Emerging Memory: The Calm After the Storm

Rev 2

Presenter: Mark Webb
MKW Ventures Consulting, LLC
www.mkwventures.com

©2023 Flash Memory Summit. All Rights Reserved

Contents

- Should we call them Emerging??
- Current status and scorecard
- Markets for Alternative Memories
- CXL Impact
- Summary

Emerging Memories are now Alternative Memories

- They have emerged and are not too disruptive.
- From Last year: Two Models of memory
- UNIVERSAL MEMORY: This is not the application to look for
- Todays “Alternative Memory” addresses a major weakness of existing memory
  - DRAM -> NV DRAM … even if slower
  - NAND -> Faster …. Even if more expensive
  - NOR -> faster, more scalable, integration into logic

Memory Technologies Reviewed

NEW Scorecard: We don't need Universal … Lots of Yellow!

	Latency	Density	Cost	HVM ready
DRAM	*****	***	***	*****
NAND	*	*****	*****	*****
SRAM	*****	*	*	*****
NOR	***	**	**	*****
MRAM 1T1R	****	**	**	****
RRAM 1T1R	***	**	**	****
PCM 1T1R/nR	***	***/??	****	****
FE RAM	***	**	***	*
Other	***	**	**	*

Flash Memory Summit

The Storm

- New Universal Memories, Replace NAND, Replace DRAM
  - If you cannot replace memory move on to next project
- Memory speed is limiter to compute. New memory required
- Show roadmap to billions in sales to get funding and sell IP
- PCM/Optane sells more bits than ALL other Emerging memory
  - 10x+ higher density than any other non-nand memory
  - Spends more on infrastructure, marketing than other memories
  - It is not the game changer or the solution customers want for the money
  - Other crosspoint memories proposed, none are in production
- CXL helps give us other options to the problems we face

The Calm After

Flash Memory Summit

- Memories are finding markets where they add value and the revenue and development cost align
- STT-MRAM, ReRAM, and 1T1R PCM all have similar applications and trade offs on cost and speed
  - Physics and reliability tradeoffs are very different. Process integration chemistry with CMOS also different
  - Note: No chips are infinite endurance despite single cell claims.
- Embedded Memory is always useful
  - Replace NOR E-Flash, gives options for alternative to SRAM, eDRAM, NVM on Chip
- SRAM Cache replacement, NVM, Cost reduction
  - SRAM may or may not scale, but MRAM may be more cost effective. No major implementation yet
- Discrete DRAM cache/high speed replacement
  - DRAM cache for NAND or NOR is replaced with more cost effective model. Reallocate costs on phone.
- Implementation are typically in metal layers with cell size options (FE FET has other options).
- People will continue to look for “Miracle memory”. It may appear… but I probably will be retired. Research is good!
  - From 2022: Only cost effective replacement would have to be close to DRAM process. Less than 20% of steps are different. Example: FerroElectric DRAM Capacitor.
  - Other grounds up processes cannot afford development. Link to 3D DRAM timing is optimal

Embedded Alternative Memory

- Embedded SOCs/Controllers have SRAM, eDRAM and eFlash
- SRAM is available on all logic processes
  - Scales to .02u^2 in TSMC 3nm and 5nm processes
  - Alternative memories don't currently scale more that SRAM
- eDRAM is used by some processes as a cheaper alternative to SRAM for volatile memory. Potentially smaller and cheaper (depends on node)
- eFlash is workhorse for embedded NVM memory
  - Used primarily on larger nodes (>65nm) when benefit of embedded NVM is clear
  - Typically adds 10+ mask layers and significant cost to process
  - Scaling unclear with Finfet
- Alternative Memories are excellent selections in these applications
  - Cheaper than eFlash in many applications even above 28nm
  - Allows re-allocation of memory bits and applications to have different memory budgets

