IBM Almaden Research Center

Storage Class Memory and the data center of the future

Rich Freitas

HPC System performance trends

- **System performance requirement has historically double every 18 mo and this trend is likely to continue**

System	Year	TF	Nodes	Cores	Memory	Storage	GB/s	Disks
Blue Pacific	1998	3	1464	5856	2.6 TB	43 TB	3	5040
White	2000	12	512	8192	6.2 TB	147 TB	9	8064
Purple/C	2005	100	1536	12288	32-67 TB	2000 TB	122	11000
HPCS (rough est.)	2011	4000	20000	300000	2000-4000 TB	80000 TB	4000	80000+

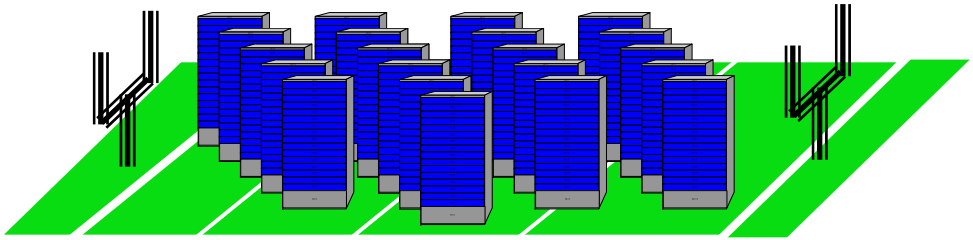
Memory power trends

- **DRAM: 2X size / chip / 3 year**
  - Growth rate lower than the requirement for system performance growth
- **Numerically, more memory chips will be needed in a system to match the requirement for growth in system performance**
- **Memory chip power is unlikely to decrease**
  - P = active power + leakage power + refresh power
  - DRAM can be put in standby mode, i.e., no active power, but refresh and leakage power are still present albeit they may be at reduced levels

Storage power extrapolation → 2020

- **Disk power is power to motor plus power for seeking and power for interface and control electronics**
- **Motor, interface and control power are always present**
- **Active power is power for seeking and transferring**
- **Disk total drive power may decrease a little in the future,**
  - Disks will get somewhat smaller, physically probably dropping to 1.8”
  - Rotational speed unlikely to decrease

	Compute centric	Data centric
Devices	**1.3 M Disks**	**5 M Disks**
space	**4500 sq.ft.**	**16,500 sq.ft.**
power	**6,000 kW**	**22,000 kW**



