How MRAM Fits in AI Applications

Tom Coughlin

Coughlin Associates,
www.tomcoughlin.com

Artificial Intelligence

- Artificial Intelligence is the simulation of human intelligence and includes learning, reasoning and self-correction
- Many different algorithms are used
- Artificial intelligence tools can be trained in a data center and then the resulting model used in the field.

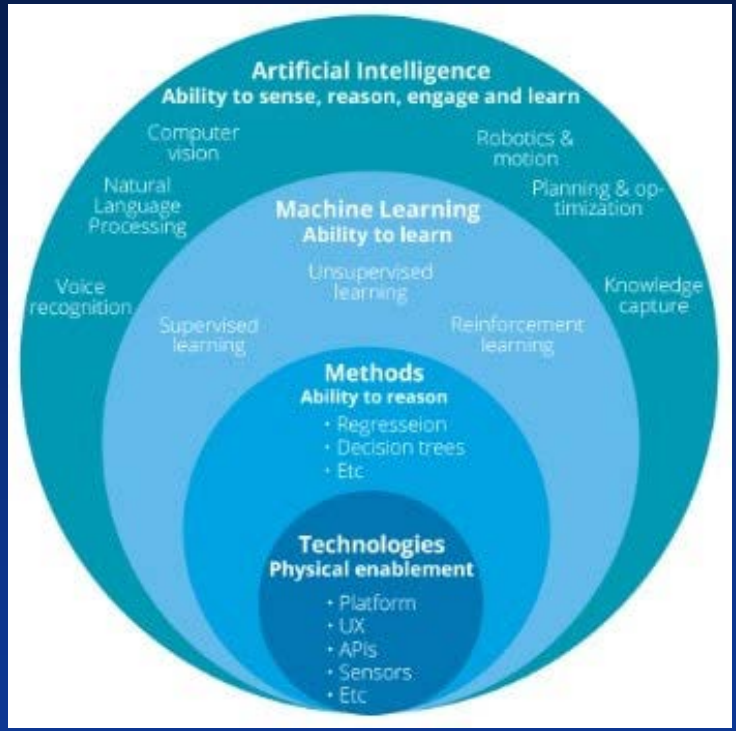
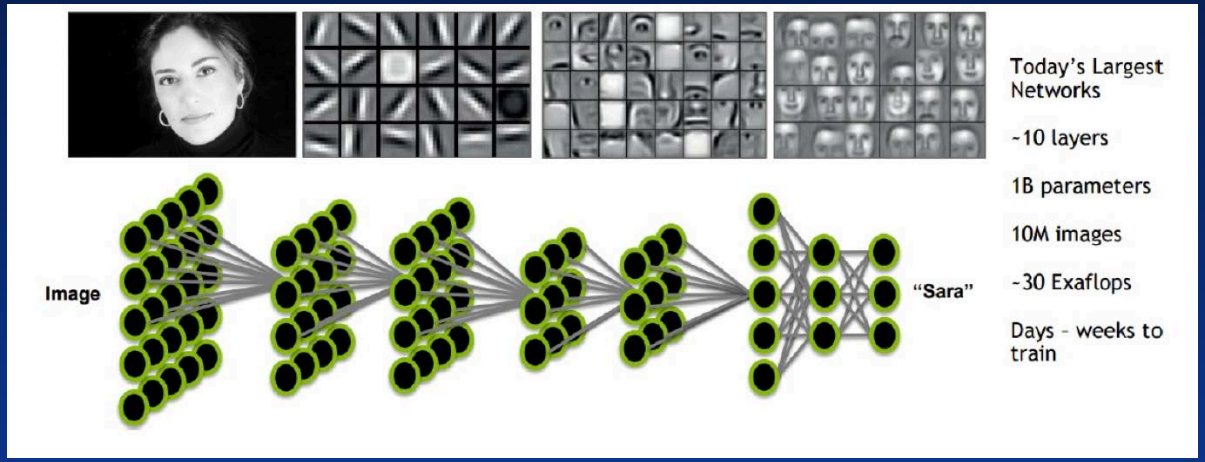


