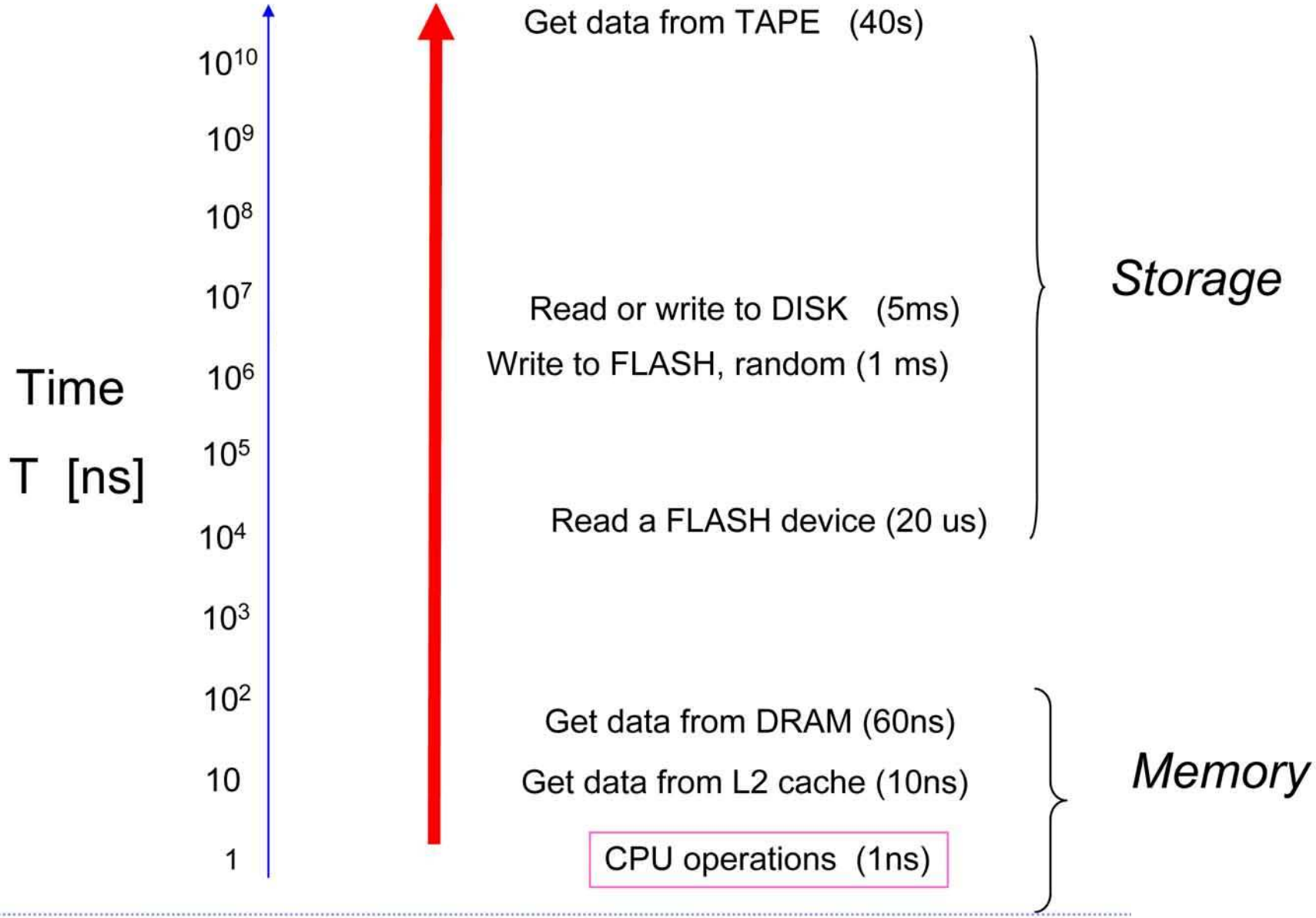
Exascale* Data Analytics
(*era of 10^{18} flops/s machines; circa 2018)

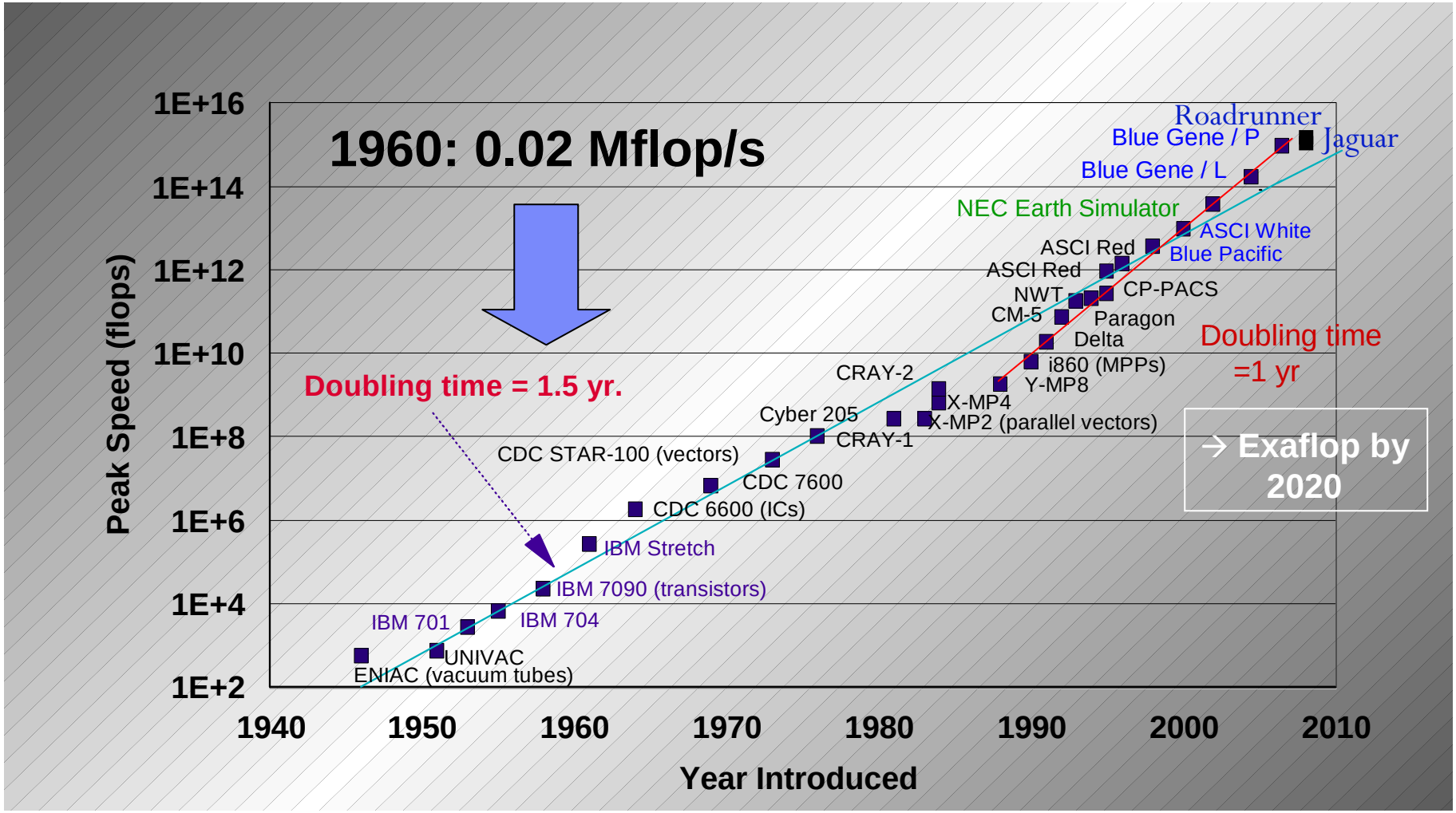
Bruce Elmegreen, Blake Fitch
IBM Research Division
Yorktown Heights, NY 10598
Green Bank May 3-5, 2011

