The Computer **IS** the Memory

Tom Coughlin, Coughlin Associates

Jim Handy, Objective Analysis

Coughlin Associates
Data Storage Consulting

Outline

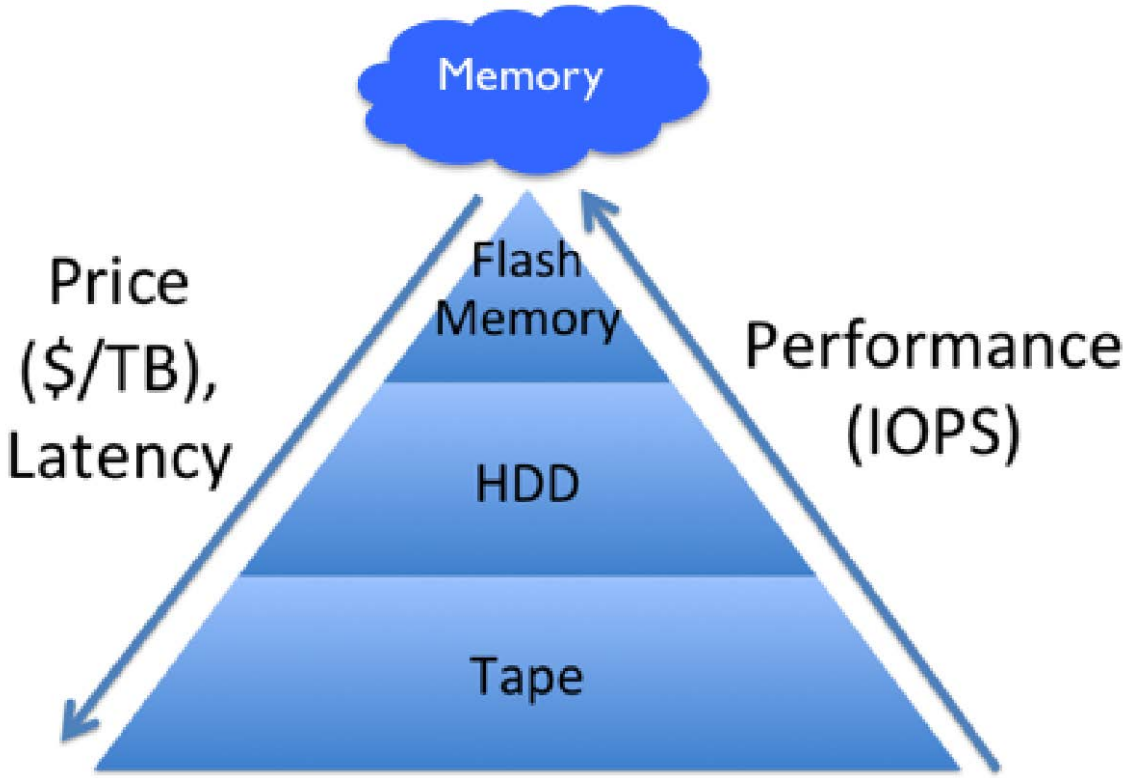
- The Shape of Things to Come
- Getting Memory & Processing Closer
- Spin-Based Memory and Processors
- Conclusions

