Annual Update on Emerging Memories

Mark Webb

MKW Ventures Consulting, LLC

mark@mkwventures.com

Topics

- Why New Memories
  - Scaling (new memories or new architecture)
  - The search for universal memory, New applications
- Historical and Recent Technologies compared
- PCM/3D Xpoint, MRAM, ReRAM
- Additional Technologies
- Equipment/Market development challenges and solutions

Memory Technologies Reviewed
whats new?

	Latency	Density	Cost	HVM ready
DRAM	*****	***	***	*****
NAND	*	*****	*****	*****
MRAM	*****	*	*	***
3DXP	***	****	****	****
ReRAM	***	****	****	**
NRAM	***	**	**	*
Other	***	**	**	*

Why New Memories? (the "Pull")

- Existing Memories will stop scaling
  - Or will they?

- Existing memories are not meeting needs
  - Speed
  - Quality

- We want universal memory
  - NAND and DRAM are far from model techs

NAND Scaling

- Once upon a time, we thought NAND was ending and we needed a replacement
  - This "time" was at 50nm, 20nm, 15nm Planar
    - 2006: 40nm is impossible: "we review the critical barriers in further scaling down NAND flash to 40nm technology node and beyond. Then, breakthrough technologies are addressed to overcome the barriers Future Outlook of NAND Flash Technology for 40nm Node"
  - We can always scale, its just a matter of choices
- 3D NAND reset the game
  - Seemed impossible with huge tradeoffs
  - Samsung jumped first for good reason and with lots of money
  - Now we have postponed the end. New architecture, not tech.

Why New Memories? NAND Scaling

- NAND flash has line of sight to 128L+ and QLC.
  - And no….. 256L does not take 6 months
- This takes us to 2023 timeframe with a ~75% bit cost reduction
- We can do a 1Tbit Chip today, 1-2TB in a package
  - Sweet spot is 256Gb so density is not issue
- String stacking, CMOS under array allows a lot of circuit options
- If cost goes down, people will make the trade offs because it benefits them.
- NAND is not done anytime soon

DRAM Scaling

- DRAM scaling has slowed down
  - Technical reasons
  - Business reasons
- This "failed scaling" is providing record margins, revenue and well managed growth
  - Exactly ZERO DRAM manufacturers are unhappy with this
- It is limiting industry, but the small supplier base is ok with that
- Plan is to get to 10nm region <u>very slowly</u> only if ROI is there
- New architectures could work… if needed
- DRAM is fine for now

Why New Memory:
Not Meeting Needs

- DRAM and NAND are not meeting needs

- NAND is slow, block oriented, wears out
  - But man is it cheap!
- DRAM is volatile, density is not growing, its getting more expensive… If you can get it
  - But at least its RAM

Summary

- NAND Scaling is going to continue
  - With unfortunate quality and speed trade offs
- DRAM scaling is going to continue
  - Much slower than we wish!
  - No incentive to risk anything
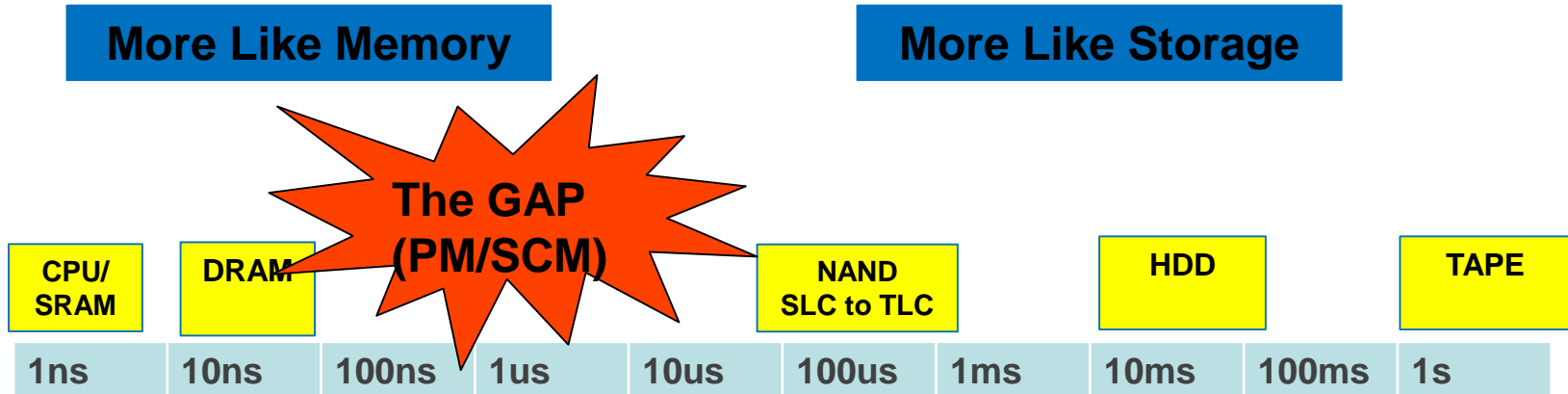- <u>Not meeting needs IS the really compelling reason for new memory</u>

