IBM Almaden Research Center

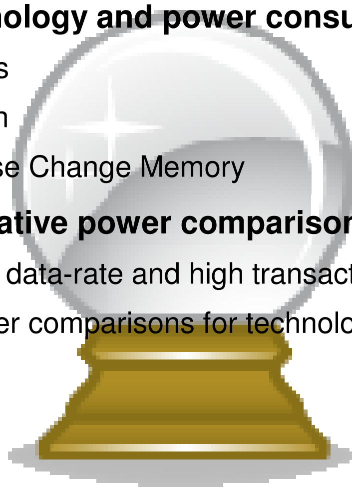
Storage Class Memory: A Low Power Storage Opportunity

Rich Freitas

Motivation

- **Trends**
  - –Demand for storage continues to be robust
  - –Storage performance gain has not kept pace with that of servers so proportionally more disks will be needed in the data center
  - –The amount of power consumed in data centers is becoming an issue
  - –New storage technologies: Flash, PCM, …
- **Questions**
  - –Where will disk storage be in 2020?
  - –Will new storage technologies help?

Agenda



- **Technology and power consumption**
  - –Disks
  - –Flash
  - –Phase Change Memory
- **Illustrative power comparison -- 2020**
  - –High data-rate and high transaction-rate scenarios
  - –Power comparisons for technologies

HDDs

- **Invented in the 1950s**
- **Mechanical device consisting of a rotating** magnetic **media** disk **and** actuator **arm w/ magnetic** head

HUGE COST ADVANTAGES

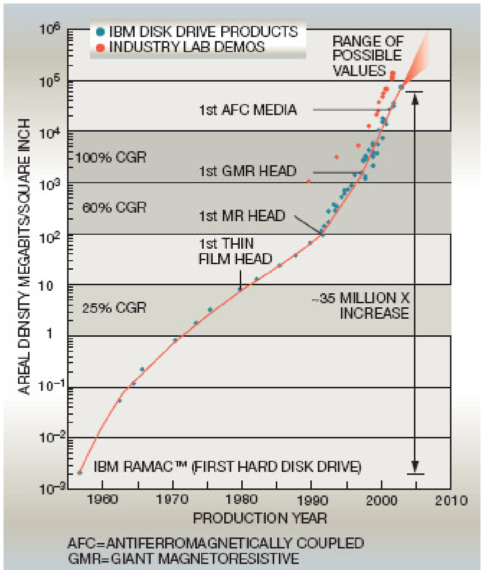
$ **High growth in** disk areal density **has driven the HDD success**

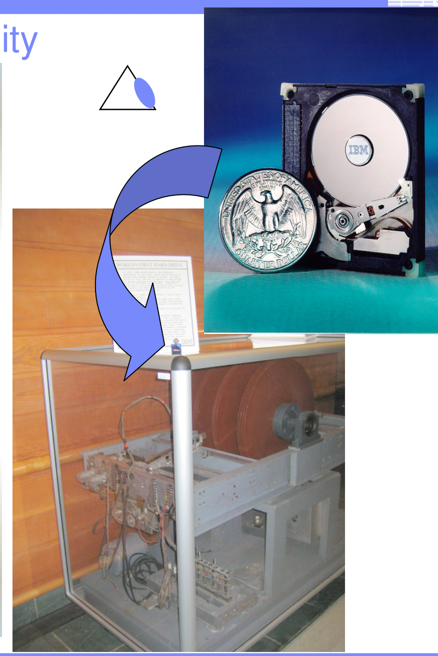
$ **Magnetic thin-film head wafers have very few critical elements per chip (vs. billions of transistors per semiconductor chip)**

$ **Thin-film head (GMR-head) has only one critical feature size controlled by optical lithography (determining track width)**