Image from Doloitte, Artificial Intelligence Defined, 2017

Artificial Intelligence (2)

- Machine learning is a technique commonly used for AI applications
- The actual learning is normally done with curated representative data sets that emulate what would be seen in the field
- An inference engine would be used by devices at network edges or at endpoints to implement a trained ML AI for application such as voice or image recognition

Artificial Intelligence (3)

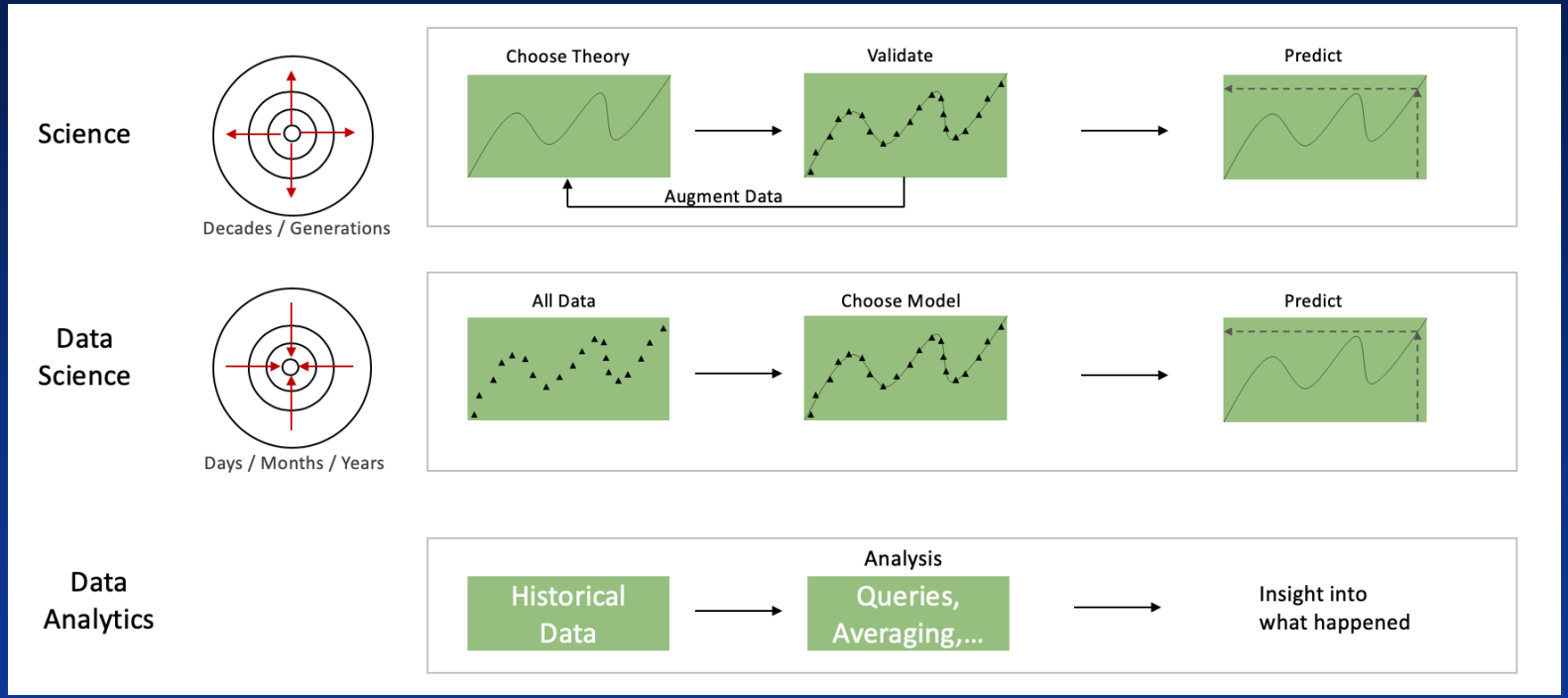


Andy Steinbach, Understanding Artificial Intelligence and Its Ramifications, 2018 Storage Visions Conference

