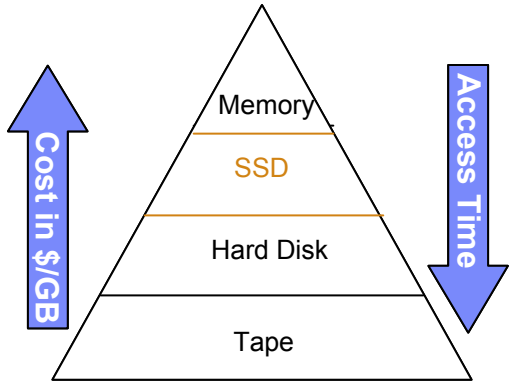
IBM Research - Zurich

Trends in Storage Technologies

Evangelos Eleftheriou, IBM Fellow

CERN Presentation
July 8, 2010

Memory/Storage Stack



CPU cycles

10^{10}	century
10^{9}	decade
10^{8}	year
10^{7}	month
10^{6}	week
10^{5}	day
10^{4}	
10^{3}	hour
10^{2}	minute
10	
1	second

Get data from TAPE (30 s)

Read or write to DISK (5 ms)

Erase FLASH, block (1.5 ms)

Write a Flash page (200 us)

Read a FLASH page (25 us)

Storage

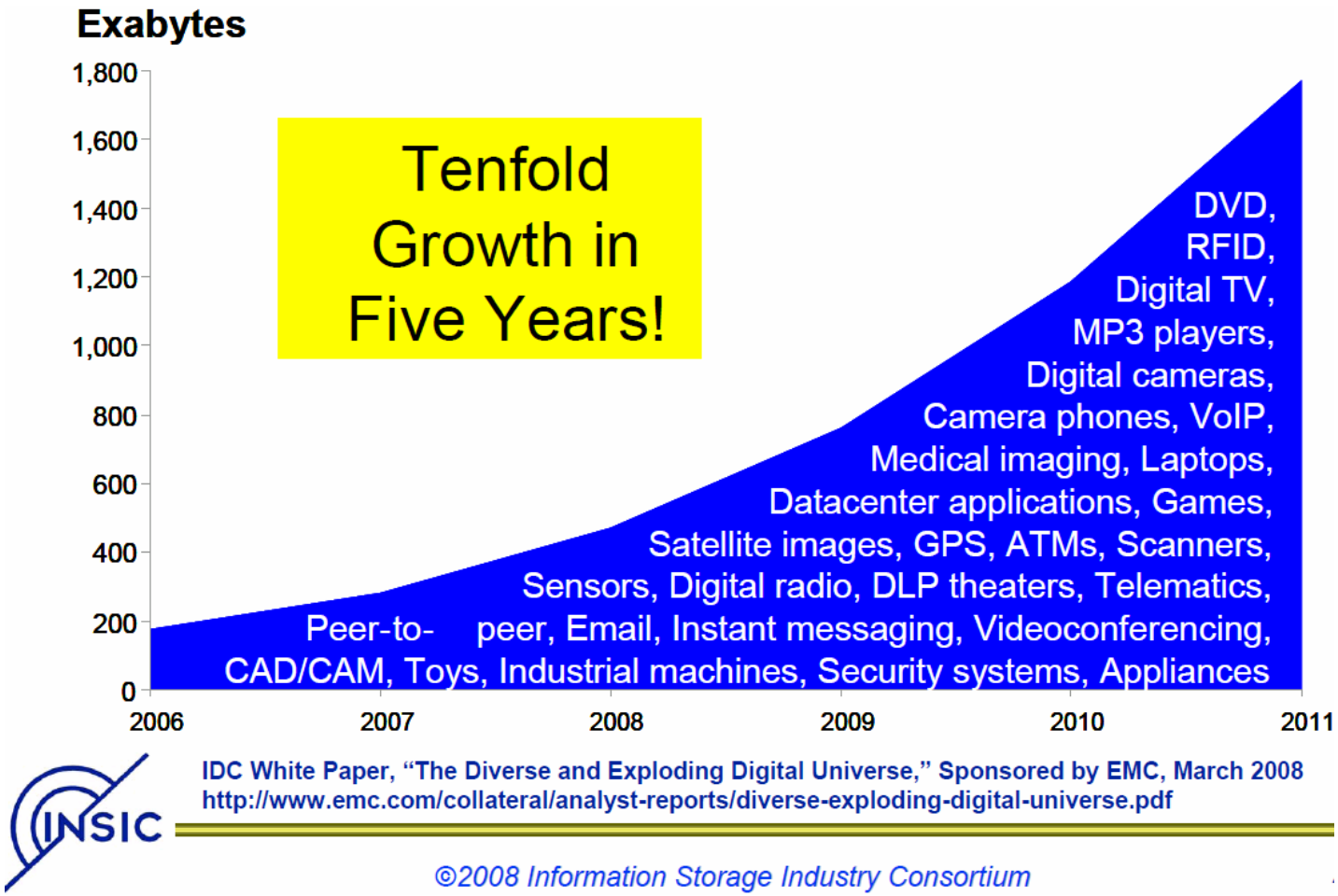
Get data from DRAM (60 ns)

Get data from L2 cache (10 ns)

CPU operations (1 ns)

Memory

Digital Information Created, Captured & Replicated Worldwide



Outline and Motivation

- Disk technology: only moderate improvements in speed and latency
  - Growing needs for streaming analytics, requires high IOPS
  - Growing need for massive digital data archival, requires low TCO

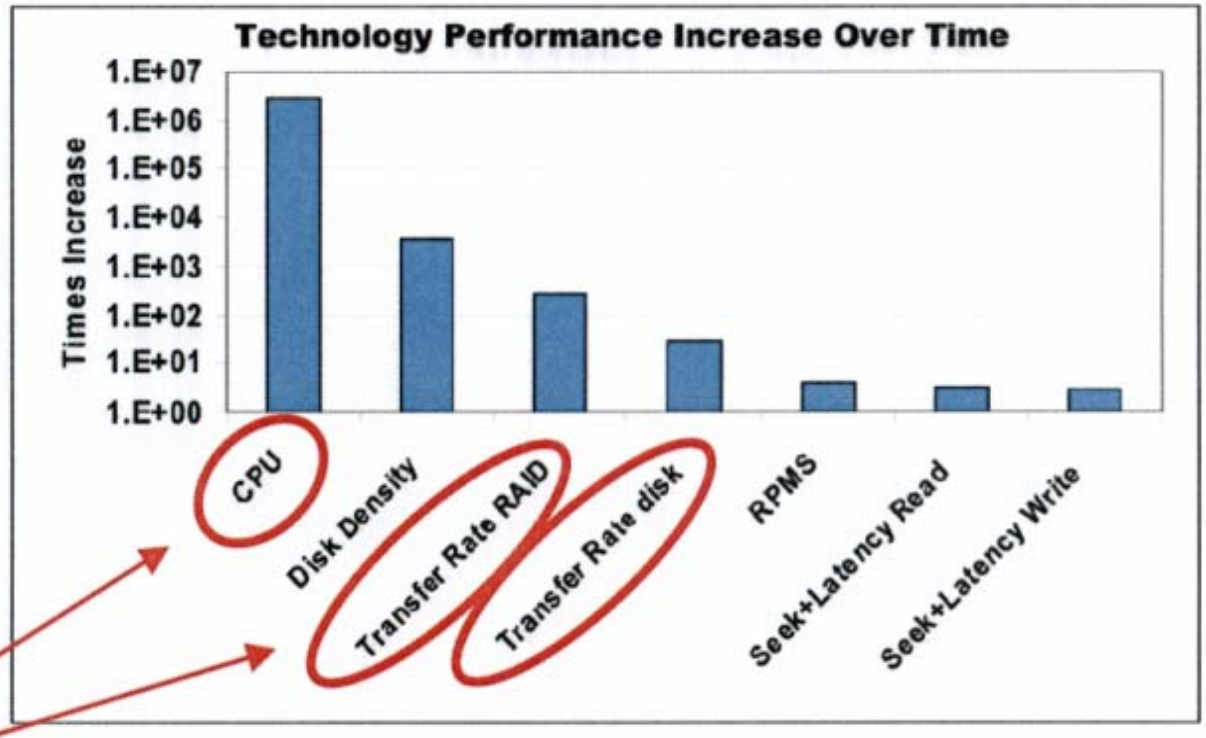
- Solid-state memory technology is addressing the high IOPS segment more efficiently than disk

  Currently Flash technology: What will come after Flash?

- Tape has addressed, so far, the low TCO segment more efficiently than disk

  Can it maintain its substantial cost/GB advantage over disk?

Performance Increases from 1977 - 2006

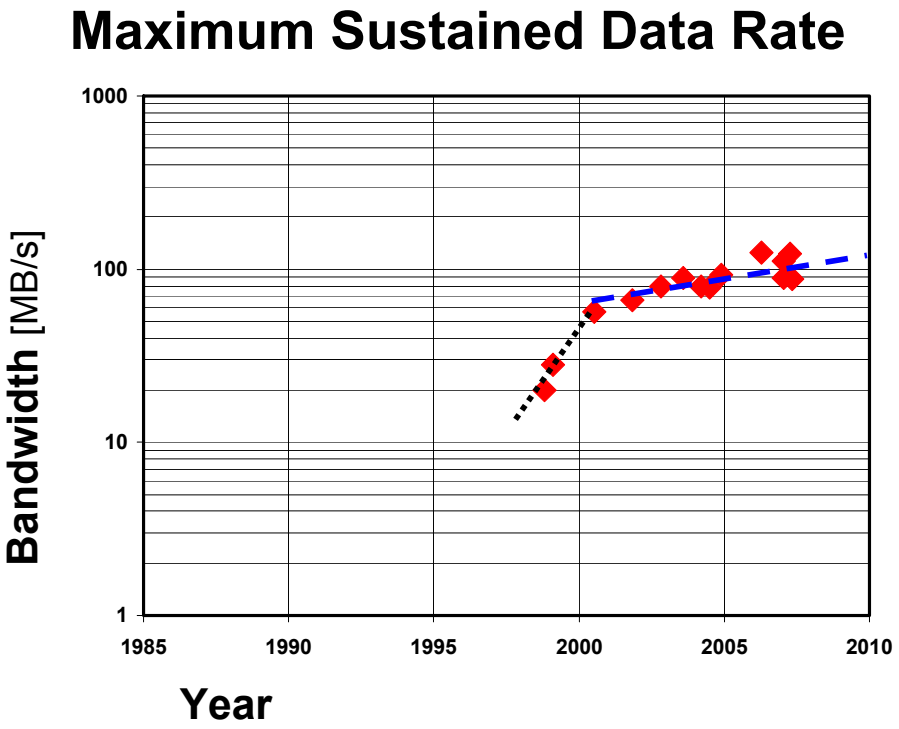


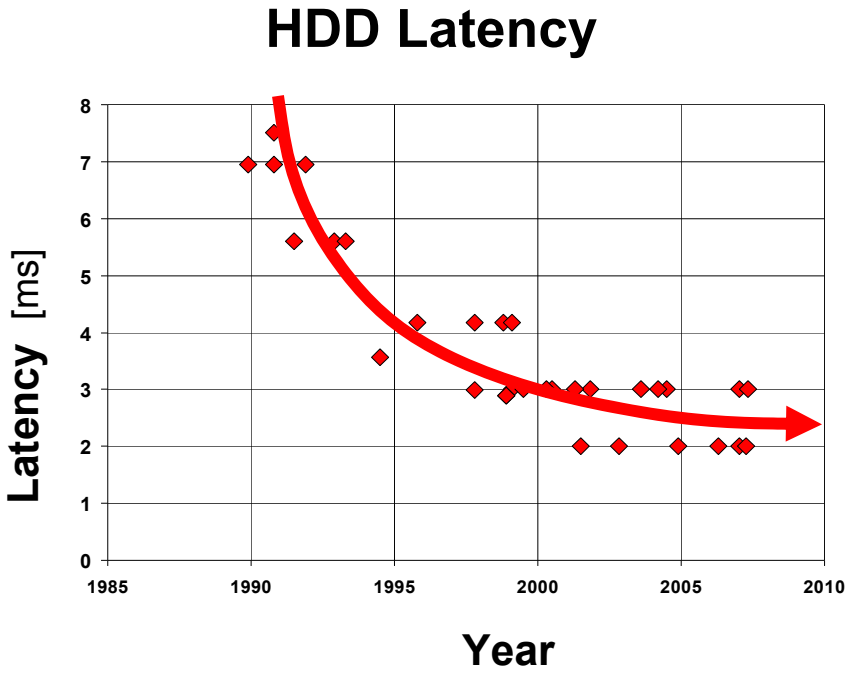
At least 4 orders of magnitude difference between improvements in speed of CPU and disk transfer rates

- Disk density has grown at same pace as processor speed
- Disk data transfer rates have grown modestly
- Disk access times have hardly improved

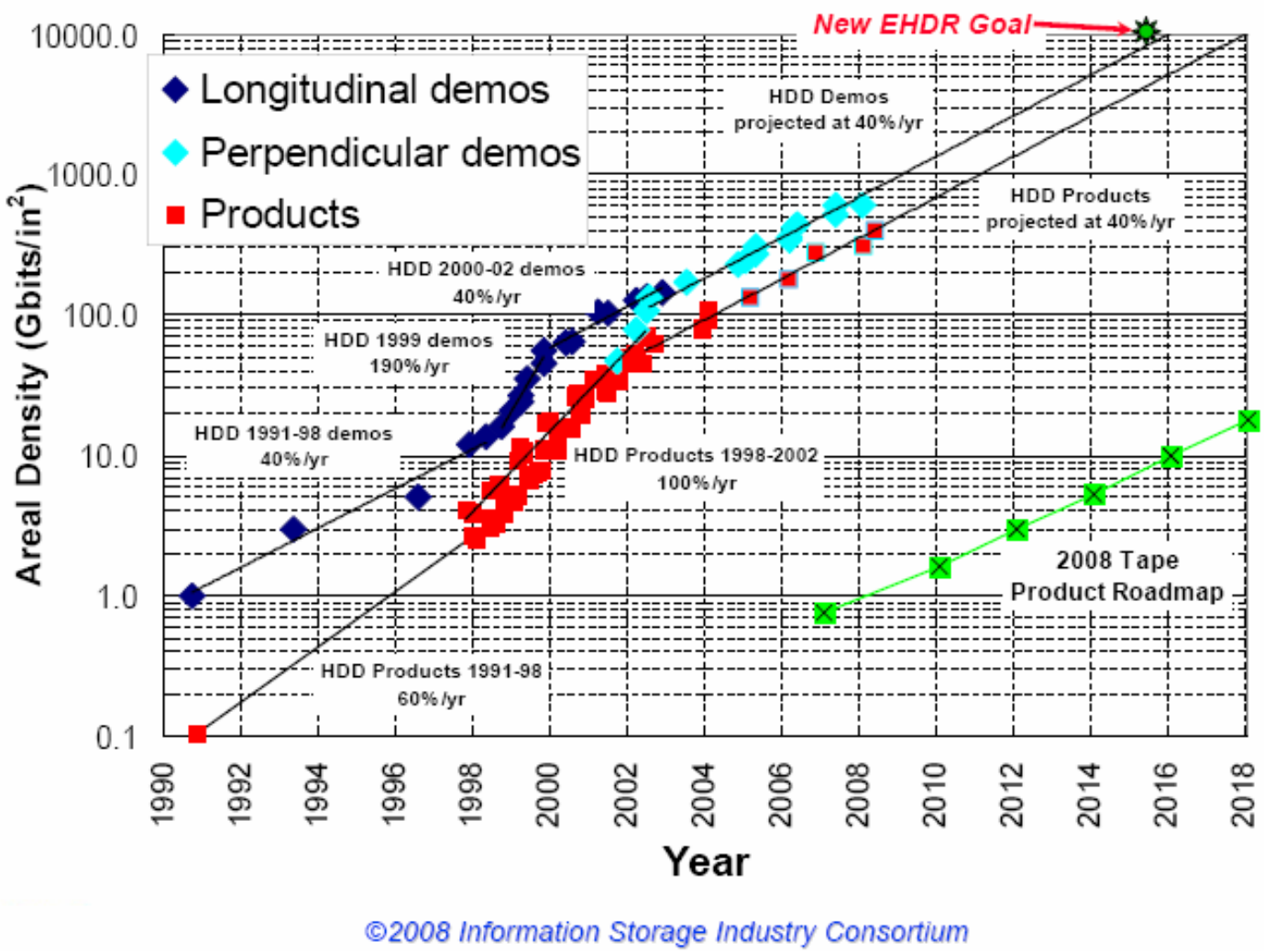
Source: High End Computing Revitalization Task Force (HEC-RFT), Inter Agency Working Group (HEC-IWG) File Systems and I/O Research Workshop

HDD Performance – Little progress…





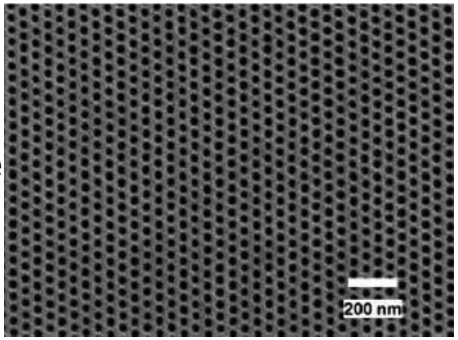
Areal Density Trends



What is Next in HDD Technology?