What is NEXT Technology

- **Ideal Universal Memory**
  - Fast as DRAM, Non-Volatile, Infinite cycles, Cheaper than NAND
  - Unicorn Memory … It's a nice dream … it doesn't exist (There are people committed to prove me wrong)
- **Reality of what will happen**
  - Combination of tradeoffs will need to be made
  - Compute architecture designed around what is available
    - Phone, PC, Server designed around DRAM+SCM or SCM+ cloud
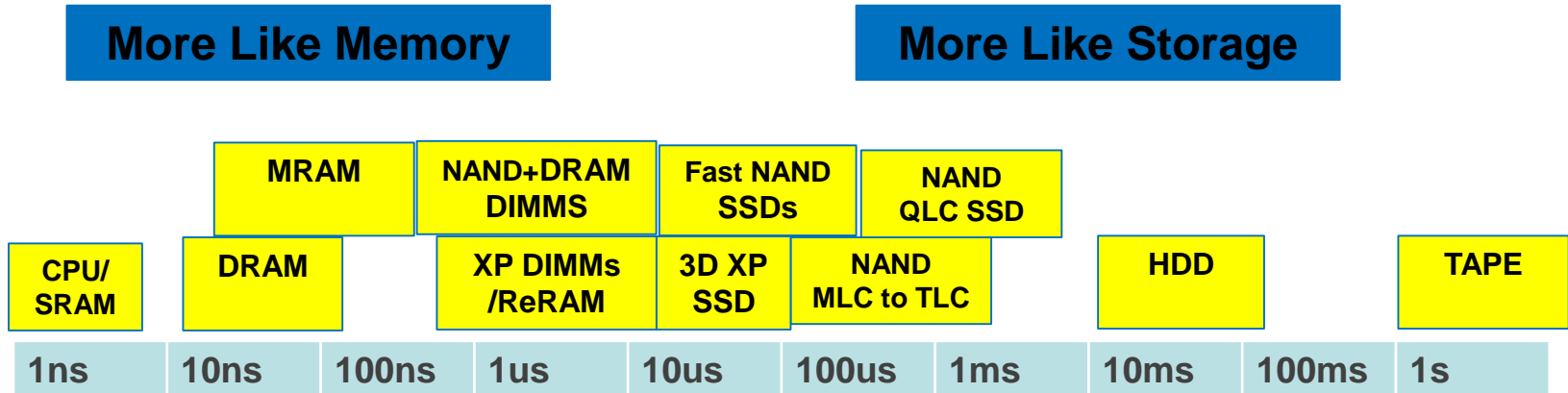- **We are making REAL progress here!**

The Latency Spectrum and Gaps ~2015

More Like Memory

More Like Storage

The GAP (PM/SCM)

| CPU/ SRAM | DRAM | | NAND SLC to TLC | HDD | TAPE |

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

Increasing Density

Increasing Cost

Mark Webb, MKW Ventures Consulting

The Latency Spectrum and Gaps
Future

More Like Memory				More Like Storage	

		MRAM	NAND+DRAM DIMMS	Fast NAND SSDs	NAND QLC SSD		
CPU/ SRAM	DRAM		XP DIMMs /ReRAM	3D XP SSD	NAND MLC to TLC	HDD	TAPE
1ns	10ns	100ns	1us	10us	100us	1ms	10ms	100ms	1s