$ **Areal density is control by track width times (X) linear density…**

HDD strong suit is Areal Density





Wikipedia

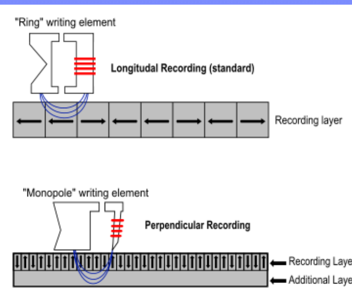
Future of HDD

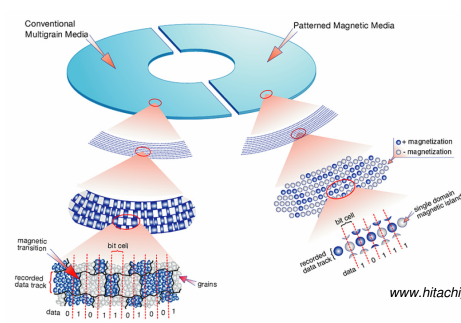
Higher densities through

- perpendicular recording

Jul 2008
610 Gb/in^2 → ~4 TB

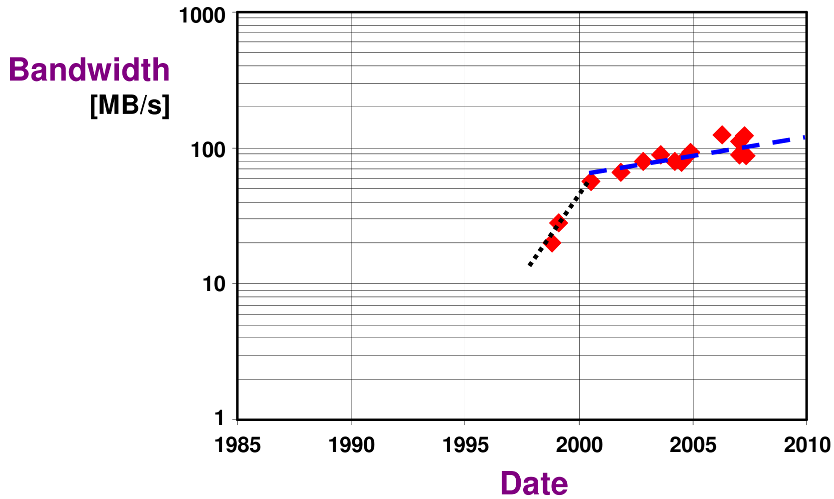
- patterned media



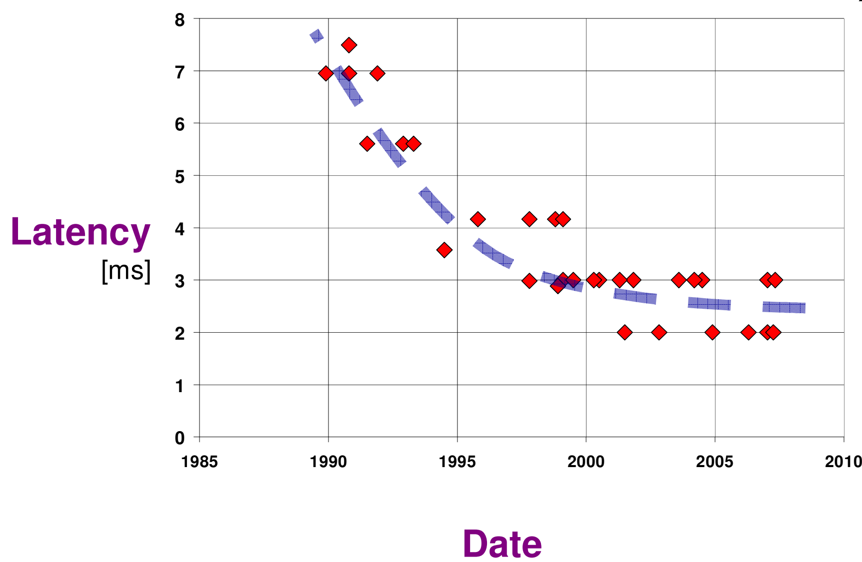


www.hitachigst.com/hdd/research/images/ pm images/conventional_pattern_media.pdf

But, problems with Disk Drive Maximum Sustained Data Rate



And Disk Drive Latency



So,

☒ **Bandwidth Problem** is getting much harder to hide with parallelism

☒ **Access Time Problem** is also not improving with caching tricks

☒ Power/Space/Performance Cost

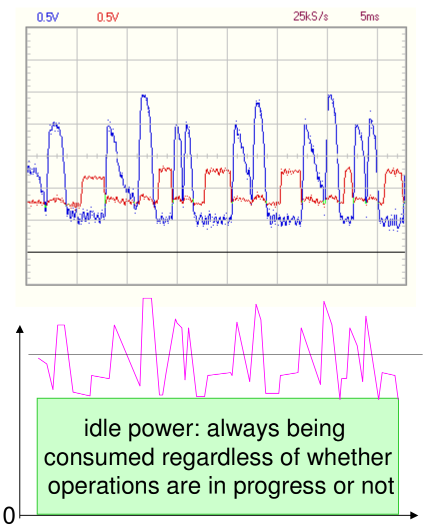
Typical Storage Device Power

		Watts			
Drive type	**size**	**stand-by**	**idle**	**typical**	**start-up**
3.5" 15K RPM FC/SAS	**300 GB**	**2.0**	**13.0**	**18.0**	**28.0**
3.5" 10K RPM FC/SAS	**300 GB**	**2.0**	**10.0**	**16.0**	**17.0**
3.5" 10K RPM FC	**300 GB**		**9.2**	**13.4**	
3.5" 7200 RPM SATA	**500 GB**	**2.0**	**10.0**	**13.0**	**40.0**
3.5" 7200 RPM SATA NL	**500 GB**		**7.4**	**9.2**	
3.5" 7200 RPM SATA	**500 GB**		**9.6**	**13.4**	
2.5" 15K RPM SAS	**73 GB**	***30% less than 15K 3.5" SAS***			
2.5" 10K RPM SAS	**73 GB**	**2.0**	**5.0**	**11.0**	**13.0**