The Shape of Things to Come

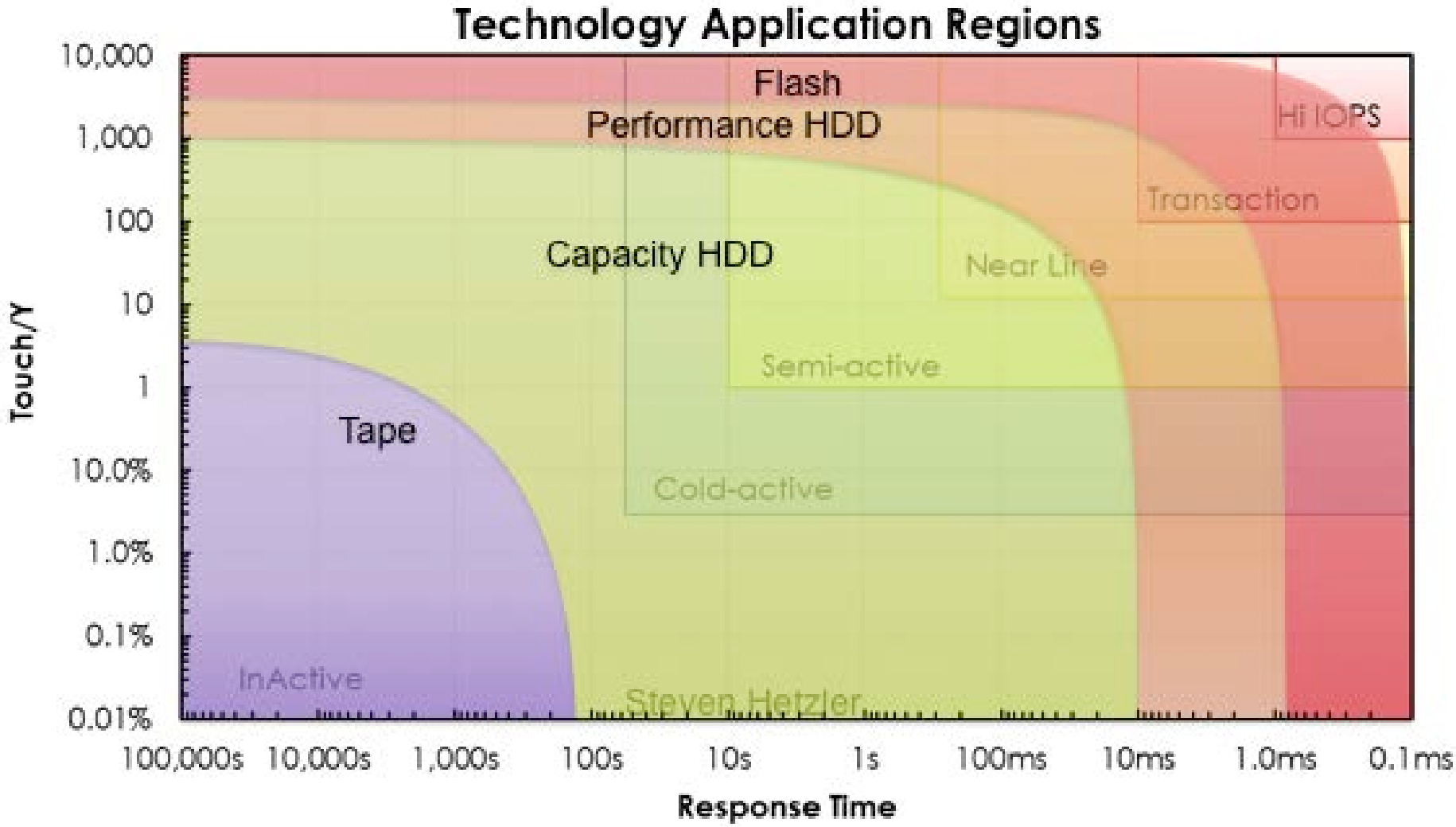
Memory/Storage Hierarchy

NVM/Storage trade-offs

- Storage costs ($/TB)
- Performance
  - IOPS
  - Data rates

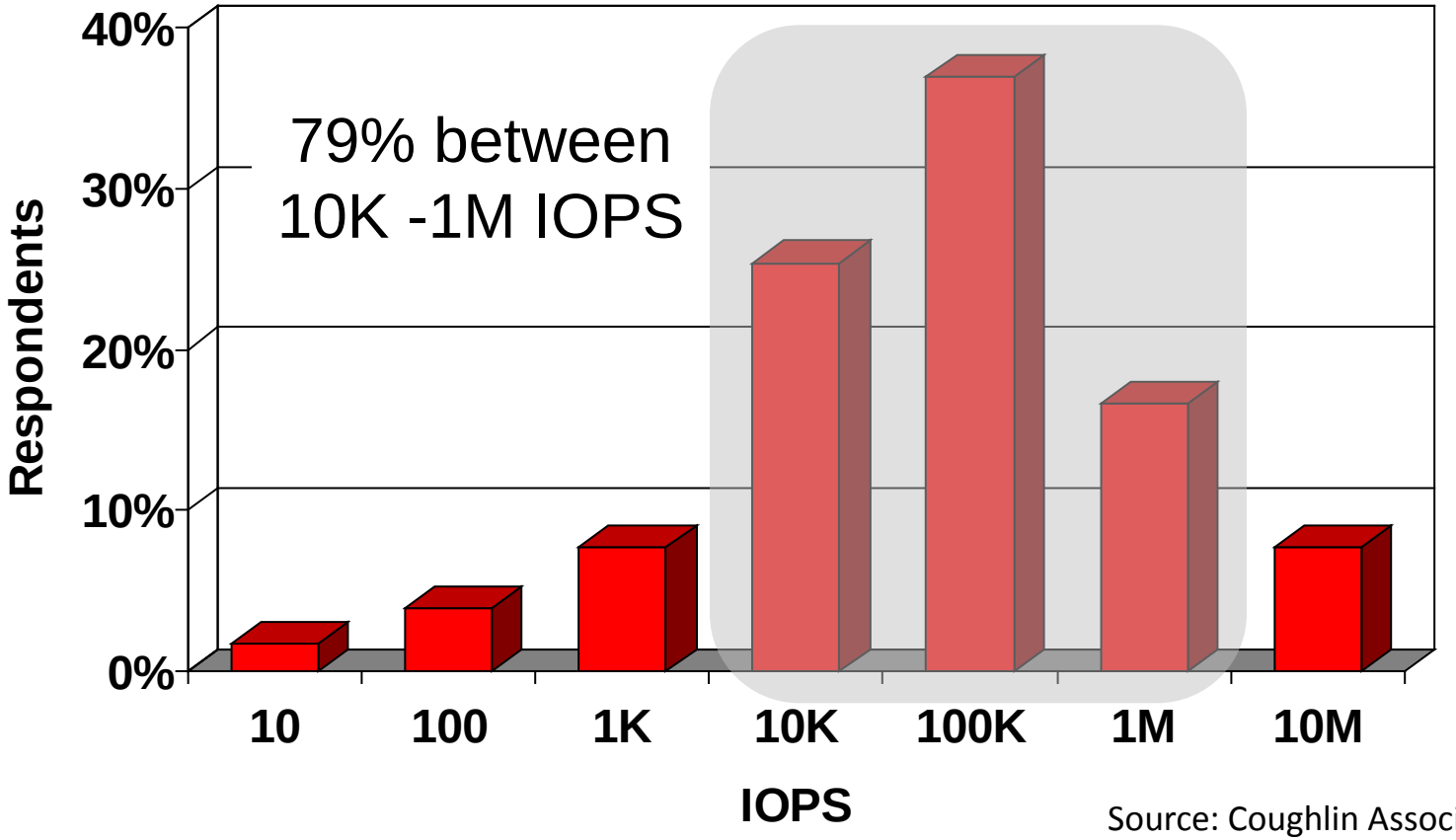


Touch Rate



Source: Steve Hetzler, IBM

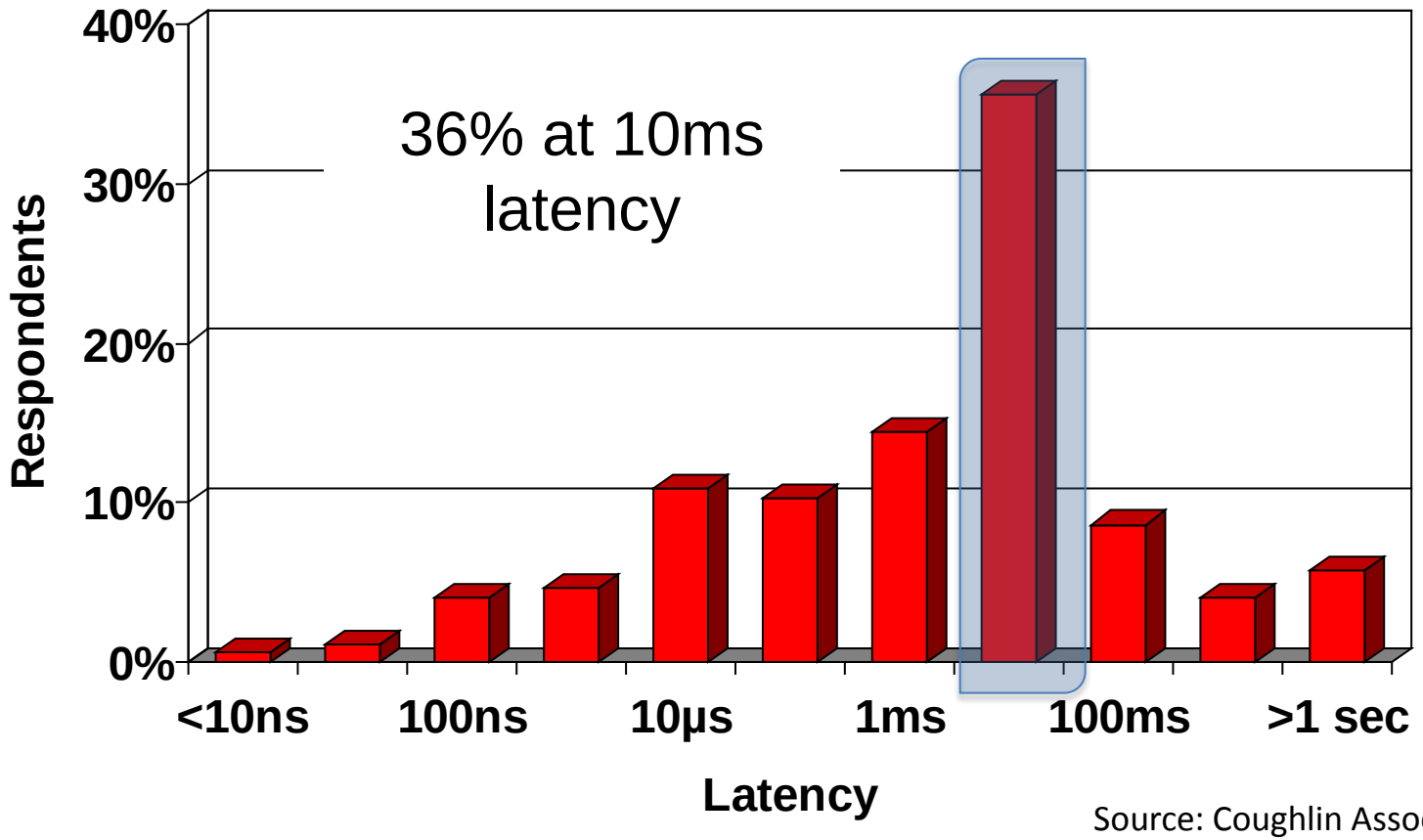
IOPS Required



Source: Coughlin Associates

How Many IOPS Do You Really Need?

