Lecture 21

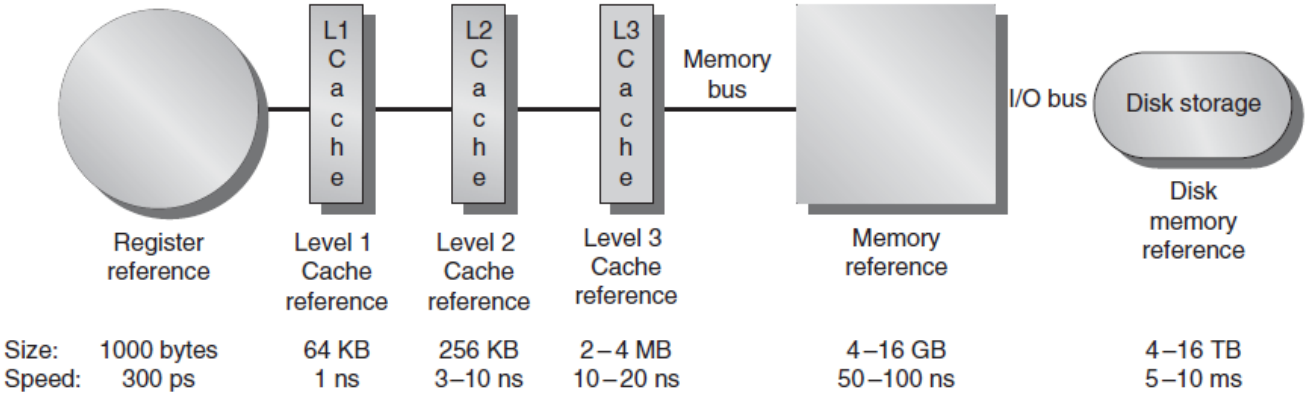
Computer Architecture
(Spring 2020)

Memory Hierarchy Design

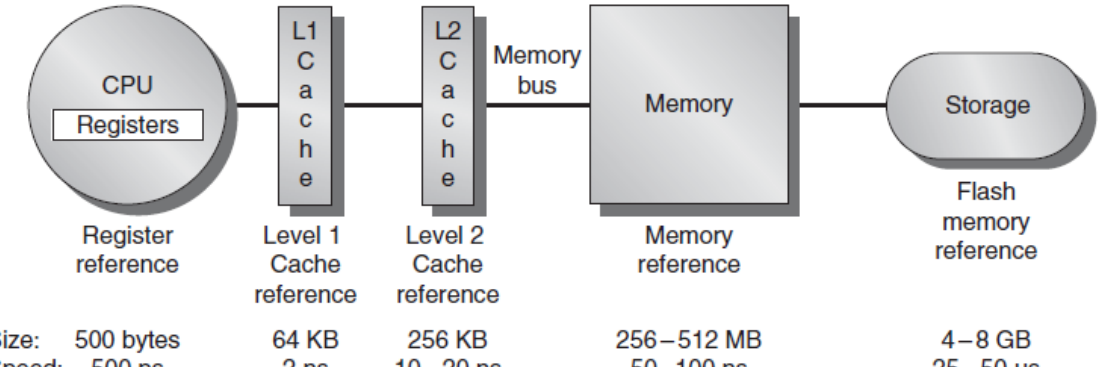
Dr. Duo Liu (刘铎)
Office: Main Building 0626
Email: liuduo@cqu.edu.cn

The Principle of Locality

- Most programs do not access all code or data uniformly
  - Locality occurs in time (temporal locality) and in space (spatial locality)
- Hardware limitations: Faster → Smaller → More Expensive.