- **Largest of current HPC systems extrapolated to 2020**

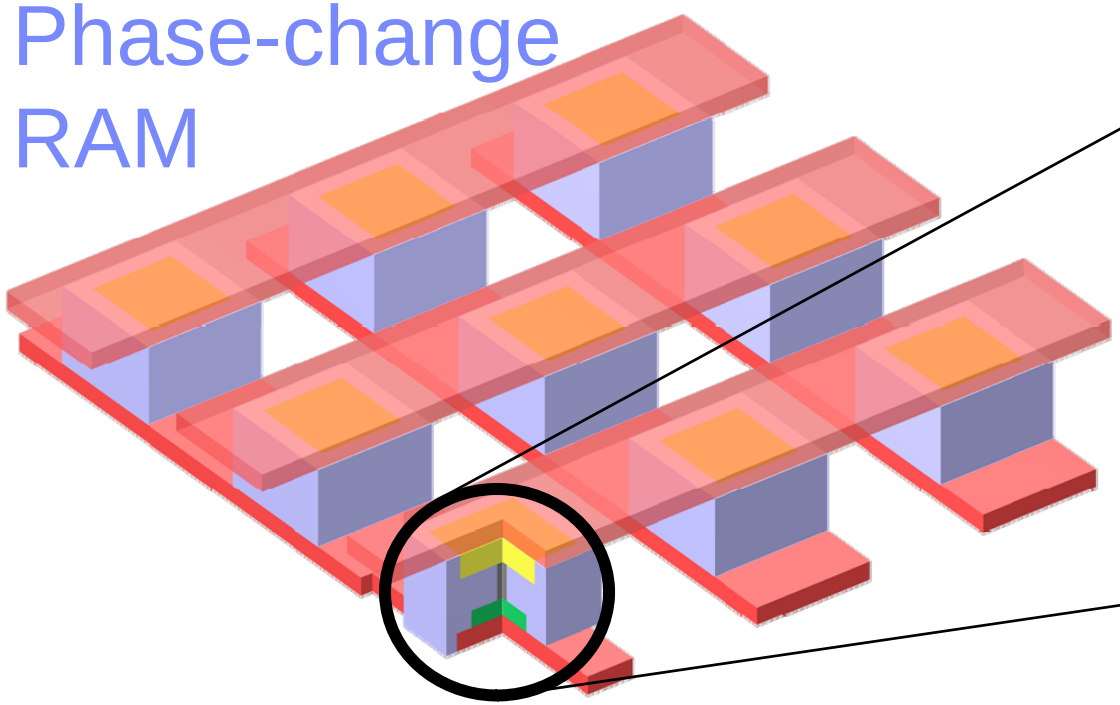
Net of memory and disk situation

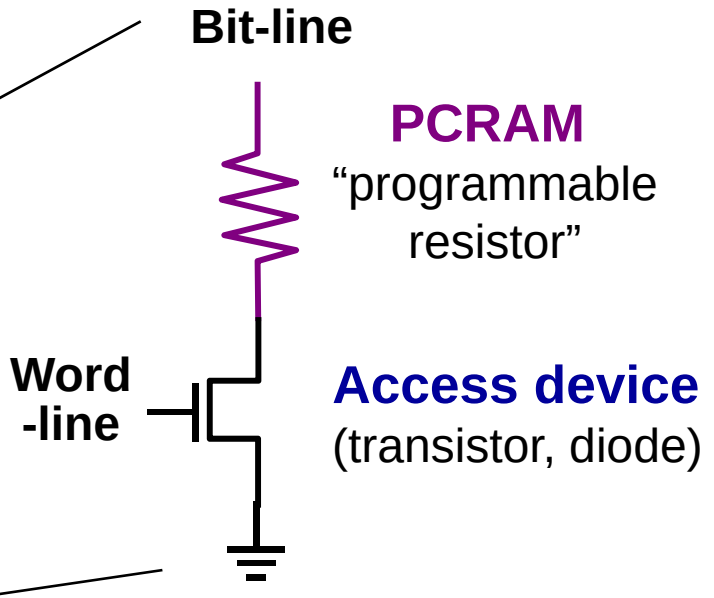
- **Memory power will increase significantly over the next few years**
- **Storage power will increase significantly over the next few years**
- **Data center power and cooling capabilities unlikely to increase significantly over the same interval**
  -
- **So, → NVRAM to the rescue**

Definition of **S**torage **C**lass **M**emory **SCM**

- **A new class of data storage/memory devices**
  - many technologies compete to be the 'best' SCM
- **SCM features:**
  - Non-volatile (~ 10 years)
  - Fast Access times (~ DRAM like)
  - Low cost per bit more (DISK like – by 2015)
  - Solid state, no moving parts
- **SCM *blurs the distinction* between**
  - MEMORY (*fast, expensive, volatile*) and
  - STORAGE (*slow, cheap, non-volatile*)

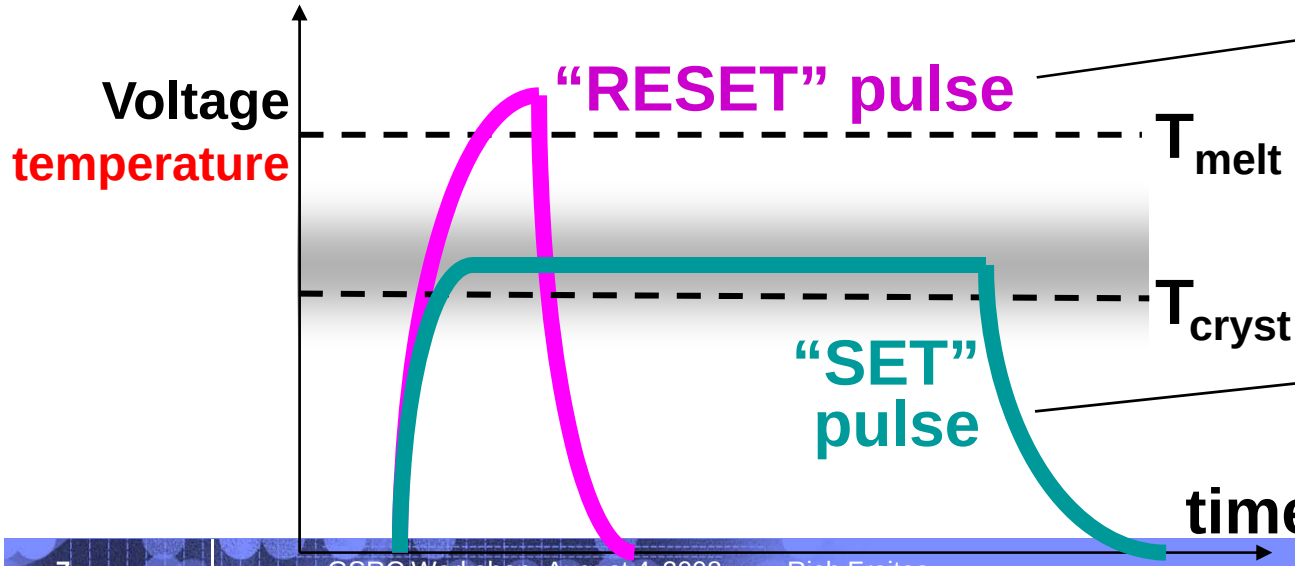
Phase-change RAM





PCRAM
"programmable resistor"

Access device
(transistor, diode)

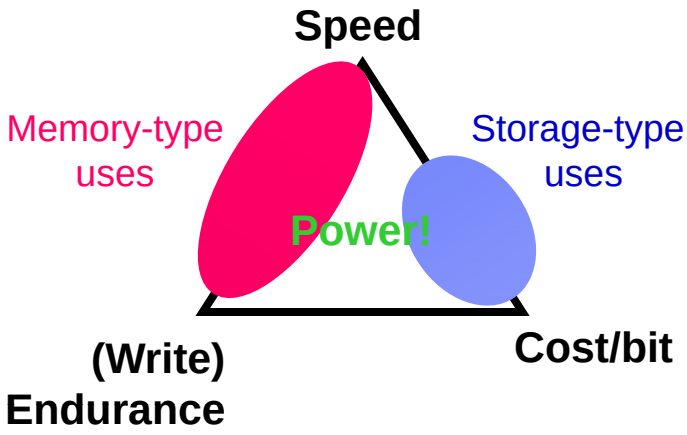


Potential headache:
High power/current
→ affects scaling!