Times for Processing to/from Memory to/from Storage

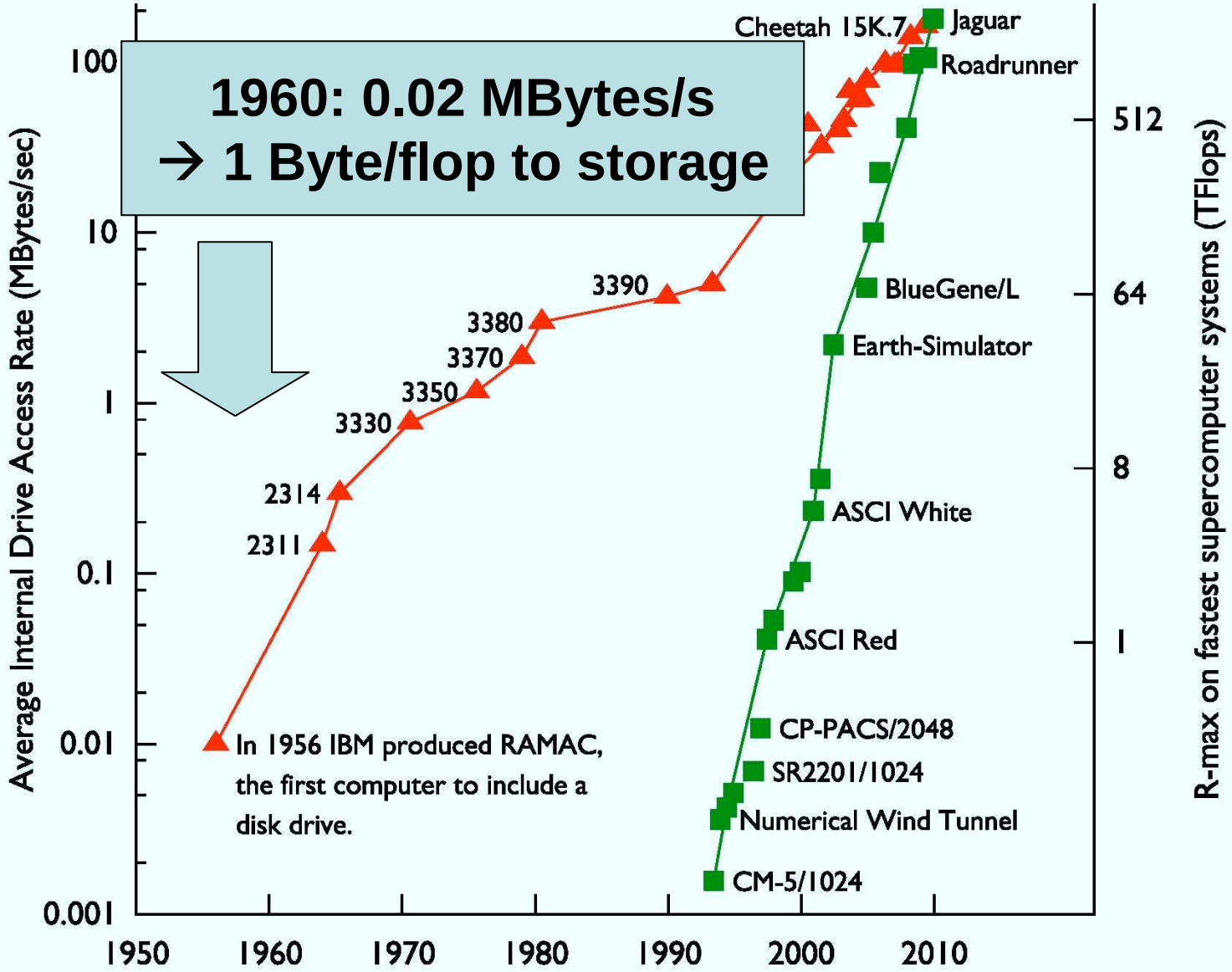


Analogy to Human Scales

Supercomputer Peak Speed

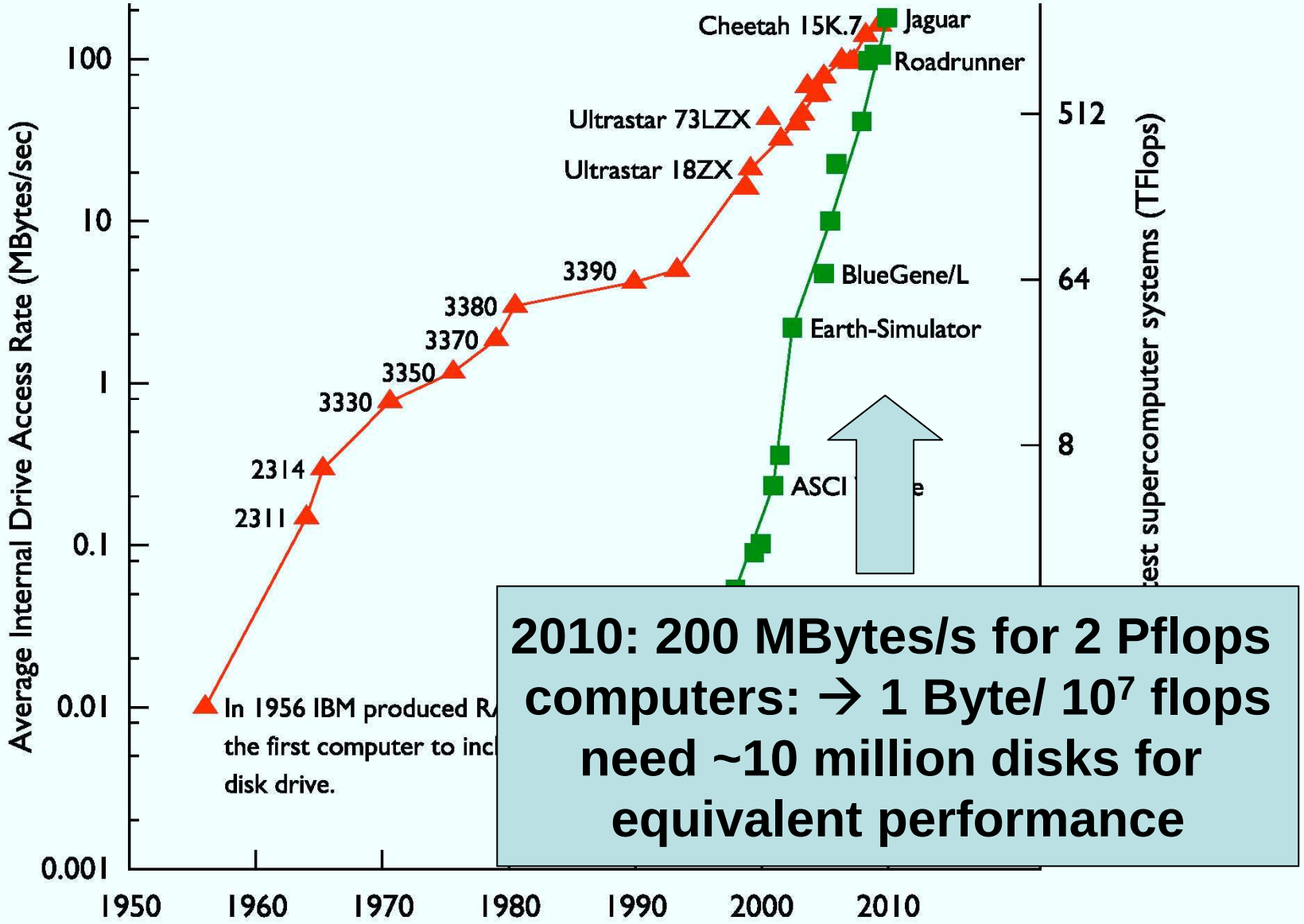


Performance Crisis