Minimum Latency Requirement



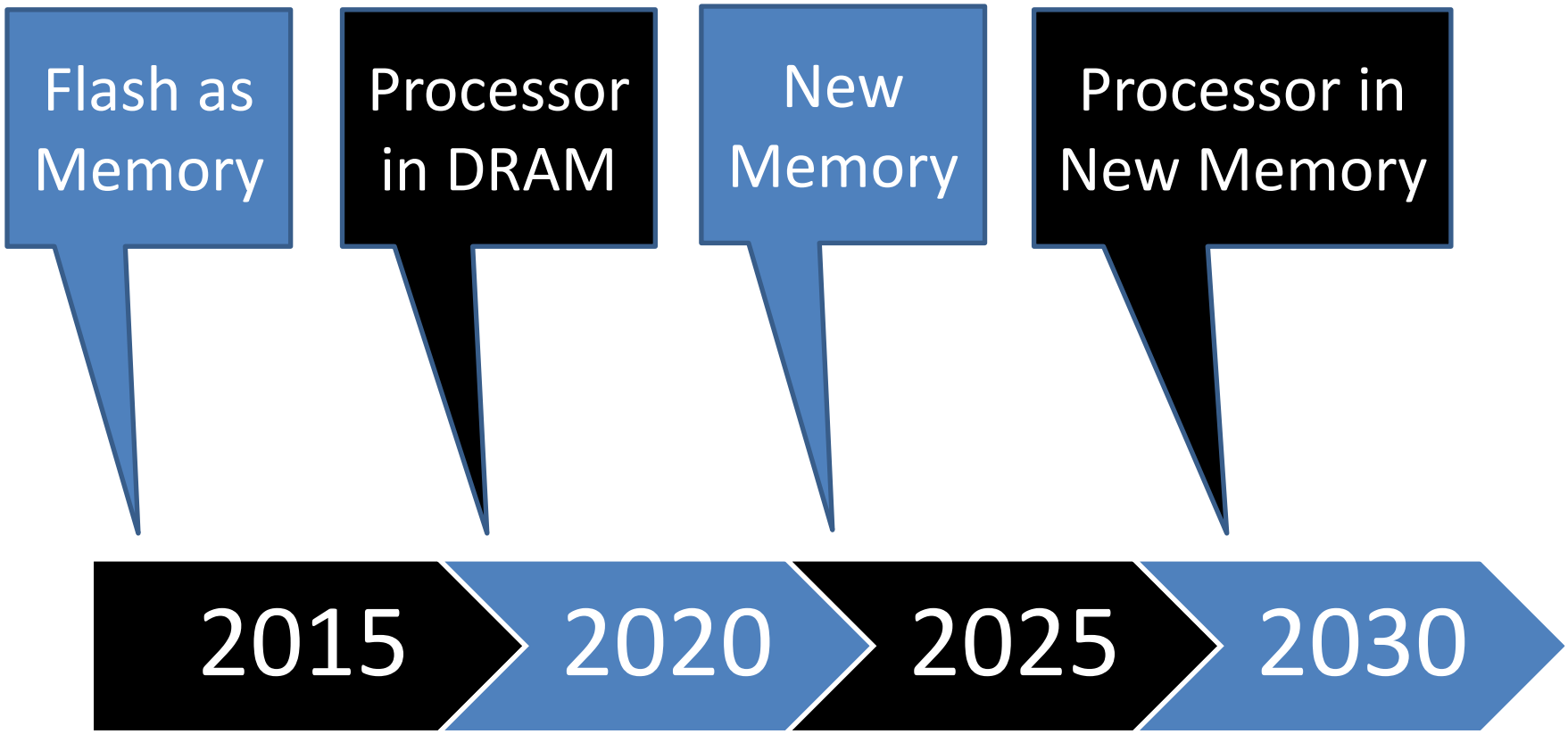
Source: Coughlin Associates

How Many IOPS Do You Really Need?

Persistent Memory Implications

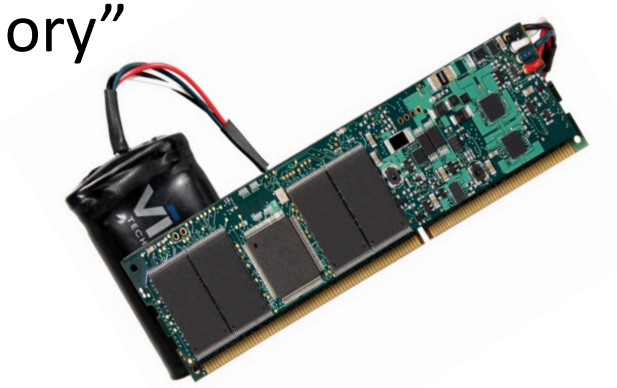
- Retains data during a power loss
  - Instant recovery of state before power down
- Lower latencies than disk
- Lower power than DRAM
- Allows persistent states for Remote Direct Memory Access (RDMA)
- Supports “logic-in-memory architecture”
  - Could lead to new distributed computer architectures