Potential headache:
If crystallization is slow
→ affects performance!

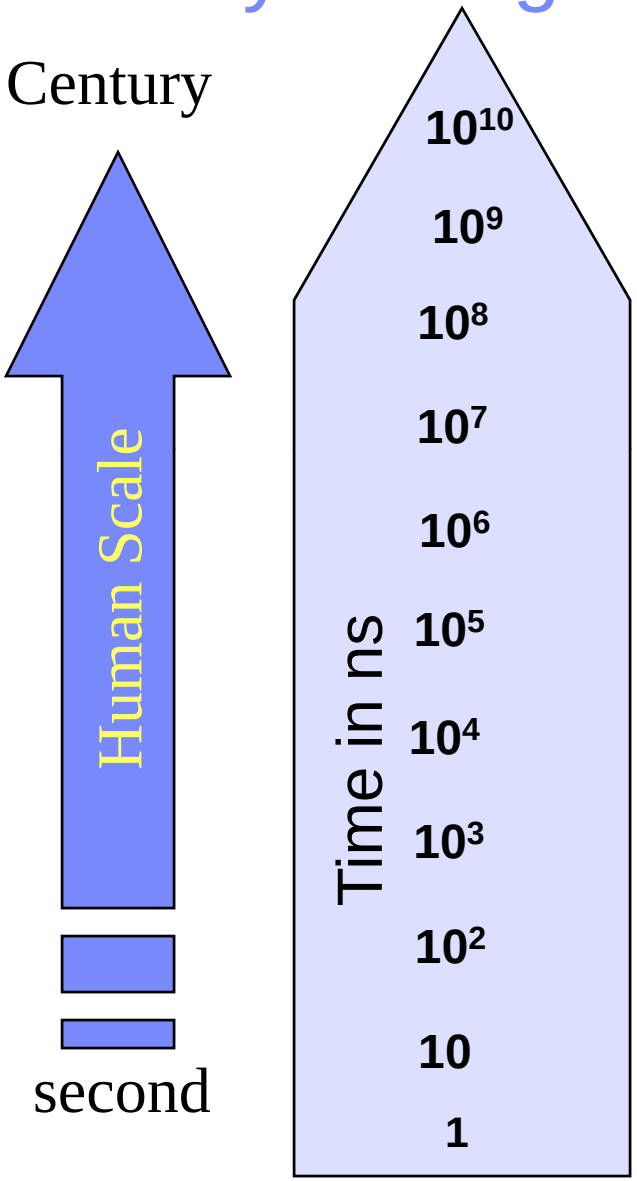
Storage Class Memory

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



- **SCM system requirements for Memory (Storage) apps**
  - No more than 3-5x the **Cost** of enterprise HDD (< $1 per GB in 2012)
  - **<200nsec (<1 μsec) Read/Write/Erase time**
  - >100,000 **Read I/O operations** per second
  - **>1GB/sec (>100MB/sec)**
  - **Lifetime** of 10^8 – 10^{12} write/erase cycles
  - 10x lower **power** than enterprise HDD

Memory/Storage Stack Latency Problem



Get data from TAPE (40s)

Access DISK (5ms)

Storage

Access FLASH (20 us)

SCM

Access PCM (100 – 1000 ns)

Get data from DRAM *or* PCM (60ns)

Get data from L2 cache (10ns)

Memory

CPU operations (1ns)

If you could have SCM, why would you need anything else?