Rob Ross' talk from SNAPI 2010 (http://personals.ac.upc.edu/toni/SNAPI2010/snapi2010_Program.htm)

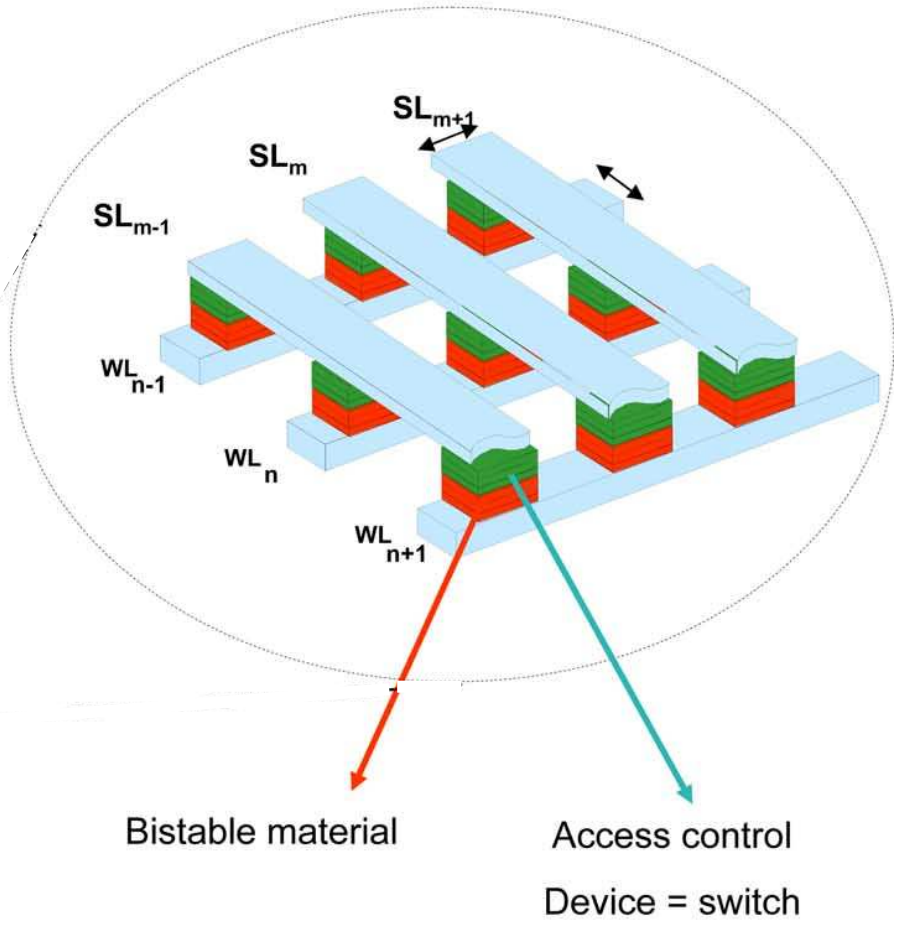
Performance Crisis



Rob Ross' talk from SNAPI 2010 (http://personals.ac.upc.edu/toni/SNAPI2010/snapi2010_Program.htm)

