CROSSBAR

ReRAM for Energy Efficient AI Inference at the Edge

Artificial intelligence (AI) is already driving many applications at Google, Facebook, Amazon, Apple, Microsoft and others, and is now poised to move from the datacenter and the cloud to the network edge, where applications can perform more analysis and provide more intelligence and understanding of the environment. Helping drive this migration is the emergence of new low-power AI technologies – hardware and software – targeting mainstream consumer mobile and IoT devices.

Now, the focus is on edge and end devices, the smartphones, tablets, personal computers and IoT devices that are increasingly being tasked with AI-based applications. And many future AI applications, from recognizing faces and places to navigating driverless cars and landing drones, will need to perform much more complex tasks, including many that must operate in real or near-real-time despite connectivity interruptions, cloud latencies, bandwidth limitations and privacy concerns.

To achieve this, much more powerful AI processing is occuring on these end devices without relying on cloud connectivity. Several semiconductor companies are already starting to offer AI processors, but with the need to more efficiently store the AI knowledge base, the trained model, locally, these companies must move beyond the power-hungry embedded SRAM and external DRAM in use today.

The greatest energy efficiencies and speeds can be achieved by storing data and computing on chip, using the same die, with fast, low-power embedded non-volatile memory. The makers of AI processors now see next-generation memory technologies as the best way to integrate all the necessary storage on-chip with the AI algorithms to further reduce overall energy consumption and system cost while boosting performance.

CrossBar non-volatile Resistive RAM (ReRAM) IP cores offer AI processor developers the best path to successfully implementing this approach through significantly lower power consumption, faster read and write times, more robust non-volatility, higher density and easier manufacturing compared to other competing memory solutions.

Distributed AI

Machine learning (ML) and neural networks are playing a key role at the center of the AI revolution today. Machine learning applications are already equaling or even surpassing expert human performance for modest tasks such as image or speech recognition. More challenging tasks like natural language comprehension and complex games, previously considered to be extremely difficult, are also being successfully implemented.

Machine learning comprises two basic phases: training and inference. An artificial neural network, designed to mimic how the brain works, is first exposed to a large amount of known data – pictures of dogs and cats, for example – so it can learn to recognize what each looks like and how they are different. This trained neural network, or trained model, is then put to work using what it has learned to infer things about new data it is presented with, in this case determining if an image is of a dog or a cat.

Today, most training occurs in the datacenter, with some on the edge. Hyperscale companies like Google, Facebook, Amazon, Apple, Microsoft and others have massive amounts of consumer data they can feed their huge cloud server farms to perform industrial-scale training operations for AI and improve their AI algorithms. The training phase requires very fast processors, such as NVIDIA GPUs or Google Tensor Processing Units, and massive amounts of data.

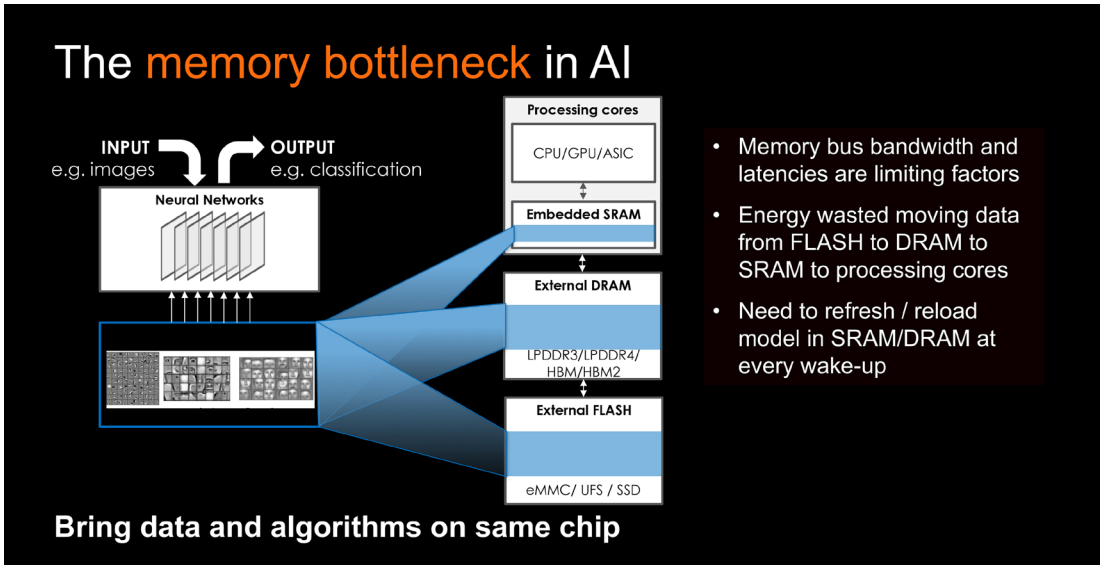
Inference occurs when data is collected by an edge device – a photo of a building or a face, for example, and is sent to an inference engine for classification. A cloud-based AI model with its inherent built-in delay would be unacceptable for many inference applications. A self-driving car that needs to make real-time decisions about objects it sees does not make sense with only a cloud-based AI architecture.