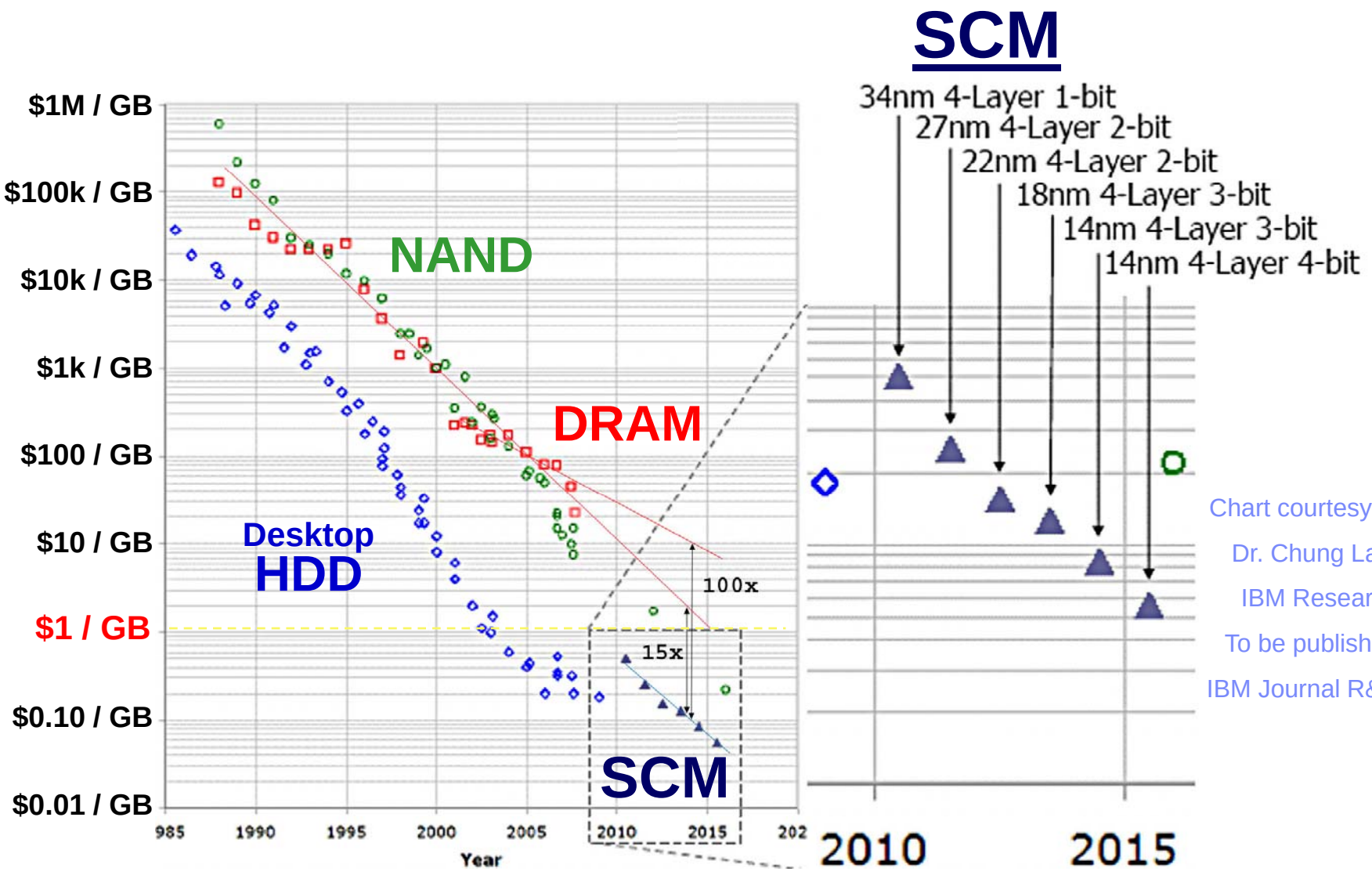
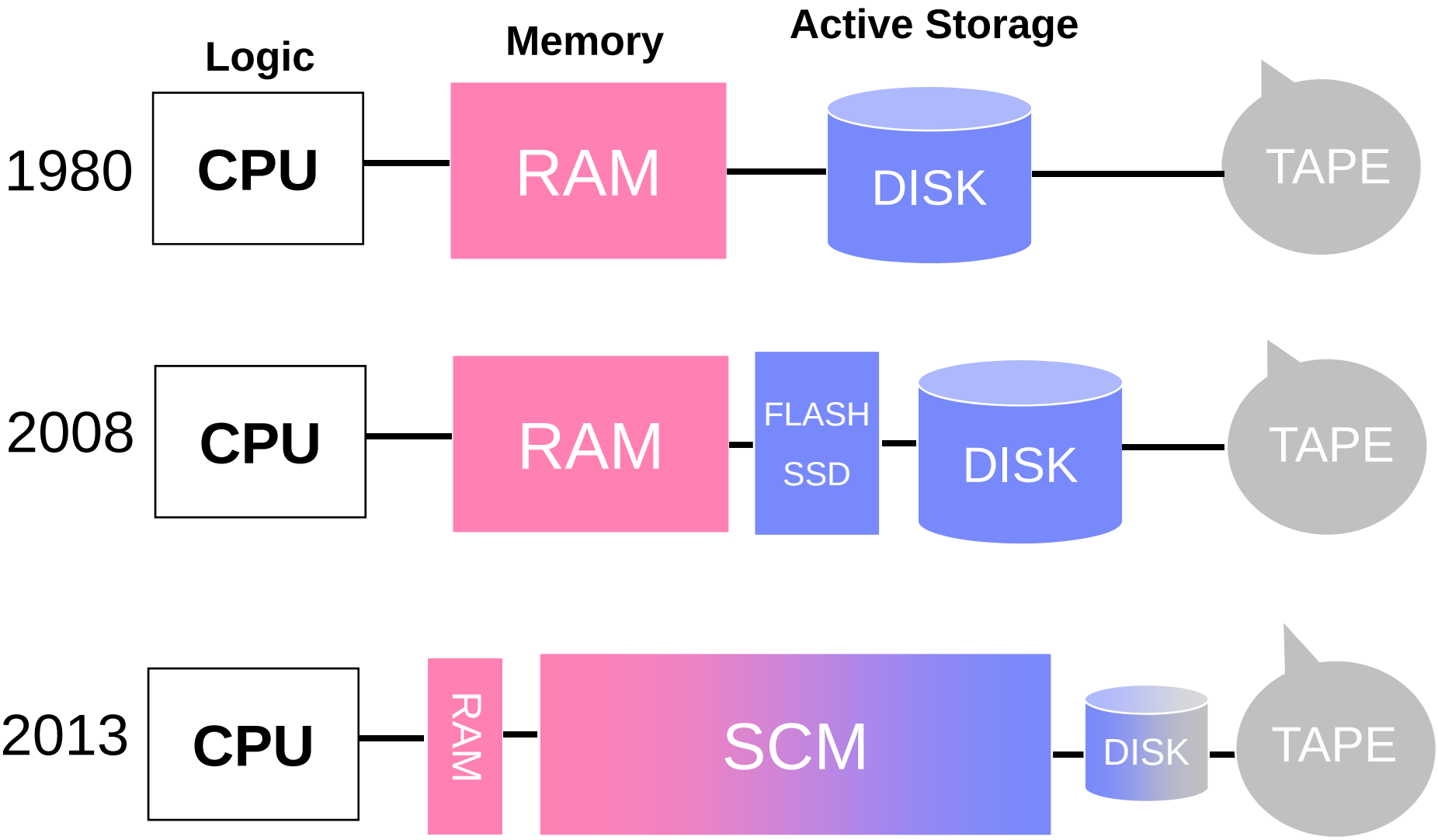