- Much of what is called AI is effectively "curve fitting"
- This can be applied to levels of complexity in the observed data

Deep learning breaks down observations, to smaller and smaller parts, looking for the best matches for all those parts

Artificial Intelligence (4)



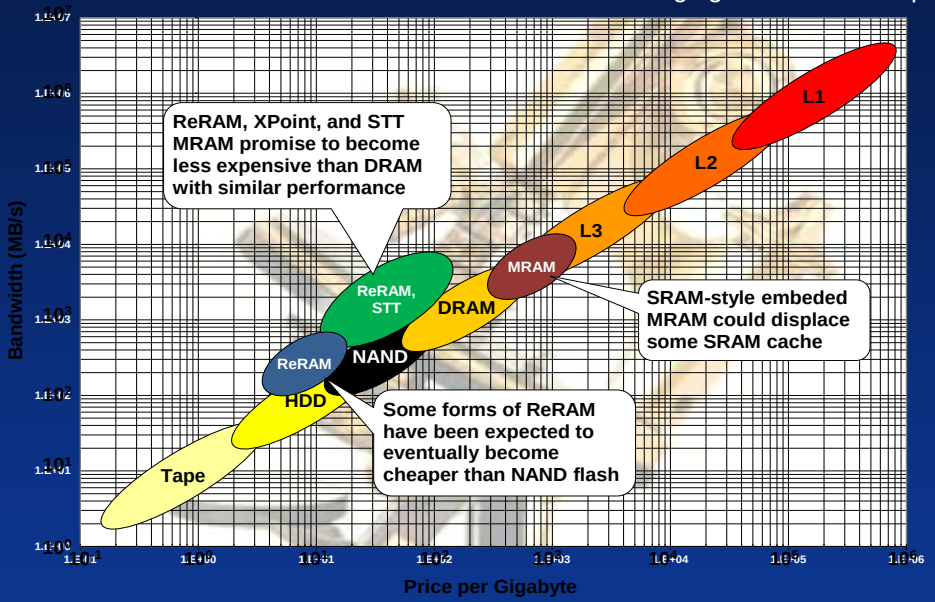
Andy Steinbach, Understanding Artificial Intelligence and Its Ramifications, 2018 Storage Visions Conference

Artificial Intelligence (5)

- ML AI applied at network edges and in embedded endpoint devices are limited by power available for operation (e.g. batteries). So energy efficiency is very important.
- ML AI models are stored on embedded device memory
- Conventional volatile memory such as DRAM/SRAM need to be refreshed or it loses memory when the power is off.
- A non-volatile memory would use less power and thus allow a ML AI model to run longer on an energy limited application