(a) Memory hierarchy for server

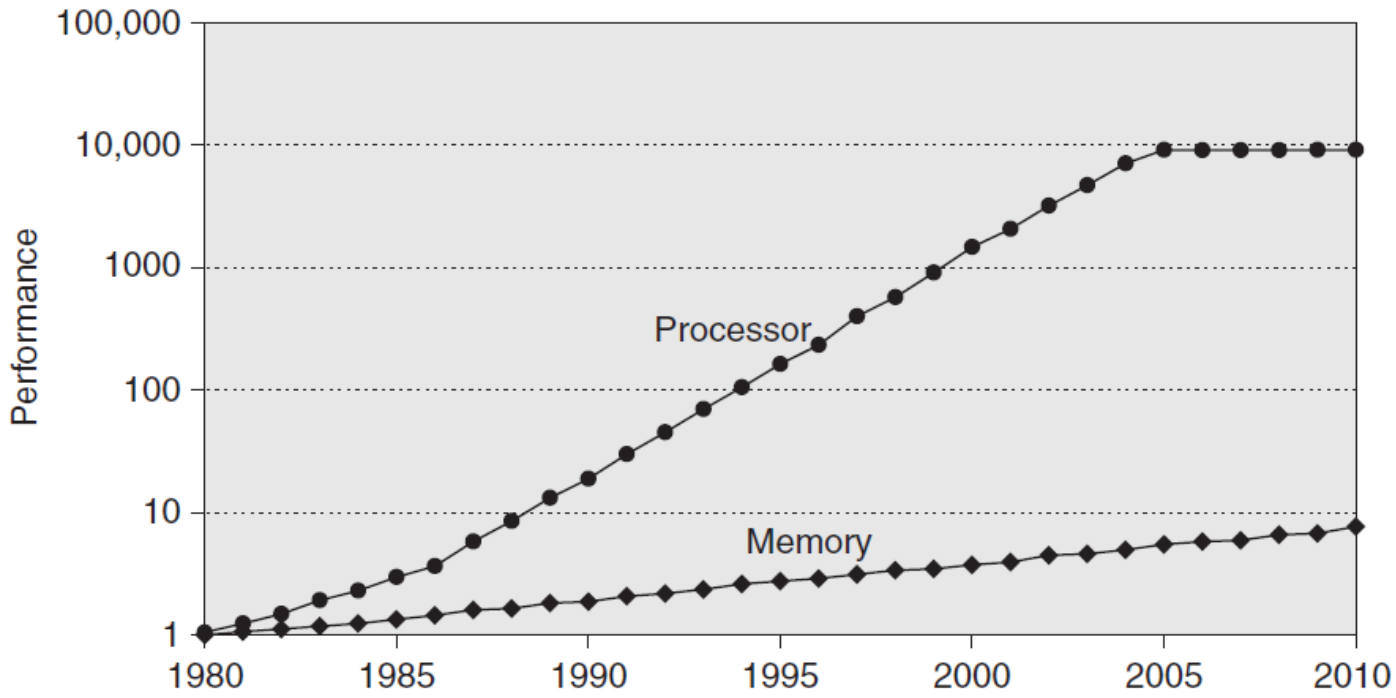


(b) Memory hierarchy for a personal mobile device

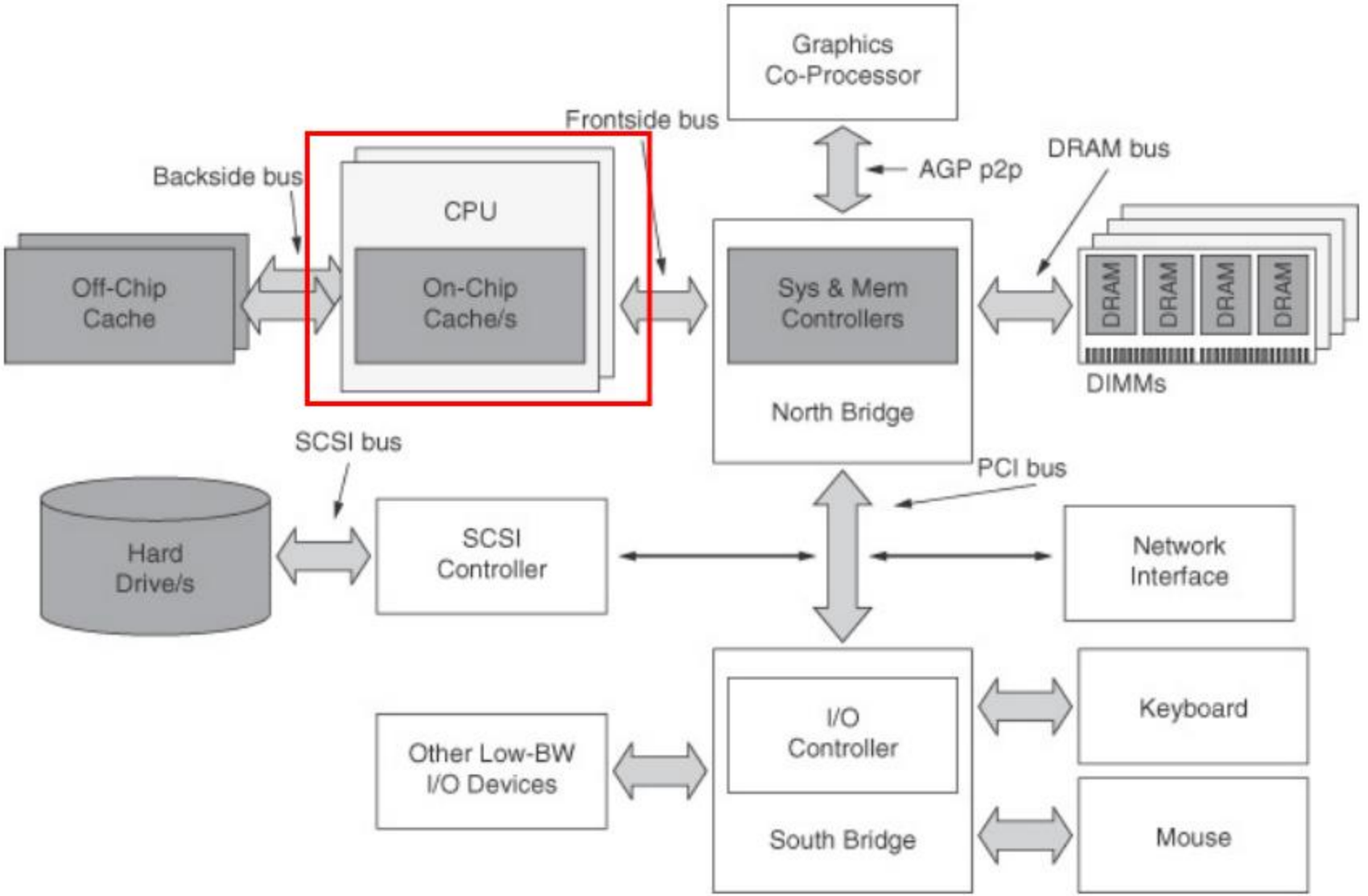
Copy
All ri

Memory Wall + Power Wall

- The increase of CPU speed is faster than Memory bandwidth.
- It is getting even worse when we are enter the world of multi-processors
- Cache consumes significant power.
  - Could be 25% to 50% of total CPU power.