Chart courtesy of
Dr. Chung Lam
IBM Research
To be published
IBM Journal R&D

SCM in a large System



Active vs passive power

- **The blue area marks active power in the power equations**
- **The red area marks passive power in the power equations**
  - Passive power is unproductive. It just causes heat
  - For memories it is leakage and refresh power, which is typically smaller than maximum active power
  - For disks it is keeping the motor spinning and the standby power of the electronics, which is typically larger than the maximum active power
  - For PCM is is the leakage and small standby power and is typically much much smaller than the maximum active power.

$$P_M = V_{dd} I_{\text{leak}} + V_{dd} I_{\text{refresh}} + \alpha C V_{dd}^2 f$$

$$P_D = \kappa d^{4.6} r^{2.8} + I_{\text{i\&c}} V + \alpha I_{\text{s\&t}} V$$

$$P_{PCM} = I_{\text{standby}} V_{dd} + \alpha I_{\text{active}} V_{dd}$$

passive active

α is the portion of time that the device is active and productive

κ is the normalized power of the disk motor

CPU's are doing fine - focus on memory/storage stack

Goal: eliminate passive power and make active power more efficient

Issues

- **Redesign of DRAM and Disks to eliminate passive power**
  - Possible but not probable
- **DRAM has fast turn off and turn on times, but it is volatile**
  - –Turning off DRAM when not active causes data loss
- **Disks are nonvolatile and turn on/off in ~ 20-30 seconds**
  - –On/off time to long for practical active storage systems
  - –Storage systems that manage power in this manner are called MAID system.
  - –So far, only used for archive systems