Increasing Density →

← **Increasing Cost**

Mark Webb, MKW Ventures Consulting

We are making REAL Progress

- We now have a volume PCM memory ramping with 100s of Millions in revenue

- We have ReRAM technologies you can pursue for your product at a foundry plus low density options

- We have MRAM being used in actual products and being offered at foundries for embedded

- We have new memories being proposed
- This IS different than past 10 years

3D Xpoint

What is 3D Xpoint

- While initial Intel and Micron were initially evasive on the technology, we now know more details
  - PCM technology, 20nm lithography
  - Cross point array, Selector in stack
  - 10+ years in the making.
  - multiple publications by Intel, Micron, Numonyx, IP providers, etc document a long history that appears to have lead to 3D xpoint
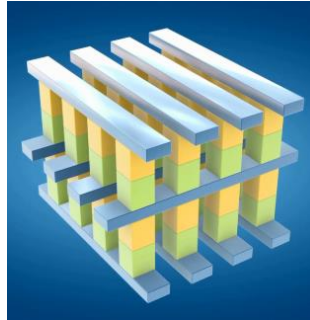
History of PCM papers

IEDM 2009 **Intel and Numonyx Achieve Research Milestone with Stacked, Cross Point Phase Change Memory Technology**
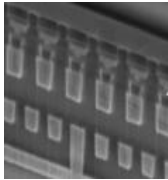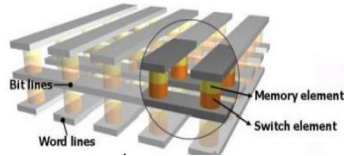
Current and Emerging Memory
Technology Landscape
Atwood FMS 2011

Speculated details on Technology based on:
- Memory and SSD modules sold
- EE-Times/Techinsights
- The register/Ron Neale/Chris Mellor Feb 1 2016
- ISS 2016 (Jan 12)
- Dave Eggleston FMS2015
- Plus multiple FMS presentations including this session from Techinsights

2015+
Source: Intel

Note: 64M PCM crosspoint arrays created and analyzed in 2009!!
Product Lifecycle timing is very informative!

3D Xpoint Hype-Reality Challenge

- Announced in July 2015 with great fanfare
  - "In manufacturing today, 1000x faster than NAND, 1000x the endurance of NAND"... will change everything ... Products in 2016
  - We now know much of that was over-hyped and reality set in.
- Reality is different but still pretty amazing
  - 3D Xpoint will have more revenue that all the other "new NVM" combined.
  - Fastest SSD, with best SSD endurance, selling millions of units
  - Optane DIMMS coming. Likely to have most bits in a NVDIMM sold in 2019 of any technology

3D Xpoint Model (what is it?)

- A model for what 3D Xpoint chips could be….
  - 128Gbit Chip with >10% overprovisioning on the chip itself
  - Read Latency: ~125ns, Write Latency: "higher"
  - Endurance: ~200K cycles with multiple management techniques
    - Next foil talks about endurance!
  - It is a very fast, high endurance, byte addressable, NVM
    - Not DRAM replacement (too slow, can't be cycled indefinitely)
    - Not pure NAND replacement (too expensive)
  - Rev 2 3D Xpoint (estimate) will be a cost reduction of ~30% released in 2019/20

Side note: Thoughts on Endurance

- Endurance metrics are FLAWED when looking at new technologies. Single cell theoretical claims (10^6, 10^8, 10^{12})

- MLC NAND is listed as "5000 cycles"
  - Most bits probably last >50,000 cycles (don't get me started on UBER and RBER
  - Today's TLC is 3000 cycles, QLC might be 1000 cycles
  - On 50nm+ SLC ("100K"). Most bits might last 1M cycles
  - Endurance management can detect and predict many failures. Wear leveling and overprovisioning can multiply endurance >2X