Solution to Problem of Memory/Storage Latency and Bandwidth: Storage Class Memory: fast access *and* large volume

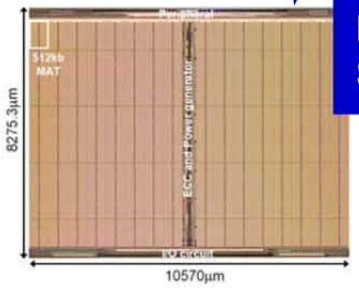
- 2D-array of memory elements
- Bistable material at array cross-points (non-volatile memory)



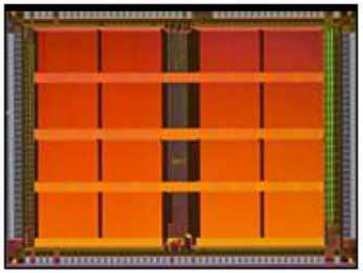
Emerging Memory Technologies

Memory technology remains an active focus area for the industry

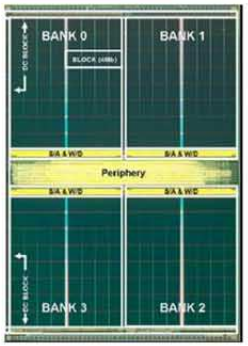
FLASH Extension	FRAM Ferro-electric RAM	MRAM Magnetic RAM	PCRAM Phase Change RAM	RRAM Resistive RAM	Solid Electrolyte	Polymer/ Organic
Trap Storage	Ramtron	IBM	Ovonyx	IBM	Axon	Spansion
Saifun *NROM*	Fujitsu	Infineon	BAE	Sharp	Infineon	Samsung
Tower	STMicro	Freescale	Intel	Unity		TFE
Spansion	TI	Philips	STMicro	Spansion		MEC
Infineon	Toshiba	STMicro	Samsung	Samsung		Zettacore
Macronix	Infineon	HP	Elpida			Roltronics
Samsung	Samsung	NVE	IBM			Nanolayer
Toshiba	NEC	Honeywel	Macronix			
Spansion	Hitachi	Toshiba	Infineon			
Macronix	Rohm	NEC	Hitachi			
NEC	HP	Sony	Philips			
Nano-x'tal	Cypress	Fujitsu				
Freescale	Matsushita	Renesas				
Matsushita	Oki	Samsung				
	Hynix	Hynix				
	Celis	TSMC				
	Fujitsu					
	Seiko Epson					



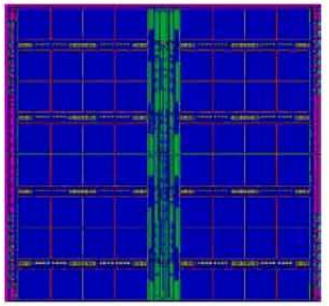
64Mb FRAM (Prototype)
0.13um 3.3V



4Mb MRAM (Product)
0.18um 3.3V



512Mb PRAM (Prototype)
0.1um 1.8V



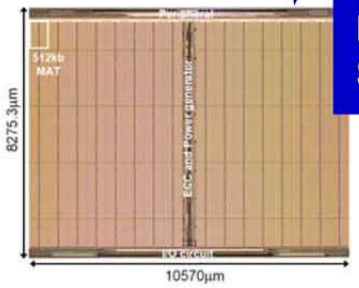
4Mb C-RAM (Product)
0.25um 3.3V

Wilcke 08

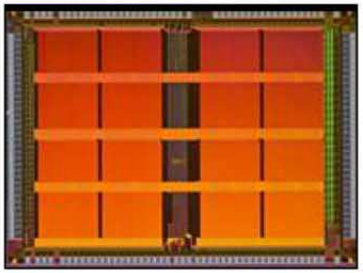
Emerging Memory Technologies

Memory technology remains an active focus area for the industry

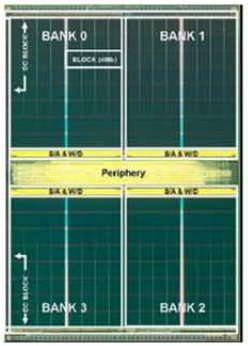
FLASH Extension	FRAM Ferro-electric RAM	MRAM Magnetic RAM	PCRAM Phase Change RAM	RRAM Resistive RAM	Solid Electrolyte	Polymer/ Organic
Trap Storage	Ramtron	IBM	Ovonyx	IBM	Axon	Spansion
Saifun *NROM*	Fujitsu	Infineon	BAE	Sharp	Infineon	Samsung
Tower	STMicro	Freescale	Intel	Unity		TFE
Spansion	TI	Philips	STMicro	Spansion		MEC
Infineon	Toshiba	STMicro	Samsung	Samsung		Zettacore
Macronix	Infineon	HP	Elpida			Roltronics
Samsung	Samsung	NVE	IBM			Nanolayer
Toshiba	NEC	Honeywel	Macronix			
Spansion	Hitachi	Toshiba	Infineon			
Macronix	Rohm	NEC	Hitachi			
NEC	HP	Sony	Philips			
Nano-x'tal	Cypress	Fujitsu				
Freescale	Matsushita	Renesas				
Matsushita	Oki	Samsung				
	Hynix	Hynix				
	Celis	TSMC				
	Fujitsu					
	Seiko Epson					