Opportunities

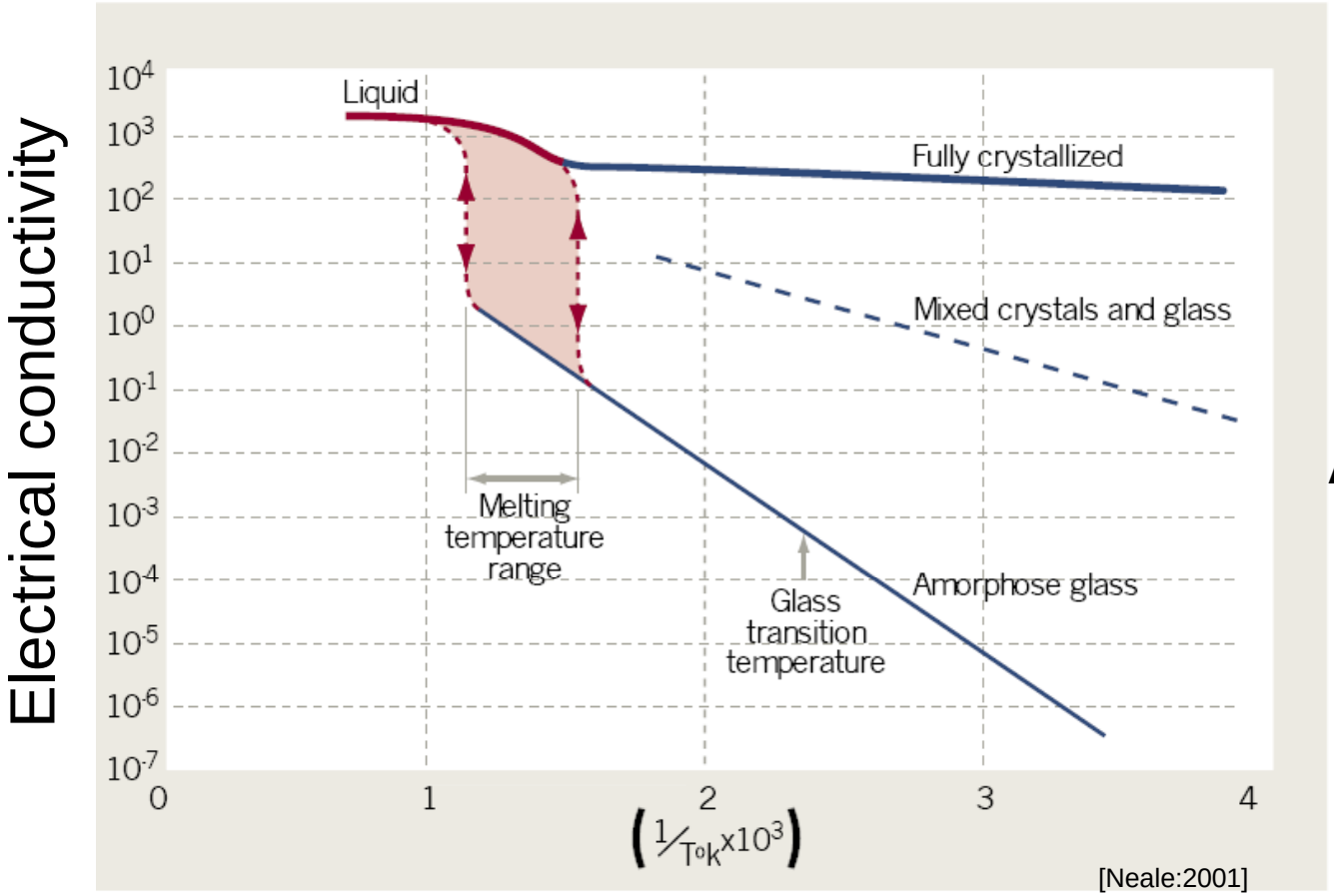
- **How can PCM be used to virtually eliminate passive power?**
  - –Active power is much greater than passive power
  - –Turn on/off time ~50us
- **How can data be laid out to minimize active power?**
  - –Memory/storage pools
  - –hierarchy
- **How can active power be used more efficiently?**
  - –Device design
  - –System architecture
  - –exploit virtualization (management challenge)
  - – exploit accelerators

Questions

History of Phase-change memory

History of Phase change memory

- late 1960's – Ovshinsky shows reversible electrical switching in disordered semiconductors
- early 1970's – much research on mechanisms, but everything was too slow!

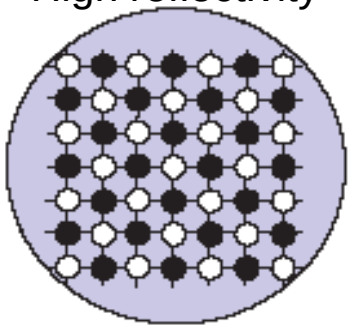


[Neale:2001]