- "NAND" cycling capability is either 1000 or 1M cycles!
  - Only the manufacturer can quantify and manage endurance, What's is the spec?
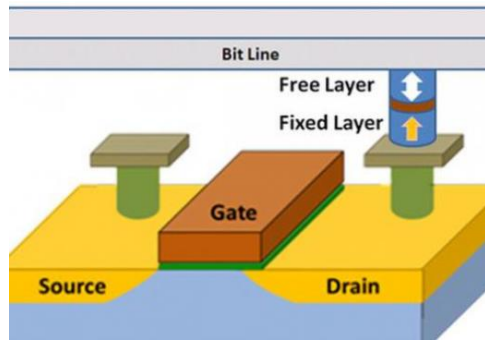
New Memories

MRAM

- Storage in Magnetic Element
- Toggle is historical tech
- STT-MRAM provides scalability
- One of the oldest "new NVM"
- 1T1R. Cross point options are not clear

MRAM Compared to Other Technologies

- **MRAM Advantages**
  - Speed. It is the fastest new NVM
  - Maturity compared to other "new NVM"
  - Multiple manufacturers joining development
    - IP providers, Foundries, Memory Companies
    - This gets equipment suppliers engaged and spending money on development
    - Example: Samsung being involved pulls everything forward.

Mark Webb, MKW Ventures Consulting

MRAM Compared to Other Technologies

- **Disadvantages**
  - Cost/Density
    - $8F^2$ theory, $10\text{-}20F^2$ planned, $50F^2+$ is more typical today
  - Very limited shipments of STT (Spin-Transfer Torque)
    - Toggle ships today but doesn't lead to cost effective applications
    - STT needed to achieve this and volume needed to make it mature
  - Even with STT, density projections cannot match ReRAM, 3D Xpoint
    - No clear crosspoint or transistor-less plan

Technology Status

- **Companies shipping measurable volumes of MRAM**
  - Toggle today as it is a mature technology with sales
  - STT-MRAM provides higher density and is the future
- **Everspin partnering with Global Foundries to ship stand alone and embedded**
  - 28nm MRAM standalone being planned (40nm Shipping now)
  - 256Mb Shipping, Plan to Sample 1Gb 2018
  - 22nm embedded MRAM available in upcoming GF designs
- **Multiple Companies licensing IP to improve performance and reliability**
- **Avalanche announced partnership with UMC**
- **All major logic companies and foundries are committing to MRAM**
- **IMO: MRAM future growth confidence is a "Tale of Two Markets"**

Embedded Market

- Embedded Market is very attractive for MRAM!
  - E-Flash scaling issues limit density and cost reduction
  - SRAM scaling is slowing as finfet SRAMs require large F^2
  - MRAM power in embedded is better than SRAM
  - Merging NVM and SRAM is more efficient
  - Densities needed are near sweet spot MRAM density
- Multiple vectors all pushing for MRAM
 - MRAM is solution to a present problem

MRAM Stand Alone Memory

- Standalone MRAM memory has more challenges than embedded
  - No short term path to MRAM being able to match DRAM on cost or density
  - NOR flash is a viable execute-in-place NVM in 256MB and below
  - Higher density NVM (>1Gbit) will use NAND due to extremely low cost
- ReRam and 3D Xpoint are structurally lower cost and more dense for "NAND-DRAM Latency gap" applications
- Therefore MRAM is best applied to 256-1Gbit where DRAM like speed is desired, NVM needed, and cost is not a major issue
  - Recently announced IBM Storage system
- MRAM Replaces battery/capacitor backed DRAM, Low density DRAM, NOR applications.

Mark Webb, MKW Ventures Consulting
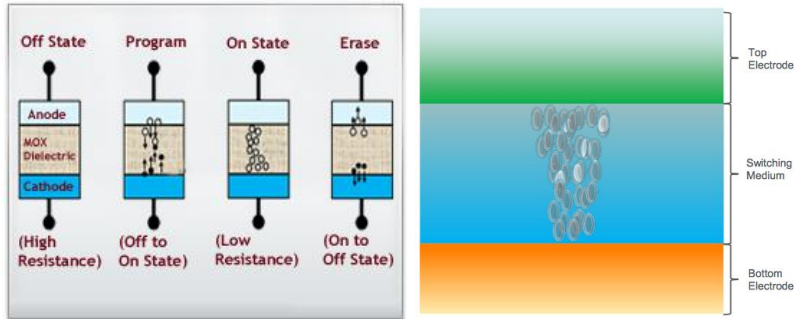
ReRAM