2.5" Mobile 7200 RPM SATA	**100 GB**	**0.3**	**1.0**	**2.5**	**5.5**
2.5" Mobile 5400 RPM SATA	**100 GB**	**0.2**	**0.8**	**2.0**	**5.0**
USB Flash Disk	**32 GB**		**0.1**	**0.5**	
2.5" laptop SSD	**73 GB**		**2.4**	**3.2**	
3.5" Enterprise SSD	**155 GB**		**5.0**	**8.0**	

IBM Almaden Research center

Illustrative Disk Drive Power Profile

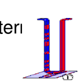
- **Standby power**
  - Both spindle and actuator Motors off
  - Base electronics powered
- **Idle**
  - Spindle motor running
  - Actuator motor off
  - Most electronics powered
- **Typical**
  - Spindle motor running
  - Actuator motor in use periodically
  - All electronics powered
- **Startup – spindle motor starting**
  - ~30 seconds for startup



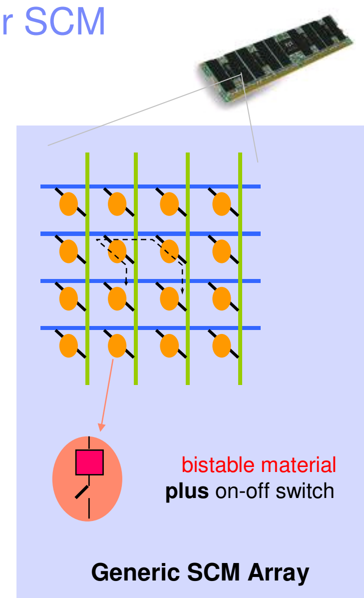
IBM Almaden Research center

Many Competing Technologies for SCM

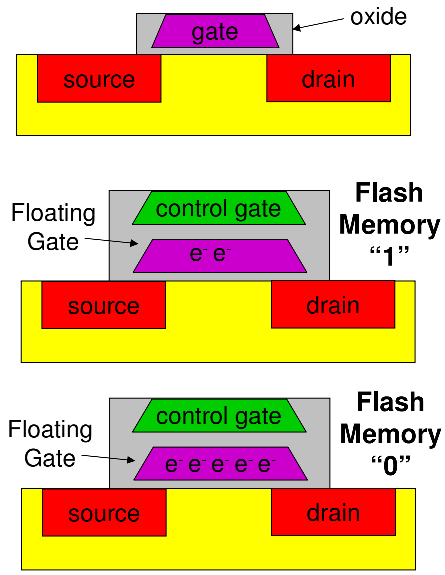
- **Phase Change RAM**
  - most promising now (scaling)
- Magnetic RAM
  - used today, but poor scaling and a space hog
- Magnetic Racetrack
  - basic research, but very promising long ter



- Ferroelectric RAM
  - used today, but poor scalability
- Solid Electrolyte and resistive RAM (Memristor)
  - early development, maybe?
- Organic, nano particle and polymeric RAM
  - many different devices in this class, unlikely
- **Improved FLASH**
  - still slow and poor write endurance



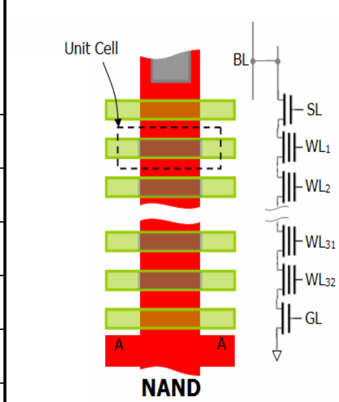
What is Flash?