Crystalline phase

Low resistance

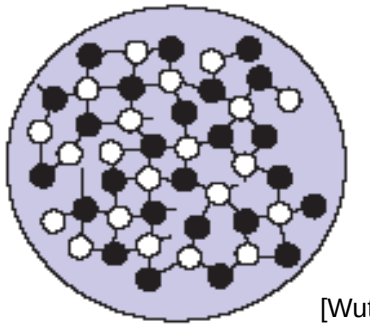
High reflectivity



Amorphous phase

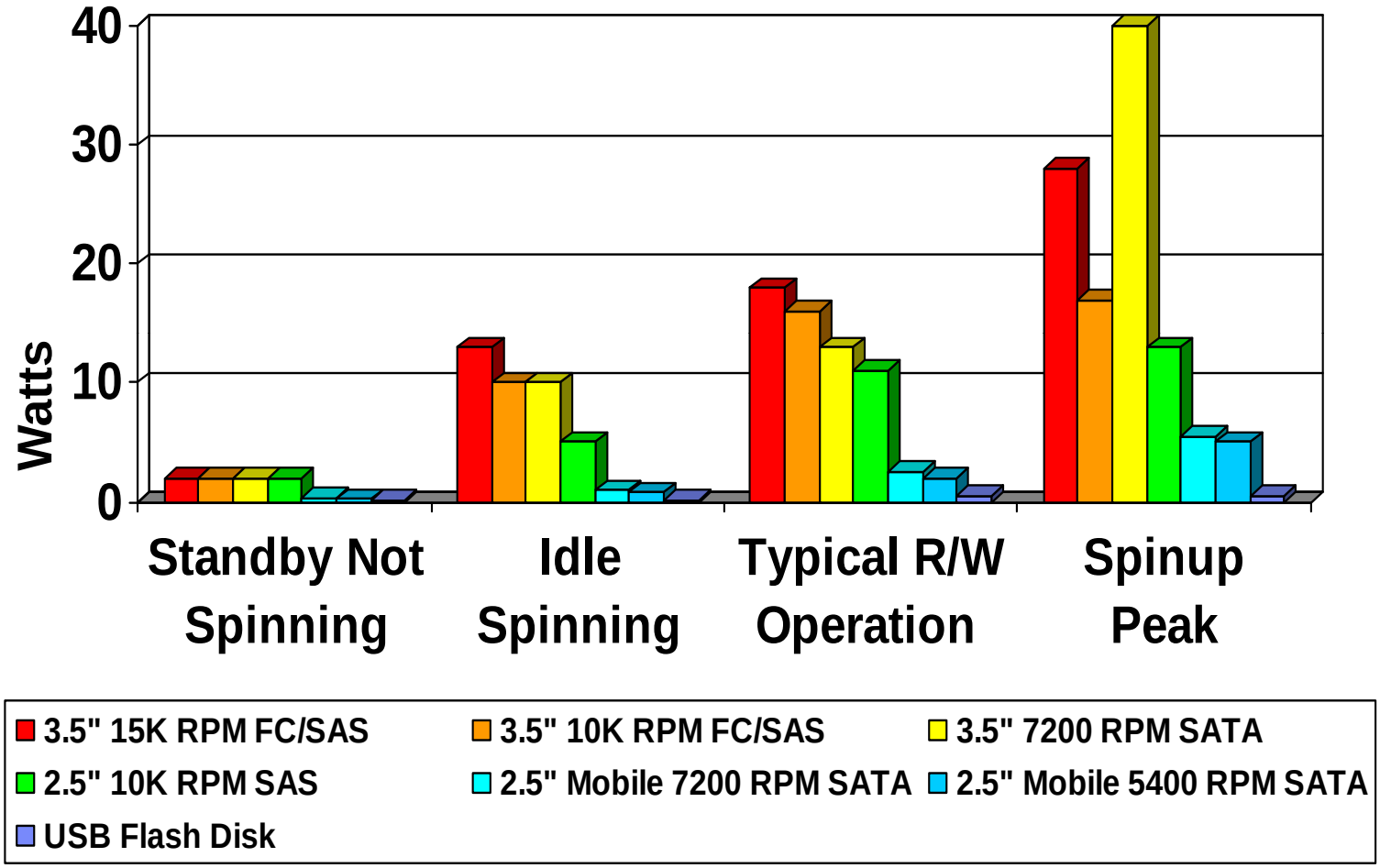
High resistance

Low reflectivity



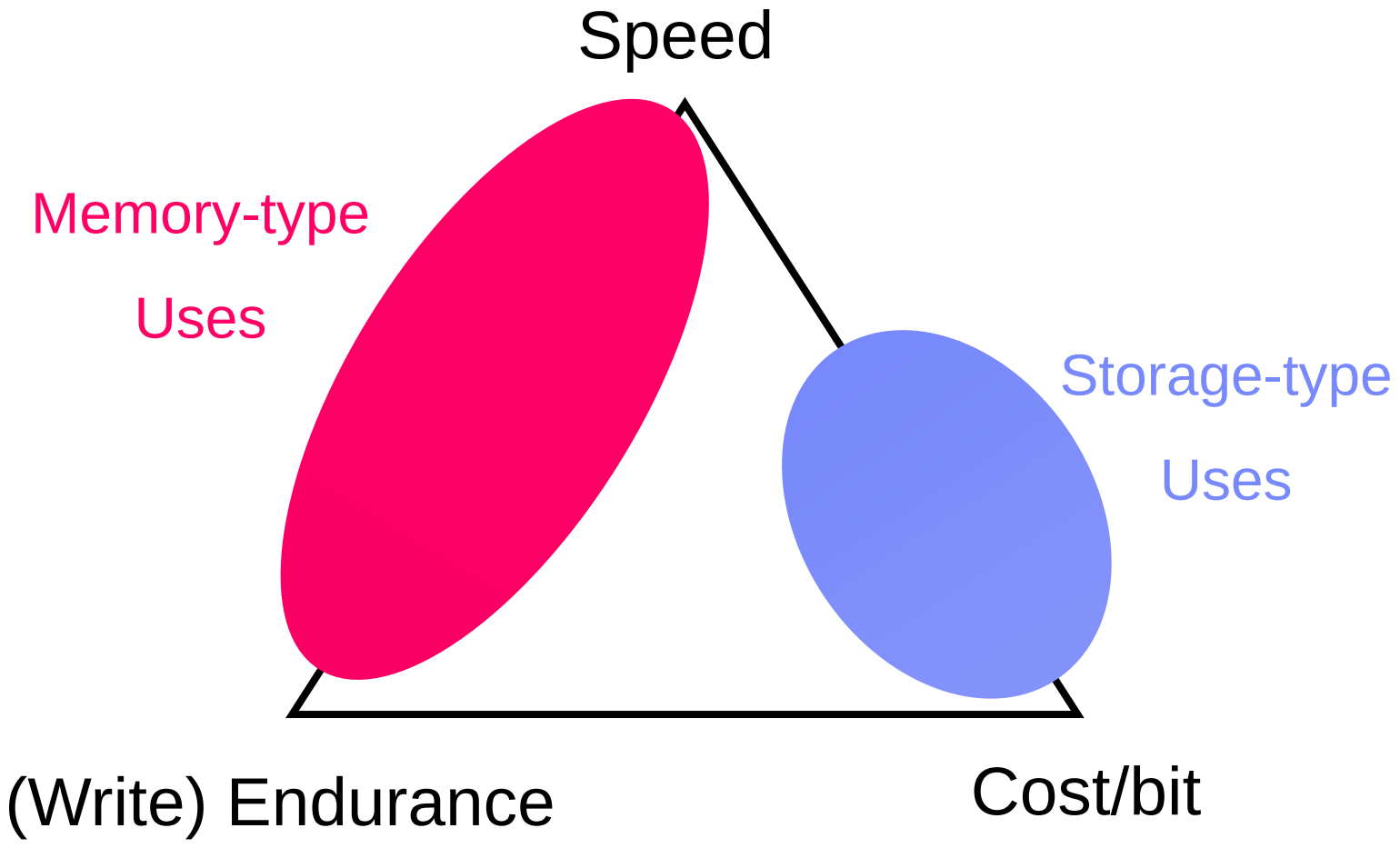
[Wuttig:2007]

