Thinking Fast & Slow: Intuition, Reasoning and Emerging Memory

Dave Eggleston
Intuitive Cognition Consulting
Abstract

- Our human brain can be modeled as two distinctly different systems: a real time intuition system, and a background reasoning system.

- As we move into the AI compute era, Emerging Memory technologies play an increasingly important role in overcoming the limitations of DRAM and NAND.

- The commercialization of Emerging Memory will therefore accelerate our realization of powerful AI systems.
“A groundbreaking tour of the mind, and explains the two systems that drive the way we think.”

“System 1 is fast, intuitive, and emotional; System 2 is slower, more deliberative, and more logical.”

Daniel Kahneman is professor emeritus of psychology and public affairs at Princeton University.
Time for a brain quiz! Ready?
$17 \times 24 = ?$
Intuition

17 x 24 = ?

Reasoning
Intuition

System 1
- Lightning fast
- Automatic
- Real time
- Effortless
- Approximate

Reasoning

System 2
- Slow
- Interrupt driven
- Background
- Energy inefficient
- Precise

Edge

Datacenter
Edge

- Non-von Neumann architecture

Intuition

Datacenter

- von Neumann architecture

Reasoning
Edge

Store Dr. Moda for now; we’ll come back to discuss Edge & Intuition

Intuition

- Non-von Neumann architecture
- von Neumann architecture

Datacenter

Reasoning
Once upon a time, long ago…
And for a while we were happy! 😊 😊 😊
Moore’s Law is slowing – but still need low cost bits

- Cost gap between DRAM and NAND continues to increase
- Need cost-effective emerging memory to fill this gap
- Trajectory for DRAM prices for the next 5 years uncertain

Source: IDC
Moore’s Law is slowing – but still need low cost bits

- Cost gap between DRAM and NAND continues to increase
- Need cost-effective emerging memory to fill this gap
- Trajectory for DRAM prices for the next 5 years uncertain

Source: IDC

Emerging memory

$/GB
## The Latency Spectrum and Gaps ~ 2015

Source: Mark Webb, MKW Ventures Consulting, FMS 2018

<table>
<thead>
<tr>
<th>Source</th>
<th>CPU/ SRAM</th>
<th>DRAM</th>
<th>NAND SLC to TLC</th>
<th>HDD</th>
<th>TAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1ns</td>
<td>10ns</td>
<td>100ns</td>
<td>10us</td>
<td>100us</td>
</tr>
<tr>
<td></td>
<td>1us</td>
<td></td>
<td></td>
<td>1ms</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10ms</td>
<td>100ms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1s</td>
<td></td>
</tr>
</tbody>
</table>

**Increasing Density**

**Increasing Cost**

More Like Memory

More Like Storage

The GAP (PM/SCM)
The Latency Spectrum and Gaps ~ 2015

Source: Mark Webb, MKW Ventures Consulting, FMS 2018

NAND is too slow!

More Like Memory

More Like Storage

Increasing Density

Increasing Cost

SRAM

1ns 10ns 100ns 1us 10us 100us 1ms 10ms 100ms 1s

SLC to TLC

Increasing Density
Limitations

- DRAM Cost
- NAND Latency
Emerging Memory Targets

Cost

1/3\textsuperscript{rd} DRAM

Latency

Read <1\mu s
Emerging Memory Targets

Does such an Emerging Memory even exist?

Cost
1/3\textsuperscript{rd} DRAM

Latency
Read <1\mu s
Switching Mechanisms

Source: M. Jurczak, imec, ISSCC 2015 Memory Forum
Switching Mechanisms

Source: M. Jurczak, imec, ISSCC 2015 Memory Forum

(Too) many switching mechanisms!
Classifying the What

- **ST-MRAM**
  - Standalone Memory
  - Embedded Memory
  - Increasing Latency
  - Increasing Cost and Endurance

- **PCM (3DXP)**
  - Standalone Memory
  - Embedded Memory
  - Increasing Latency
  - Increasing Cost and Endurance

- **ReRAM**
  - Standalone Memory
  - Embedded Memory
  - Increasing Latency
  - Increasing Cost and Endurance
WHO is doing what?