As AI capabilities move to the edge, they will drive more AI applications, and increasingly these applications will require ever more powerful analysis and intelligence to begin to allow systems to make some operational decisions locally, possibly semi- or even completely autonomously.

For ML algorithms to become pervasive, increased computational resources on edge devices will be needed. But traditional computing CPUs are not very good at these tasks, and high-end GPUs consume too much energy and are too expensive. Inference at the edge demands more affordable, lower power chips that can quickly traverse the neural network to recognize (classify) an animal, identify a face, pinpoint a tumor or translate German to English.

Today, more than 30 companies are focusing on developing dedicated AI hardware to achieve the greater speeds and efficiencies required for these specialized computing tasks in datacenters and in a variety of smartphones, tablets and other IoT edge devices. Huawei has introduced a Neural Processing Unit, the iPhone X employs the A11X Bionic processor, the Qualcomm Snapdragon 835 can accelerate TensorFlow, Caffe, Caffe2, MxNet and Android NNAPI across its CPU, GPU and DSP. Intel Movidius is bringing machine vision to drones. Arm® Holdings recently announced their Project Trillium, IP and software for machine learning that speed AI algorithms across CPU, GPU and ML processors of smartphones, tablets, smart cameras, smart home devices and other "AI-enabled devices" at the network edge.

Analysts at both Research and Markets and TechNavio have predicted the global AI chip market to grow at a compound annual growth rate of about 54 percent between 2017 and 2021. The need for high-power hardware that can handle the demands of machine learning is a key driver to this growth.