Use Small Form Factor and Solid State disks to help lower average and peak disk storage power consumption



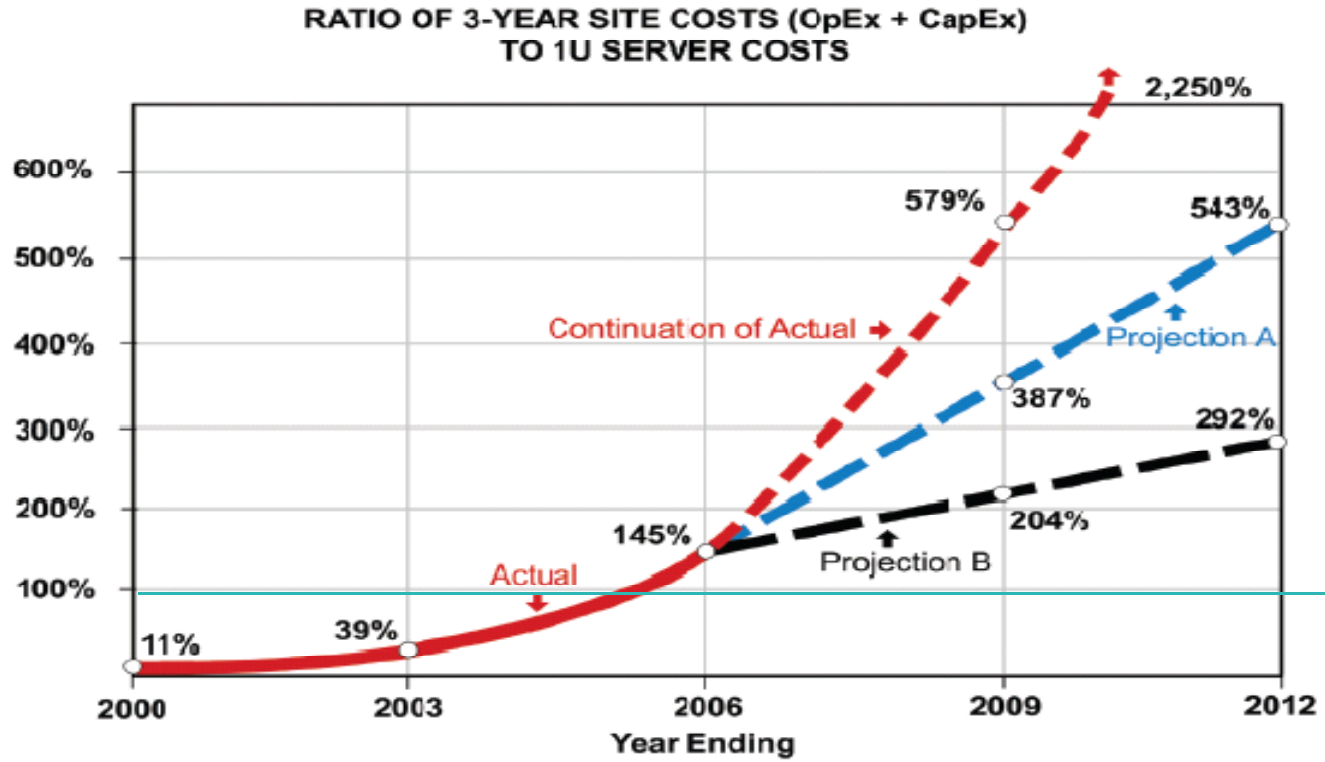
Note: Data for disks from datasheets/specs on Seagate website
3.5” FC/SAS disks are 300MB capacity, 3.5” 7200 RPM SATA disk is 500 GB
2.5” 10K RPM SAS disk is 73GB, 2.5” Mobile SATA disks are 100GB

SCM Design Triangle



Power!

Energy & Infrastructure Cost exceed Server Cost *



* for cheap 1U Servers

Figure 1: Site infrastructure costs (OpEx + amortized CapEx) for data-center power and cooling are a growing percentage of the cost of buying a server

Source: Uptime Institute