- Based on MOS transistor
- Transistor gate is redesigned
  - Charge is placed or removed near the “gate”
  - The threshold voltage V_{th} of the transistor is shifted by the presence of this charge
  - The threshold Voltage shift detection enables non-volatile memory function.
- Single Level vs Multi-level cell
- Scaling issues

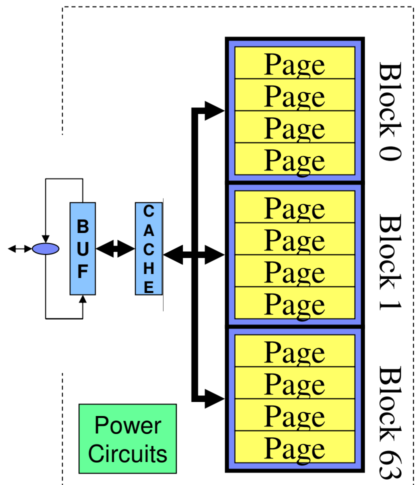
Feeds and Speeds for typical NAND Flash

	NAND
Cell Size	4 F^2 (2 F^2 virtual x 2-bit MLC)
Read Access Time	20-50 us
Read	15-25 MB/s
Write	5-8MB/sec
Erase	2ms
Start Up Time	50-100 us
Market Size (2007)	$14.2B
Applications	Multimedia



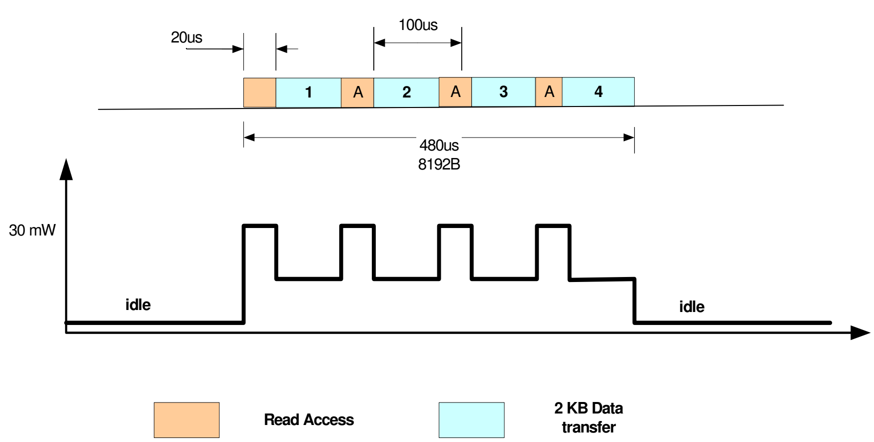
Representative NAND Flash Device

- **Interface: one or two bytes wide**
  - Transition to ONFI for some vendors
- **Data accessed in pages**
  - **2112, 4224 or 8448 Bytes**
- **Data erased in blocks**
  - Block = 64 - 128 Pages
- **Power circuits**
  - Charge Pumps
  - Clock drivers
  - Etc.

