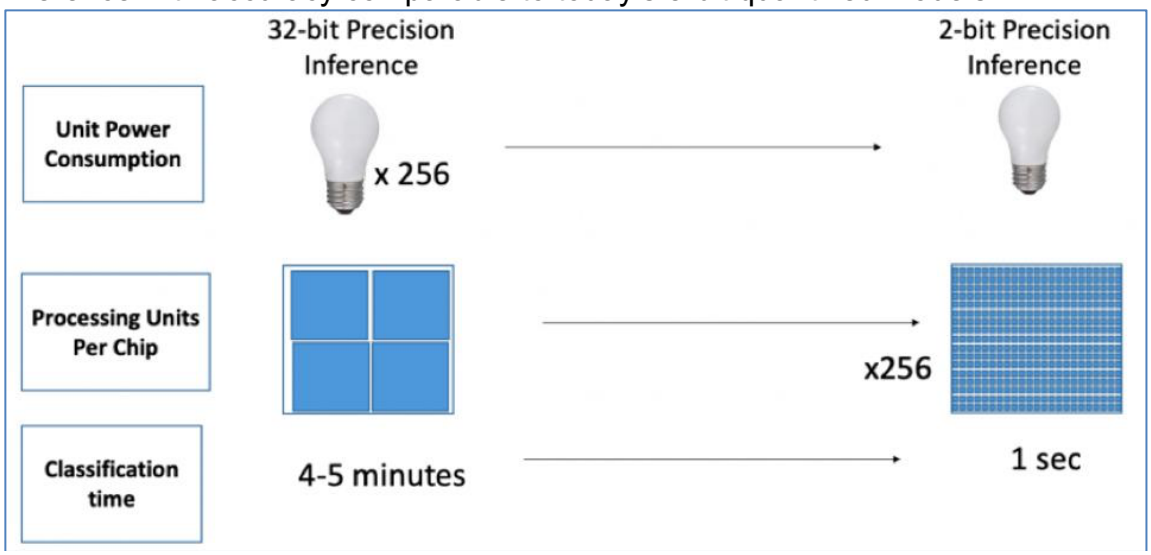


The AI Hardware Center envisions three phases of research: Accelerated Digital AI Cores for training and inference, Analog AI Cores for in-memory computation, Heterogeneous Integration to enable high-speed component interconnections and an AI Technology Testbed, a customizable platform for testing the accuracy and efficiency of new AI technologies for training and inferencing on image, speech, and text workloads. Significantly, the number of AI Research Center participants has expanded from six to sixteen academic and industry partners.

In the Center’s first contributions last year, IBM announced reduced and variable precision for training deep neural networks. The challenge with reducing precision, or quantization as it is known, is to maintain adequate model accuracy while reaping the benefits of using smaller numbers. Lower precision significantly reduces cost and increases performance, improving as the square of the number's bit-length. Given the proper hardware implementation, reducing accuracy from 32- to 16-bit can speed up calculations by 4x. Implementing 8- or 4-bit hardware could increase performance by 8-16x or 32-64x, and the new 2-bit approach could boost performance by up to 256x.

While using 16-bit math for training is becoming a bit more common in the industry, the IBM Center took the concept a step further, exploiting "hybrid" precision that tailors the number of bits used for the exponent mantissa differently for forward- and backward-pass training computations. With this technique, IBM trained a neural network for vision, speech and language processing using only 8 bits for the weights and activations. For inference processing, IBM pioneered two techniques, parameterized clipping activation (PACT) and statistics-aware weight binning (SAWB), which have demonstrated 2-bit

inference with accuracy comparable to today's 8-bit quantized models.



Recently, IBM Research announced its third-generation digital AI core, unveiled at the ISSCC 2021 conference. The new 4-core design increases the performance efficiency for training and inference by sixfold, significantly outperforming the goal of delivering 2.5X annual improvements.

In the analog area, IBM Research has begun to build a foundation for analog neural network computing with the newly announced AI Hardware Composer tool. The Composer provides access to IBM's open-source analog libraries with an easy-to-use interface, allowing both novices and experienced developers to tune analog devices to create accurate AI models. AI researchers can test neural network optimization tools to design analog hardware-aware models. Rensselaer Polytechnical Institute is designing coursework for students to learn how to exploit such models using IBM's Composer.

CONCLUSIONS

Many may be unaware of the extensive hardware research being conducted today at IBM. However, the AI Hardware Center has already accomplished a great deal in the realm of digital AI cores and has now launched its first development tool to begin familiarizing developers with the workflow to design in-memory analog compute devices. IBM has attracted a high-caliber roster of participants to help develop and productize the technology arising from IBM Research. We believe that IBM is creating an IP portfolio that could impact the industry for years to come while providing the consortium's membership with advanced and highly differentiated technologies to accelerate AI at the edge and in the data center.