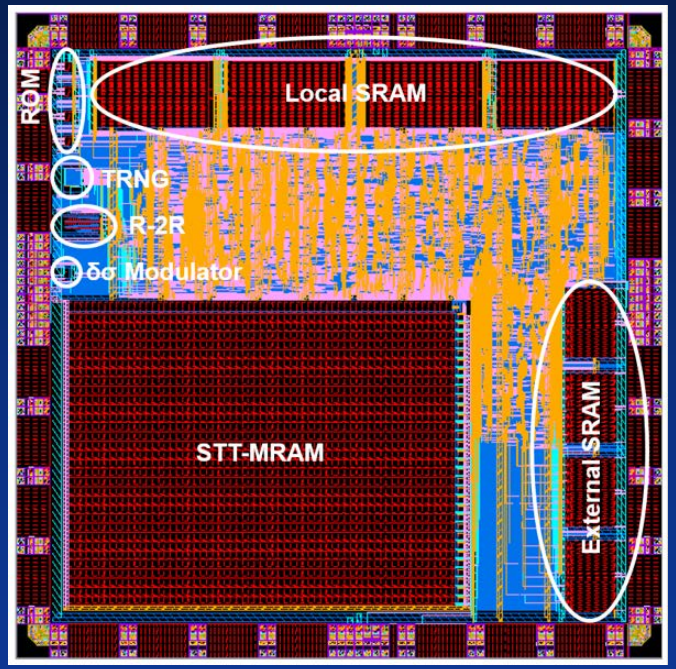
Future Memory/Storage Hierarchy

2019 Emerging Memories Ramp Up, Coughlin Associates and Objective Analysis, 2019



- MRAM is a good option to to replace NOR Flash and some SRAM for AI applications

MRAM In Embedded Devices (2)



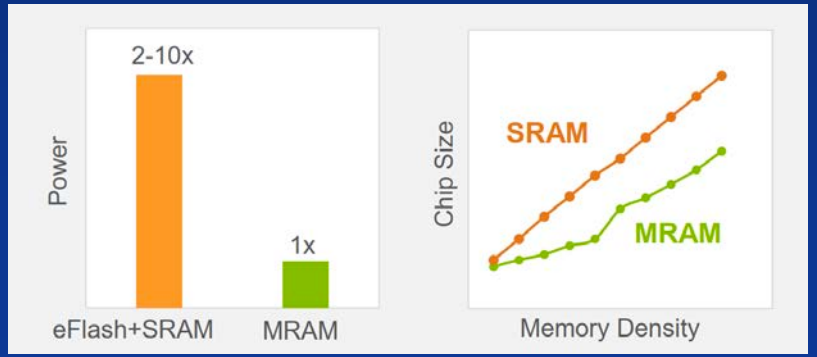
CEA-LETi, September 2018

- MRAM has a clear path to shrink to smaller geometries
- Memory retention, endurance and capacity can be adjusted in MRAM to provide memory for various applications environments, such as in automobiles

MRAM In Embedded Devices

- MRAM is smaller than SRAM (with 5-6 transistors/cell)
- NOR Flash cannot shrink beyond about 22-28 nm
- MRAM could replace SRAM and NOR flash memories in embedded AI devices consuming less power with higher memory density

Images from Kevin Moraes, AMAT, 2019 Semicon



MRAM AI Developments

- All the major semiconductor foundries have said they will provide embedded MRAM in SoC products including Samsung, TSMC, Global Foundries UMC, etc.
- Many of these foundries are also looking to move MRAM integration away from BEOL to earlier in the chip production process to reduce cost
- New tools are needed for MRAM, driving capital equipment spending

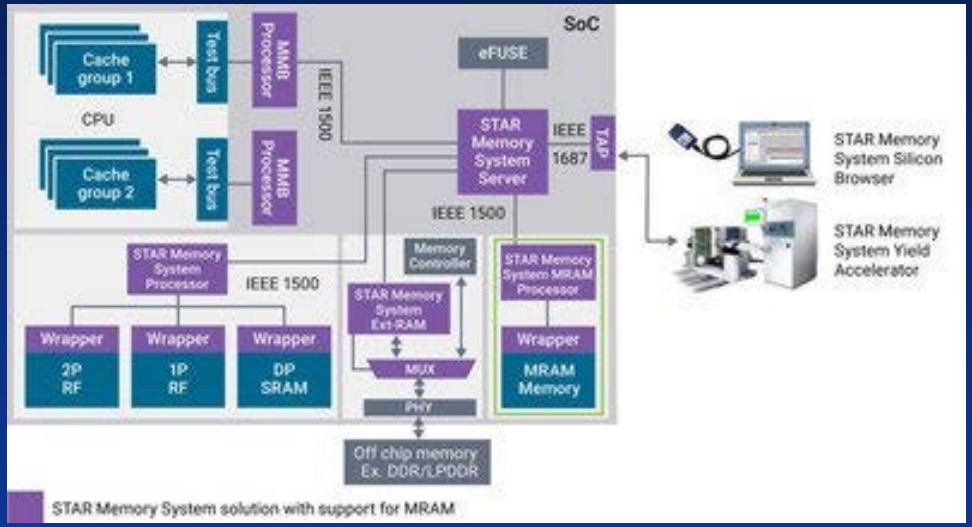
Lightspeeur® 2802M, *Production AI Accelerator Chip with MRAM (2019 CES)*