Storage by changing resistance of a element

Example from Filament ReRAM

High resistance State
Bias applied: Causes filament to form (low resistance)
Reset to eliminate filament to return to high bias state

What We Know About ReRAM

- Lots of technology papers on storage and select elements
- Historic Revenue has come from embedded applications in Kbits or Mbits
    - Unity (Rambus), Adesto, Panasonic, etc
    - Higher Density Needed
- 2018: Crossbar,Inc ReRAM is available today as embedded memory (1T1R) from Foundry. Standalone memory planned
- 2017: 4DS announced functionality on 1000+ cells, DRAM like speeds. Now partnering with IMEC for development
    - Interface switching is different from historical filament base
- Multiple Memory companies are investigating ReRAM

Why ReRAM

- Scalable, Crosspoint solutions are low cost ($4F^2$) relatively simple and allow foundry utilization
- No significant new Fab tools needed (model)
- Selector elements allow for high density 1TnR
- Latency is far better than NAND, 100nS range
- Cycling past 1M cycles on cells is achievable
- All of these are in same category as 3D Xpoint.
  - Faster than NAND, slower than DRAM
  - Cheaper than DRAM, More expensive than NAND
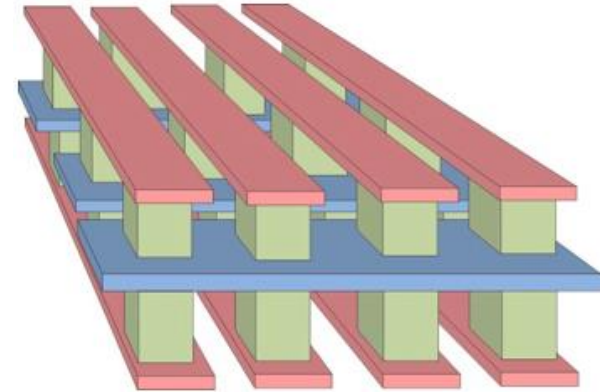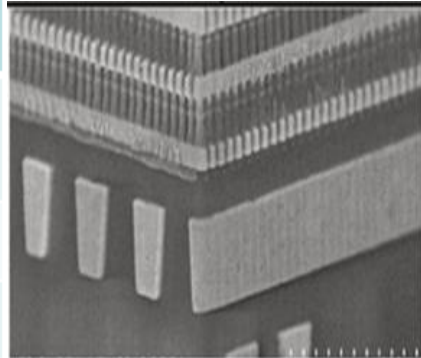- Scalable, stackable, high density, lots of research

Previous Technologies Reviewed

Device	COST $/GB
2017 DRAM COST	1X
27nm	1.1x
20nm	0.6x
14nm	0.4x
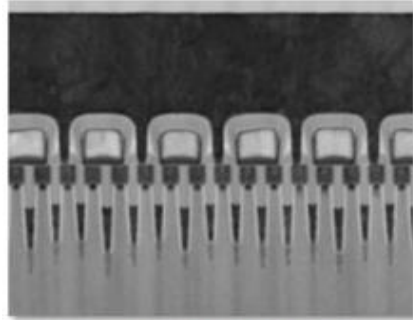
Advanced 27nm RRAM Cell