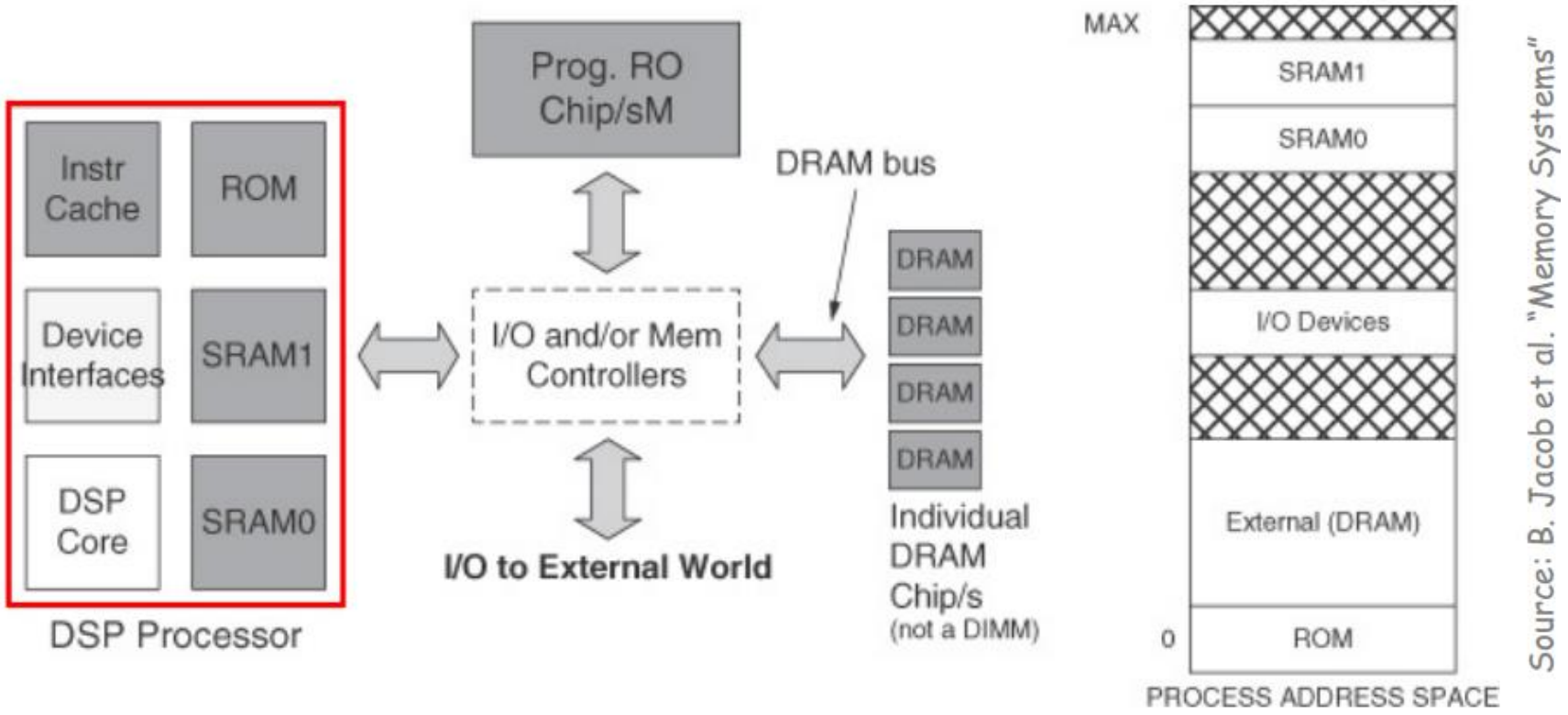


Typical PC Organization



Source: B. Jacob et al. "Memory Systems"

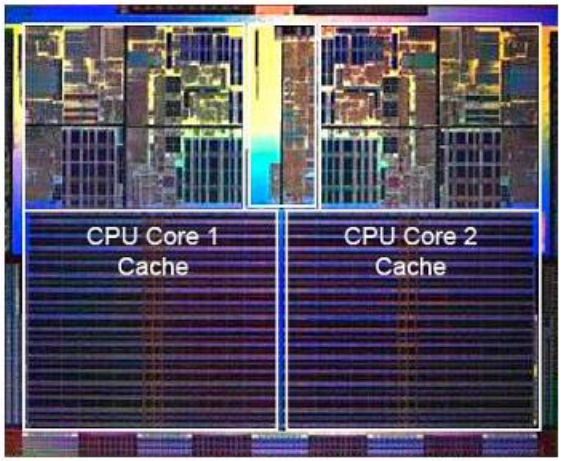
DSP-Style Memory System:
Example based on TI TMS320C3x DSP family