- Includes: The GME (Gyrfalcon MRAM Engine)
- 9.9 TOPS/W in a 22nm ASIC
- Produced via TSMC Collaboration
- Industry leading features, like Non-Volatile Memory

~ 40 MB of Memory	Large embedded models	
	Multiple AI models :	
	Image Classification	Facial recognition
	Voice identification	Voice Commands
	Text to speech	And others…..
Power Savings	20-50% when compared to SRAM or “other MRAM”	
Custom Designs	One Time Programmable Memory	
	up to 10 ns Read Speed (~30 TOPS/W)	
	Non-Power Leakage	

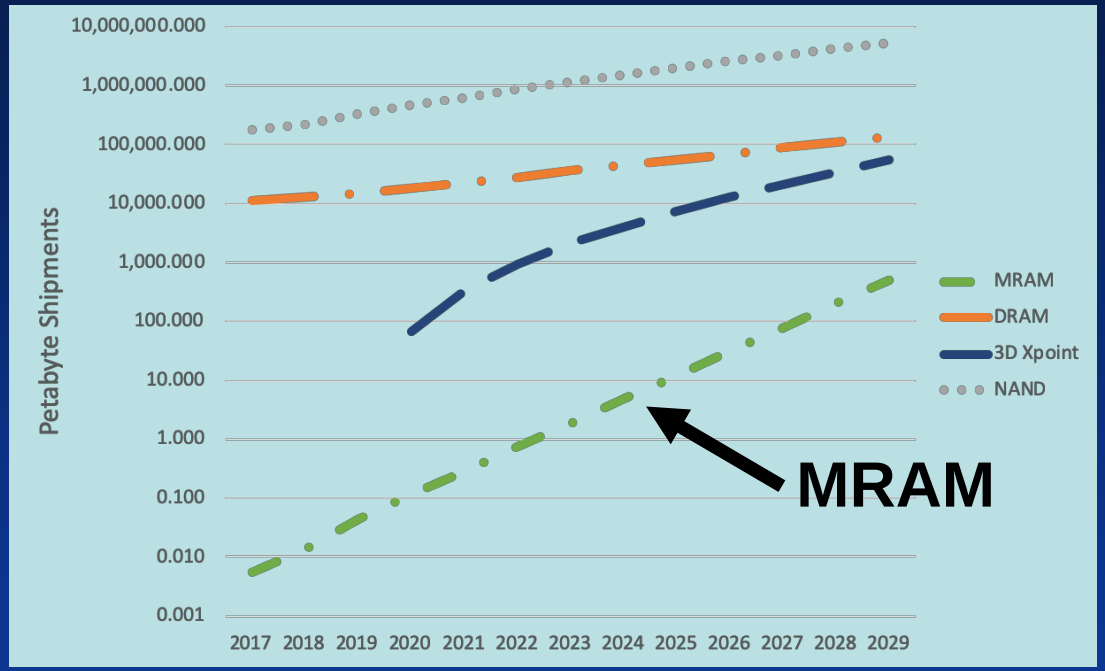
Design Tools MRAM Options

- MRAM integration into SoC's needs support from EDA tools
- Synopsis announced MRAM design support for Q2 2019