Device	COST $/GB
2017 DRAM COST	1X
24nm	0.6x
14nm	0.3x

- Multiple ReRAM technologies reviewed
- **Stacked Crosspoint (~4F^2) is optimum for next 5 years**
- Allows memory optimized or foundry
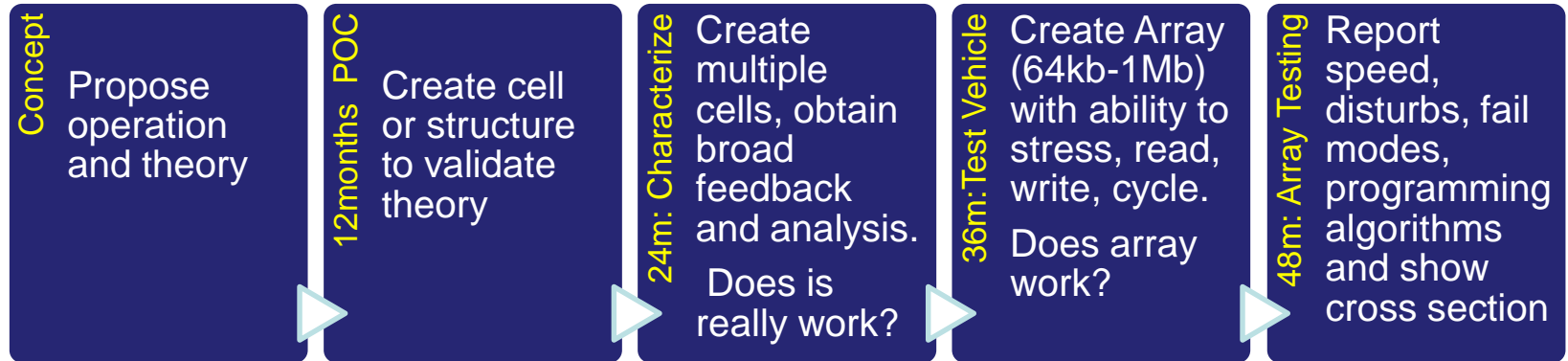- 3D NAND like vertical later

Additional Memories

- 3D Xpoint, MRAM, ReRAM are in some stage of availability today.

- All have been in development for years

- Previous versions of PCM memory from Intel/Micron were published in detail with 64M arrays almost 10 years ago
  - MODEL: IT IS 10 years from tested array to revenue
  - That is with 100s of Millions in investment

- Other memories show promise but have limited chip level data so ramp in next 5 years is not clear

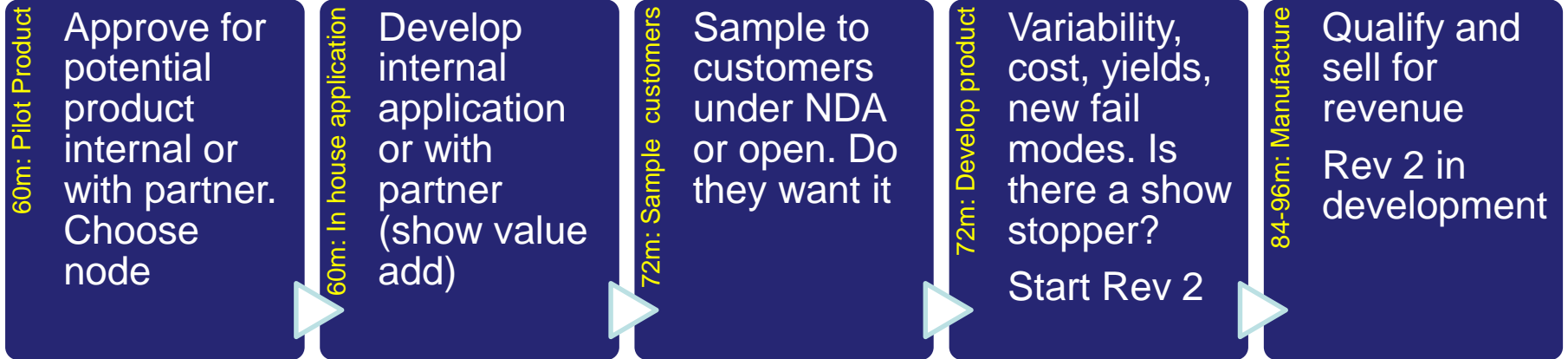# NEW: Product Lifecycle for Memory Part 1: Technical Proof/Open Communication

Concept — Propose operation and theory

12months POC — Create cell or structure to validate theory

24m: Characterize — Create multiple cells, obtain broad feedback and analysis. Does is really work?

36m:Test Vehicle — Create Array (64kb-1Mb) with ability to stress, read, write, cycle. Does array work?

48m: Array Testing — Report speed, disturbs, fail modes, programming algorithms and show cross section

Summary: " A Novel xRAM technology built in 90nm Lithography With pictures!