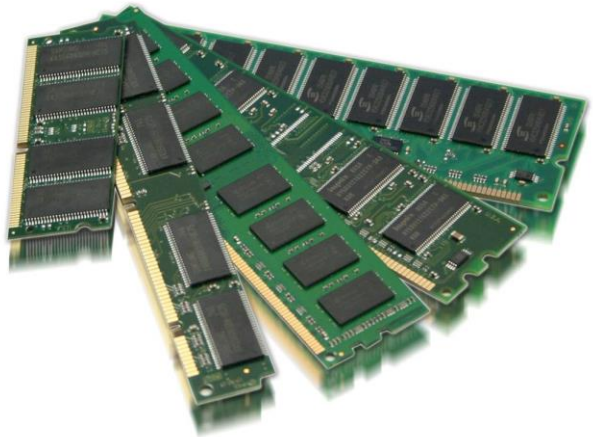


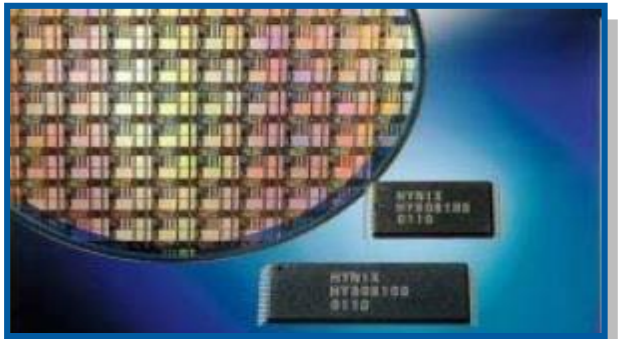
- **dual tag-less on-chip SRAMs (visible to programmer)**
- **off-chip programmable ROM (or PROM or FLASH) that holds the executable image**
- **off-chip DRAM used for computation**

Memory Technology

- **At the core of the success of computers**
- **Various types of memory**
  - **– most common types**
    - **Dynamic Random-Access Memory (DRAM)**
    - **Static Random-Access Memory (SRAM)**
    - **Read-Only Memory (ROM)**
    - **Flash Memory**
- **Memory Latency Metrics**
  - **– Access time**
    - **time between when a “read” is requested and when the desired word arrives**
  - **– Cycle time (>Access time)**
    - **minimum time between two requests to memory**
    - **memory needs the address lines to be stable between accesses**

