Introduction

The pace of deep machine learning and artificial intelligence (AI) is changing the world of computing at all levels of hardware architecture, software, chip manufacturing, and system packaging. Two major developments have opened the doors to implementing new techniques in machine learning. First, vast amounts of data, i.e., “Big Data,” are available for systems to process. Second, advanced GPU architectures now support distributed computing parallelization. With these two developments, designers can take advantage of new techniques that rely on intensive computing and massive amounts of distributed memory to offer new, powerful compute capabilities.

Neuromorphic computing-based machine learning utilizes techniques of Spiking Neural Networks (SNN), Deep Neural Networks (DNN) and Restricted Boltzmann Machines (RBM). Combined with Big Data, “Big Compute” is utilizing statistically based High-Dimensional Computing (HDC) that operates on patterns, supports reasoning built on associative memory and on continuous learning to mimic human memory learning and retention sequences.

Emerging memories range from “Compute-In-memory SRAMs” (CIM), STT-MRAMs, SOT-MRAMs, ReRAMs, CB-RAMs, and PCMs. The development of each type is simultaneously trying to enable a transformation in computation for AI. Together, they are advancing the scale of computational capabilities, energy efficiency, density, and cost.

Challenges of Selecting Memory Architectures for ML/AI Computing

There are several challenges facing system designers in choosing the optimal computing architecture and the associated combination of memories supporting their objectives for an ML/AI application. Although traditional embedded SRAM, Caches, and Register Files are heavily utilized today, there is no generic nor exotic memory solution that satisfies the newly required AI loads being developed. However, as machine learning is projected to consume a majority of the energy consumed, optimizing memories for machine learning helps designers hit their power budgets. This has major implications for system design.

OpenAI, a research institution devoted to AI’s benefit to humans and funded by Microsoft and Khosla Ventures, observes that the compute demand by deep learning has been doubling every three months for the last 8 years (Source: “AI is changing the entire nature of compute,” https://www.zdnet.com/article/ai-is-changing-the-entire-nature-of-compute/). Tiernan Ray, ZDNet, retrieved 03/22/2020). That observation has pushed memory technology development in modularity, packaging (stacking) and cooling (thermal management).
Designers balance the requirements of their designs as they determine which of the nine major challenges are most critical at a given time:

1. Throughput as a function of energy (peta-ops per watt)
2. Modularity and scalability for design reuse
3. Thermal management to lower costs, complexity, and size
4. Speed supporting real-time AI-based decision making
5. Reliability especially for human life-sensitive applications
6. Processing compatibility with CMOS for components constituting a system. As an example, STT-MRAM can be easily integrated with a CMOS-based processor.
7. Power delivery
8. Cost; best expressed in the “sweet spot” node for a function and with integration (packaging) cost
9. Exhibiting analog behavior mimicking human neurons

Each of these memory challenges can be addressed in multiple ways, as usually there is more than one alternative for the same objective. Each alternative will have pros and cons, including further scalability implications for architectural decisions. For example, designers must choose between using SRAMs or a ReRAM array for compute-in-memory. The power and scalability implications of these two options are at extreme opposites. The SRAM option is the right choice when the size of the memory block is relatively small, the required speed of execution is high, and the integration of the in-memory compute within an SoC comes naturally as the most logical option (although SRAM is costly in area and in power consumption – both dynamic and leakage). On the other hand, a highly parallelized matrix multiplication typical of deep neural networks requiring a huge amount of memory makes the argument for using ReRAM, because of the density advantages.

Multi-port SRAMs play a special and unique role in Compute-in-memory (CIM) architectures because Boolean logic functions are operations involving multi-inputs and require the ability to simultaneously read data from multiple addressable locations and write the results back in desired memory locations. Multi-port SRAMs and Register files offer that precise flexibility. Also, multi-port SRAMs can be used to construct register files for GPUs crucial for efficient multi-threading.

Understanding the Range of Emerging Memories

The most prominent emerging memories are STT-MRAM, SOT-MRAM, ReRAM, CB-RAM, FeRAM, and PCM. Rather than detailing the make-up of each particular memory, understanding the main features that make them major candidates for neuromorphic computing architectures (as well as candidates for universal memory) is helpful when selecting memories.

**STT-MRAM / SOT-MRAM**

STT-MRAM and SOT-MRAM are members of the spintronics-based class of MRAMs characterized by a bit-cell with a thin tunneling junction (magnetic tunnel junction MTJ) between two magnetic electrodes, one of fixed magnetization (pinned layer) and the other is of a “free” magnetization. Figure 1 shows the progression of MRAMs since the first commercial MRAM dating back to 2007.
MRAMs introduced a class of non-volatile ultra low power memories (FIMS and TAS) fully compatible with CMOS processing. But, they have traditionally been relatively slow, susceptible to noise-induced flipping and not scalable compared to SRAMs, embedded Flash, & other technologies. Field induced magnetic SRAMs (FIMS) were programmed through a magnetic field induced by a flowing current. Thermally assisted switching (TAS) was an enhanced MRAM where a flowing current generated "heat" that lowered the threshold for free layer flipping. But the true breakthrough was in the STT-MRAM where a "spin orbit torque" is induced in the free magnetization layer through a current flowing through the tunneling junction. The direction of the current determined if the free layer magnetization aligned with the fixed magnetization layer (low resistance) or the free layer magnetization progressed to anti-alignment with the magnetization of the fixed layer (high resistance). The "1" or "0" was measured through sensing the tunneling-magneto-resistance (TMJ) of the junction through a current flow across the junction and the access transistor; a relatively small current not high enough to program the bit.