Product Lifecycle for Memory Part 2: Usually Stealth
"We are going to build a product"

| 60m: Pilot Product	60m: In house application	72m: Sample customers	72m: Develop product	84-96m: Manufacture
Approve for potential product internal or with partner. Choose node	Develop internal application or with partner (show value add)	Sample to customers under NDA or open. Do they want it	Variability, cost, yields, new fail modes. Is there a show stopper? Start Rev 2	Qualify and sell for revenue Rev 2 in development

Problem with Hype/New tech

- Jump to page two before completing page one. Fine to speed flow, but we know that 90% of technologies don't make it.

- It rarely makes sense to be stealth in part one
  - You want to show research
  - You need funding
  - The work is either patented or is not novel
  - You need to get feedback

- Open and closed companies follow this (FinFet, 3D NAND, PCM, MRAM, ReRAM, etc

- RECOMMENDATION: Evaluate all technologies based on where they are on lifecycle.
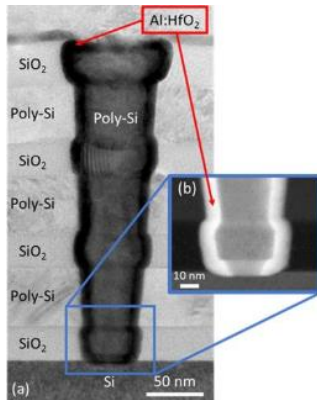
NRAM

NRAM Cross-Section

NRAM CELL
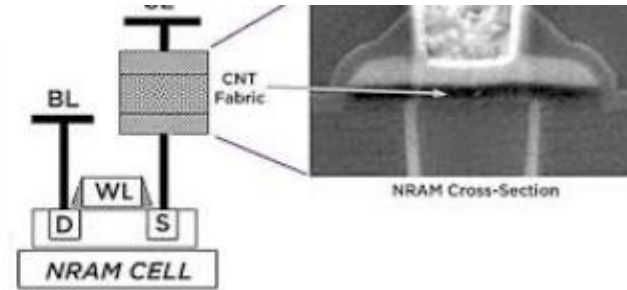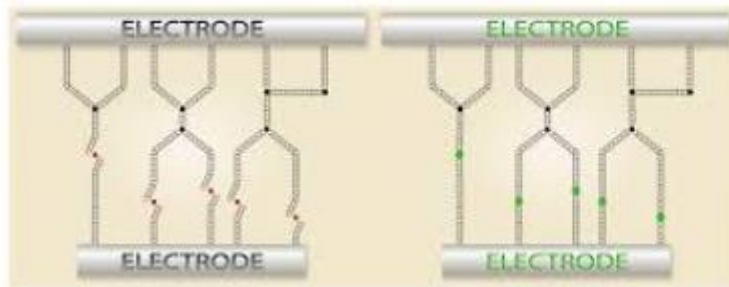
- Carbon Nanotube Memory
- Nanotubes aligned to conduct or not based on bias
- Reported to be 10ns latency, infinite endurance, scalable to 16Gb DDR4 memory and beyond (Nantero presenting this week)
- This week: Low density at Fujitsu 2019, high density samples 2019, production 2020
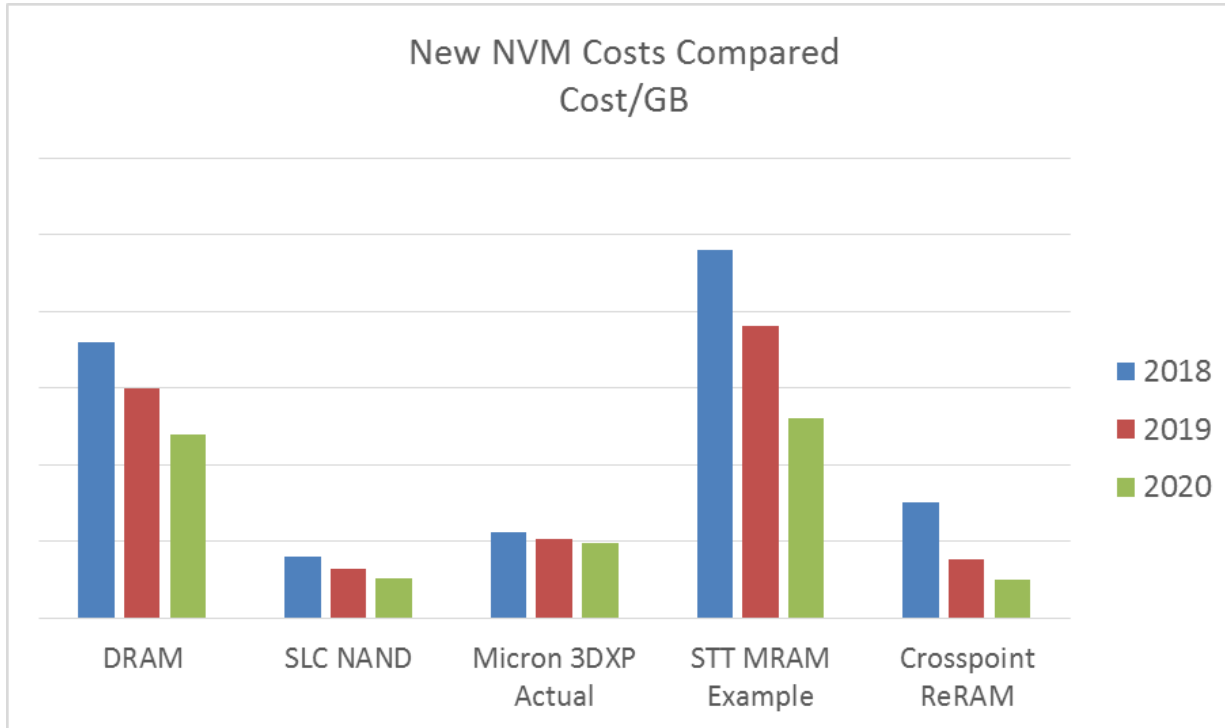
FeRAM