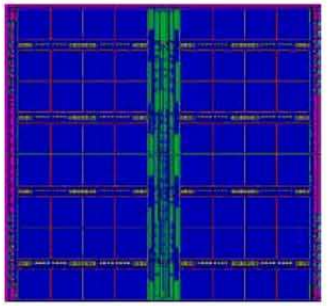
64Mb FRAM (Prototype)
0.13um 3.3V



4Mb MRAM (Product)
0.18um 3.3V



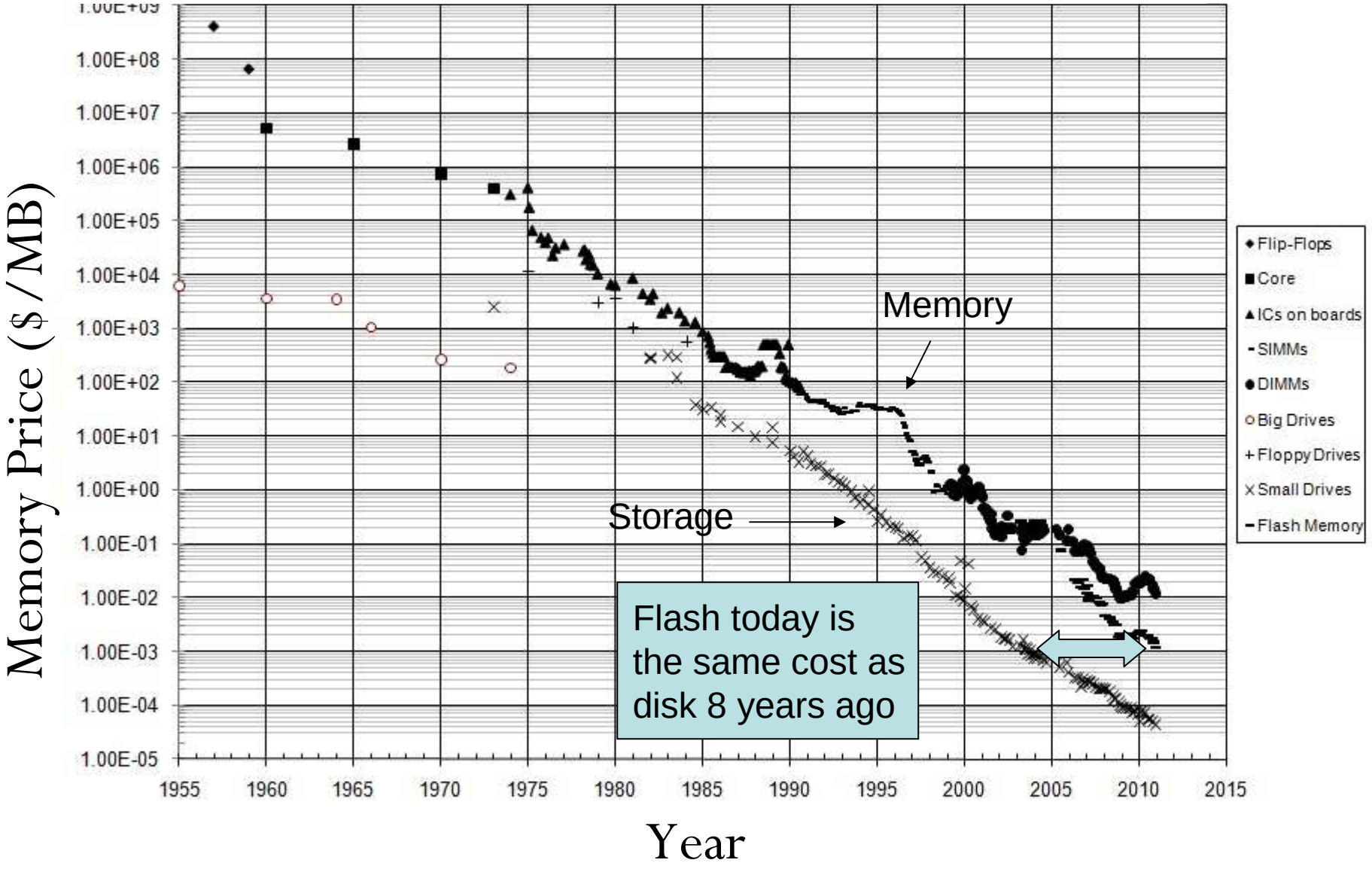
512Mb PRAM (Prototype)
0.1um 1.8V



4Mb C-RAM (Product)
0.25um 3.3V

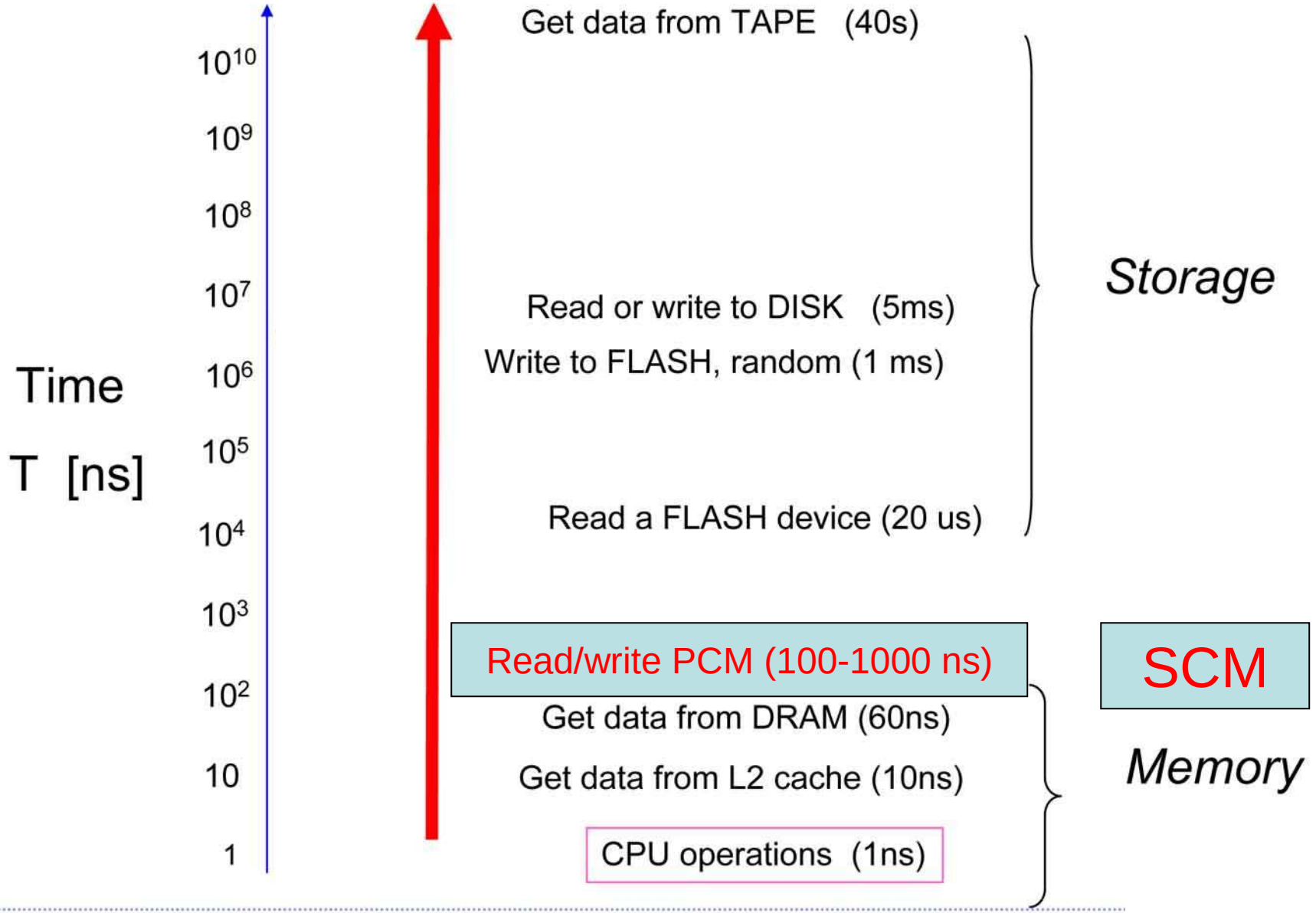
Wilcke 08

Historical Cost of Computer Memory and Storage

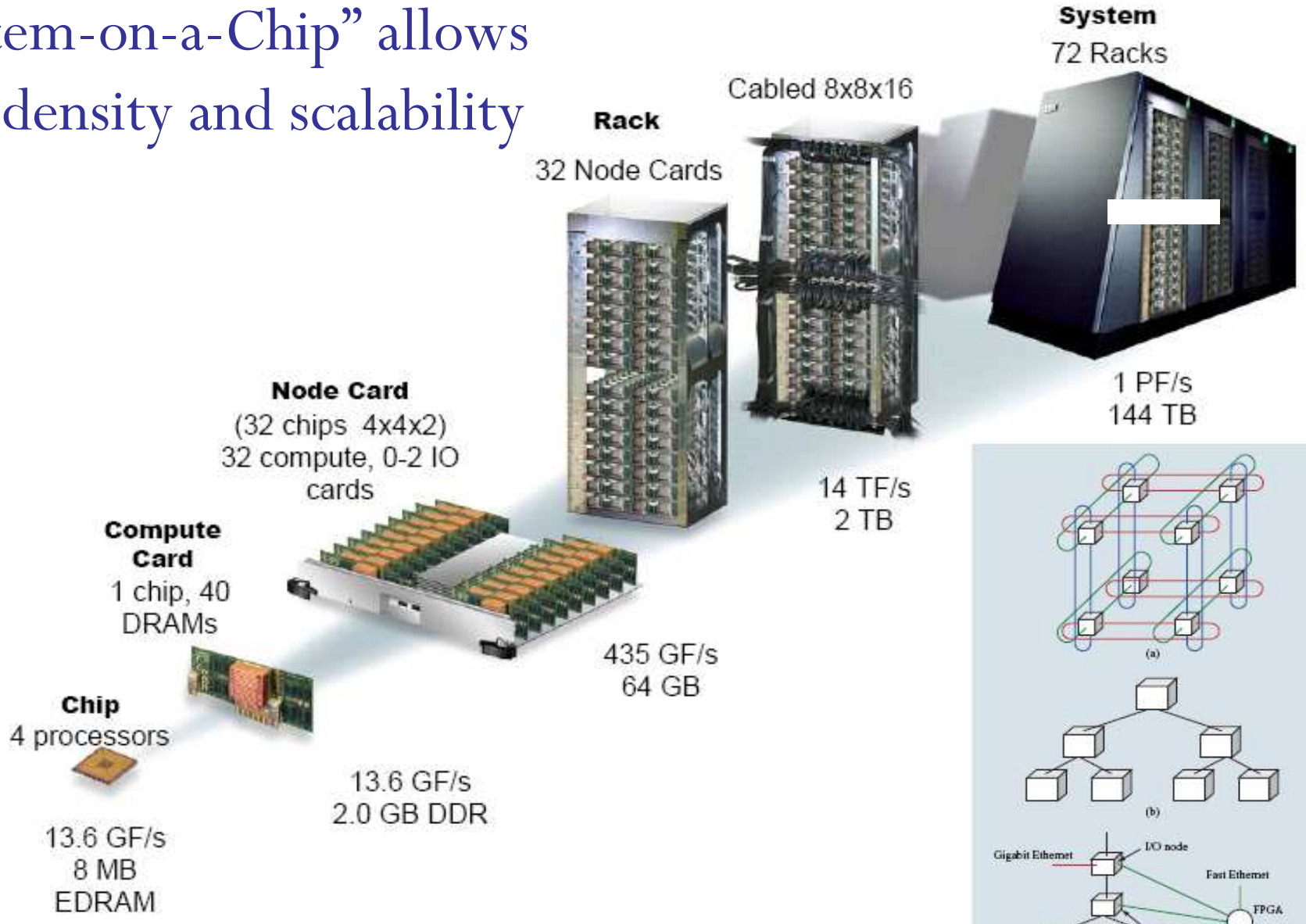


http://www.jcmit.com/mem2010.htm