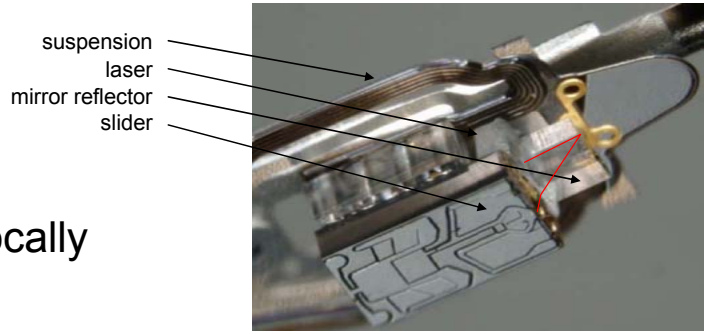
- In 2012-2013 HDD requires 25 nm features in the head for 1 Tb/in² densities
- After 2013 HDD must exceed the IC lithography roadmap to support areal densities > 1 Tb/in²

300 Gbit/in²,
25 nm islands on 50 nm pitches



- **Patterned media**
  - Relies on unproven nano-imprint technology
  - Embodies significant chalanges in “locating” the bit on disk surface
  - Requires major capital investment in factories ($10 B to $20 B)
  - Sustainability/extendibility of roadmap not well projected

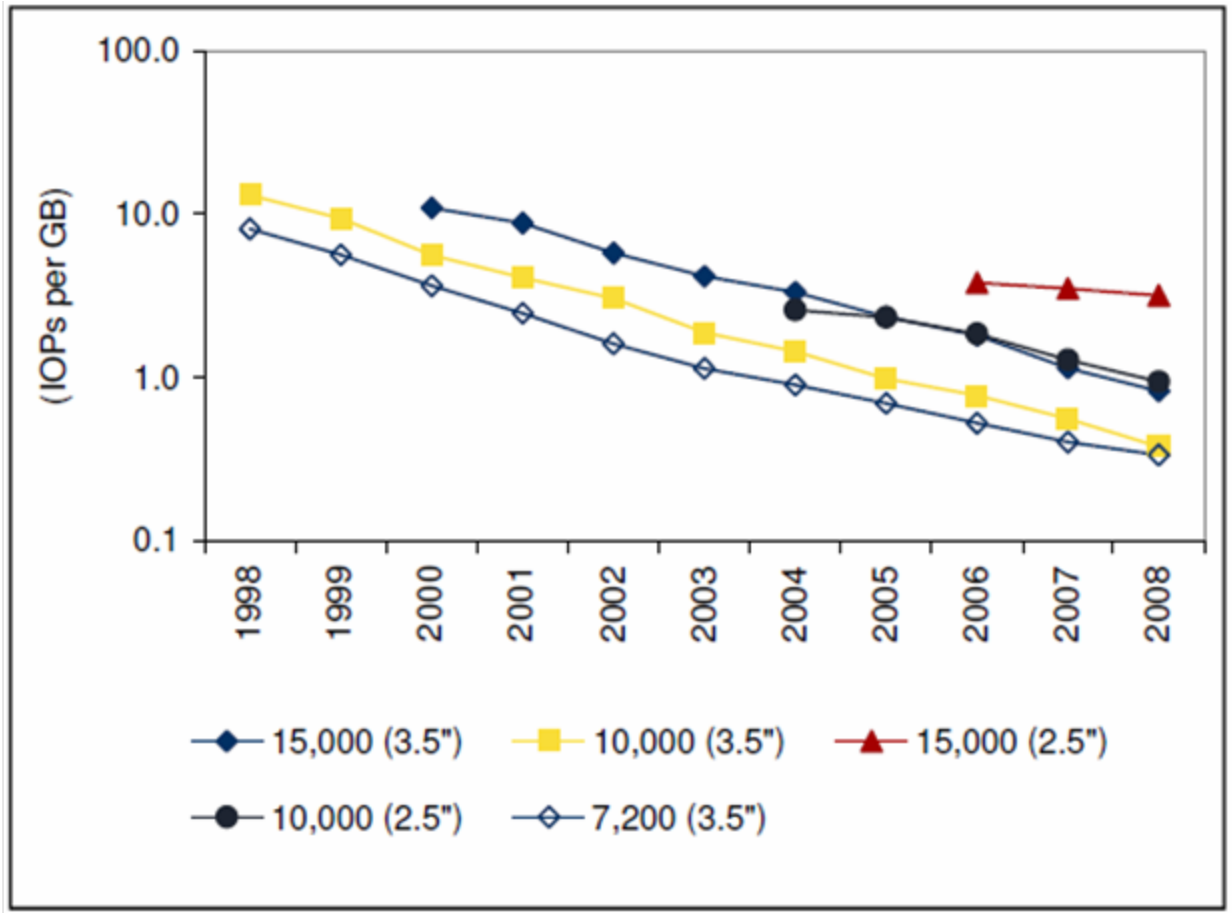
- **Heat assisted magnetic recording (HAMR)**
  - Use high H_k material to stabilize magnetic transition in smaller bit cells
  - Write on media using energy to heat and reduce H_k locally
  - Head complexities and serious cost issues



Unlikely that HDD will be leading the IC industry in lithography !

Source: Gary Decad & Bob Fontana, IBM

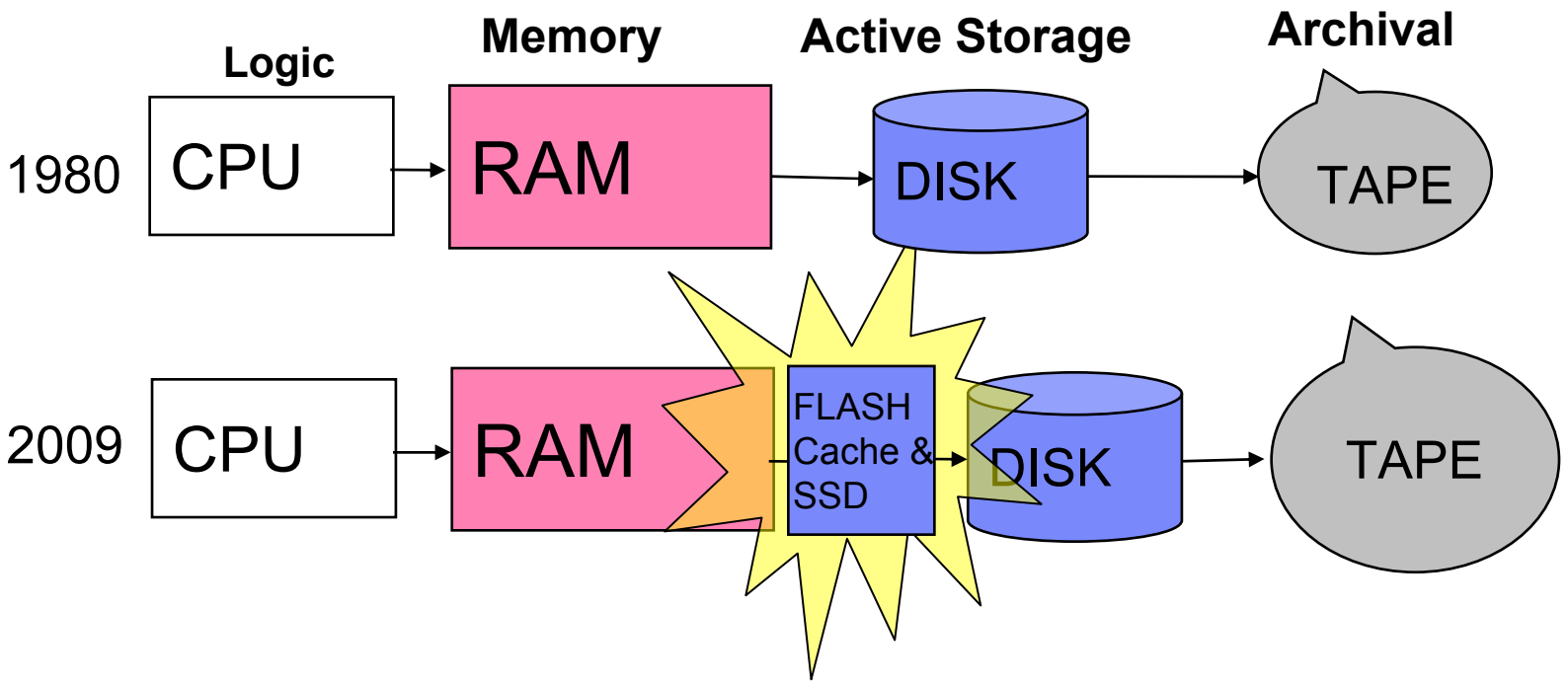
HDD IOPS per GB



More spindles are needed, but not the capacity!

Source: Enterprise Strategy Group

Evolution of Memory/Storage Stack



Flash shakes out the memory/storage hierarchy

Flash – a Disruptive Technology

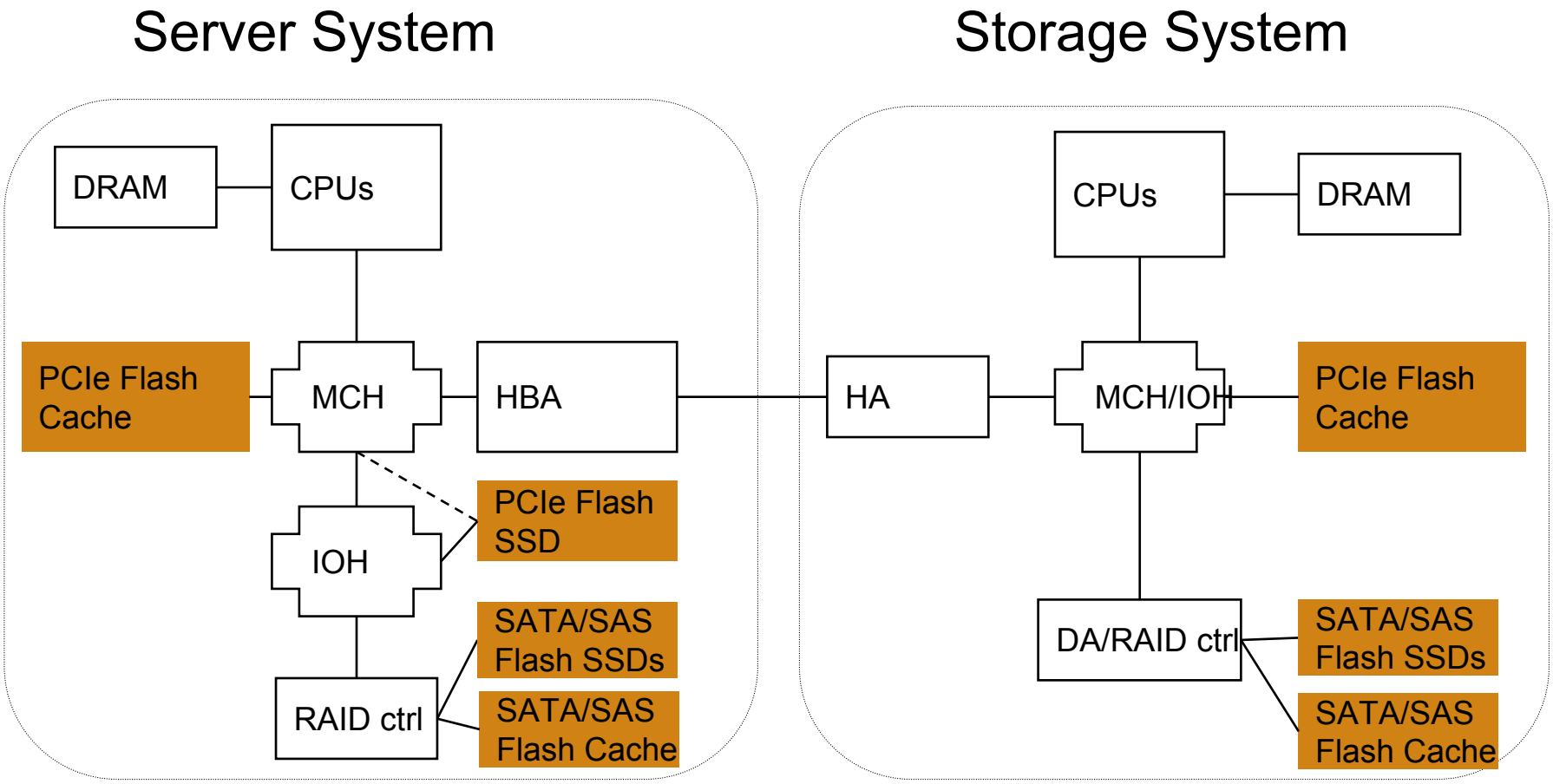
- Solving the I/O bottleneck will have profound system implications
- Order of magnitude better I/O performance
  - Desktop Flash : 25,000 read IOPS, up to 8000 write IOPS (Intel, Samsung, etc)
  - Enterprise Flash : 120,000 read IOPS, 50,000 write IOPS (TMS, STEC, etc)
- Energy efficiency and form factor improvement make it more attractive
  - Two orders of magnitude better energy efficiency and form factor

	DRAM	NAND Flash	Enterprise Disk
Package	chip	chip	2 ½" - 3 ½" disk
Read access time (us)	.05	25	5000
Device Capacity (GB)	.5	64-512	500 - 2000
Device BW (MB/s)	>400	>100	100-150
Endurance	10^{15}	10^{4} - 10^{6}	10^{12}
Life time (years)	10	10	5
Device Power (W)	.2	<.2	20
Power On/Of time	<100us	<100us	15-30s
Vibration /shock	>15G	>15G	<1G

SSD is 100 – 200 times more efficient in IOPS/Watt !