- FRAM have been used in low density (Kb) for years
- Recent IMEC focus has targeted higher density.
- Potential for speed and endurance is driver, tradeoffs TBD

Cost for Multi-Gbit Technologies

New NVM Costs Compared
Cost/GB

2018
2019
2020

DRAM   SLC NAND   Micron 3DXP Actual   STT MRAM Example   Crosspoint ReRAM

Summary of Emerging Memory

	Pros	Cons	Summary
3D XPoint	Fast SCM, in production, relatively high density and low cost	Can't cycle like DRAM. Slower than DRAM. Scaling	Highest Volume new NVM
MRAM	Fastest NVM… Multiple products in production, embedded apps	high cost, relatively low density	In volume with multiple applications
ReRAM	High density, low cost, Fast SCM, available today	Can't cycle like DRAM. Slower than DRAM. Low volume	Will compete with 3DXP apps with volume release
NRAM	Fast, high density, no cycling issues, stable	No manufacturing data, limited reports	Planning sample in 2019, volume in 2020
FRAM	Fast, potential for endurance	5+ years	Need data
Other	In Concept mode	10+ years	Need data

Mark Webb, MKW Ventures Consulting

Equipment/Investment Challenges

- It takes 100s of millions to properly develop a new memory to a volume market
  - The revenue while in development is an order of magnitude less
  - Equipment vendors only support high volume products
  - Must overcome this paradox or development takes 10-20 years
- Samsung and Intel both overcame this, Can others?
- Do we need to actually hit the brickwall to prioritize new memories
- Do we require deep pocket partner for all work?

Future NVM Scenario 2018

- NAND and DRAM stay through 2023. But we have options!

- PCM/3D Xpoint will initially dominate the NAND/DRAM gap. Ramp in 2018-2020

- Competing PCM chips proposed by other companies with similar architecture to Xpoint with be release

- ReRAM develops to address similar market with similar performance and cost.

- MRAM continues to grown in lower density, higher speed applications plus very high end NVDIMMS/SSDs

- Chip Samples/technical evaluations on NRAM, FRAM reported in 2020. based on results, revenue products are proposed

Summary

- NAND has 5+ years and 3+ generations to go.

- DRAM will slowly scale for next 5 years

- There is no QLC NAND replacement… ultracheap memory

- We have solid new memories that are actually available

- These will grow and augment, not replace DRAM and NAND