A Timeline For Change

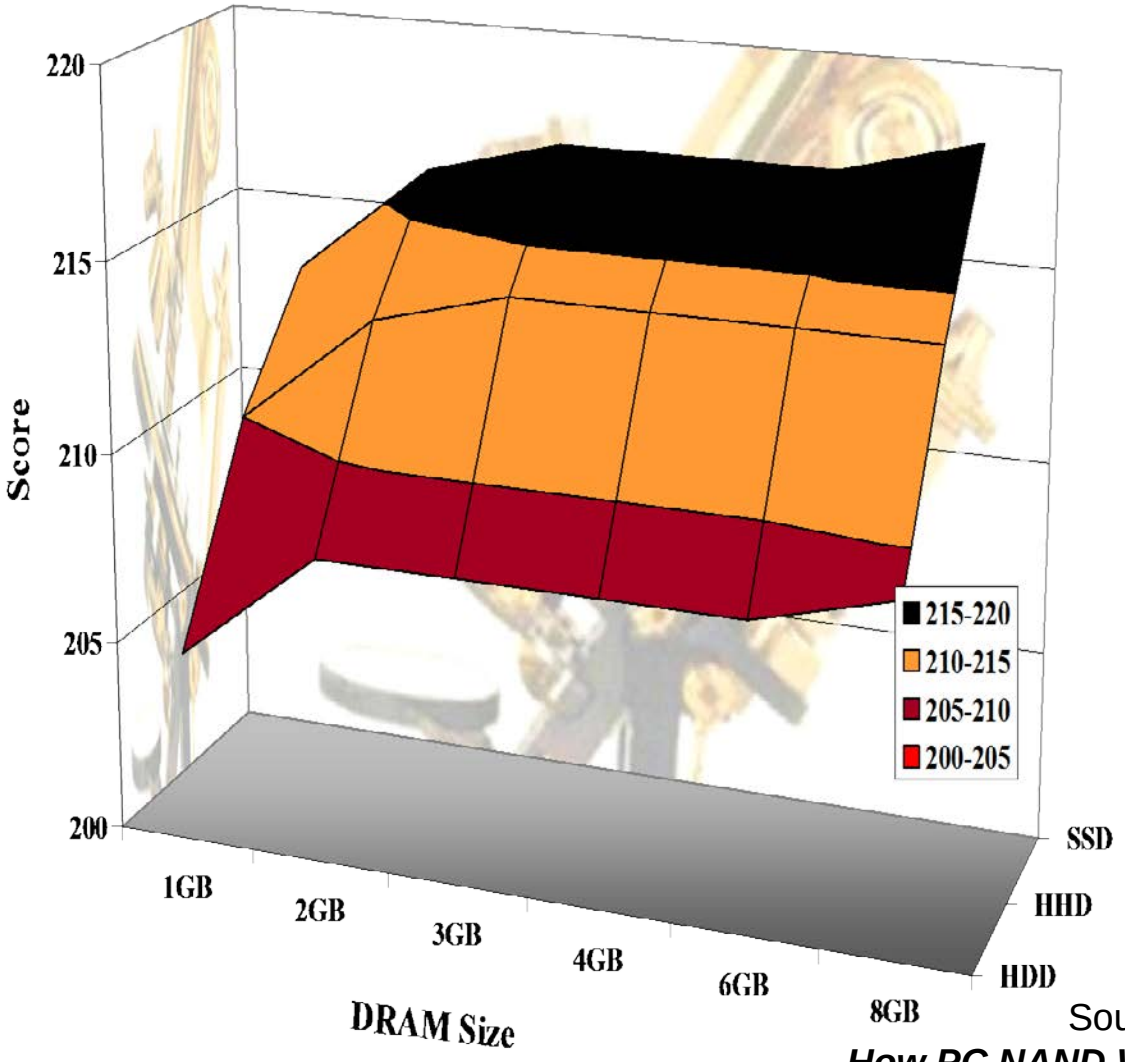


Flash as Memory

- Already exists on PCIe bus
  - Fusion-io pioneered “Extended Memory”
- Now migrating to memory bus
  - NVDIMM-N: DRAM backed by flash
  - NVDIMM-F: Flash on a DIMM
- Presents new problems to existing compute model
  - Memory is persistent
  - Flash is weird (wear, block erase, page write, etc.)
  - SNIA standards evolving to cover this



Flash Already More Economical than DRAM



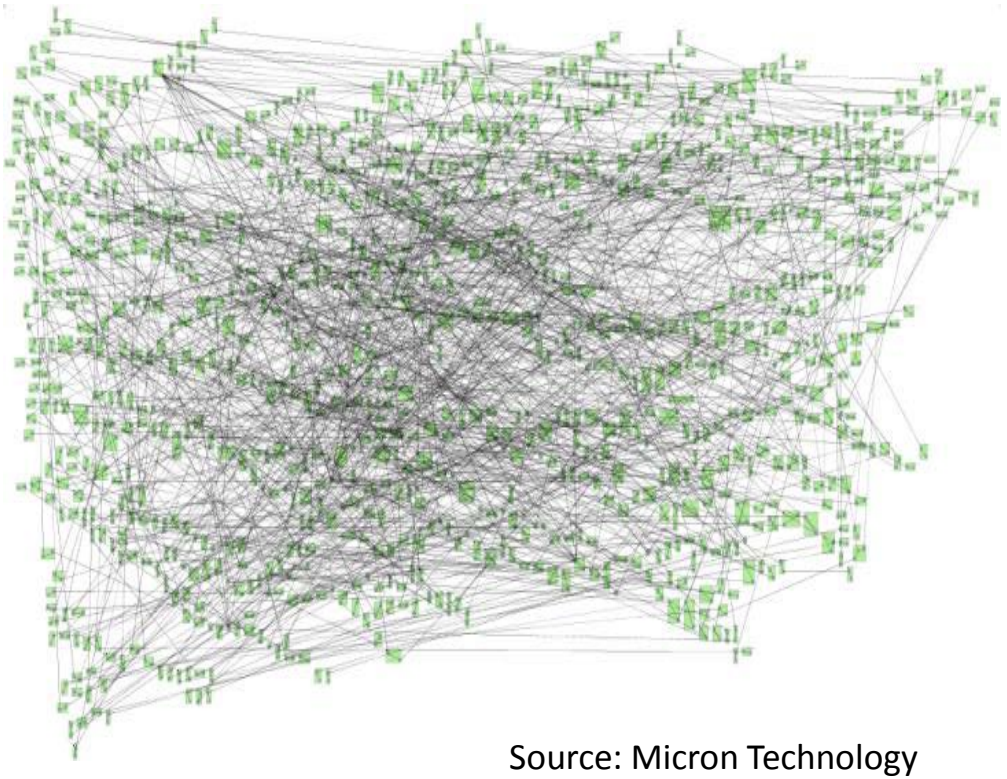
Source: Objective Analysis

How PC NAND Will Undermine DRAM

Getting Memory & Processing Closer

Processor in Memory

- Pioneering efforts already exist:
  - Micron Technology Automata Processor
  - Venray Technology TOMI
  - Berkeley iRAM (1996)
  - Notre Dame PIM (1996)



Source: Micron Technology

Trouble Mixing Logic & DRAM

Logic Process

- High speed
- Variable capacitance
- Multiple metal layers
- Weaker cost focus
- Muddies up a DRAM process

DRAM Process

- Slow speed
- High capacitance
- Few metal layers
- Cost-obsessed
- Undermines a logic process

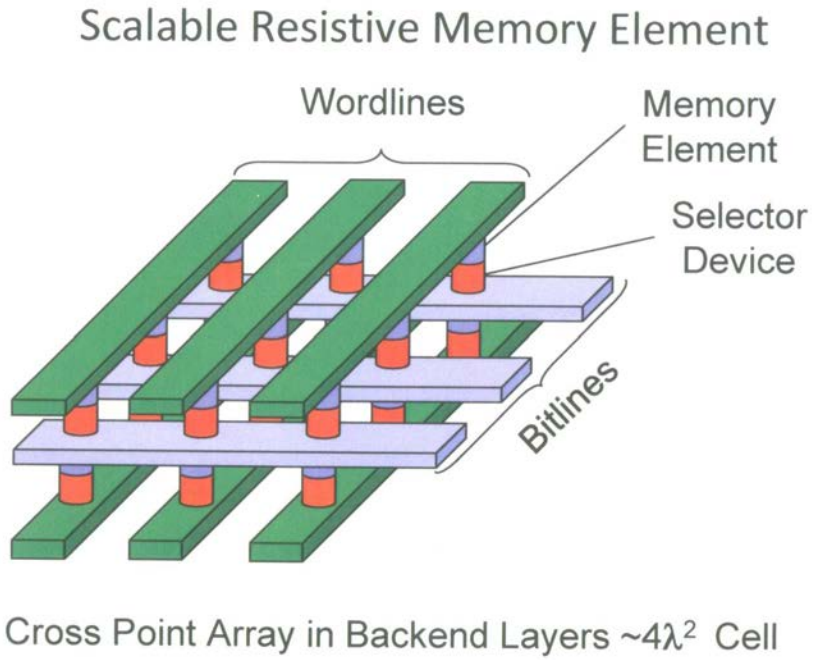
Alternative Memories to the Rescue

- Byte read/write
  - Read/write speeds roughly balanced
- No “Erase Before Write”
- Logic-friendly fabrication process
  - Supports easier integration of logic and memory
- Persistent, inexpensive, & fast

Yes, But *Which* New Memory Will Win?