- SOT-MRAM is a variation on STT-MRAM that allows for a much faster read and write times. The process associated with generating the SOT bitcell is more complex than that of an STT-MRAM. STT and SOT MRAMs offer key features such as non-volatility.
- Low leakage
- Scalability
- Ease of integration with CMOS
- Very high retention time ($10^{16}$)
- High durability (really function of read current)

As shown in Figure 2, STT-MRAM and SOT-MRAM are fully compatible with CMOS, and SOT MRAM shows a lot of promise in low voltage operation as an alternative to existing memories used in AI/ML (Figure 3).

![Figure 2: MTJ Array TEM Cross-Section](source)

Source: O. Golonzka, et al., "MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology", Intel Corp., IEDM 2018

Cache memory scaling: SOT-MRAM

![Figure 3: MTJ Array TEM Cross-Section](source)

STT-MRAMs are good substitutes for L2 and L3 caches. STT-MRAMs can offer fast write times, as low as 1ns, when the tunnel junction is thinned to as low as 20A, which is comparable to L2 and L3 caches. The 1ns write time comes at the expense of retention time which drops to as low as few days. But a lot of L2 and L3 caches require a retention of only a few milliseconds and therefore a retention time of a few days is more than sufficient. In addition, as lower write currents are required for thinner junctions, the power dissipation is reduced. Engineering the STT-MRAM can vary from slow-write, ultra low power and very long retention time most suited for IoT applications to fast, much lower retention but still favorable to the use of classical L2/L3 caches. There is a wide range of compromise between retention time and write time and write current.

![Graph showing power consumption with duty cycle for different memory types](image)

Figure 4: STT-MRAM use in IoT
Source Qualcomm & TDK, IEDM, 2015

It is worth noting that STT-MRAMs are widely deployed for IoT applications. Figure 4 illustrates the use of STT-MRAM as a universal memory for IoT applications with extremely tight power budgets and low duty cycle (typical of sensor circuits). STT-MRAM can act as a universal embedded memory, replacing or eliminating the need for external Flash, embedded SRAM, and embedded Flash to save cost, power consumption, and area.

**SOT-MRAM**

The basic principles and operation for SOT-MRAMs are the same as STT-MRAM. The minor difference is purely in the technology used to make SOT-MRAM. SOT-MRAMs use high spin-orbit coupling materials such as heavy metals or topological insulators for generating a high spin-coupling layer. SOT relies on an in-plane charge current going through the high spin-orbit coupling layer to alter the magnetization of the free layer. SOT-MRAMs have a separate read path than the write path. Therefore SOT-MRAM avoids passing a high write current through the tunnel barrier. For system designers, its functionality is identical to STT-MRAM and the ease of CMOS integration is the same. While there have been many variants of SOT-MRAM, but the concept and end result of each is the same.

**Phase-Change Memories: PCM, ReRAMs, RRAMs, and CB-RAMs**

Phase-change memories are all non-volatile memories characterized by two distinct states of low resistance and high resistance based on the direction of the current applied between the two electrodes forming the memory.

Each of these memories has commonalities and differences. It is not entirely wrong to “lump” all of PCM, ReRAMs, RRAMs, and CB-RAMs into a “class” of memories under the classification of “phase change memories” although the mechanism of the phase change varies widely. Also, common to all of them is the desire to have as large a set-reset resistance ratio as possible for acceptable sensing in a noisy environment.
Most video, audio and image data are analog in nature. Storing them in a digital memory requires costly digital-to-analog converters. A better option is to store this data in non-volatile memory arrays that exhibit analog behavior (though non-ideal analog behavior, but rather "good enough"). PCMs, ReRAMs and CB-RAMs fall in that category of memories.

Phase-change memories (PCM, ReRAMs, and CB-RAMs) offer an additional benefit when used as a synapsis weight in the training phase of a neural network. The rapid incremental switching leads to incremental conductance states, which enforces the connectivity (weight) of a synapse as more pulses are applied. This class of memories (PCM, ReRAMs, CB-RAMs) offer that capability.

It is worth noting that variants of all these memories lend themselves to easy integration with CMOS as well as for 3D stacking.

**PCM**

PCM is a good candidate for a non-volatile low power, high density, simple process complexity memory.

**ReRAM**

ReRAM cross-bar arrays are ideal for in-memory computing. ReRAMs can be utilized in neuromorphic computing as synapses which connect the neurons of different layers of a network which provides a much more brain-like architecture than current embedded SRAMs provide today.