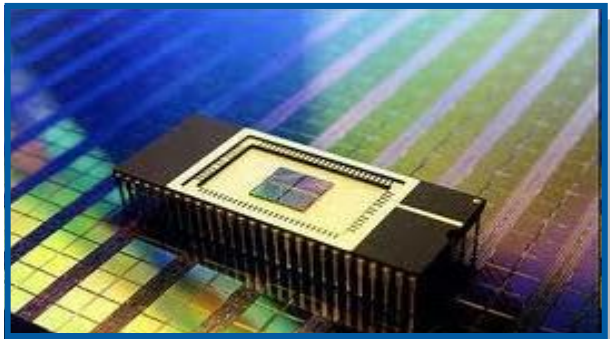




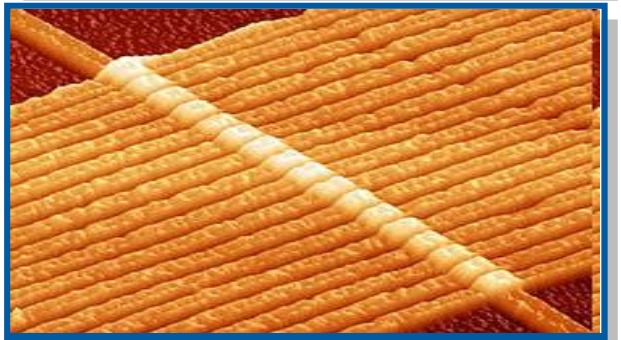


Ferroelectric RAM (FeRAM)
Toshiba, 2009

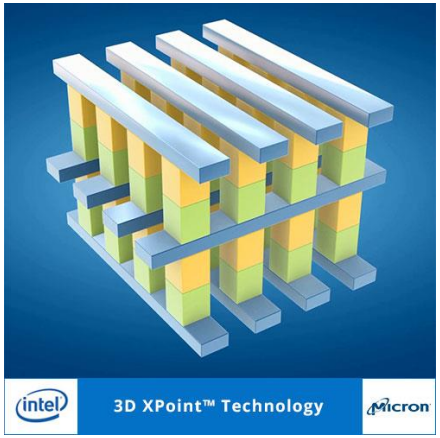
Magnetic RAM (MRAM)
EverSpin, 2008



Phase Change Memory (PCM)
Samsung, 2008



Resistive RAM (Memristor)
HP Lab, 2009



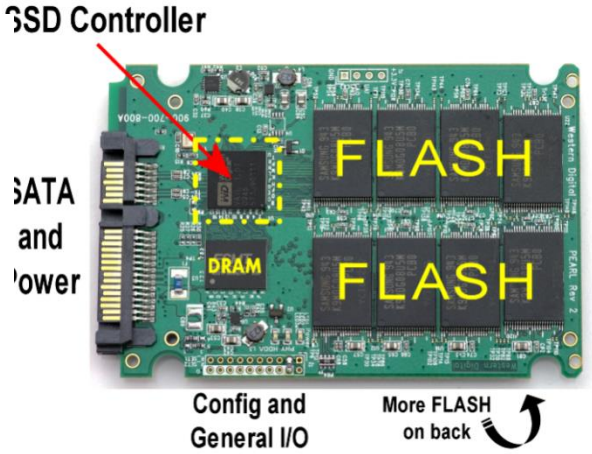

1.44MB
INDEX
Floppy Coaster


boedem
FLOPPY DISK
5.25 HD DS

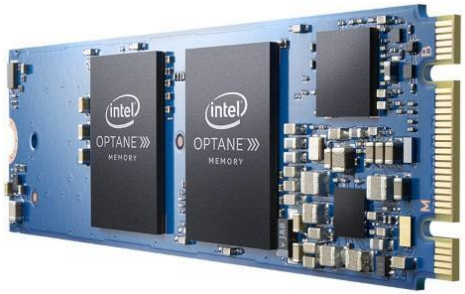

SanDisk


1.0TB

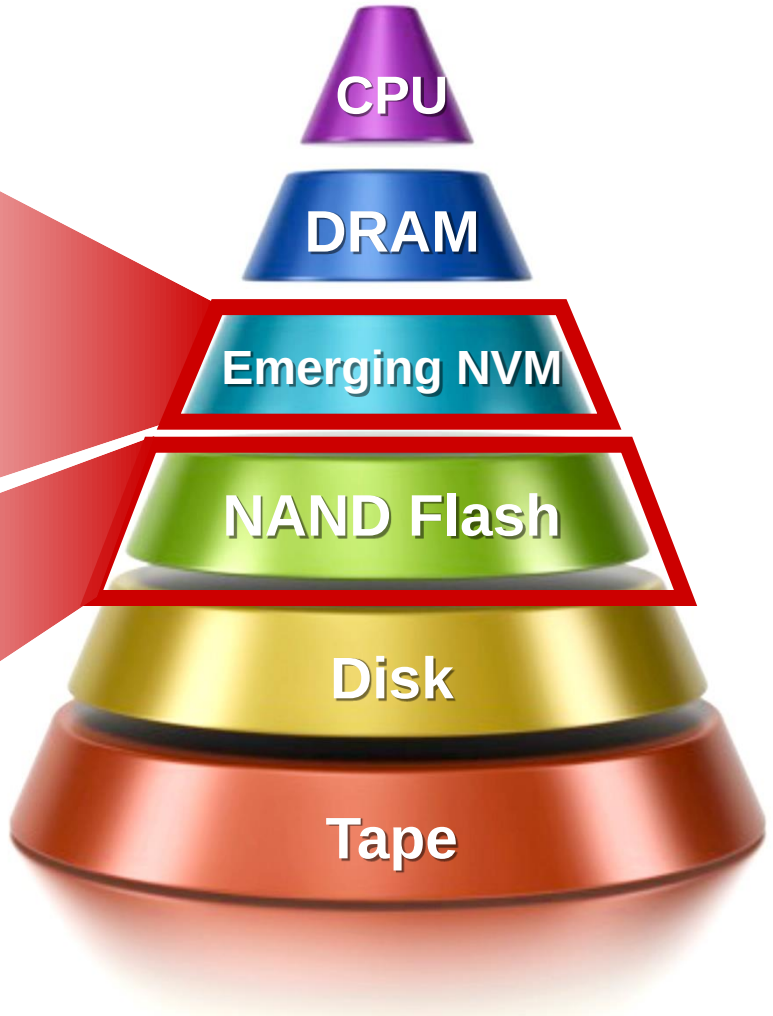

SanDisk
SSD
SATA 5000 2.5"
32GB


SSD Controller
SATA
and
Power
FLASH
DRAM
FLASH
Config and
General I/O
More FLASH
on back