A New Paradigm for Storage: Storage Class Memory

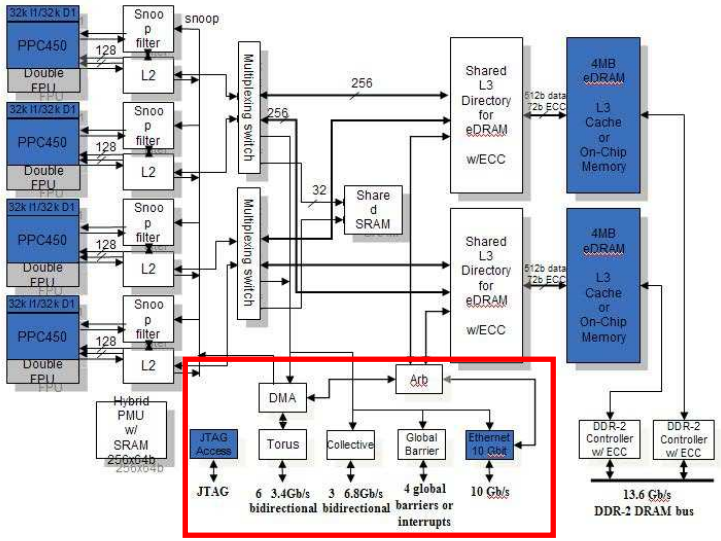


The Blue Gene Design: "System-on-a-Chip" allows high density and scalability

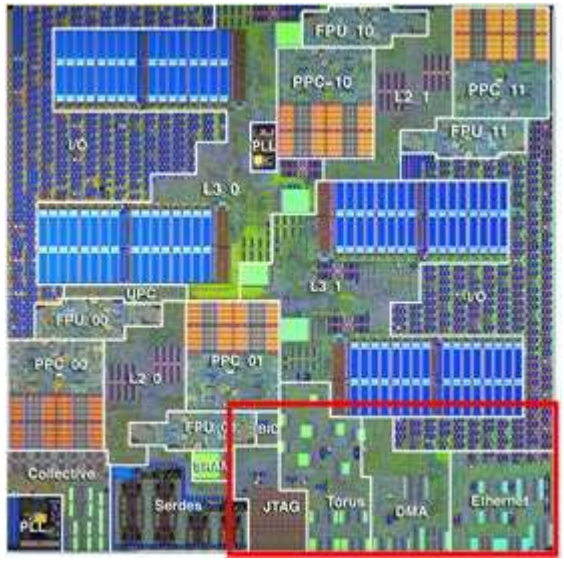


BG/P System-on-a-Chip Networks

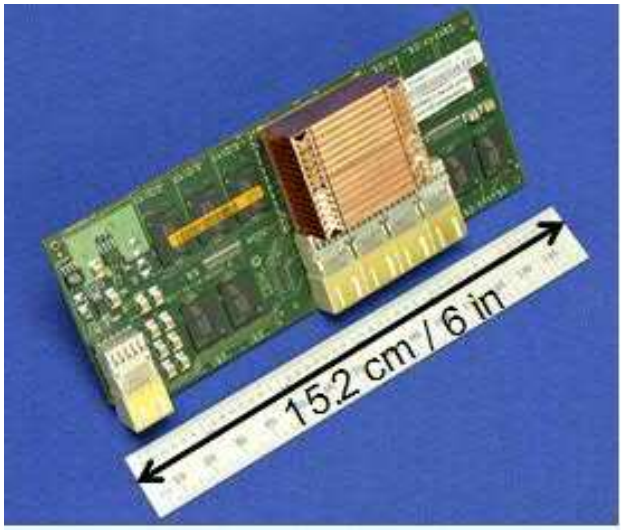
Logic Diagram



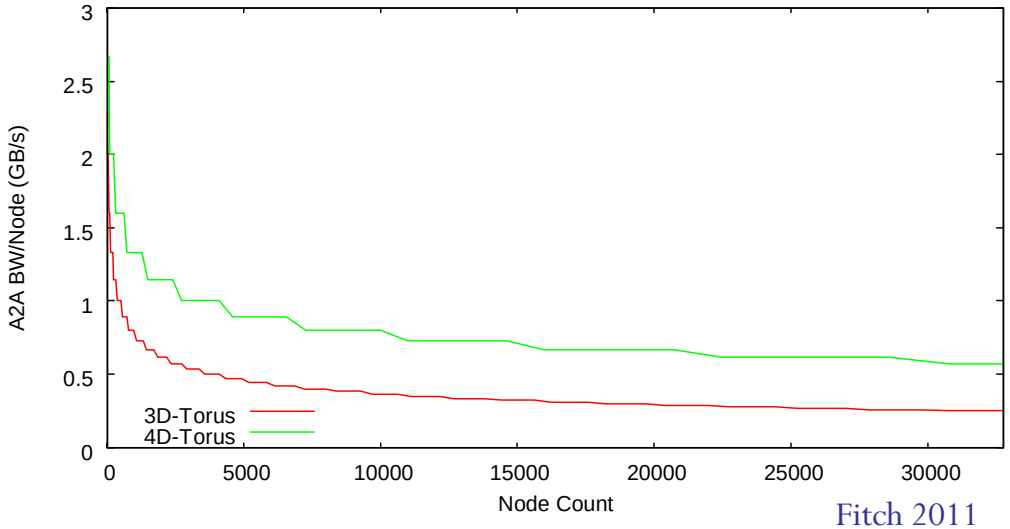
Physical Layout



Packaged Node

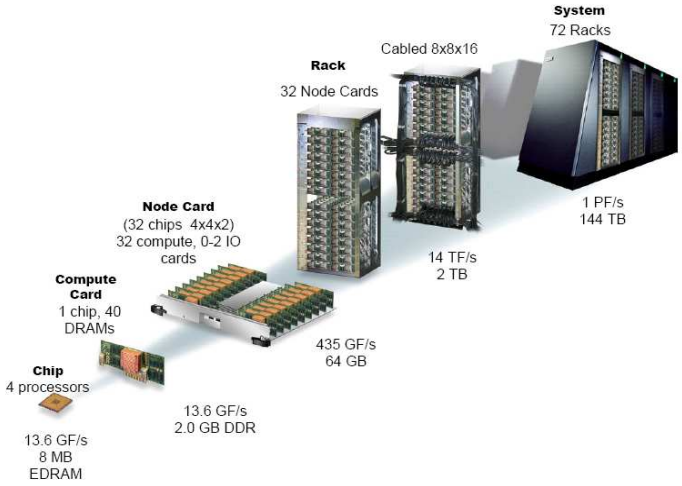


- Cost scales linearly with number of nodes
- Bisectional BW rises with system size
- A Torus with 1GB/s links yields theoretical peak all-to-all bandwidth of:
  - 0.5-1GB/s at 4k nodes
    - (3D:16x16x16; 4D:8x8x8x8)
- 1 GB/s is feasible input to a SCM card

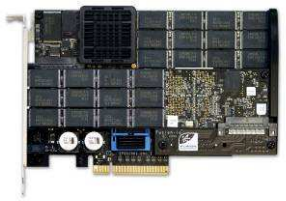


Fitch 2011

Blue Gene + SCM → BG Active Storage Concept



PCIe Flash Board Repackaged
(example: Fusion-io)