**CB-RAM**

Conductive-bridge RAM (CB-RAM) is a member of the resistive RAM family. It is an energy efficient non-volatile memory that lends itself to integration with CMOS and is significantly useful in realizing in-memory computing and in the implementation of neural networks.

The CB-RAM technology relies on the electrochemical making and breaking of a conductive link. This process changes the resistance of the CB-RAM storage element which is used to represent data as shown in Figure 5. The left side of the figure shows the cross-section of a CB-RAM and the CMOS compatibility. The right side shows the architecture of a 1K of RRAM integrated into a CMOS chip.

![Figure 5: Cross section of CBRAM integrated with CMOS](source: Adesto website [www.adesto.com](http://www.adesto.com))
Table 1 summarizes the comparison between the emerging memories often considered for neuromorphic computing utilized as off-chip solutions and compared to traditional on-chip SRAM & Register Files. It reflects the "latest" most representative numbers for each technology and it is simply sampled from one of many in most cases.

<table>
<thead>
<tr>
<th></th>
<th>SRAM</th>
<th>Register Files</th>
<th>FLASH (NAND)</th>
<th>PCRAM</th>
<th>RRAM</th>
<th>MRAM (STT-MRAM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Device</strong></td>
<td>Planar</td>
<td>Planar</td>
<td>&quot;Monolithic&quot; 3D</td>
<td>Planar</td>
<td>&quot;Discrete&quot; 3D</td>
<td>Planar</td>
</tr>
<tr>
<td><strong>Feature Size</strong></td>
<td>6T</td>
<td>Multi-port</td>
<td>1T</td>
<td>1T</td>
<td></td>
<td>1BJT/1R</td>
</tr>
<tr>
<td><strong>Cell Size</strong></td>
<td>7nm</td>
<td>7nm</td>
<td>19nm</td>
<td>20nm</td>
<td>28nm</td>
<td>22nm</td>
</tr>
<tr>
<td><strong>Capacity</strong></td>
<td>16Mb</td>
<td>256KB</td>
<td>1Tb/Die</td>
<td>16Gb</td>
<td>16Gb</td>
<td>4Gb</td>
</tr>
<tr>
<td><strong>Retention</strong></td>
<td>∞</td>
<td>∞</td>
<td>10⁵</td>
<td>10⁹</td>
<td>10⁹</td>
<td>10⁻⁹</td>
</tr>
<tr>
<td><strong>Leakage</strong></td>
<td>1X</td>
<td>1X-3X</td>
<td>~ 0.8X</td>
<td>O(0.01X)</td>
<td>Sneak Current Paths</td>
<td>O(0.01X)</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>1X</td>
<td>1X</td>
<td>O(0.1X)</td>
<td>O(0.1X)</td>
<td>O(0.1X)</td>
<td>1X</td>
</tr>
<tr>
<td><strong>Stacking</strong></td>
<td>No</td>
<td>No</td>
<td>3D Stack Embedded ok</td>
<td>3D Stack Embedded ok</td>
<td>3D Stack Embedded ok</td>
<td>3D Stack Embedded ok</td>
</tr>
</tbody>
</table>

Table 1: Comparison between emerging memories for neuromorphic computing shows that no single memory type can be the "perfect" memory for all AI chips, but each have their advantages.

**Synopsys Addressing Memories for Neuromorphic Computing**

The non-volatile memories addressed in this paper are mainly for none Von-Neumann architectures and are the back-bone of emerging neuromorphic computing. But the list of memories involved in neuromorphic computing is not complete without addressing the classical SRAM memories.

SRAMs and register files remain the backbone of AI/ML architectures for neuromorphic computing with their un-matched latency in all memory categories. However, the overriding theme of maximizing the TOPS/W metric for neuromorphic computing and for Von-Neumann architectures dictates the use of parallelism that can be accomplished with multi-port memories with utmost degrees of configuration flexibility to accommodate compute-in-memory (CIM) and near memory computing. Synopsys actively supports research in CIM area while supporting near-memory computing as the most energy efficient yet versatile form of computing. Towards that end we offer a whole family of multi-port SRAMs that are energy efficient and that can operate at ultra-low voltages.
Figure 6 shows a sample of all the SRAM compilers and register files Synopsys supports at the 7nm node, which has been pivotal in AI chip development.

![Figure 6: Synopsys SRAM compilers and register files offered at the 7nm node](image)

**Summary**

The era of Big Data and Big Compute is here. Per OpenAI, compute demand by deep learning has been doubling every three months for the last 8 years. Neuromorphic computing with deep neural networks is driving AI growth however it is heavily dependent on compact, non-volatile energy efficient memories with various attractive features to suit different situations. These include STT/SOT-MRAM, SOT-MRAMs, ReRAMs, CB-RAMs, and PCMs.

Neuromorphic computing relies on new architectures, new memory technologies and more efficient than current processing architectures, and it requires compute-in-memory and near memory computing as well as the expertise in memory yield, test, reliability and implementation.