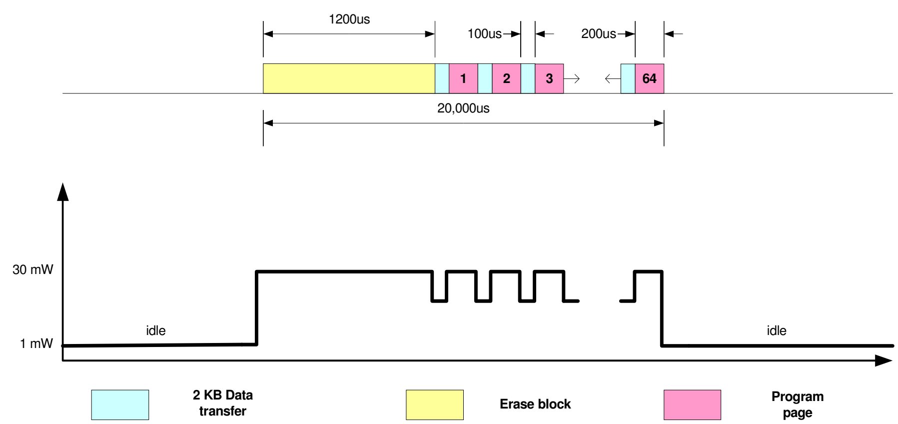


ONFI → Open NAND Flash Interface

Illustrative Flash Read Power Profile



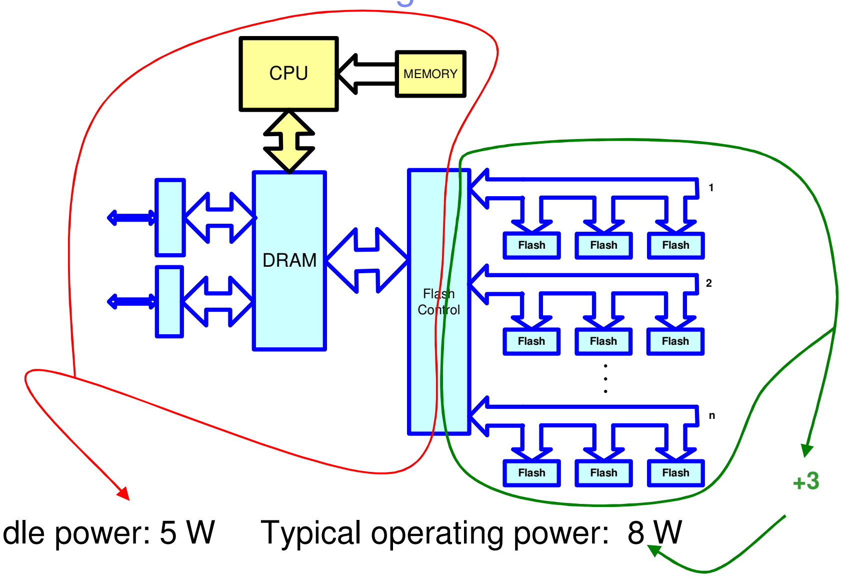
Illustrative Flash Write Power Profile



Typical Storage Device Power

		Watts			
Drive type	size	stand-by	idle	typical	start-up
3.5" 15K RPM FC/SAS	300 GB	2.0	13.0	18.0	28.0
3.5" 10K RPM FC/SAS	300 GB	2.0	10.0	16.0	17.0
3.5" 10K RPM FC	300 GB		9.2	13.4	
3.5" 7200 RPM SATA	500 GB	2.0	10.0	13.0	40.0
3.5" 7200 RPM SATA NL	500 GB		7.4	9.2	
3.5" 7200 RPM SATA	500 GB		9.6	13.4	
2.5" 15K RPM SAS	73 GB	*30% less than 15K 3.5" SAS*			
2.5" 10K RPM SAS	73 GB	2.0	5.0	11.0	13.0
2.5" Mobile 7200 RPM SATA	100 GB	0.3	1.0	2.5	5.5
2.5" Mobile 5400 RPM SATA	100 GB	0.2	0.8	2.0	5.0
USB Flash Disk	32 GB		0.1	0.5	
2.5" laptop SSD	73 GB		2.4	3.2	
3.5" Enterprise SSD	155 GB		5.0	8.0	

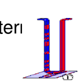
Illustrative Flash SSD Design



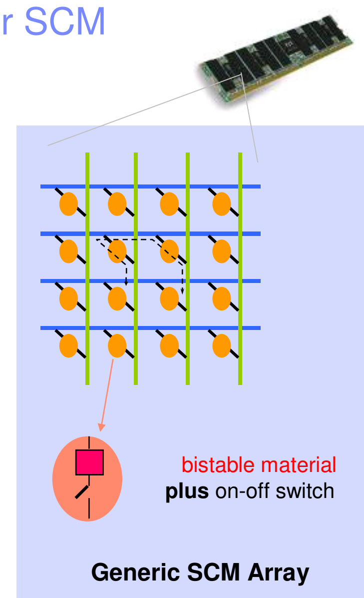
Idle power: 5 W   Typical operating power: 8 W