ST-MRAM

ST-MRAM

PCM (3DXP)

PCM (3DXP)

ReRAM

ReRAM

Standalone

Global Foundries

Samsung

TSMC

UMC

Embedded

Everspin

Intel

Micron

Adesto

Sony

Crossbar

Panasonic
Let’s focus on the two shipping technologies!

ST-MRAM

PCM (3DXP)

ReRAM

Standalone

Embedded

2018 Storage Developer Conference. © Intuitive Cognition Consulting All Rights Reserved.
WHO is doing what?

<table>
<thead>
<tr>
<th>ST-MRAM</th>
<th>PCM (3DXP)</th>
<th>ReRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standalone</td>
<td>Embedded</td>
<td>Other “SCM”</td>
</tr>
</tbody>
</table>

Logos are the property of their respective trademark owners.
Does such an Emerging Memory even exist?

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCM (3DXP)</td>
<td>1/3rd of DRAM</td>
<td>Read &lt;1us</td>
</tr>
</tbody>
</table>

Uh oh…
That’s not good.
A Word to the Wise: Replacements

• There are no 1:1 memory replacements.
• Stop. looking. for. Them.
• Moving from DRAM+NAND world into one with Combinations of memory.
Let’s use memory combinations!

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>Latency</th>
<th>Winning Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/3rd of DRAM</td>
<td>Read &lt;1us</td>
<td></td>
</tr>
<tr>
<td>PCM (3DXP)</td>
<td>+DRAM DIMMs for reduced latency!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST-MRAM</td>
<td>+NAND for reduced SSD cost!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2018 products featuring Emerging Memory
FMS18: Intel Optane DIMM (3DXP on DRAM bus)

Source: Mark W. Henderson, Intel, FMS 2018

- Big and Affordable Memory
- High Performance Storage
- Direct Load/Store Access
- Native Persistence

- 128, 256, 512GB
- DDR4 Pin Compatible
- Hardware Encryption
- High Reliability

Now shipping samples
broad developer engagement
FMS18: Intel Optane DIMM (3DXP on DRAM bus)

- Arch, software, hardware total effort
- Reduces Optane read latency to a few μs
- DRAM operates as “near memory”, Optane operates as “far memory”
- Intel controlled

Source: Mark W. Henderson, Intel, FMS 2018
FMS18: IBM Flash Core Module (MRAM+NAND)

Introducing The IBM FlashCore Module

IBM FlashCore™ technology delivers key differentiators

- Built in, performance neutral hardware compression and encryption
- Using 64 layer 3DTLC NAND
- Enterprise data reliability
- Cognitive Algorithms for Wear Levelling, Health binning, Heat segregation and media management
- Intelligent media management that keeps settings ideal to keep performance consistent.
- Endurance without latency penalty
- FIPS 140 certification

MRAM = Fast persistent write cache in SSD

4.8TBu, 9.6TBu, 19.2TBu capacity options with up to 3:1 compression
The Latency Spectrum and Gaps ~ Now

Source: Mark Webb, MKW Ventures Consulting, FMS 2018

More Like Memory
- MRAM
- NAND+DRAM DIMMS
- Fast NAND SSDs
- NAND QLC SSD

More Like Storage
- CPU/SRAM
- DRAM
- XP DIMMs/ReRAM
- 3D XP SSD
- NAND MLC to TLC
- HDD
- TAPE

Latency:
- ns
- 10ns
- 100ns
- 1us
- 10us
- 100us
- 1ms
- 10ms
- 100ms
- s
The Latency Spectrum and Gaps ~ Now

Source: Mark Webb, MKW Ventures Consulting, FMS 2018

More Like Memory
- MRAM
- NAND+DRAM DIMMS
- Fast NAND SSDs
- NAND QLC SSD

More Like Storage
- CPU/SRAM
- DRAM
- XP DIMMs/ReRAM
- 3D XP SSD
- NAND MLC to TLC
- HDD
- TAPE

Several memory combinations reduce latency!
Other interesting 2018 “stuff”
FMS18: Toshiba XL-Flash (Low Latency SSD)