- Many good alternatives:
  - STT MRAM
  - ReRAM
  - Memristor
  - Racetrack
  - PCM
  - FRAM
- Hard to know which will cross over DRAM & NAND pricing

ReRAM or RRAM Cross Point Array

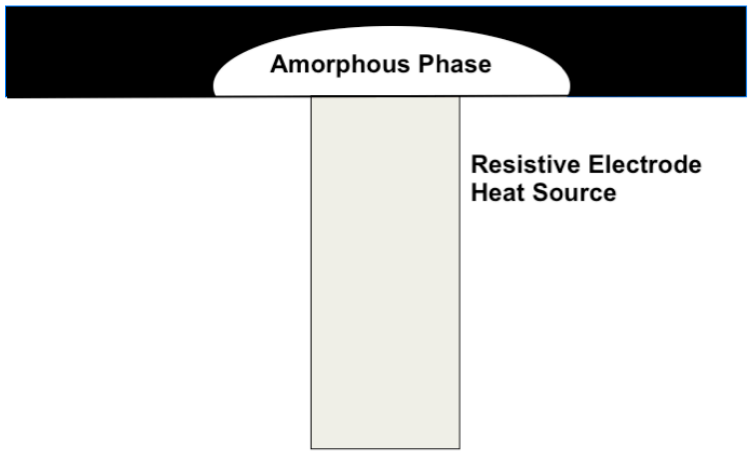


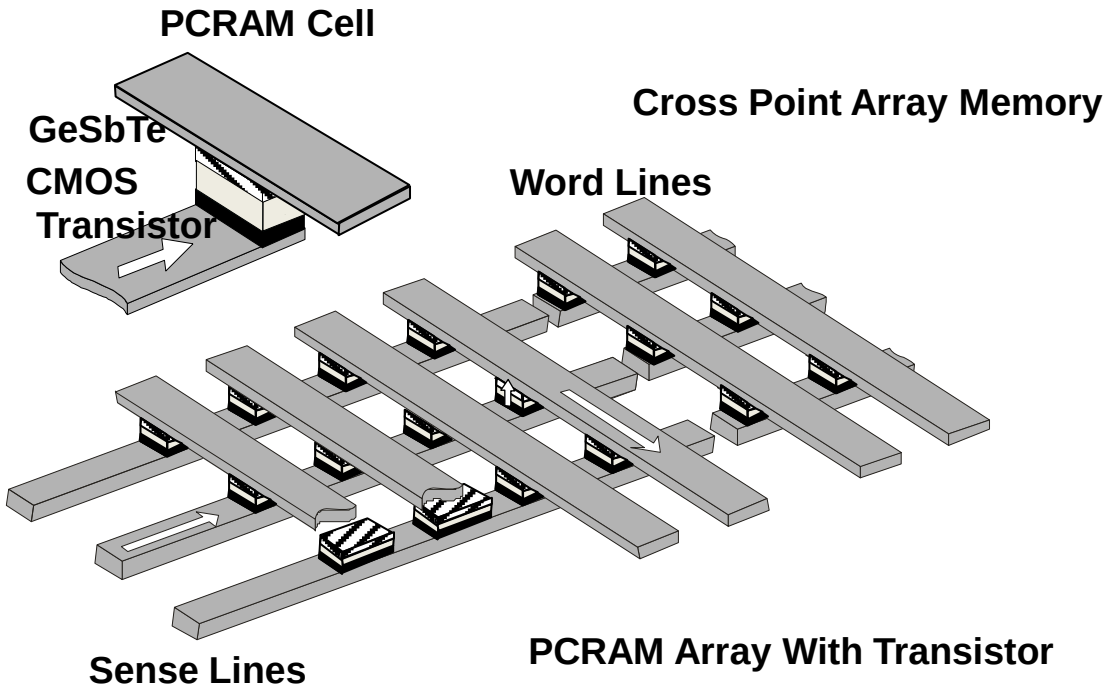
Source: Flash Memory Summit 2013

- Likely NAND replacement
- Multi-film diodes eliminate need for larger CMOS selector

Phase Change Memory (PCRAM)

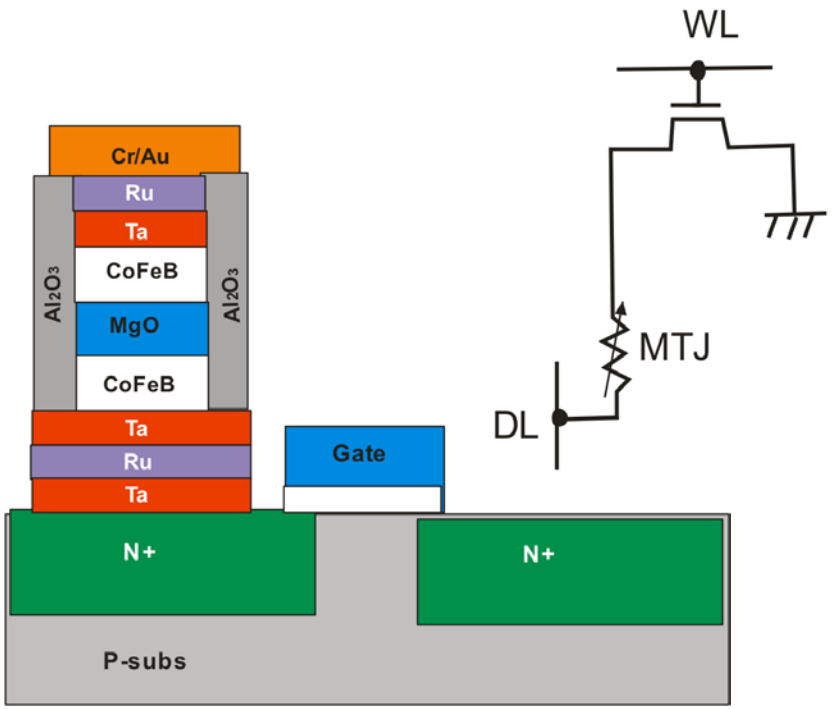
- Bit addressable, high-density arrays.
- The phase change causes resistance change:
  - Crystalline: Low resistance
  - Amorphous: Higher resistance.