SSD is 30 – 50 times more efficient in IOPS/$ per GB !

Flash in Enterprise Storage and Servers



Flash – a Disruptive Technology

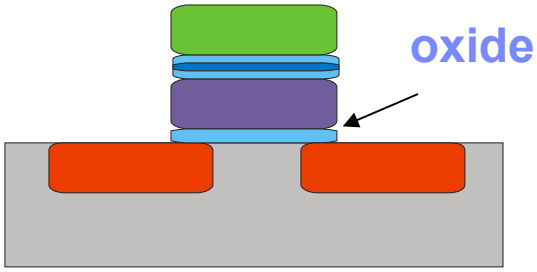
But …..

- Can endurance be substantially improved, i.e., can the maximum number of program/erase cycles be substantially increased, to meet the enterprise-class storage requirements?
- Can low-cost, less reliable MLC penetrate the enterprise-class storage segment?
- Would the scaling to 3x and 2x nm lithography nodes affect endurance and latency?

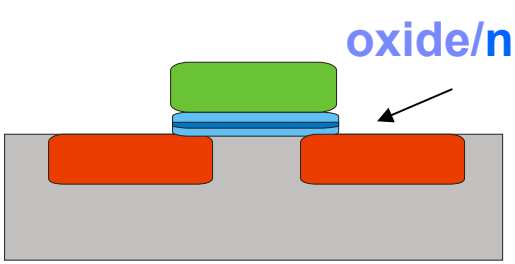
Innovative techniques to improve endurance – comparable to HDD ?

Can Flash Continue Scaling?

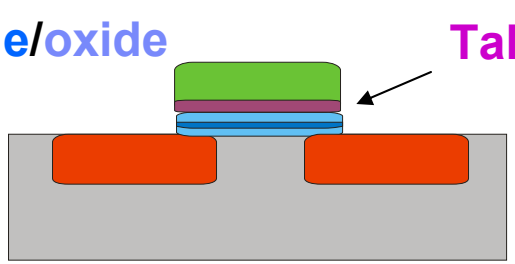
Litho. Node: **40nm** → **30nm** → **20nm**



Floating Gate
<40nm ???



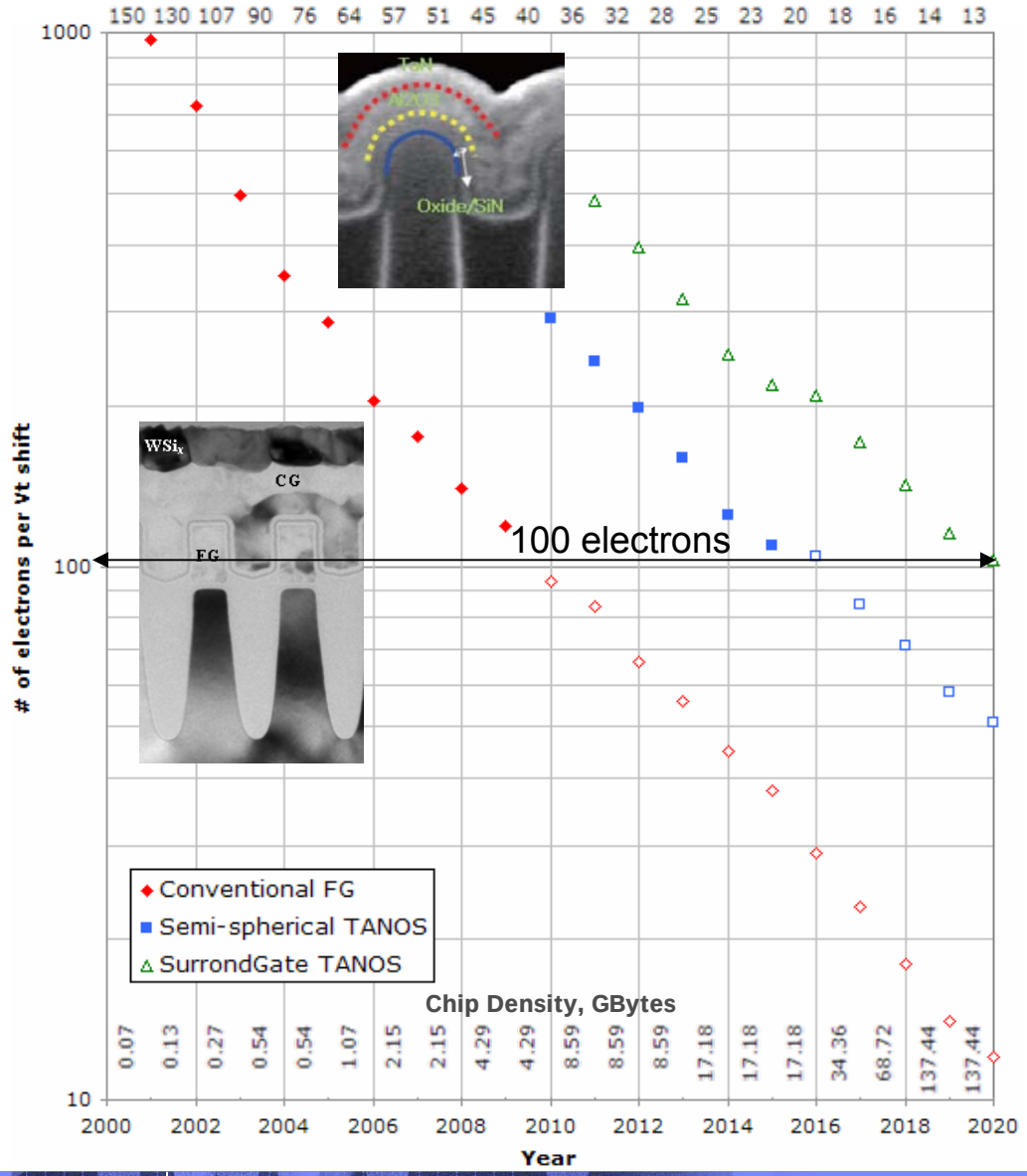
SONOS
Charge trapping
in SiN trap layer



TaNOS
Charge trapping
in novel trap layer
coupled with
a metal-gate (TaN)

Challenge will be to maintain same performance, write endurance, and retention specs

Flash-memory Scaling Challenges



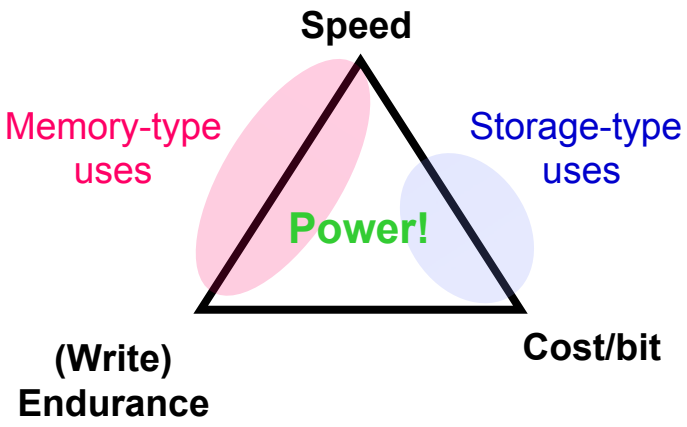
FLASH Challenges

- not lithography itself, but…
- not enough electrons
  - ***Only ~100 e- per level today***
  - Can't even loose one electron per month!
- Electric field stress on gate during programming too high
  - Write endurance drops
    - as number of levels increases
    - as device geometry gets smaller

Source: Chung Lam, IBM

Storage Class Memory (SCM)

A solid-state memory that blurs the boundaries between storage and memory by being low-cost, fast, and non-volatile.

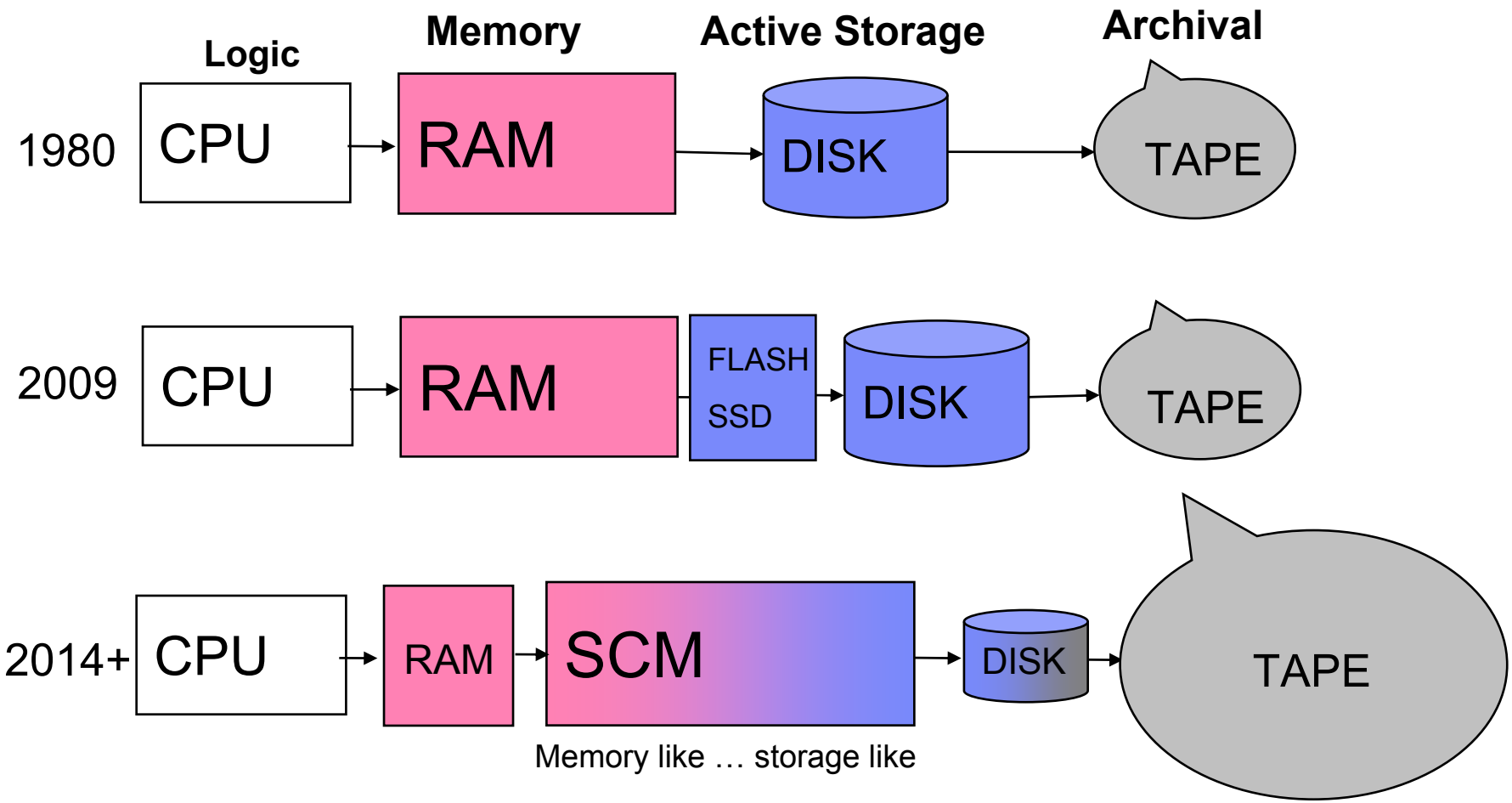


- SCM is not a technology, but rather a new class of data storage/memory devices that fulfill the following criteria:
  - SCM system requirements for Memory (Storage) applications
    - Solid state, no moving parts
    - < 200nsec (<1 μsec) Read/Write/Erase time
    - High write endurance, of 10^8 – 10^{12} write/erase cycles
    - 10x lower power than enterprise HDD
    - No more than 3-5x the Cost of enterprise HDD (< $1 per GB in 2012)
    - High retention time (2 – 10 years)

SCM Impact on Workloads

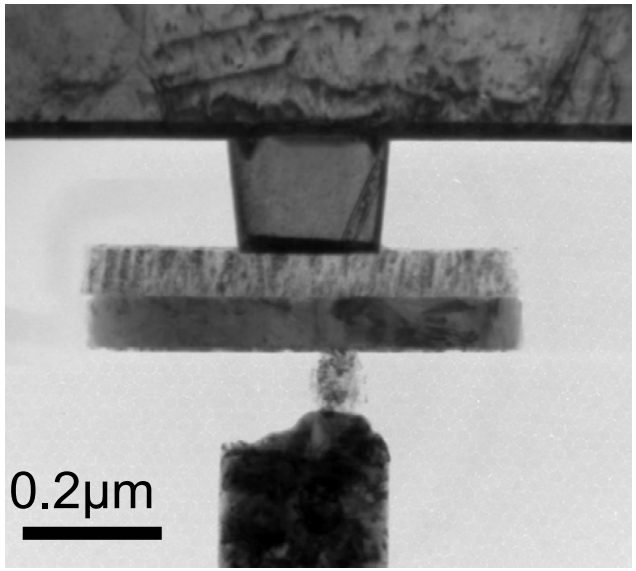
- The benefits of larger (cheaper) memory and faster storage cut across all workloads
- Benefit of SCM-based storage
  - Simplify substantially current enterprise SW used to hide I/O latencies (e.g., databases)
  - Faster storage will alleviate the major problem of shortening time windows available for batch processes
  - High IOPS will facilitate the growing needs for streaming analytics
  - Larger data caches would accelerate response to transactional workloads
- Benefit of SCM-based memory
  - Memory cost and power pressures have severely degraded memory-capacity to CPU performance ratios by a factor of 10
  - Accepted memory-capacity to CPU performance ratios:
    - Commercial: 8 Bytes (RAM) / IPS (IPS = Instruction per second)
    - Sci/Eng: 1 Byte (RAM) / FLOPS (FLOPS = Flt. Point operation / second)
  - SCM can potentially reverse this trend (of decreasing ratios)
  - SCM persistence could enable new applications that leverage byte-addressable, ultra-fast “storage”

Evolution of Memory/Storage Stack

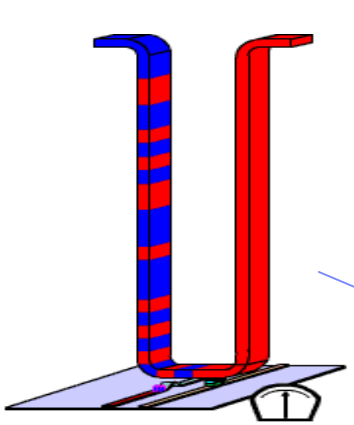


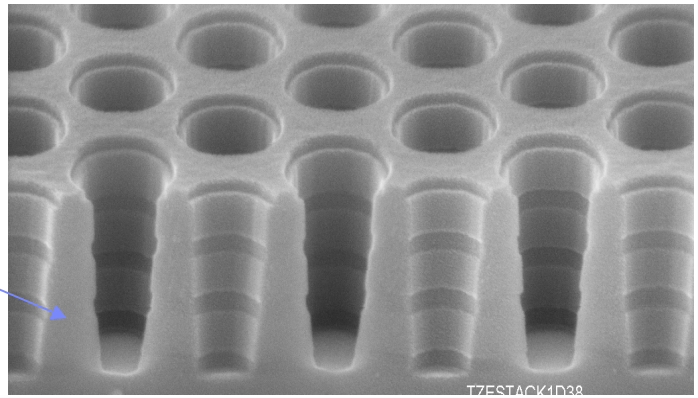
SCM Candidate Device Technologies

- MRAM (Magnetic RAM)
- FeRAM (Ferroelectric RAM)
- Magnetic Racetrack memory
- Organic & polymer memory
- RRAM (Resistive RAM)