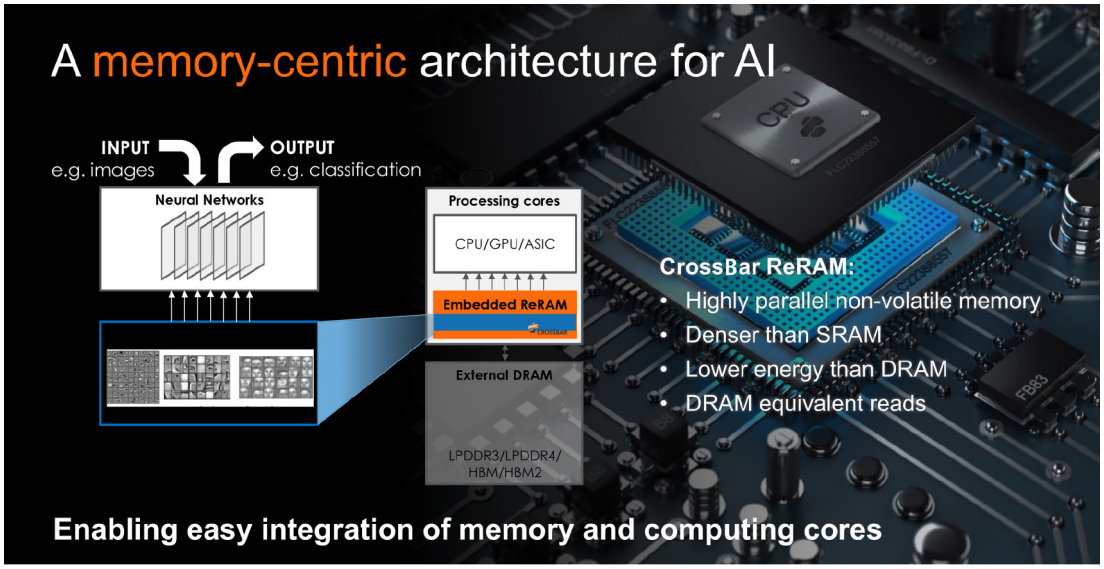
Removing the Memory Bottleneck

All AI processors rely upon data sets, which represent models of the "learned" object classes (images, voices, etc.), to perform their recognition feats. Each object recognition and classification requires multiple memory accesses. The biggest challenge facing engineers today is to overcome the memory speed and power bottleneck in the current architecture to get faster data access while lowering the cost in terms of energy for that access.

The greatest speed and energy efficiency can be gained by placing this model data as close as possible to the AI processor core, as the convolutions in neural networks commonly in use today take 90 to 99 percent of computation and runtime to perform their tasks. But the storage architecture employed by today's designs, created several years ago when there were no other practical solutions, is still the traditional combination of fast but small embedded SRAM with slower but large external DRAM. When trained models are stored this way, the frequent and massive movements of data between embedded SRAM, external DRAM and the neural network increase energy consumption and add latencies.

Further, the SRAM and DRAM are volatile memories, limiting the ability to achieve power savings during sleep periods by having to ensure data retention and requiring all the model data to be reloaded at startup, wasting even more energy and time.

Much greater energy efficiencies and speeds can be achieved by storing the entire trained model directly on the AI processor die with low-power, embedded non-volatile memory that is dense enough and fast enough. By enabling a new, memory-centric architecture, the entire trained model or knowledge base could then be on-chip, connected directly to the neural network with the potential to achieve massive energy savings and performance improvements, resulting in greatly improved battery life and a better user experience. Today, several next-generation memory technologies are competing to accomplish exactly this, which represents a critical step in achieving successful AI applications at the edge.



CrossBar ReRAM

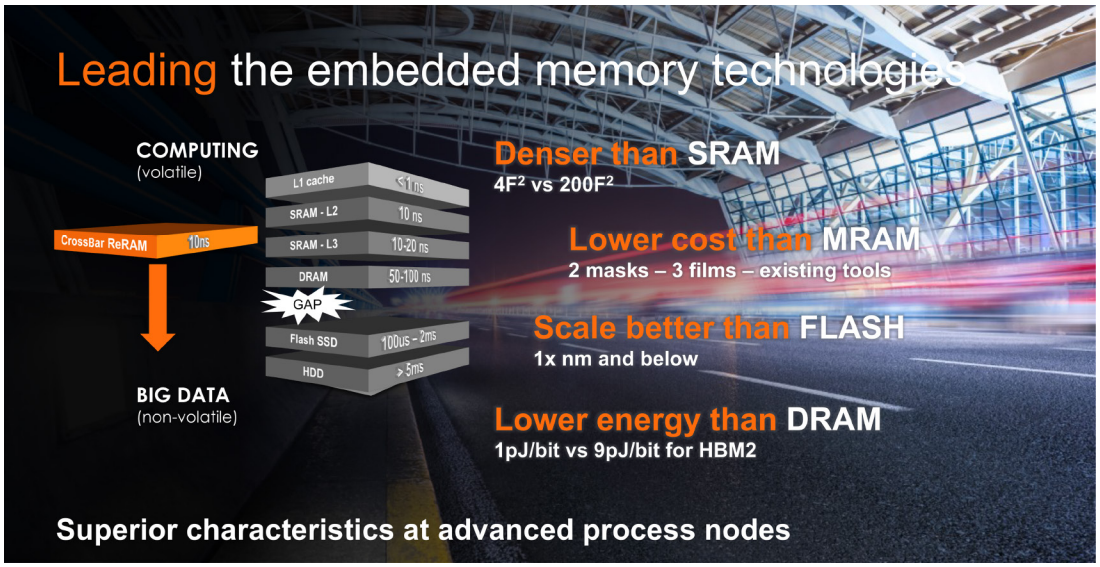
The ideal non-volatile embedded memory for AI applications must be very simple to manufacture, easy to integrate in the back-end-of-line of well-understood CMOS processes, easily scaled to advanced nodes and available in high volume, in addition to delivering the energy and speed performance levels required for these applications.