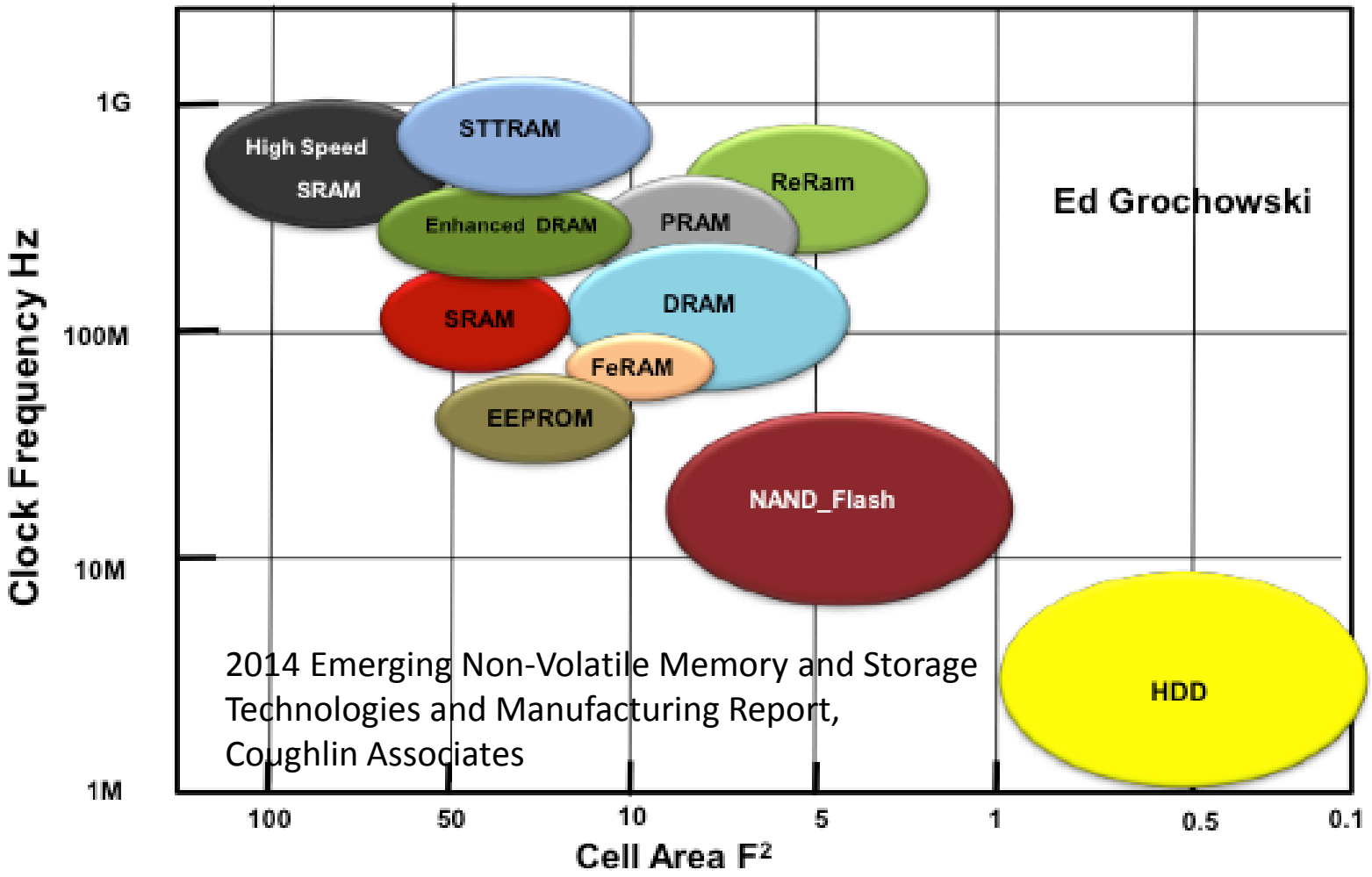
Magnetic RAM (MRAM)

- 3 Major types
  - Toggle Mode (field driven)
  - Spin Torque Transfer (STT)
  - Magnetothermal MRAM (Heat Assisted)