- 10x reduced latency vs. TLC
- Still not 1μs 😞
- For low latency SSDs, attached to compute nodes
- Samsung Z-NAND, Intel Optane SSDs are competitors

Source: Jeff Oshima, Toshiba, FMS 2018
FMS18: JEDEC NVDIMM-P

- Emerging memory and DRAM on the same DDR bus
- Open standard
- Non-deterministic behavior allowed
- Will compete with Intel Optane DIMMs
- Backed by all major memory companies

Source: Bill Gervasi, Nantero, FMS 2018
Summary: Your 2018 SSD/DIMM Watch List

- Everspin MRAM in low latency SSDs
- Intel 3DXP in lower cost server Optane DIMMs
- Toshiba XL NAND in lower latency SSDs
- JEDEC NVDIMM-P in open standard DIMMs
Back to Intuition at the Edge!
Let's talk about Intuition at the Edge now!

- Non-von Neumann architecture

Intuition
Edge

- Intuition system in the human brains use <20 watts of power
- Using von-Neumann architecture for Intuition >20 Gigawatts!
- Highly networked, local compute nodes
- Trained neural nets (NN) perform lightning fast intuition
- Embedded Emerging Memory used to hold weights inside NN
- Sum the weights using analog combine
- Most efficient implementation is analog memory (6+ levels per cell)
- Doesn’t require 7nm/5nm process!

Non-von Neumann architecture
Mythic currently uses embedded NOR Flash in analog mode to hold weights.

Source: Dave Fick, Mythic, Hot Chips 2018
Hot Chips 2018: Edge Intuition SoC

Mythic currently uses embedded NOR Flash in analog mode to hold weights.

Source: Dave Fick, Mythic, Hot Chips 2018

Lightning fast, low power, trained NN, analog memory
FMS18: Future Edge Intuition SoCs use ReRAM?

Simultaneous Processing with Deterministic Performance

- Parallel comparison against all identities
- If no match, new identity created (learning)
- Classification performed in one cycle independent of number of identities

Source: Hagop Nazarian, Crossbar, FMS 2018
Classification of 100,000s Identities stored in ReRAM in one iteration

Simultaneous Processing with Deterministic Performance

- Parallel comparison against all identities
- If no match, new identity created (learning)
- Classification performed in one cycle independent of number of identities

Source: Hagop Nazarian, Crossbar, FMS 2018

Embedded ReRAM as analog memory in NN

FMS18: Future Edge Intuition SoCs use ReRAM?
Why do I care? What is Emerging Memory enabling for AI?

Edge
- Analog memory in NN
- Lightning fast intuition at low power

Datacenter
- TBs of memory
- Better reasoning, based on more data
Why do I care? What is Emerging Memory enabling for AI?

Emerging Memory accelerates AI!

- Lightning fast intuition at low power
- Faster reasoning, based on more data
Take-away points:

- Your brain has **two distinct systems**
- Reasoning system needs **reduced latency and cost**
- There are **NO** 1:1 memory **replacements**
- Emerging memory **combos** with NAND and DRAM
- Intuition system needs **new architecture** and low power
- Emerging memory utilized as **analog weight** in neural nets
- Emerging memory **accelerates AI** systems
Register NOW for SNIA 2019 PM Summit!

Dave Eggleston
Intuitive Cognition Consulting
Technology & Business Strategy

Email: dave@in-cog.com
Twitter: @NVM_DaveE
LinkedIn: linkedin.com/in/deggleston/
Backup
Talk Outline

1. Demonstrate the human brain has two distinct systems
2. Discuss the best fit compute architecture to model each AI system
3. Articulate how DRAM and NAND are applied to the compute architectures
4. Identify the key limitations of DRAM and NAND
5. Present and classify some Emerging Memory alternatives
6. Discuss the Emerging Memory system enhancements
7. Identify who is doing what (by when) in the Emerging Memory landscape
8. Articulate the unique challenges in realizing an AI intuition system
9. Propose how Emerging Memory may solve some intuition problems
10. Point to the future of AI systems based on Emerging Memory