Embedded Benefits Today

Flash Memory Summit

- 28nm and below have memory options available from major foundries who have spend years and millions developing technology. Many offer multiple eNVM options
- Smaller foundries have partnerships to deliver solutions. Customers work with foundry and NVM IP providers. Options available today
  - Great opportunity for smaller Memory IP companies
- Potentially cheaper than eFLASH, although maturity of eFLASH is challenge
- If the memory can replace SRAM/eDRAM as well as NVM, Significant cost reductions are possible or capabilities can be increased
  - 4M of MRAM/ReRAM vs 1M of SRAM+4M of eFLASH
- Ability to scale to Finfet and GAA architecture helps provide future solutions
- Different technologies (MRAM, ReRAM, PCM) can be used with similar results
  - Experience has a lot of impact for each foundries choice.
  - 1T1R is the architecture
- Initial model is for 10% of 16nm and below foundry wafers to provide embedded NVM by 2030
  - We do not track revenue impact of this as it is not meaningful to us as growth metric. Others do.

MRAM Cache for SRAM replacement

- This has been a project for Last level cache for over a decade
- SRAM is expensive, volatile, and potentially running out of scaling room
- In 2016 timeframe, we were predicting no SRAM scaling below 16nm
  - TSMC 5nm vs 3nm has minimal scaling, its still 4 generations beyond believed limit
- On pure performance and scaling, MRAM still has not met goal of replacing large last level caches. ReRAM/PCM are even less likely.
- Maturity of implementations today at foundries could change this AND provide NVM capability
  - There is question on the pros and cons of NVM capability
  - AI Server modules/HBM/advanced packaging may change this

Discrete Applications

Flash Memory Summit

- While large markets are possible, initial use is in smaller “niche” markets
- Niche: A market where the features match perfectly, with minimal competition, and the pricing/margin matches product support.
- MRAM for NV cache for Storage or fast storage for specific applications
  - Companies focusing on this and able to target performance and cost
  - High densities for Storage (Gbit Chips, Gbyte systems/devices)
- ReRAM or others for AI/Learning/Compute in memory
  - DRAM, NAND are not optimal, ability to develop specific memory characteristic is perfect
  - Scalable for future and potential as discrete or embedded
  - Many people propose that this is the largest future market
- Low density for markets below NAND/DRAM density
  - Potentially all NOR markets. Boot, start up, code
  - Select small DRAM/SRAM markets Kbit to low Mbit

How Does Chiplets Change This?

- Simple model: Chiplets are an extension of processes we have used forever
  - Stacking chips so they are directly connected, not on a bus (Smartphones)
  - Connecting chips through substrate, interposer or board
  - Direct bonding speeds and tightens the designs
- If you want Alternative memories close to the processor, you can get it.
  - Chiplet MRAM cache, fast NVM for compute, ReRAM next to SSD controller
- It changes the need to embed the memory in a logic process
- Chiplets add flexibility in module and connect speed but not capability
  - CPU/APU/xPU + MRAM + ???
  - Needs to meet capability
- The end result is that These memories can be integrated, or chiplet, or on bus.