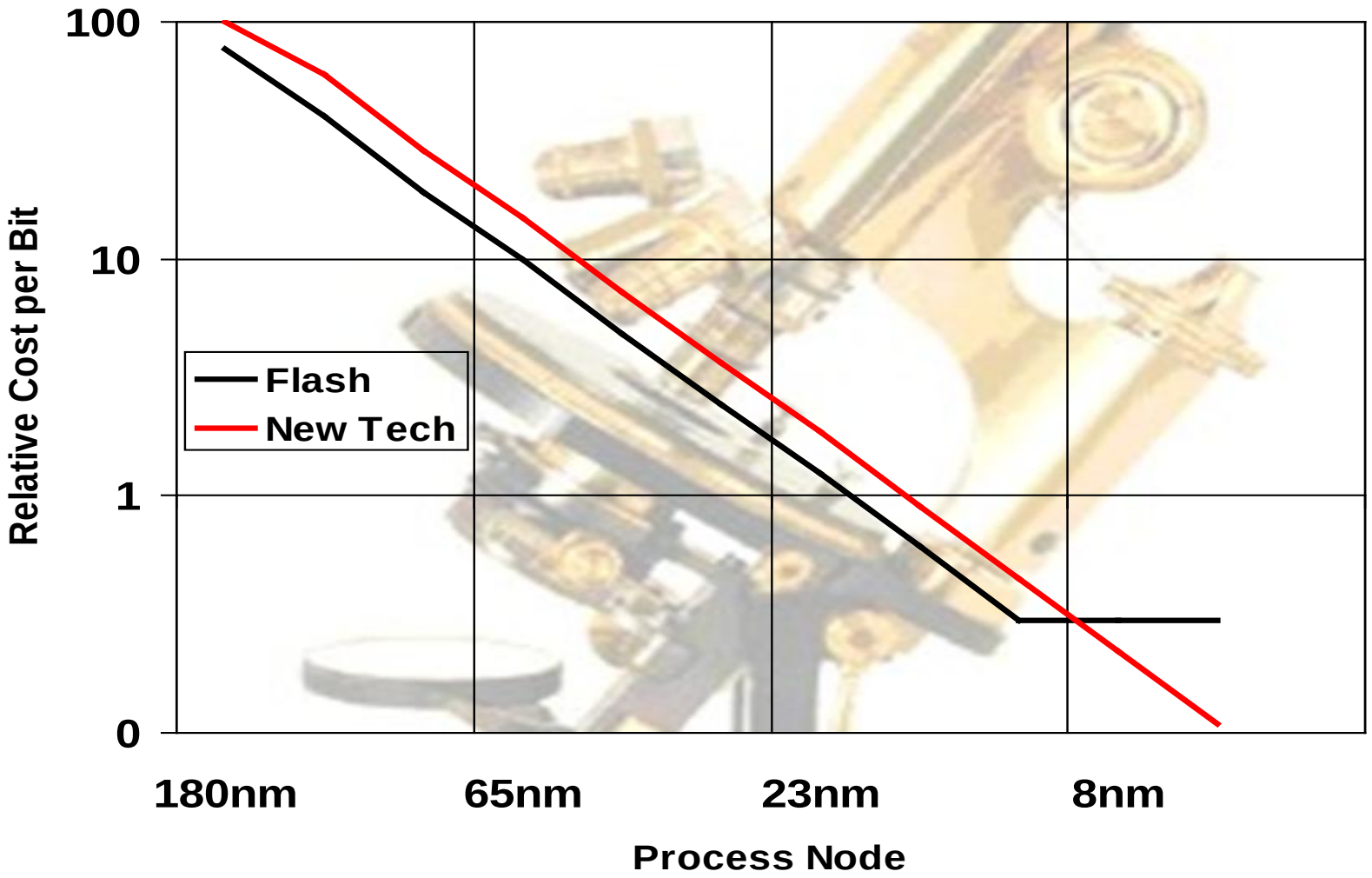


STT MRAM Stack on Single CMOS Transistor

Memory/Storage Performance/Density Roadmap



How Alternatives Will Win



Source: Objective Analysis

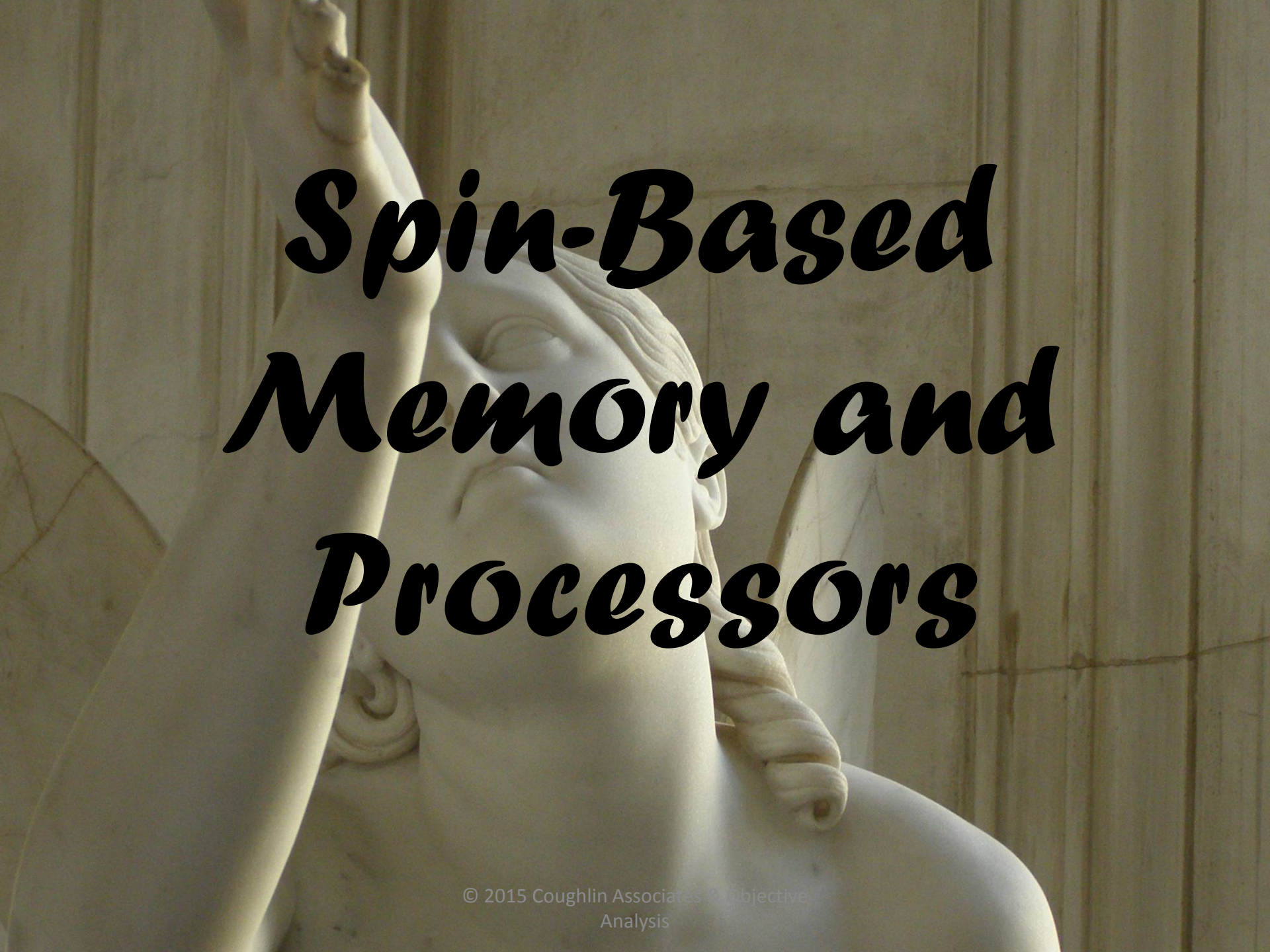
The Impact of Moore's Law

- "Doubling transistors per chip every year or two"
  - Put differently: "Making two chips into one every year or two"
- Using the 18-month rule of thumb
  - 2015 single-socket server with 96GB of RAM & an SSD
    - 200-300 chips
  - 2030 – 15 years later
    - About 1/5th to 1/3rd of one chip
- Today's 10,000-server data center could fit in 10 server slots!
  - Computing structures will adapt to this evolution
- Will Moore's Law stop? Stay tuned!

Summary

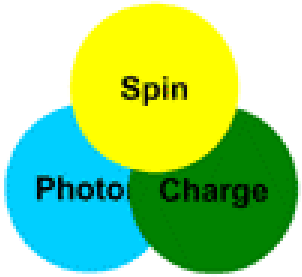
- First step: Persistent Memory
- Second Step: Processor in Memory
- Third Step: Processor in Persistent Memory
- Finally: Data Center in a Rack

Architectures and code will have to adapt!



Spintronics

Spintronics (Spin + Charge)



Semiconductor (Charge)

Magnetic materials (Spin)

SPINTRONICS

- From Tohoku University in Japan

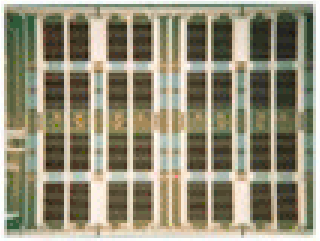
Applications)

HDD (Hard Disc Drive)
Read head

GMR

Large TMR + Low R
Large CPP-GMR

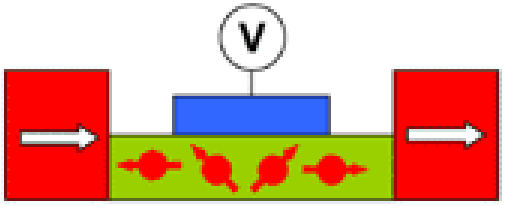
MRAM (Magnetic Random Access Memory)



M. Johnson, *IEEE Spectrum* **37**, 33 (2000).

Huge TMR

Spin-FET (Spin - Field Effect Transistor)



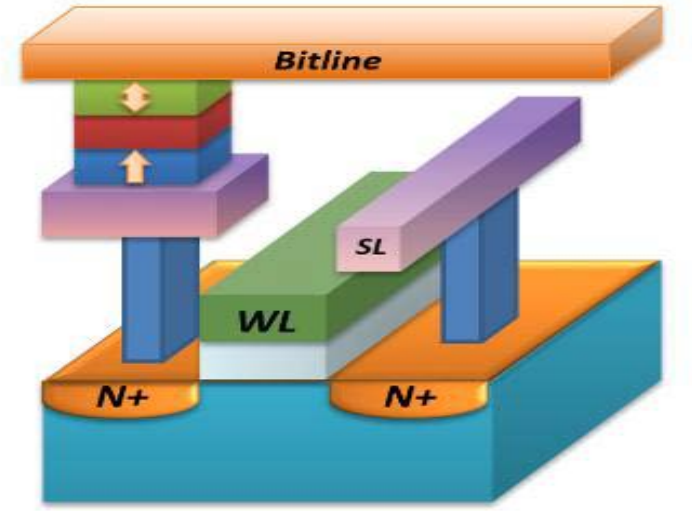
FM1 Semiconductor FM2

S. Datta and B. Das, *Appl. Phys. Lett.* **56**, 665 (1990).

High spin injection efficiency into semiconductor

•TMR : Tunnel Magnetoresistance
•CPP-GMR : Current Perpendicular to Plane- Giant Magnetoresistance