Synopsis Press Release, October 2018

Growth in MRAM Memory Shipments

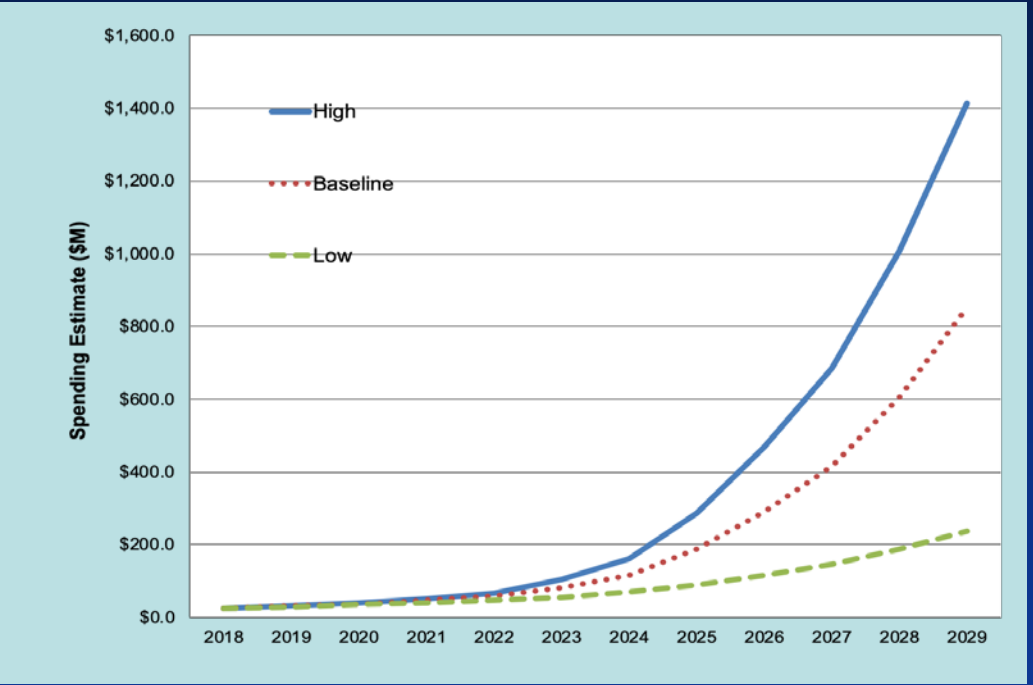


2019 Emerging Memories Ramp Up, Coughlin Associates and Objective Analysis, 2019

- MRAM will replace the bulk of embedded NOR and SRAM in SoCs, mostly for AI apps
- The chart shows projected baseline petabyte memory shipments from 2018-2029.

MRAM Capital Spending Up

The chart shows low, baseline and high MRAM manufacturing equipment spending estimates from 2018-2029.



2019 Emerging Memories Ramp Up, Coughlin Associates and Objective Analysis, 2019

For Further Information

Report on Emerging Memories and Workshop on Emerging Memories and AI

EMERGING MEMORIES RAMP UP

Available June, 2019

This report, jointly produced by Objective Analysis and Coughlin Associates, provides an exhaustive look at emerging memory technologies and their interaction with standard memories, both as discrete devices and in embedded applications (the memories within logic chips like ASICs and MCUs). The report provides a well of technical information, market dynamics, forecasts, and competitive analyses of the leading companies. Forecasts show how the markets will grow not only for the technologies themselves, but also for the capital equipment used to produce them. Read this to understand the competitive landscape and market drivers for these new memories, and to learn how to profit from tomorrow's market.

Table of Contents (Top Level):

Order at: https://tomcoughlin.com/tech-papers/
or https://Objective-Analysis.com/reports/#Emerging

EMAI 2019
Emerging Memory and Artificial Intelligence Workshop

Bechtel Conference Center at Encina Hall
Stanford University
August 29, 2019

This is a one-day workshop featuring invited experts speaking on emerging memory technology, such as MRAM, RRAM, FRAM and PCM as well as experts on applications using various types of AI, such as machine learning, talking about memory requirements for these applications. The morning will feature speakers on the foundational knowledge of emerging memory technologies and AI, with the afternoon featuring speakers on applications for AI including these applications using emerging memory technologies.

To register and for detailed event information, please visit:
https://emai19.sites.stanford.edu