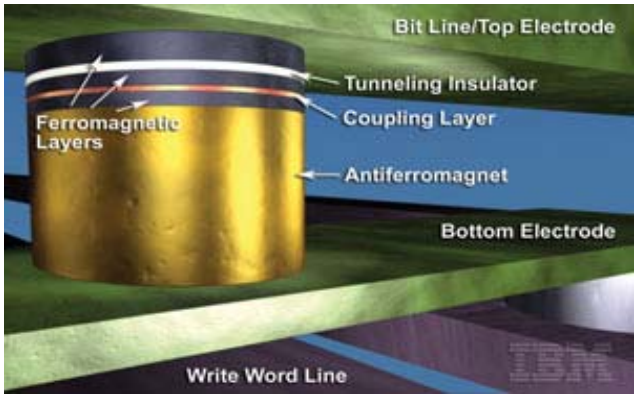


TEM image of phase-change element





Racetrack: Holds long-term promise of very high densities



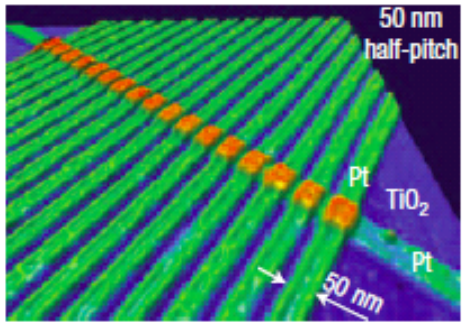
MRAM (In use today)

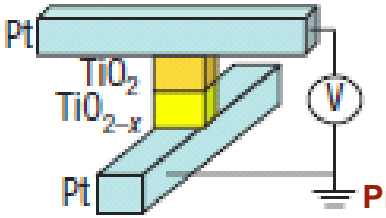
Resistive Memories

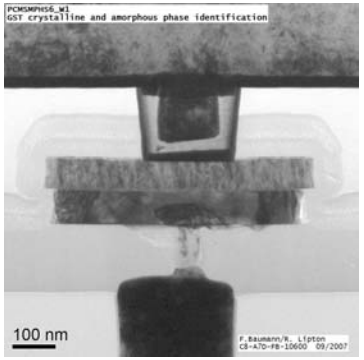
- Key advantages:
  - Non-charge based storage
  - Good potential for scaling
  - Amenable to new computing paradigms

- Common types of resistive memory
  - Phase-Change Memory
    - E.g. $Ge_2Sb_2Te_5$ based memory devices
    - Unipolar switching
    - Mechanism: Joule heating induced rearrangement
  - Transition-metal-oxide-based memory
    - Eg. TiO_2 based "memristor"
    - Bipolar switching
    - Mechanism: Field induced drift of oxygen vacancies
  - Programmable-metallization-cell memory
    - Eg. Mobile metal ions embedded in an electrolyte glass matrix
    - Bipolar switching
    - Creation and annihilation of metallic bridges between two electrodes

TiO_2 based memristor, HP

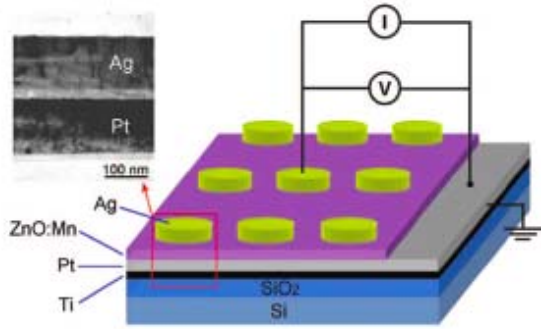




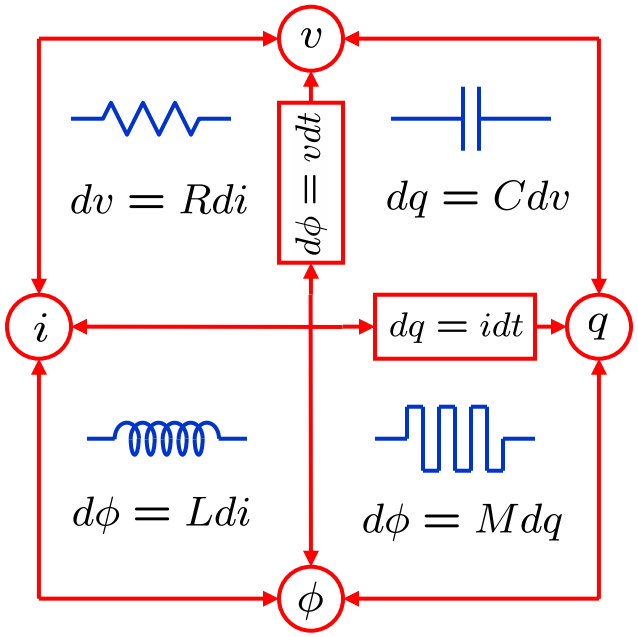


PCM Mushroom Cell, IBM

Programmable-metallization-cell



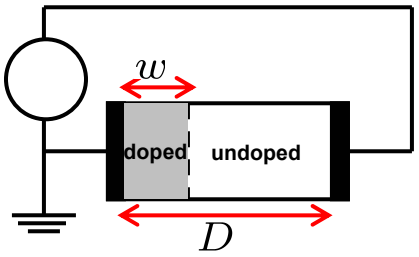
Memristor



$$d\phi = M(q)dq \Rightarrow v = M(q)i$$

A physical model of a memristive device

Strukov et al., Nature, 2008



$$v(t) = \left[R_{ON}\frac{w(t)}{D} + R_{OFF}\left(1 - \frac{w(t)}{D}\right)\right] i(t)$$

$$\frac{dw(t)}{dt} = \mu_v \frac{R_{ON}}{D} i(t)$$

$$v(t) = M(q)i(t)$$
$$M(q) = R_{OFF}\left(1 - \frac{\mu_v R_{ON}}{D^2} q(t)\right)$$

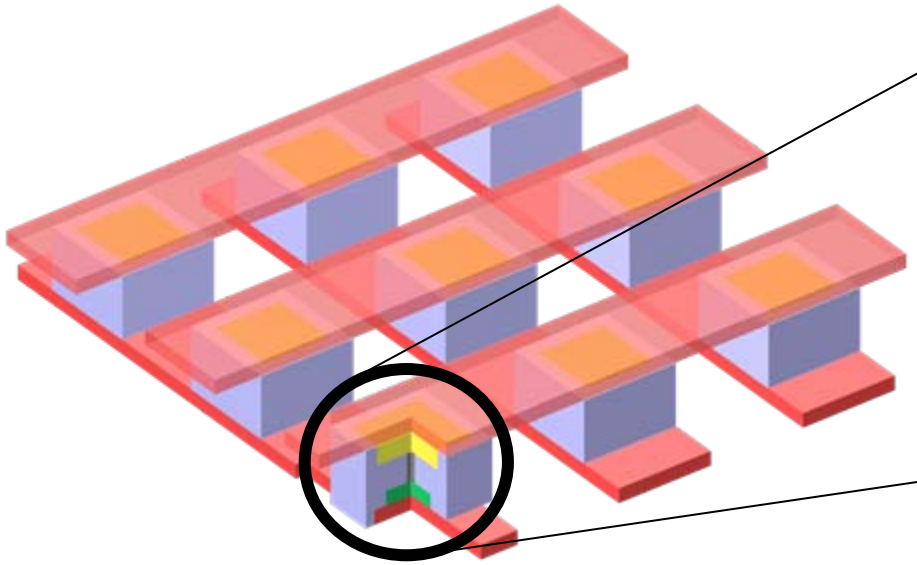
Phase-Change Memory (PCM)

- *Currently, leading contender for first true Storage Class Memory*
  - 0.5 Gigabit PCM chips being sampled now

- Use two distinct solid phases of a metal alloy to store a bit
- *Ge-Sb-Te* exists in a (quasi) stable amorphous and a stable crystalline phases
  - Phases have *very* different electrical resistances – ratio of 1:100 to 1:1000 (!)
  - They also have different optical reflectivity – used today in writable CD/DVD
- Transition between phases by controlled heating and cooling

Write '0' (RESET) :	short (10ns) intense current pulse melts alloy crystal => amorphous = high resistance
Write '1' (SET) :	longer (50ns) weaker current pulse re-crystalizes alloy => crystalline = low resistance
Read :	short weak pulse senses resistance, but doesn't change phase

Phase-Change Memory



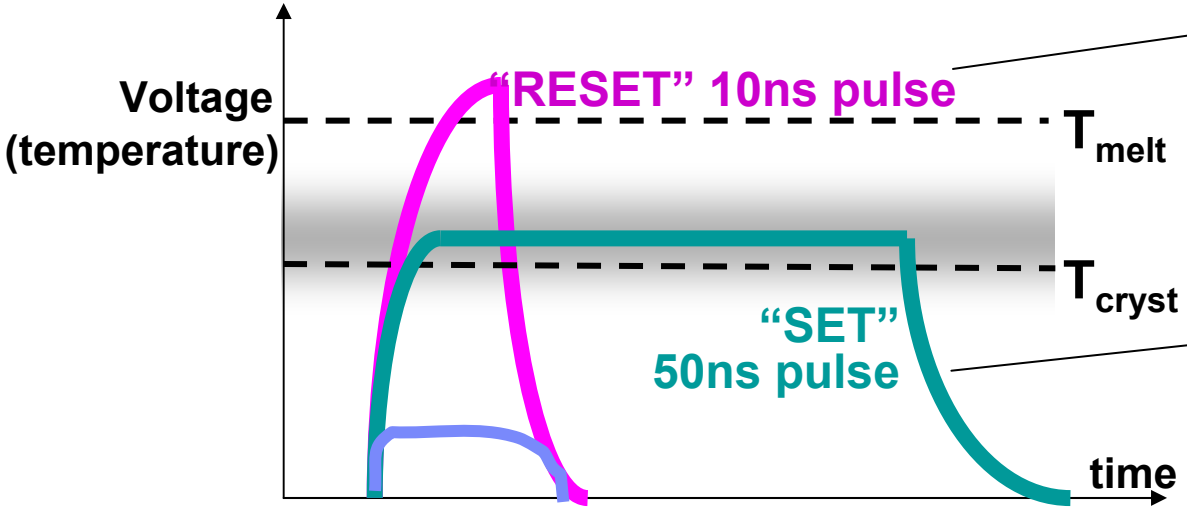
Bit-line

PCRAM
"programmable resistor"

Word -line

Access device
(transistor, diode)

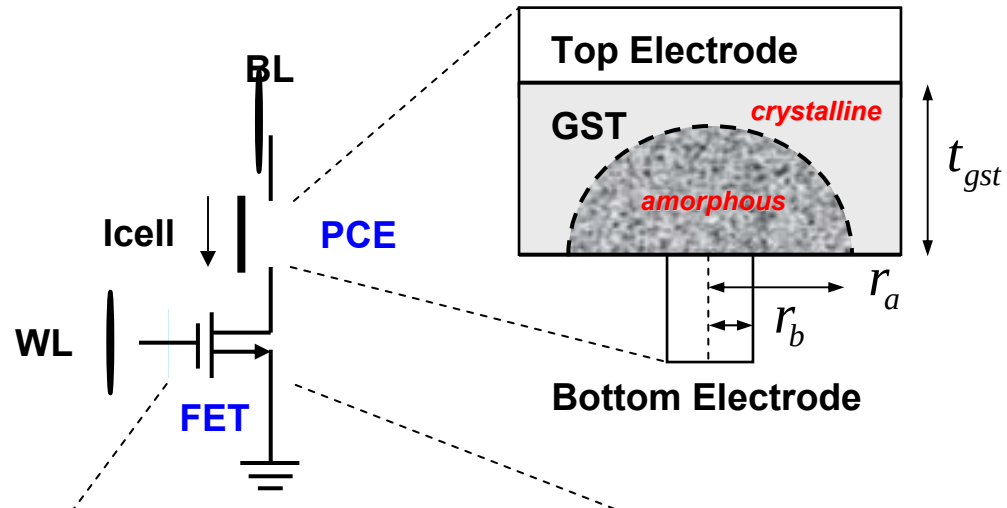
Currently determines Scaling limits ($4F^2$)

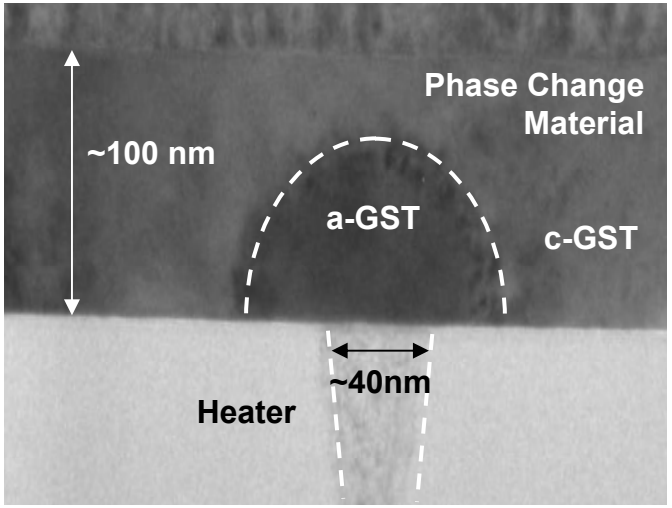


Potential challenge:
High current/power
→ affects scaling!

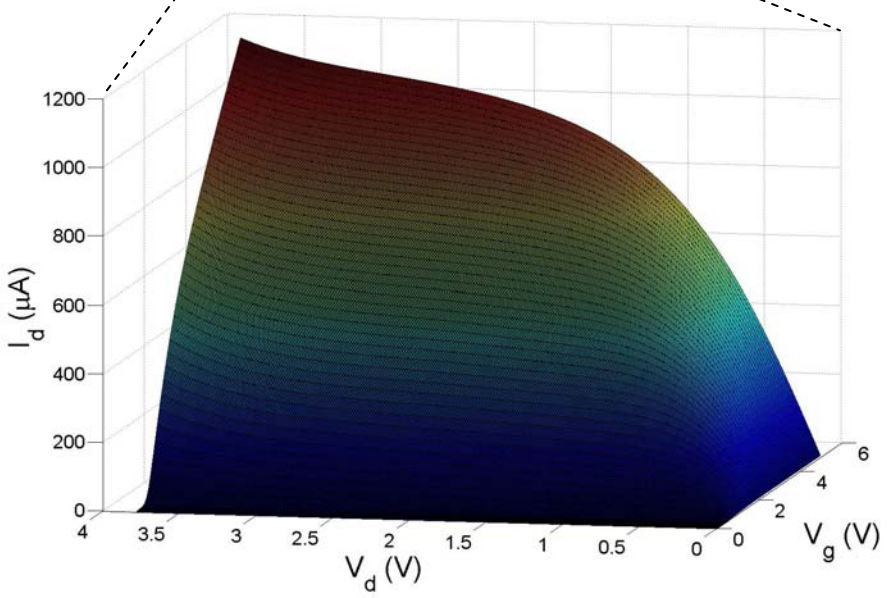
Potential challenge:
If crystallization is slow
→ affects performance!