STT MRAM

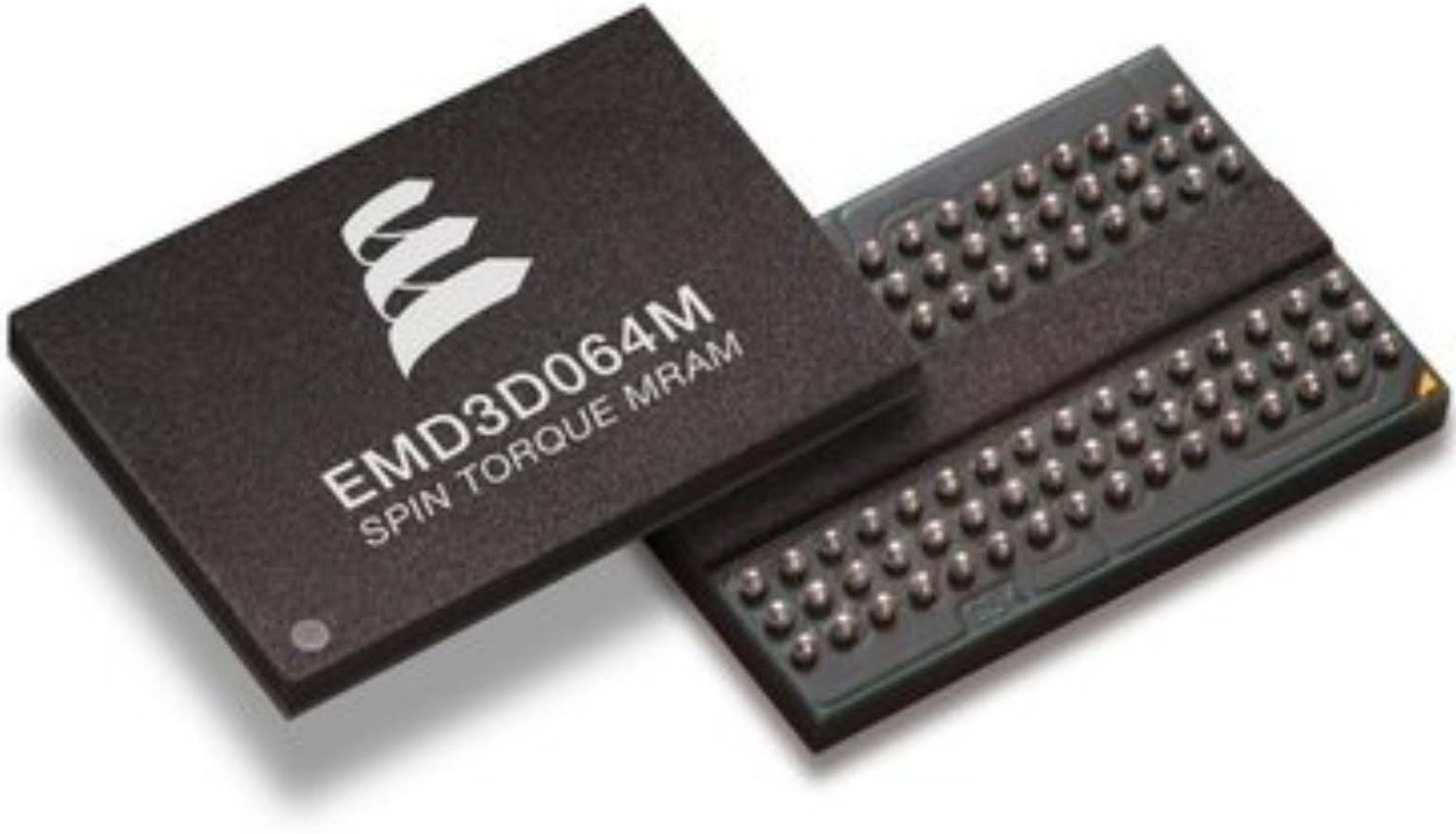


2014 Emerging NVM Report and Their Manufacture, Coughlin Associates

- STT MRAM uses the electron spin to create a memory element
- STT MRAM market could exceed $2 B by 2019

	STT MRAM	DRAM	SRAM
Read Time (ns)	3-5	30	1-100
Write Time (ns)	3-15	50	1-100
# Rewrites	$>10^{15}$	10^{16}	10^{16}
Input Voltage	1.5	2	None

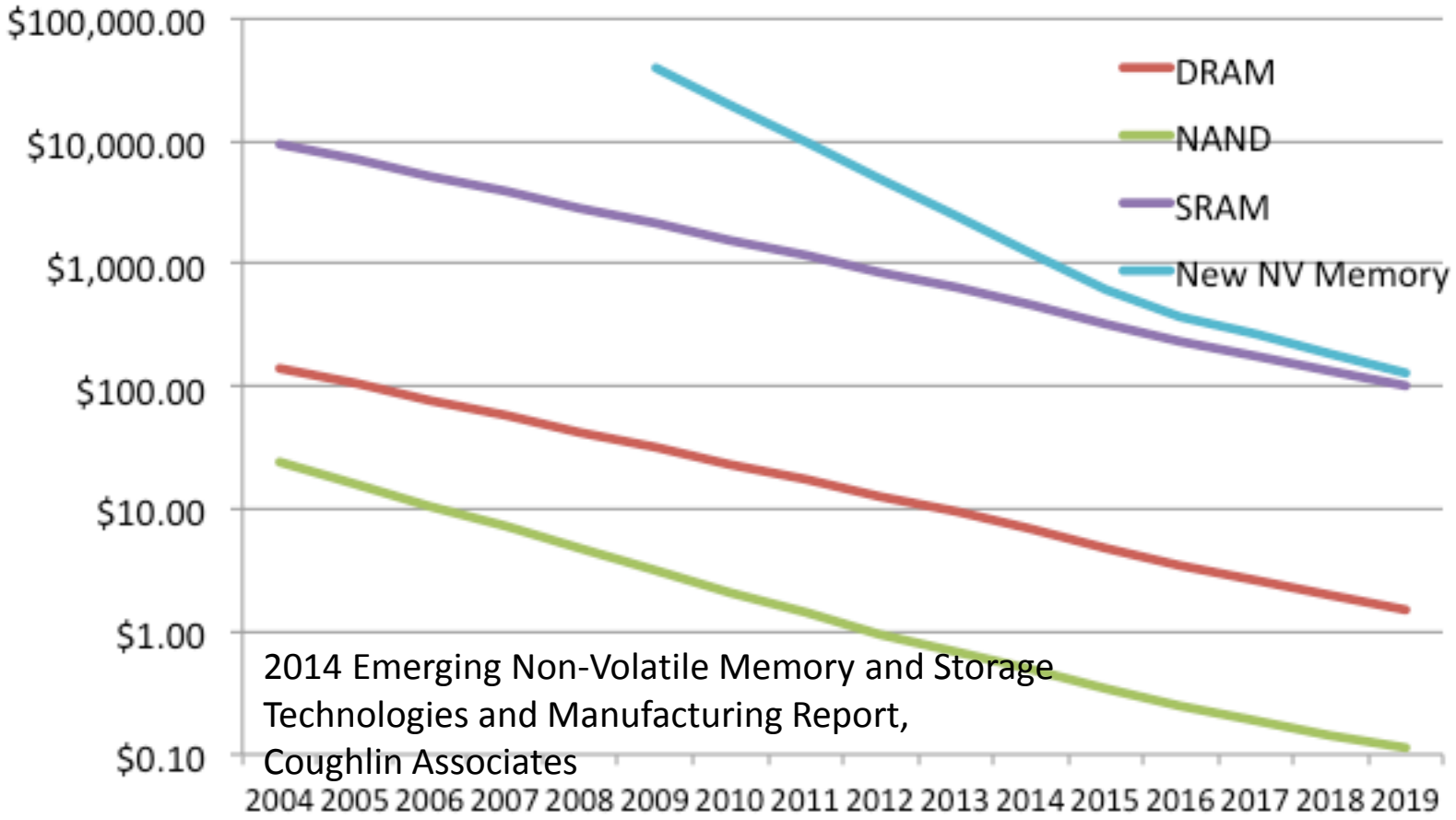
Everspin 64 Mbit STT MRAM Chip Used for Caching



Over 40 M MRAM Chips shipped

$/GB for Memory Technologies

(includes data from Jim Handy, Objective Analysis)



What if Spin Transfer Replaced Charge Transfer (Current)?

- Spin transfer would not generate the heat that electrical currents generate
- Build powerful devices in small spaces
- Spin could be used for both processing and memory/storage

Conclusions

- The storage/memory landscape has more options than ever
- Non-volatile memories will replace volatile memories
- NVM and processors want to come together
- Spin-based electronics are one example of a technology that could put memory and processing in one device


Thanks