Many Competing Technologies for SCM

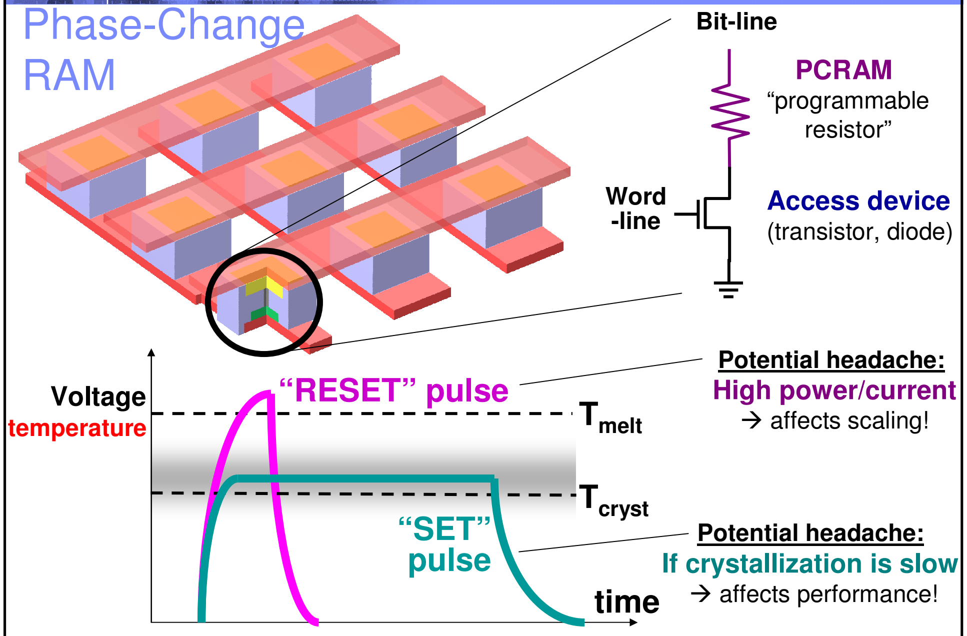
- **Phase Change RAM**
  - most promising now (scaling)
- Magnetic RAM
  - used today, but poor scaling and a space hog
- Magnetic Racetrack
  - basic research, but very promising long teri



- Ferroelectric RAM
  - used today, but poor scalability
- Solid Electrolyte and resistive RAM (Memristor)
  - early development, maybe?
- Organic, nano particle and polymeric RAM
  - many different devices in this class, unlikely
- **Improved FLASH**
  - still slow and poor write endurance

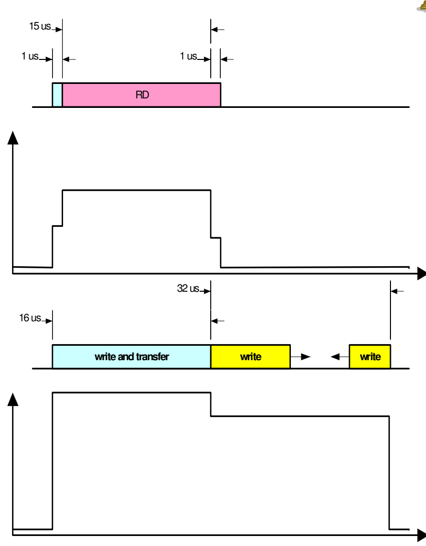


Phase-Change RAM



PCM performance and power SWAG

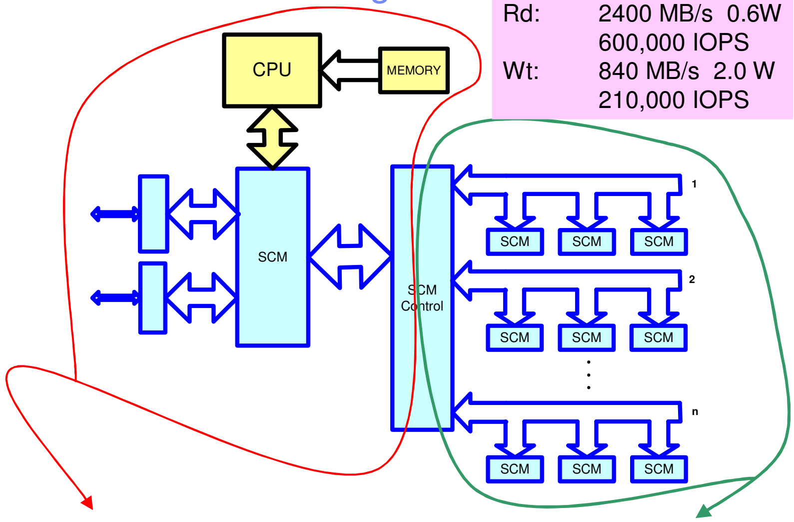
- **Performance**
  - For devices optimized for storage applications:
    - Read access time of 1 us
    - Write access time of 3-5 us
  - Interface LPDDR or similar
- **Power**
  - Transfer power 15mW
  - Read power 20 mW
  - Write: 85 mW
    - Write-in-place
    - S0, no erase