PCM Cell





t_{gst} : GST thickness
r_a : radius of amorphous area
r_b : radius of bottom electrode



- Phase-change memory cell: "Mushroom type"
  - Amorphous material forms a "dome" around the heater contact
- Access device needed for selection of a particular cell in a cell array
- Access device may be an FET, BJT, or even diode

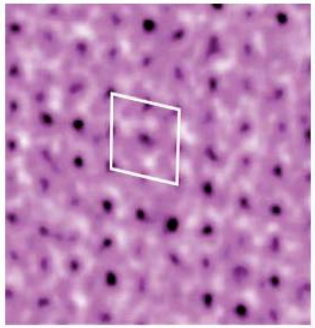
Demonstration of PCM Scalability

- Several approaches for PCM scalability demonstration
- Thin films, bridge-cell structure, nanodots
- Important implications for scaling of phase change memory devices

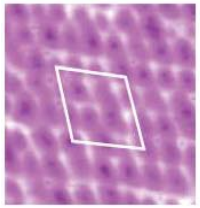
H. Hamann et. al. (2006)

Thermal recording of phase-change domains

1.6 Tbit/in²



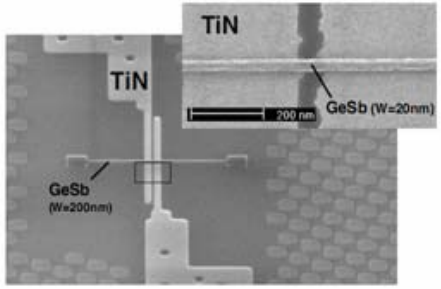
3.3 Tbit/in²



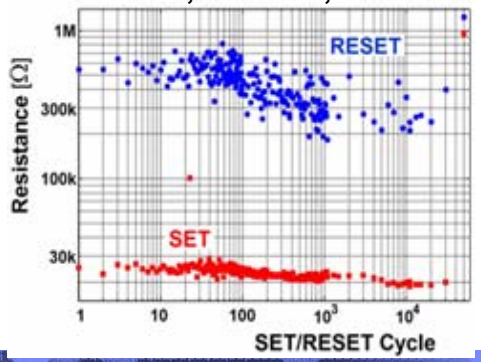
IBM/Macronix (2006/7)

Phase-change bridge cell

SEM picture of bridge cell



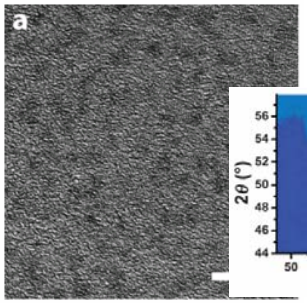
Cycling of small cell: H=3nm, W=20nm, L=50nm



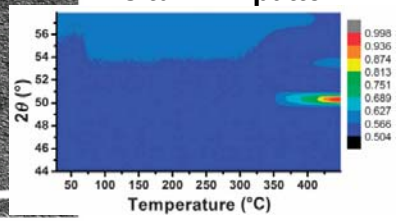
IBM ARC/YKT & Stanford (2007-9)

XRD studies of phase-change thin-films, nanodots, nanoparticles (1.8nm Ø)

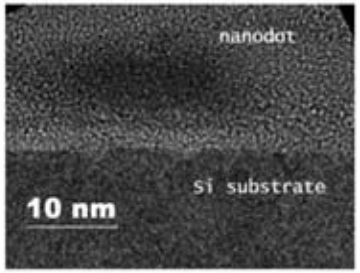
TEM picture of 1.8nm nanoparticles



In-situ XRD pattern



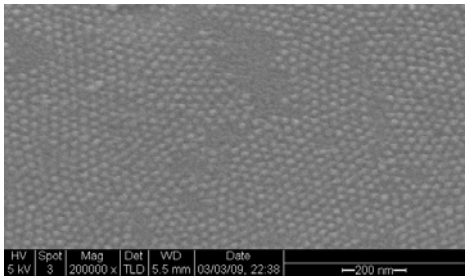
TEM image of single GeSb nanodot after XRD



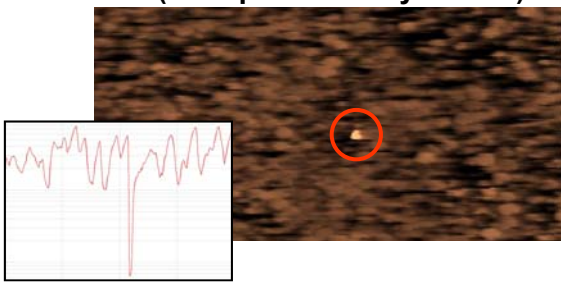
IBM-Zurich/Stanford (2009)

Joule-heating induced switching of single phase-change nanodots (~15nm Ø)

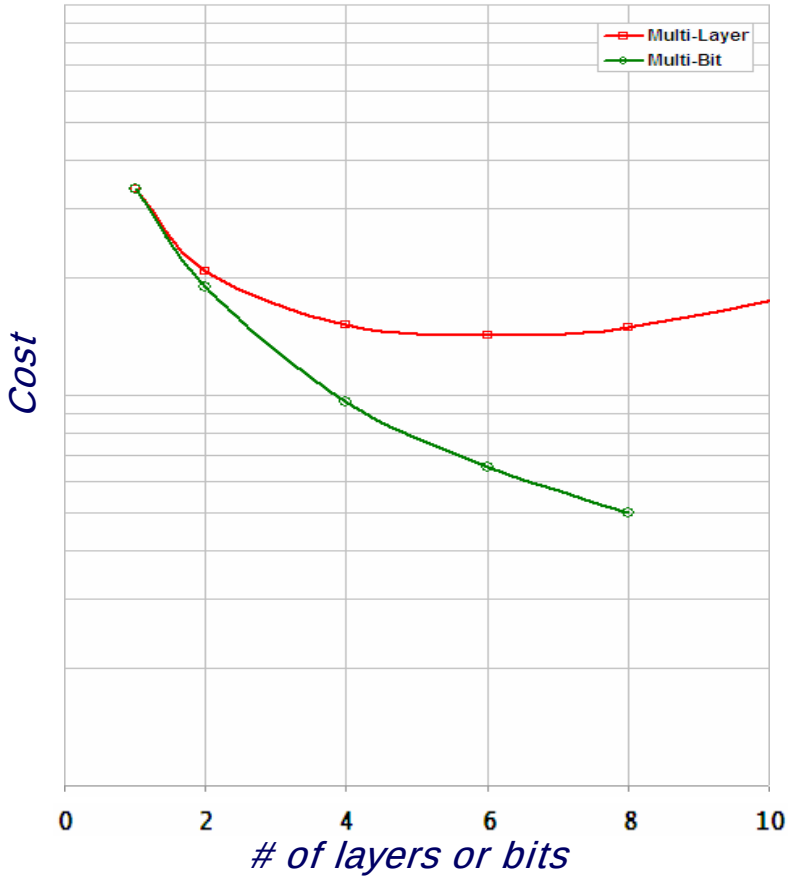
SEM picture of array of nanodots



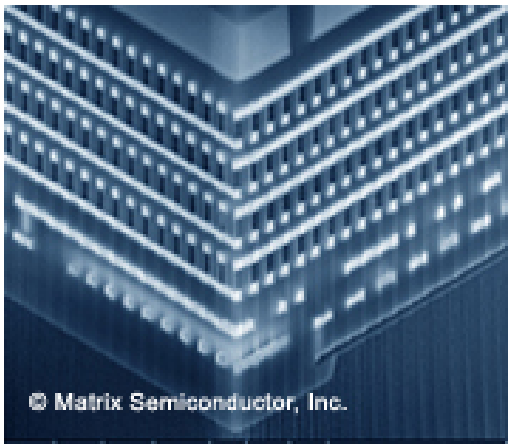
Switching of a single nanodot (amorphous → crystalline)

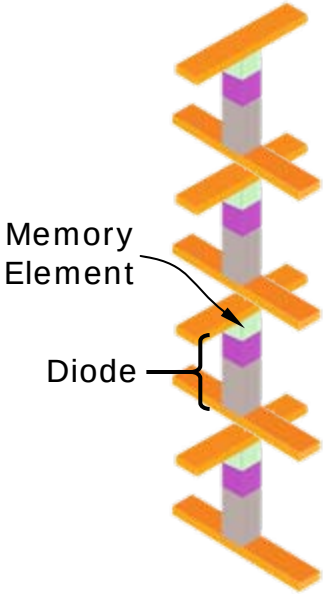


Multi-layer Integration & Multi-bit Operation

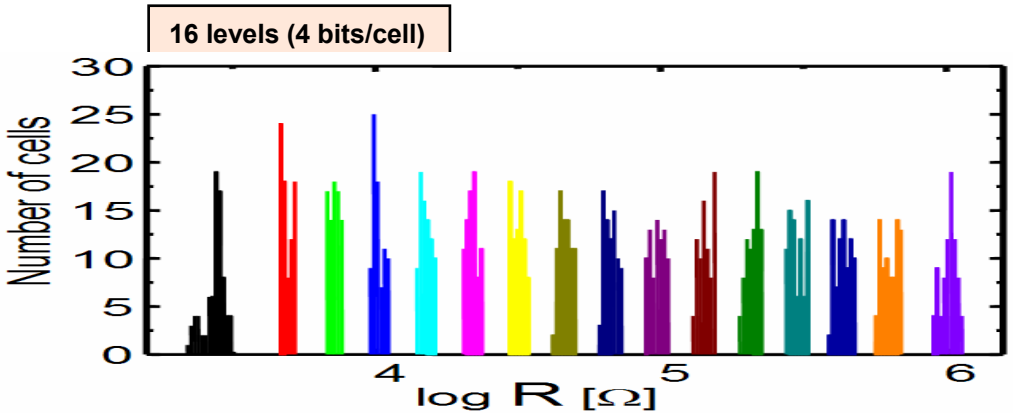


Back End Multi-Layer Integration





Iterative programming: Multi-Bit

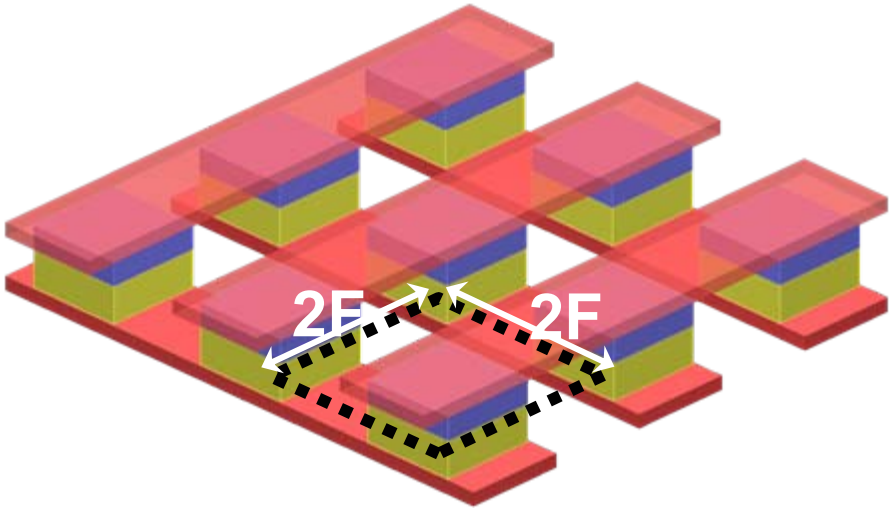


To compete against NAND Flash, PCM must be able to function as Multi-bit and/or Multi-layer technology

Towards Ultra-high Areal Density

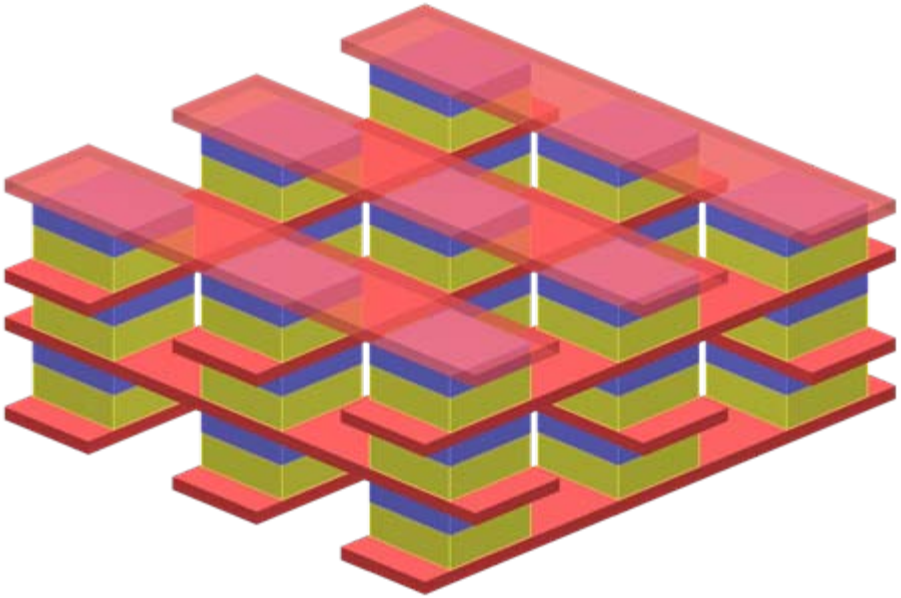
At the 32nm node

	Density	***Chip***
Base **($4F^2$)**	**24 Gb/cm²**	→ **16 GB**
2x	**48 Gb/cm²** (2 bits/cell)	→ **32 GB**