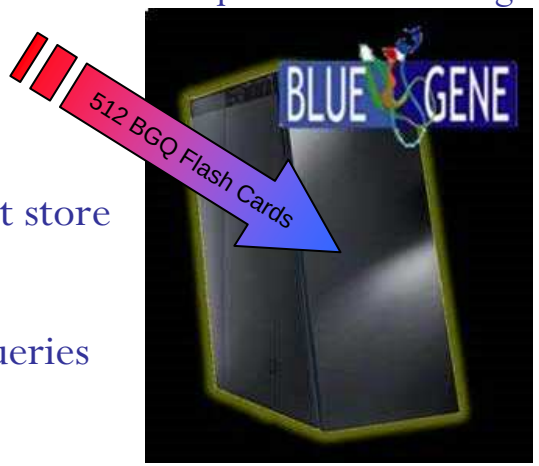


Single Fusion-io Card

ioDrive Duo SLC	Today
Flash Capacity	320 GB
I/O Bandwidth	1.5 GB/s
IOPS	207,000

(1 disk ~ 100 IOPS, 50-100 MB/s)

Integrated in BG to give scalable computation and storage



All-to-all throughput roughly Equivalent to SCM bandwidth

Blue Gene Active Storage Rack

Nodes	512
Storage Cap	640 TB*
I/O Bandwidth	768 GB/s
Random IOPS	100 Million

* Assumes a 4- fold added Flash capacity for 2012 system.

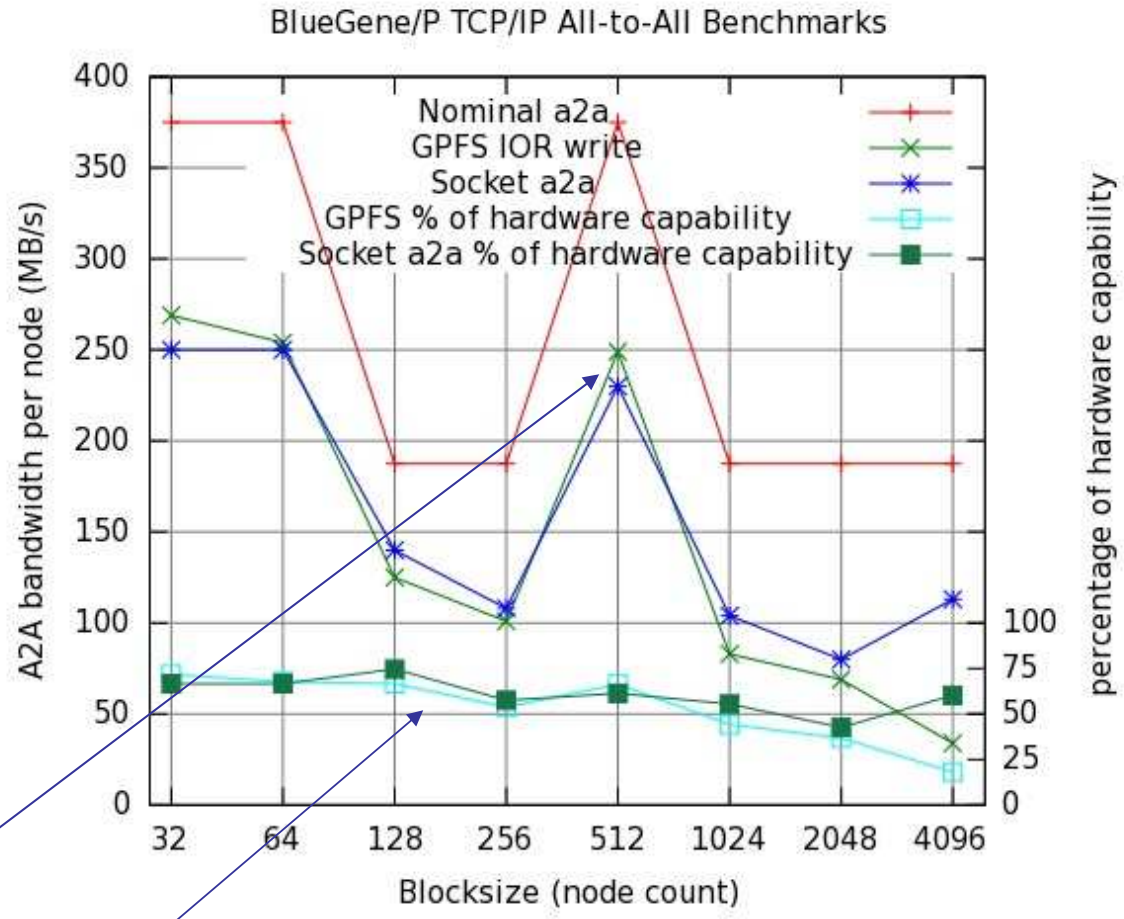
Sample Applications

- High performance shared file system or object store
- Graph-based algorithms
- Allows Join, Sort, "order by" & "group by" queries
- Map-Reduce (heavy reduce phase)
- Aggregation operations
  - count(), sum(), min(), max(), avg(), …
- Real-time analytics

Fitch 2011

Prototype using simulation with DRAM

- Environment
  - Linux, TCP/IP on Torus
  - Fraction of 4GB DRAM is simulated disk: RamDisk
  - GPFS takes ~1 GB from each node and creates a shared file system
- Tests
  - All-to-all TCP/IP socket BW
  - IOR Standard Benchmark
    - all nodes do large contiguous writes – tests A2A BW
- Results
  - IOR into GPFS RamDisk measured on ½ rack (512 nodes) yields 120GB/s write (250MB/s per node)
  - GPFS and TCP/IP socket tests yield 50-60% of network all-to-all bandwidth



SUM: By 2018, Exascale HPC and Big Data

- Disk options look infeasible:
  - Exascale I/O requirements: 60TB/s I/O Bandwidth
    - equivalent to 60,000 ports of 10 GbE
  - Data analytics require high-speed random access to memory/storage
- Storage Class Memory (PCM, etc) solution:
  - "Solid State Disks" with SCM chips provide ~ GB/s I/O bandwidth each
  - can fit on internode network for network-speed memory
    - 1 Byte / 10's of flops
  - SCM much faster, more reliable, and lower-power than disks
- Total power/reliability/speed gain with storage inside rack

THE END