Illustrative PCM SSD Design

SWAG

Rd:	2400 MB/s 0.6W 600,000 IOPS
Wt:	840 MB/s 2.0 W 210,000 IOPS



Idle power: 1 W

Typical operating power: 3 W

Device comparison: energy and power

		disk		Flash SSD today	PCM SSD projected
		Today	2020		
		3.5" 15K	1.8" 15K	3.5"	1.8"
IOPS		200	400	50,000	600,000
BW		100 MB/s	300 MB/s	250 MB/s	2400 MB/s
Idle power		9 W	4 W	5 W	1 W
Typ. power		16 W	6 W	8W	3 W
Energy	10% utilization	485,000 uJ	105,000 uJ	1,060 uJ	20 uJ
per Rd op	50%	125,000 uJ	25,000 uJ	260 uJ	7 uJ
in 1 sec	90%	85,000 uJ	16,000 uJ	171 uJ	5 uJ

- Energy is what you pay for
- In 2020 a disk will cost ~3000x more per operation than a PCM SSD
- IOPS, cost, power → power, cost IOPS

- Flash SSDs may not exist in 2020
- Power for controllers, etc. viewed as second order effects

Extrapolate Storage Performance to 2020

- **Start with ASCI Purple class machine**
- **Trend for storage performance growth: 70% CAGR**
  - Driven by application requirements and investment
  - Could be impacted by changing in architecture
- **Assume 50 yr disk trends continue and no game changing technology invention**
- **Semiconductor technologies stay on a roughly 40% CAGR**
- **PCM edges out FLASH for the solid state storage crown**

- **Bandwidth: 0.4 TB/s → 400 TB/s**
- **Transaction rate: 2 MIOPS → 2000 MIOPS**

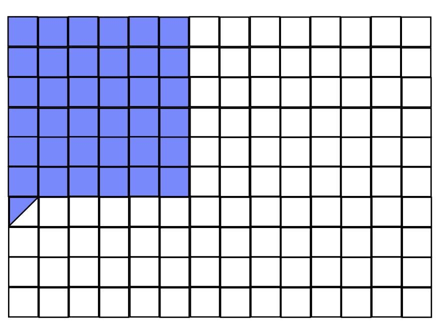
The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers

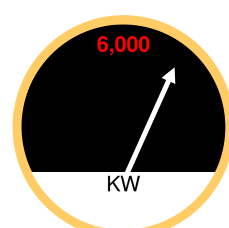
Bandwidth Driven Storage System: 400 TB/s

DISKS

Floor Space



Power



6000 Square Feet

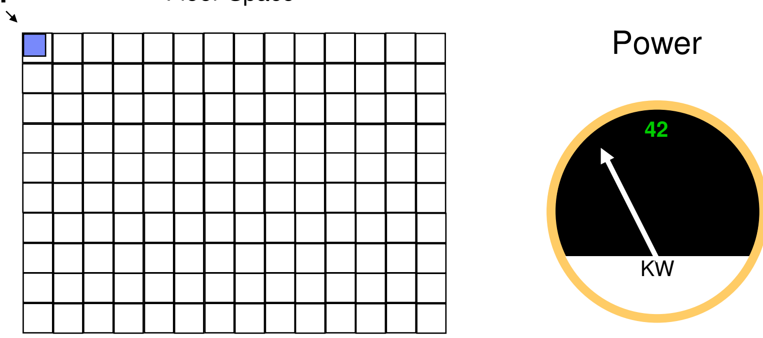
The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers

Bandwidth Driven Storage System: 400 TB/s

SCM Floor Space

Power



85 Square Feet

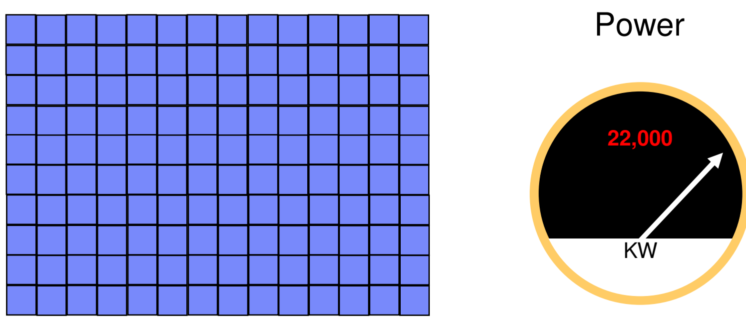
The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers

