Emerging Memories And Artificial Intelligence

Tom Coughlin

On August 29, 2019 I put on a workshop on Emerging Memories and Artificial Intelligence at Stanford University put on by the Stanford Center for Magnetic Nanotechnology and Coughlin Associates. We had several interesting speakers talking about various types of artificial intelligence and the role that new non-volatile memories will play in both training AI models and implementing them in the field using inference engines. This piece will talk about some of the material presented at this workshop.

Dr. Shan Wang, co-organizer of the event gave a introduction, talking about emerging non-volatile memories and in particular on Magnetic Random Access Memory (MRAM). He spoke about how various new memories work—in particular Resistive RAM (RRAM), Phase Change Memory (PCM), MRAM and Ferrroelectic RAM (FRAM). He said that current volatile memory power requirements, especially static power, have been increasing with smaller semiconductor lithographic features. Of all these new memories MRAM is promising for SRAM and DRAM replacement or as a complement to these memories and allowing lower power operation as shown in the table below.

	SRAM (Static) for cache	DRAM (Dynamic) for main memory	PCM (Phase change)	RRAM (Resistive)	MRAM (Magneto-resistive)
Latency	300 ps ~ ns	10 ~ 30 ns	~50 ns	< 10 ns	few ns or faster
Endurance	$> 10^{16}$	$> 10^{15}$	$10^{8} - 10^{12}$	$10^{6} - 10^{12}$	$> 10^{15}$
Retention	volatile	volatile	> 10 years	> 10 years	> 10 years

Dr. Wang's Comparison of Memory Technologies

Image from Stanford Presentation

While MRAM holds promise to replace or complement DRAM and SRAM, PCM and RRAM could provide larger capacity and slower storage and also be used as analog AI accelerators (neuromorphic computing).

Dr. Wang’s talk explored the physics of magnetoresistive technology, which is in common use for reading data in hard disk drives. He discussed the differences between the spin valves used in HDD heads and the magnetic tunnel junctions used in MRAM devices. The next generation of MRAM devices are based upon spin torque transfer and ambient thermal energy to switch the memory devices. These devices are now ramping into production as

discrete 1 Gb devices and as a memory option to replace NOR and some higher -evel SRAM cache. Future MRAM devices will use a technology called Spin Orbit Transfer (SOT), that is capable of much faster switching, allowing possibly full replacement of today's volatile SRAM memory. SOT MRAM can have switching times in the picosecond (ps) range, as fast as SRAM.

Dr. Deliang Fan, from Arizona State University spoke about energy efficient in-memory computing, especially with parallel processing in-memory accelerators. He called this Processing-in-Memory (PIM) Architecture. He also explored a very tight coupling between processing and memory, especially with Spintronic devices that can combine memory and logic functions that can perform both memory read/write and AND/OR logic operations. In-memory logic research efforts include Pinatubo, which uses non-volatile memory.

The table below compares some simulation data for processing-in-memory accelerators using non-volatile and volatile memories. The SOT-MRAM result is very promising for a high performance system, similar to the SRAM simulation. He also talked about how better AI learning models running on lower power (deep neural networks) can be made with PIM approaches.

Dr. Fan's Processing in Memory Accelerator Comparisons

Image from Stanford Presentation

Dr. Hsinyu (Sidney) Tsai, from IBM Almaden Research Center gave a comprehensive review of how deep neural network (DNN) analog memory device accelerators work, particularly using PCM as the memory technology. Note that IBM has been working for several years on PCM-based neuromorphic memory chips. The analog functions that these memories provide is called multiple-accumulate. This technique can be applied for direct neural network training and involves forward inferences based upon previous training, back-propagation of errors from these inferences to correct the matrix of weights used for the inference step. She pointed out how high DNN accuracy can be achieved despite imperfect PCM devices.

Dr. Any Steinbach from startup Paradigm Shift AI gave an interesting talk about the detailed operation of machine learning how it can be applied in the Semiconductor Industry. In particular he discussed how it can be used to fix lithographic adhesion defects as well as production lot scheduling, especially using industrial IoT sensor information.

Sylvain Dubois from Crossbar (an exhibitor at the workshop) talked about using near-memory computing accelerators for faster machine learning using Crossbar's RRAM technology in a highly parallel memory array. As a demonstration of this approach the company built a USB stick with what the company calls its Crossbar XPU to show significant improvements in object look-up speed (OLUPS) compared to and ARM processor operating at much higher operating frequency as shown below.

Crossbar Object Lookup Comparison for ARM versus Crossbar XPU

Image from Stanford Presentation

Dr. Peter Corcoran from the National University of Ireland, Galway spoke about why AI at the Edge will drive the need for storage and bandwidth. He talked about the importance of computing carried out in the memory fabric. He pointed out that AI networks could now be built into an SD card format with 2-orders of magnitude lower energy use than GPUs. Since Peter works on consumer vision systems he spoke about the improvement in consumer camera technology and the use of local visual processing to put AI at the edge. He discussed how the next generation of these cameras could include built in convolutional neural networks (CNNs).

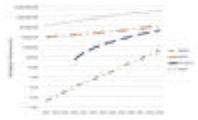
Dr. Corcoran pointed out that going even further to create truly life-like images and better image analysis (for example for driver monitoring) will require considerable processing and memory resources. The slide below shows some of his estimates for storage and monitoring resources to do this.

Dr. Corcoran AI Data and Monitoring Requirements

Image from Stanford Presentation

Mahendra Pakala from Applied Materials spoke about production equipment used to make modern emerging memories like RRAM, MRAM and PCM and Andy Walker from MRAM startup Spin Memory spoke about how STT-MRAM can make AI affordable, especially from the energy consumption view point.

My talk finished the day talking about how emerging memories enable the AI market and projections for the resulting growth of emerging memories, such as MRAM and 3D XPoint, based upon a recently finished report on emerging memory growth. The report projections are shown below.



Emerging Memory PB Shipment Projections

Image from Coughlin Associates

If you are in the San Francisco Bay Area and are interested in learning more about the role of emerging non-volatile memories in artificial intelligence application and in networking with storage professionals, storage users and other interesting parties you might be interested in the Storage Valley Supper Club meeting, September 18, 2019 at Techcode. Registration is open until September 16.

Emerging memories will play an important role in the development of AI and other applications both for learning and for applications of AI models in the real world. The 2019 Stanford Workshop gave lots of insights on how various AI operate and how new memory technologies can be integrated into AI systems and memory/storage architectures to enable the next generation of data center and client applications.