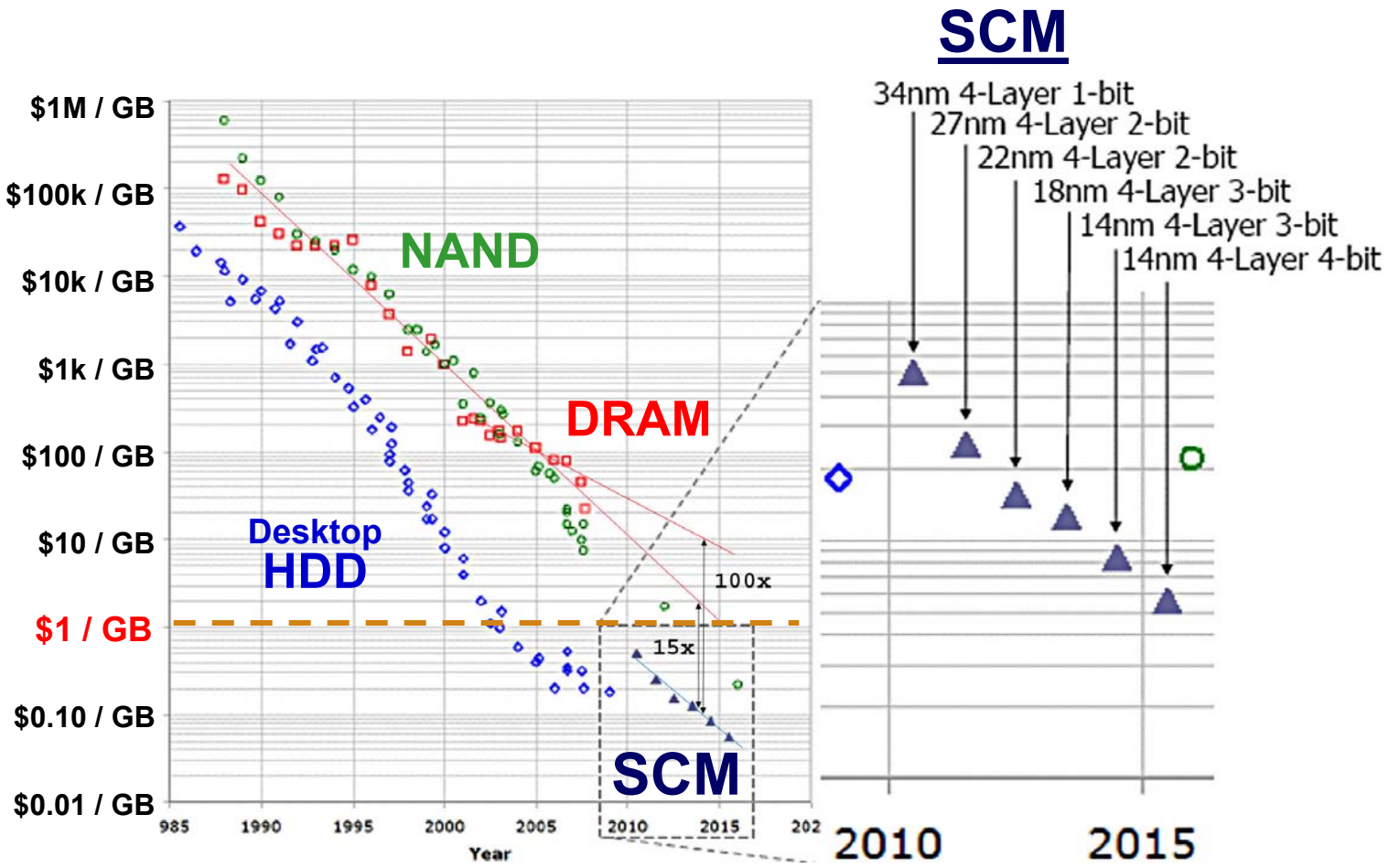
If we could shrink $4F^2$ by…

4x	**96 Gb/cm²** (2 layers of 3-D, 2 bit/cell)	→ **64 GB**
8x	**192 Gb/cm²** (4 layers of 3-D, 2 bits/cell)	→ **128 GB**
16x	**384Gb/cm²** (4 layers of 3-D, 4 bits/cell)	→ **0.25 TB**



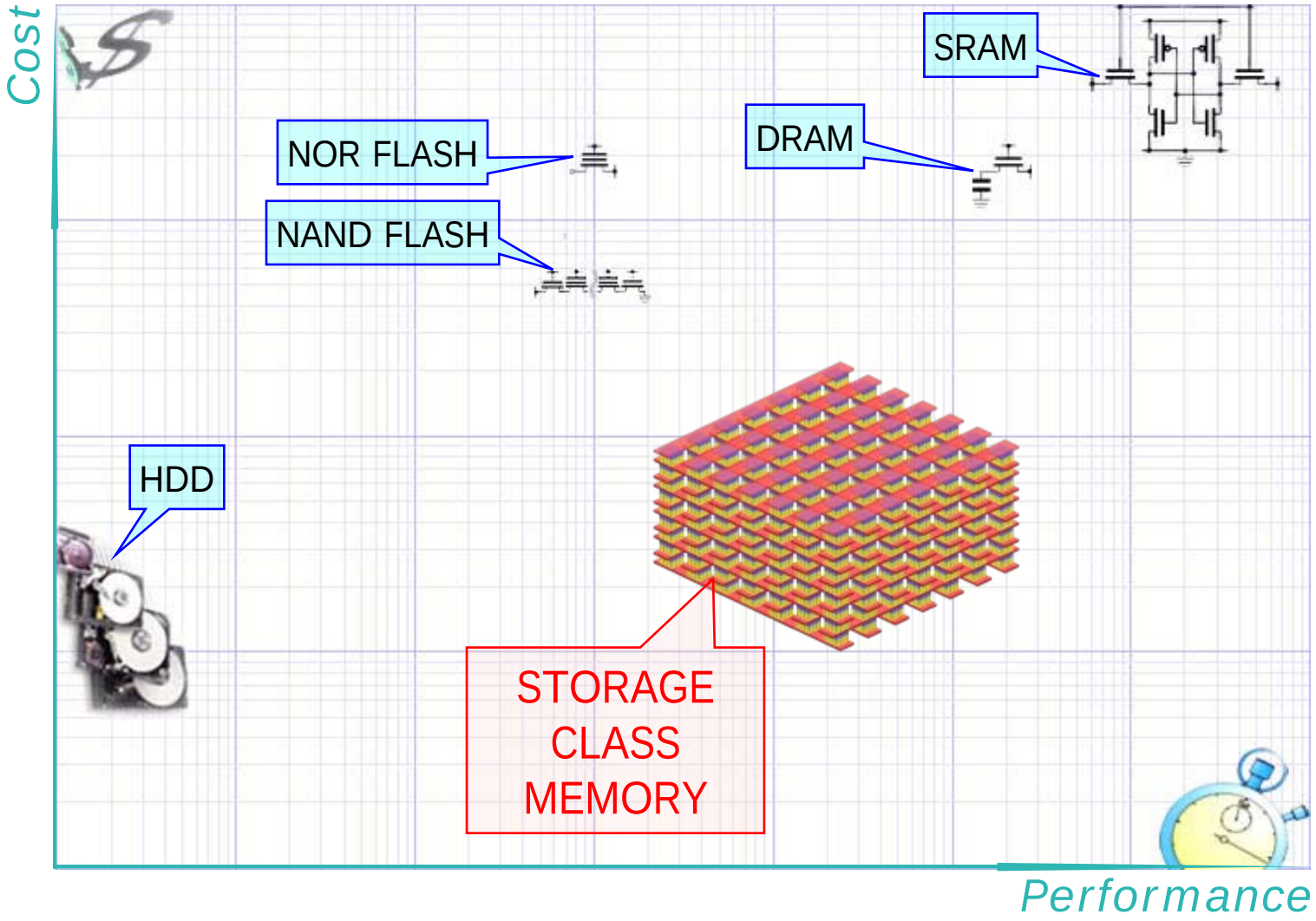
1 TB is a real challenge !

SCM: Why would you need anything else?