Transaction Rate Driven Storage System: 2000 MOP/s

DISKS Floor Space

Power



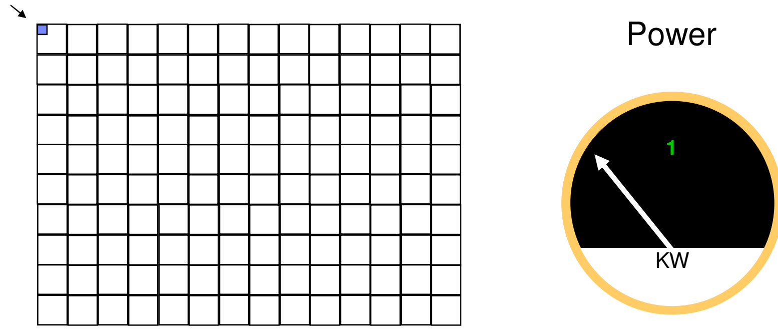
23,000 Square Feet

The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers

Transaction Rate Driven Storage System: 2000 MOP/s

SCM Floor Space



11 Square Feet

My Conclusions

- **Idle power for disks drives up their energy costs significantly**
- **Energy per op for SSDs is much better than for HDDs**
  - –SSDs gain because of their very high transaction rates and and high bandwidth
  - –Combined with generally lower idle power
- **But, even so, SSDs have a higher idle power that one would expect for solid state technology**
  - –Careful design of SSDs should alleviate much of this
- **Current trends would seem to indicate that Enterprise disks may not be widely used in 2020 and their position will be taken by Solid State Storage**

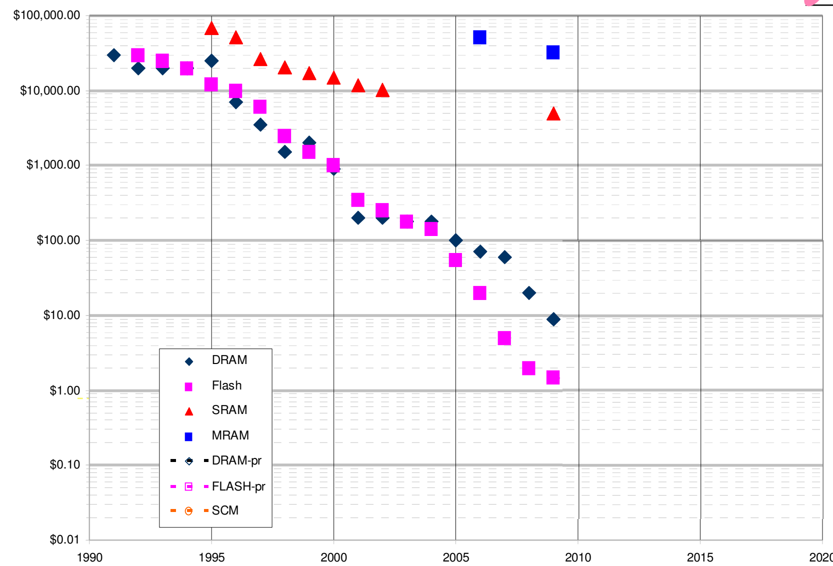
Questions?

Break Time

Further reading

- **IBM Journal of Research and Development, special issue on Storage Technologies and Systems**
  - Volume 52, Number 4/5 July/September 2008
  - R. F. Freitas and Winfried Wilcke, "Storage-class Memory: The next storage system technology", pg 439-448.
  - G. W. Burr, etal., "Overview of candidate device technologies for storage-class memory", pg 449-464.
  - S. Raoux, etal., "Phase-change random access memory: a scalable technology", pg 465-480.
- **Other papers**
  - Grupp, Laura M., etal., "Characterizing Flash Memory: Anomalies, Observations, and Applications", *MICRO'09,* December 12–16, 2009, New York, NY, USA., pg: 24-33.
  - Lee , Kwang-Jin, etal., "A 90 nm 1.8 V 512 Mb Diode-Switch PRAM
  - With 266 MB/s Read Throughput", IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 43, NO. 1, JANUARY 2008, pg. 150.
  - Chen, Feng, etal., "Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives" , *SIGMETRICS/Performance'09,* June 15–19, 2009, Seattle, WA, USA. Copyright 2009.

Input from the device cost crystal ball



Input form the Subsystem Price Crystal Ball