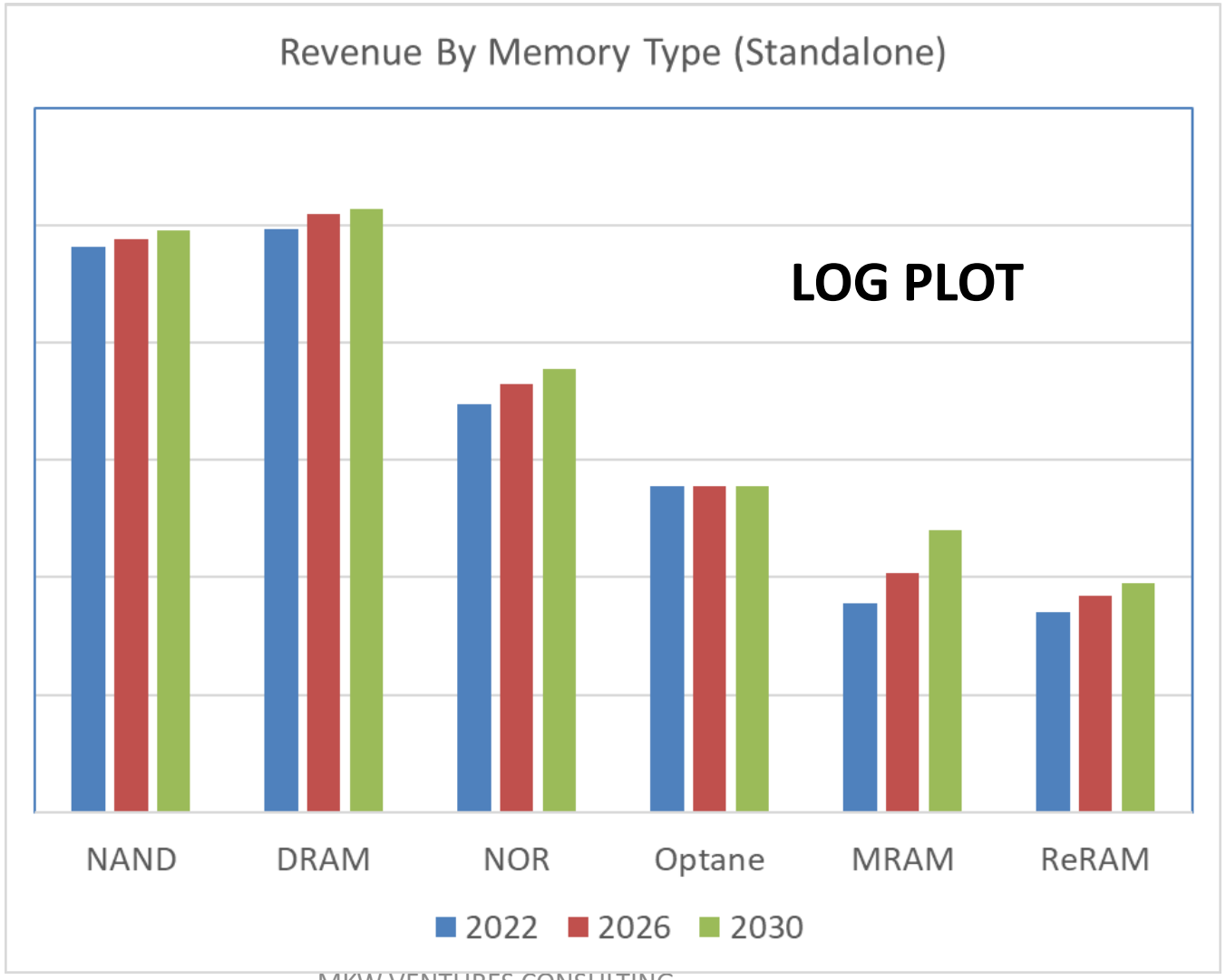
CXL and What Happened to Optane?

Flash Memory Summit

- Optane had two major values
  - LARGE memory, higher density than DRAM, large databases
  - Fast storage (fast SSDs, NVM main memory)
- While CXL doesn't replace the memory type, it allows options to get similar effects
  - Large memory is one the first CXL applications
  - Mixing storage (NAND) and memory on a bus allows effects of fast storage and NVM main memory
- Years ago we predicted that CXL could provide the tipping point for Optane/emerging NVM. Now it may be that it may be the opposite
  - Alternative NVM will end up on the CXL bus. We model DRAM and DRAM/NAND Bits on bus to be >10x higher long term.
  - NVDIMMS of all types will take advantage of moving to CXL
- CXL provide options to minimize the needs to Fast NVM in main memory

Approximate Revenue by Memory Type (2030)

Emerging Memory will NOT be $30B in revenue ever. $1B is more likely



- Yes, I have a Log plot of this
- Linear is more enlightening
- Emerging memory is operating at a different level

Flash Memory Summit

Summary

- We have entered a “Maturity” of emerging memories.
- They look for and are finding multiple applications with reasonable business models
- New technologies will continue to be proposed but are years from production IF they can overcome development cost issues
- Top Markets are Embedded, discrete NOR replacements, and AI/ML/neural chips. 10% of sub 16nm foundry wafers will have embedded NVM in 2030
- Market size is not comparable (embedded not tracked by us) through 2030 to mainstream memory revenue.

©2023 Flash Memory Summit. All Rights Reserved

Flash Memory Summit

Thank You!

Mark Webb
MKW Ventures Consulting LLC
www.mkwventures.com

Mark@mkwventures.com
505-681-7614

©2023 Flash Memory Summit. All Rights Reserved