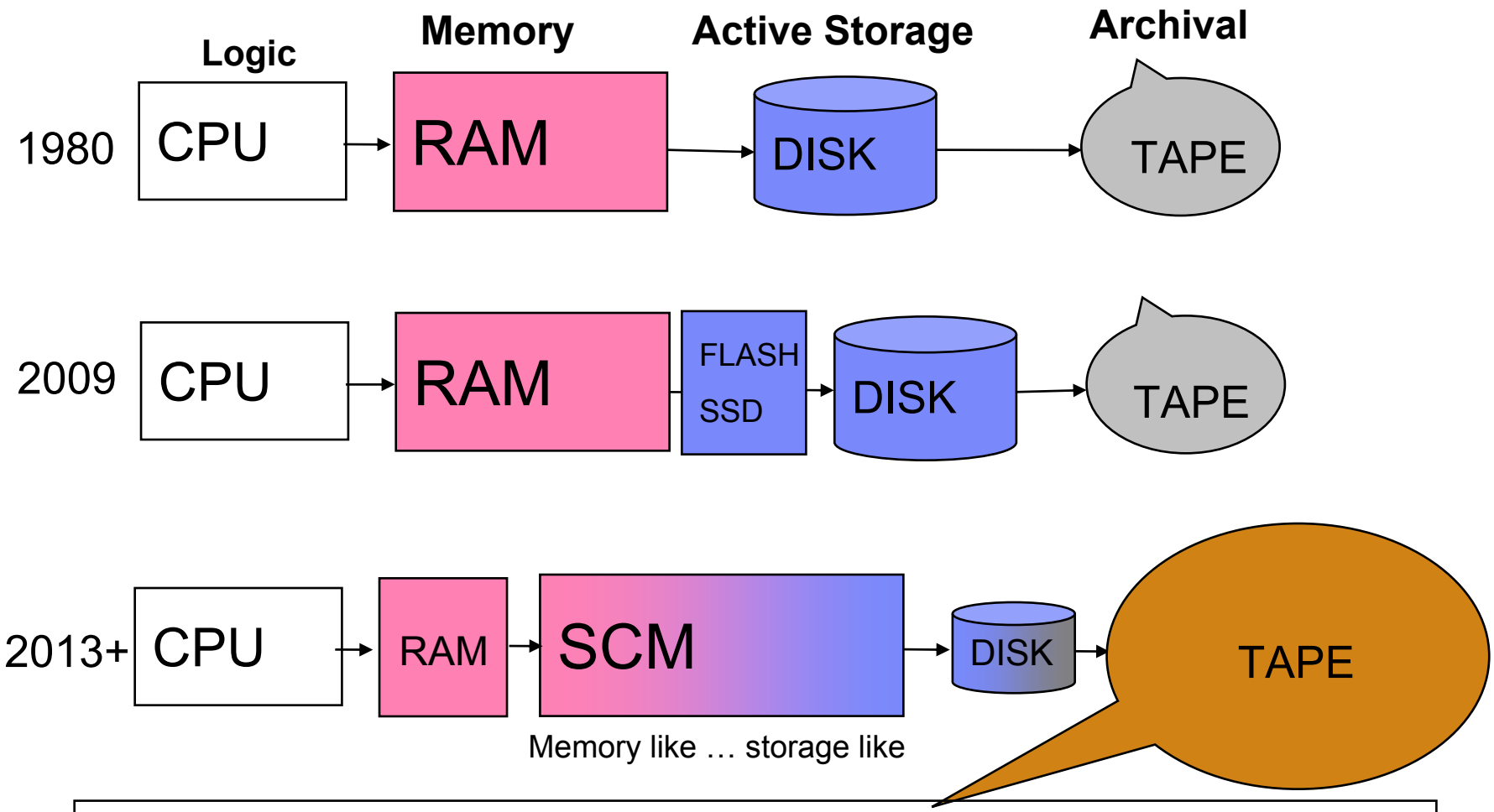


Source: Chung Lam, IBM

SCM vs. Existing Technologies

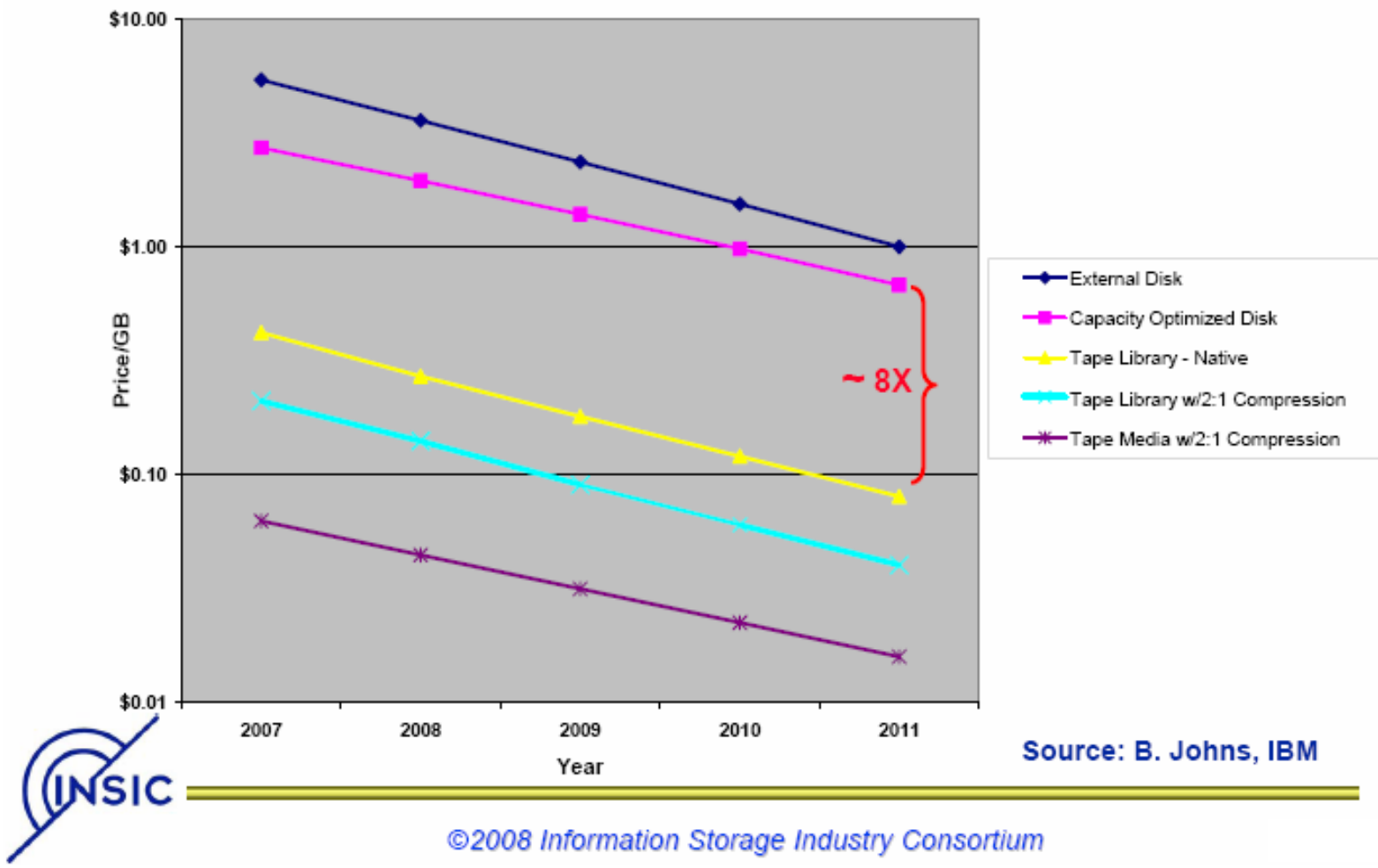


Evolution of Memory/Storage Stack

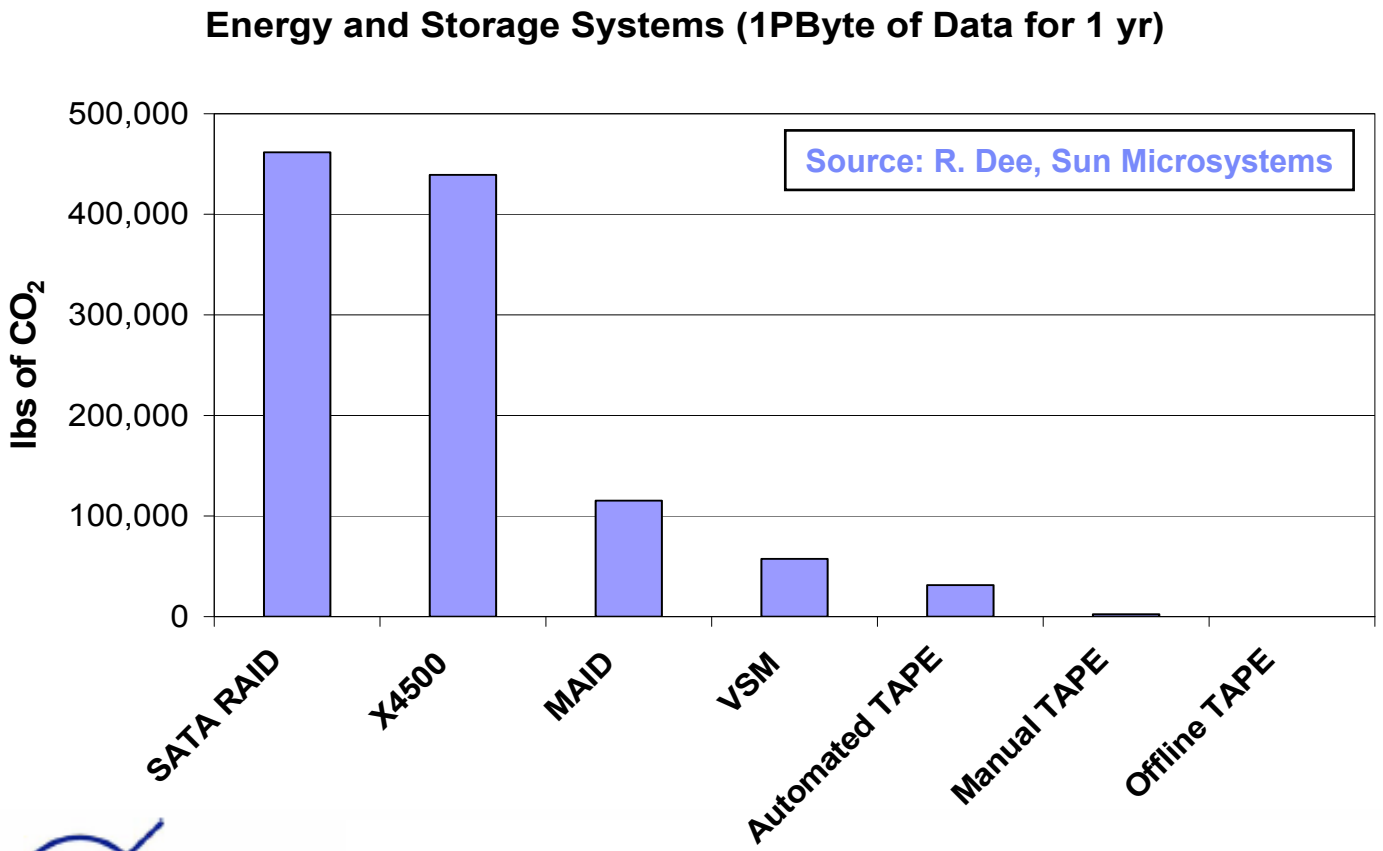


"The report of my death was an exaggeration", Mark Twain, May 1897

Price Comparison of Disk and Tape Storage

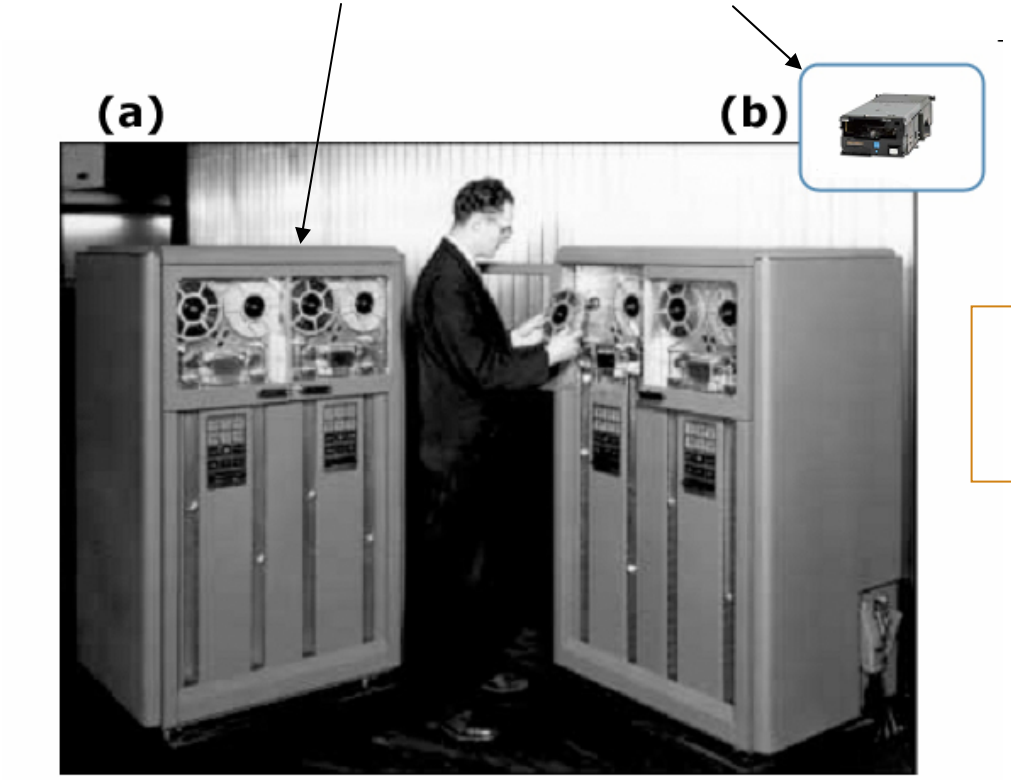


Tape is the "Greenest" Storage Technology



Magnetic Tape (R)evolution

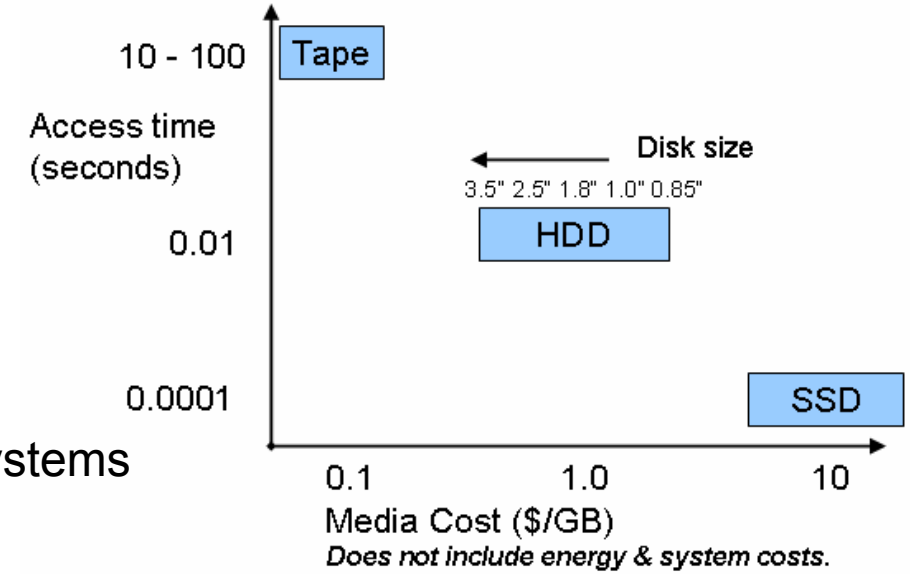
Product / Year:	IBM 726 /1952	JAG3 / 2008	LTO6 / 2012
Capacity:	2.3MByte	1TByte	3TByte
Areal Density:	1400 bit/in^2	790Mbit/in^2	1.87Gbit/in^2
Linear Density:	100 bit/in	343 kbit/in	488 kbit/in
Track Density:	14 tracks/in	2.3 ktracks/in	3.84 ktracks/inn



Track density increase will be the key contributor for future tape capacity increase

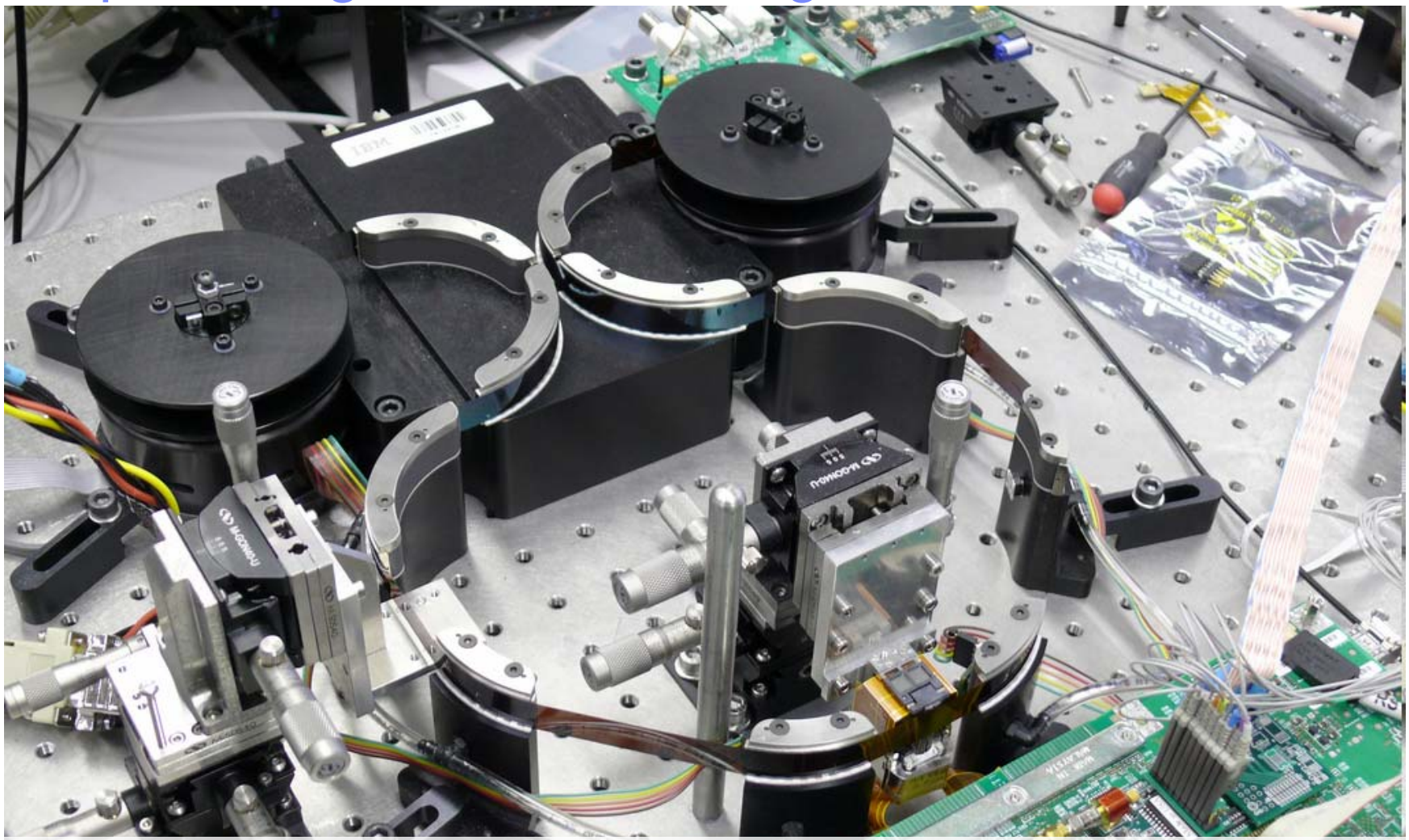
Tape Storage: Threats and Growth Opportunities

- Storage devices with removable media
  - Slowest access time, typically nearline
  - Greenest technology
  - Portable, interchangeable, archivable
  - "Infinite capacity", volumetric efficiency
  - Data Compression build into drives
  - WORM cartridges available
  - Very strong encryption @ line speed
- The *main threat* to tape in multi-user IT applications *is low-cost HDD* storage systems
  - Disk provides improved functionality
    - Data deduplication (effective increase in capacity & data rate)
    - Continuous data protection
- Optical technologies pose less of a threat (Holographic storage out of race)
- Big growth opportunity for tape in archival applications
  - New file system: Linear Tape File System (LTFS)
  - Opportunity to offer complete system level archive solutions
  - Key factors are the low energy cost and volumetric efficiency



Source: M. Sharrock, Imationm, August 2008

Tape Storage: Demonstrating 29.5 Gb/in^2



This demonstration shows that tape can sustain the roadmap for at least another decade while maintaining a cost advantage over other storage technologies.

29.5 Gbit/in2 Demo: BaFe Media Technology

- Media features:
  - BaFe particulate media technology
    - uses low cost coating technology
  - Reduced particle volume
  - reduced media noise
    - improved SNR
  - Smoother media
    - reduced magnetic spacing

- Achieved linear/track densities:

  1-sigma PES = 23.4 nm, CDF=87nm

  Linear density = 518 kbpi w/ 0.2 um reader

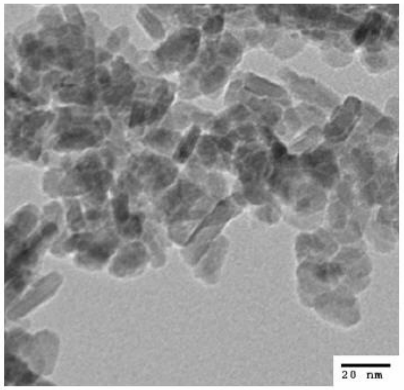
  Track density = 57 ktpi (track width = 0.446 um)

Areal density of 29.5 Gb/in^2

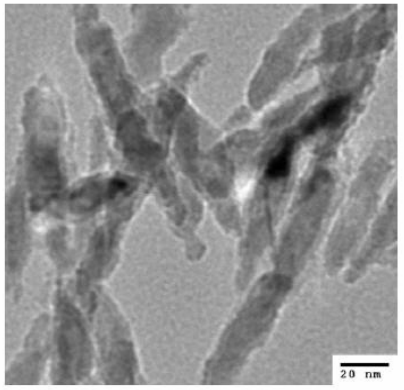
➔

35TB cartridge capacity

* TEM: Transmission Electron Microscope



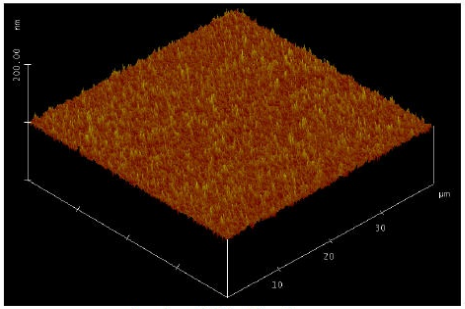
Latest BaFe particle
Volume: 1600nm^3



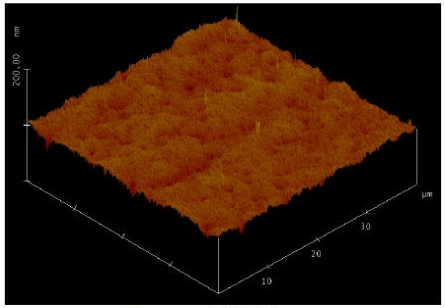
Current metal particle
Volume: 4650nm^3

- FUJIFILM succeeded in the microparticulation of BaFe particles to 1600nm^3 which is approximately one-third of current metal particle volume.