CrossBar ReRAM has a much greater ability to scale than magnetic RAM (MRAM) or phase change memory (PCM) alternatives, an important consideration when looking at 14-, 12- and even 7-nanometer process nodes. These other technologies require inherently more complex, difficult and expensive manufacturing processes and require more power to operate than CrossBar ReRAM, all of which are major concerns when addressing the highvolume, cost-sensitive requirements of mass market mobile edge devices.

The unique nanofilament technology of CrossBar ReRAM enables scaling below 10nm without impacting performance. CrossBar ReRAM technology is based on a simple device structure using CMOS-friendly materials and a standard manufacturing process that can be easily integrated with and manufactured on existing CMOS fabs. As it is a low temperature, back-end-of-line process integration, multiple layers of CrossBar ReRAM arrays can be integrated on top of CMOS logic wafers to build 3D ReRAM storage arrays.

AI needs the best performance per watt, and this is especially true when applied to extremely power-limited edge devices. CrossBar ReRAM has been shown to achieve an energy efficiency five times greater than that of DRAM, as much as 1000 bit reads per nano-Joule, while exhibiting better overall read performance than DRAM, up to 12.8 GB/s bandwidth with less than 20-ns random latency. More than 75 percent of today's AI applications – multi-layer perception, long short-term memory and convolutional neural networks – can be addressed by 32MB of embedded CrossBar ReRAM with up to a five
times greater power savings.

ReRAM achieves much greater value than any embedded SRAM and DRAM, as well as any of the other next-generation non-volatile alternatives, including MRAM, PCM and 3D flash.



Enabling Memory-Centric Architectures

Starting at 40nm and scaling below 10nm, the CrossBar high-performance memory IP can be integrated at the same process nodes of microcontrollers (MCU), systems-on-chip (SoC) and field programmable gate arrays (FPGA), enabling new memory-centric architectures that will revolutionize computing for deep-learning neural networks and AI. From machine vision, speech recognition and healthcare to robotics, energy and automotive to finance, entertainment and eCommerce, CrossBar memory-centric computing opens the door to a broad range of new energy-efficient AI applications at the edge.

Today, non-volatile memory technologies from CrossBar ReRAM are helping to address AI performance and energy challenges at the edge. In the near future, CrossBar ReRAM technology has the potential to play a significant role in some radically new approaches to AI now under consideration. Based on the pioneering work of Carver Mead, scientists are already exploring a variety of novel brain-inspired paradigms to achieve much greater energy efficiencies by imitating the way neurons and synapses of the central nervous system interact. Artificial synapses

based on ReRAM technology is considered a very promising method for enabling these high-density and ultimately scaled synaptic arrays in neuromorphic architectures. Several experimental low-power cognitive computing systems based on ReRAM have already been proposed by researchers. Neuromorphic architectures based on ReRAM have great potential in applications involving the processing of real-world signals and that operate on compact, low-power devices, including robotic applications such as brain-machine interface, sensor networks, embedded systems and portable devices.

3200 Patrick Henry Dr., Suite 110 Santa Clara, CA 95054 | (408) 884-0281 | info@crossbar-inc.com
www.crossbar-inc.com