* AFM: Atomic Force Microscope



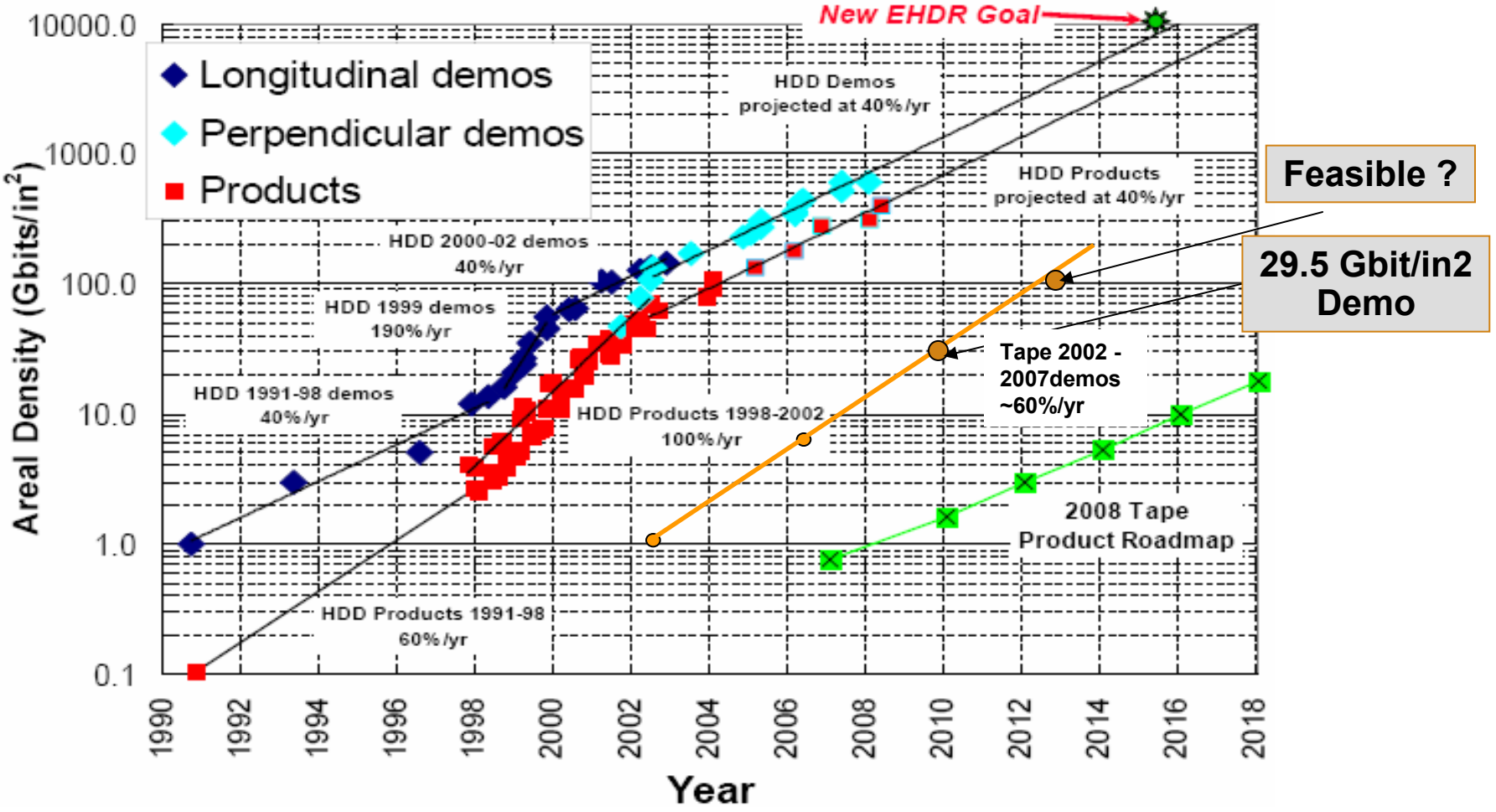
Latest BaFe tape
Ra: 2.1nm



Current MP tape
Ra: 3.6nm

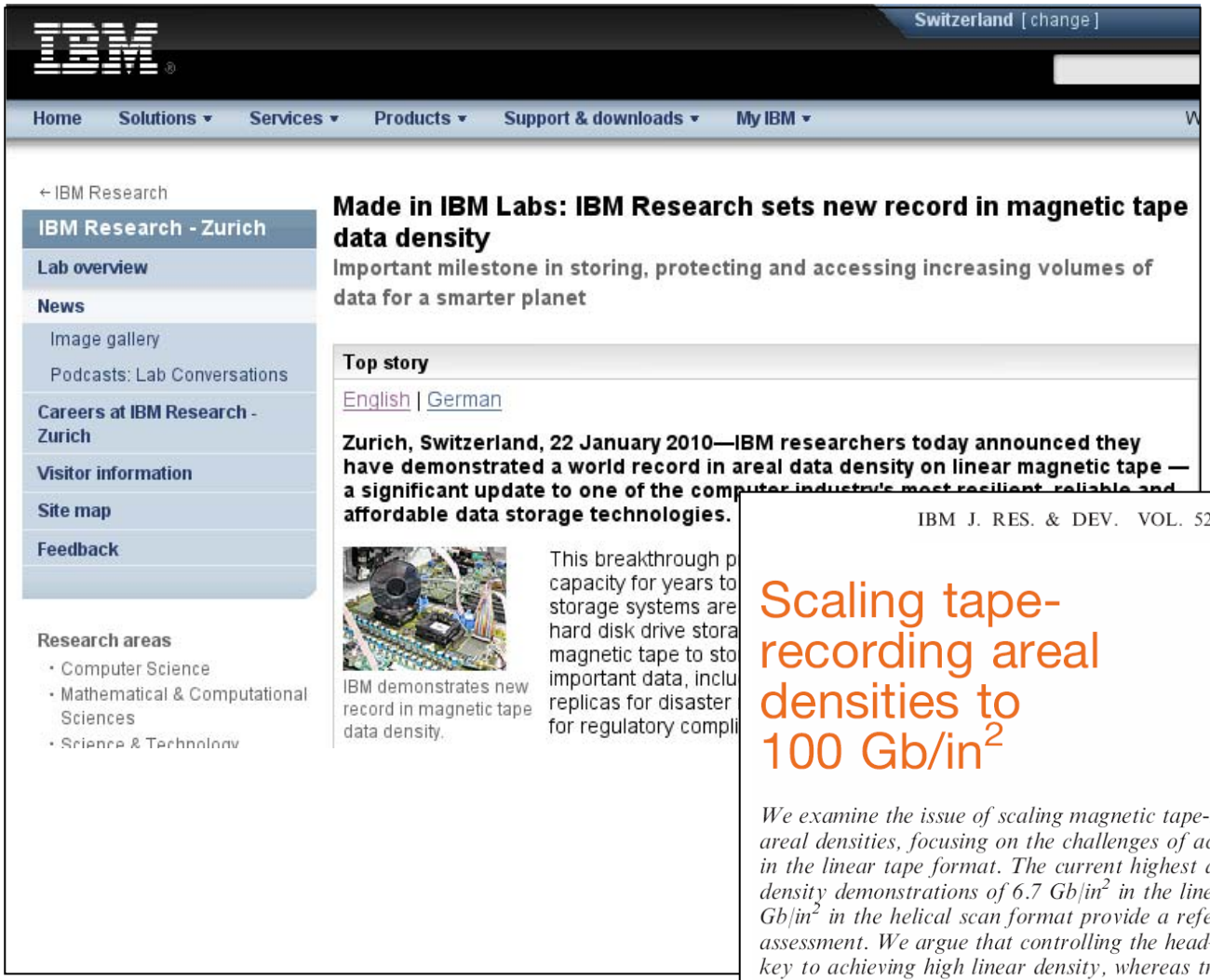
- The BaFe particle is dispersed more uniformly than the current metal particle, the surface of the latest BaFe tape is smoother than the current MP tape.

Areal Density Trends



Tape is required to grow capacity at least 40% per year to maintain the substantial cost advantage of ~10x over disk in $/GB

29.5 Gb/in^2 Demo – 100 Gb/in^2 Appears Achievable



Switzerland [change]

Home | Solutions | Services | Products | Support & downloads | My IBM

← IBM Research

IBM Research - Zurich

- Lab overview
- News
  - Image gallery
  - Podcasts: Lab Conversations
- Careers at IBM Research - Zurich
- Visitor information
- Site map
- Feedback

Research areas

- Computer Science
- Mathematical & Computational Sciences
- Science & Technology

Made in IBM Labs: IBM Research sets new record in magnetic tape data density

Important milestone in storing, protecting and accessing increasing volumes of data for a smarter planet

Top story

English | German

Zurich, Switzerland, 22 January 2010—IBM researchers today announced they have demonstrated a world record in areal data density on linear magnetic tape — a significant update to one of the computer industry's most resilient, reliable and affordable data storage technologies.

IBM demonstrates new record in magnetic tape data density.

This breakthrough p
capacity for years to
storage systems are
hard disk drive stora
magnetic tape to sto
important data, inclu
replicas for disaster
for regulatory compli

IBM J. RES. & DEV. VOL. 52 NO. 4/5 JULY/SEPTEMBER 2008

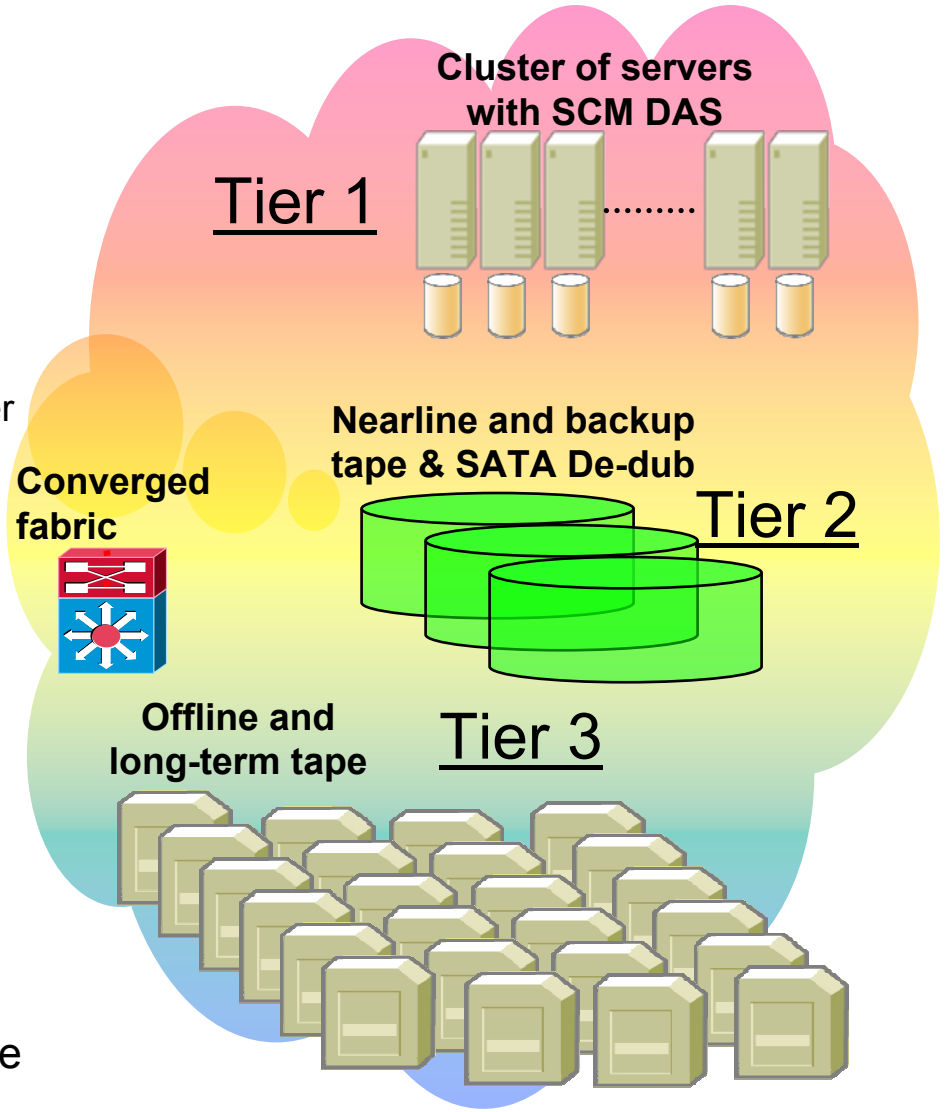
Scaling tape-recording areal densities to 100 Gb/in^2

A. J. Argumedo
D. Berman
R. G. Biskeborn
G. Cherubini
R. D. Cideciyan
E. Eleftheriou
W. Häberle
D. J. Hellman
R. Hutchins
W. Imaino
J. Jelitto
K. Judd
P.-O. Jubert
M. A. Lantz
G. M. McClelland
T. Mittelholzer
C. Narayan
S. Ölçer
P. J. Seger

We examine the issue of scaling magnetic tape-recording to higher areal densities, focusing on the challenges of achieving 100 Gb/in^2 in the linear tape format. The current highest achieved areal density demonstrations of 6.7 Gb/in^2 in the linear tape and 23.0 Gb/in^2 in the helical scan format provide a reference for this assessment. We argue that controlling the head–tape interaction is key to achieving high linear density, whereas track-following and reel-to-reel servomechanisms as well as transverse dimensional stability are key for achieving high track density. We envision that advancements in media, data-detection techniques, reel-to-reel control, and lateral motion control will enable much higher areal densities. An achievable goal is a linear density of 800 Kb/in and a track pitch of 0.2 μm, resulting in an areal density of 100 Gb/in^2.

Impact of Disruptive Technologies on Tiered Storage

- Tier 1: online transactional
  - DAS model (back to the future!)
  - Clustering of SSD DAS
    - High performance for I/O-limited workloads
    - leverage huge local bandwidth by co-locating processing and data
    - cluster servers to enable virtualization of server and storage over converged fabric and link
    - Clustered software for advanced functions and protection
- Tier 2: nearline and backup
  - Large capacity SATA for nearline data and backup of de-duplicatable data
  - Power-efficient storage (VTL with tape)
  - Linear tape file system LTFS
- Tier 3: offline and long-term archival
  - Power-efficient, high-capacity, low-cost tape Linear tape file system LTFS



Conclusion

- Storage Class Memory (SCM) has potential to fundamentally change the design of future information processing systems
- Flash memory today, Storage Class Memory soon
  - Phase Change Memory prime contender for SCM
  - PCM superior to FLASH in latency and write endurance
  - Need to completely rethink the memory/storage stack
- Solid-state storage and tape help to reduce power consumption
  - IT is already consuming 2% of world's energy, annual growth rate 16-23%
- Tape is the greenest and lowest TCO technology providing ultimate insurance policy for the enterprise
  - Big growth opportunity for tape in archival applications
  - 100 Gb/in^2 appears achievable, thus tape is poised to maintain 10x cost advantage over disk

Acknowledgement

Robert Haas, “Storage Systems” Group, IBM Research-Zurich

Haris Pozidis, “Phase Change Memory” Group, IBM Research-Zurich

Jens Jelitto, “Tape Technologies” Group, IBM Research-Zurich

Mark Lantz “Exoloratoty Tape” Group, IBM Research-Zurich

Richard Freitas, IBM Almaden Research Center

Winfried Wilcke, IBM Almaden Research Center

Glen Jaquette, IBM STG Tucson

Vincent Hsu, IBM STG Tucson

Barry Schechtman, INSIC

www.zurich.ibm.com/sto/