MONDAY
TUESDAY
WEDNESDAY
THURSDAY
FRIDAY
SATURDAY
SUNDAY


intel
OPTANE
MEMORY
intel
OPTANE
MEMORY

新一代非易失性存储技术



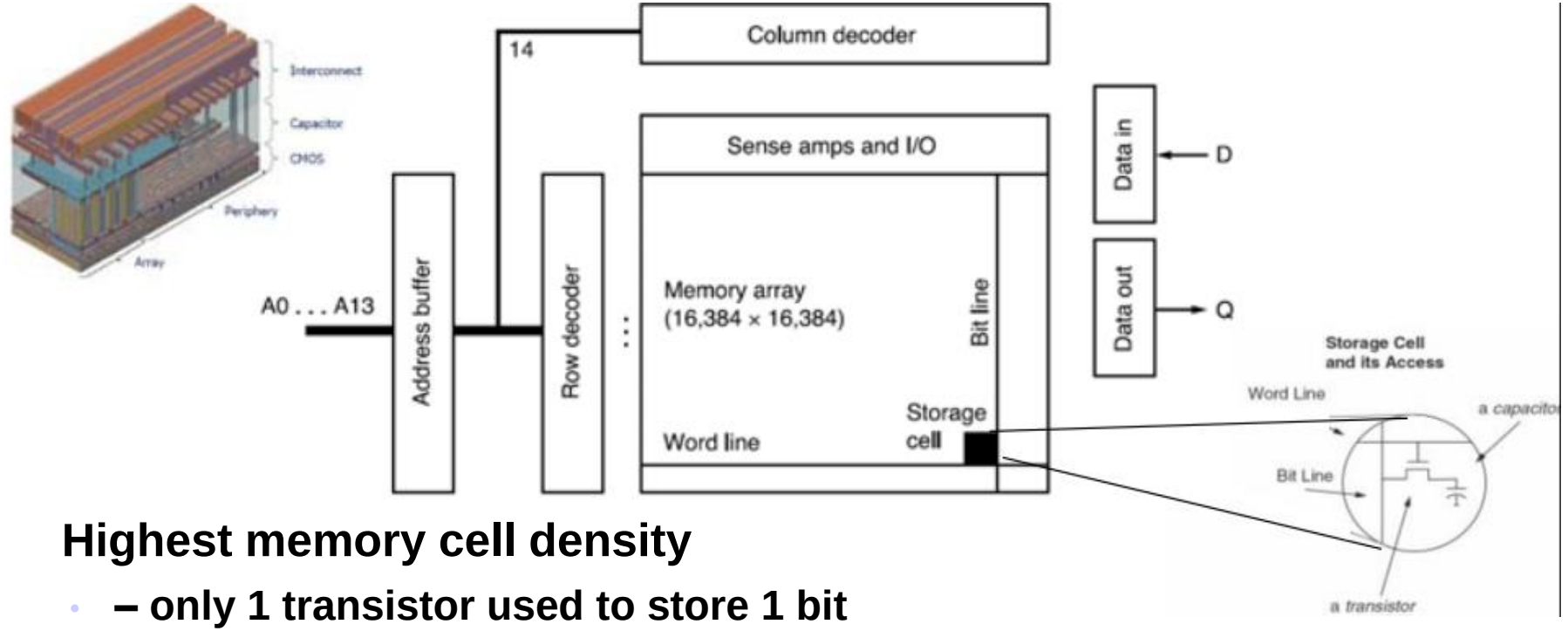
[Courtesy: Yuan Xie (PSU), Glen Hawk (Micron)]

新型存储器件性能比较

	SRAM	DRAM	NOR	NAND	MRAM	PCM	STT-RAM	R-RAM
Data Retention	N	N	Y	Y	Y	Y	Y	Y
Memory Cell Factor (F2)	50-120	6-10	10	2-5	16-40	6-12	4-20	<4
Read Time (ns)	1	30	10	50	3-20	20-50	2-20	<50
Write /Erase Time (ns)	1	50	10^5-10^7	10^6-10^5	3-20	50-120	2-20	<100
Endurance	10^{16}	10^{16}	10^5	10^5	10^{15}	10^6 -10^{10}	10^{15}	10^{15}
Power Consumption – Read/Write	Low	Low	High	High	Med/ High	Low	Low	Low
Power Consumption – Other than R/W	Leakage Current	Refresh Power	None	None	None	None	None	None

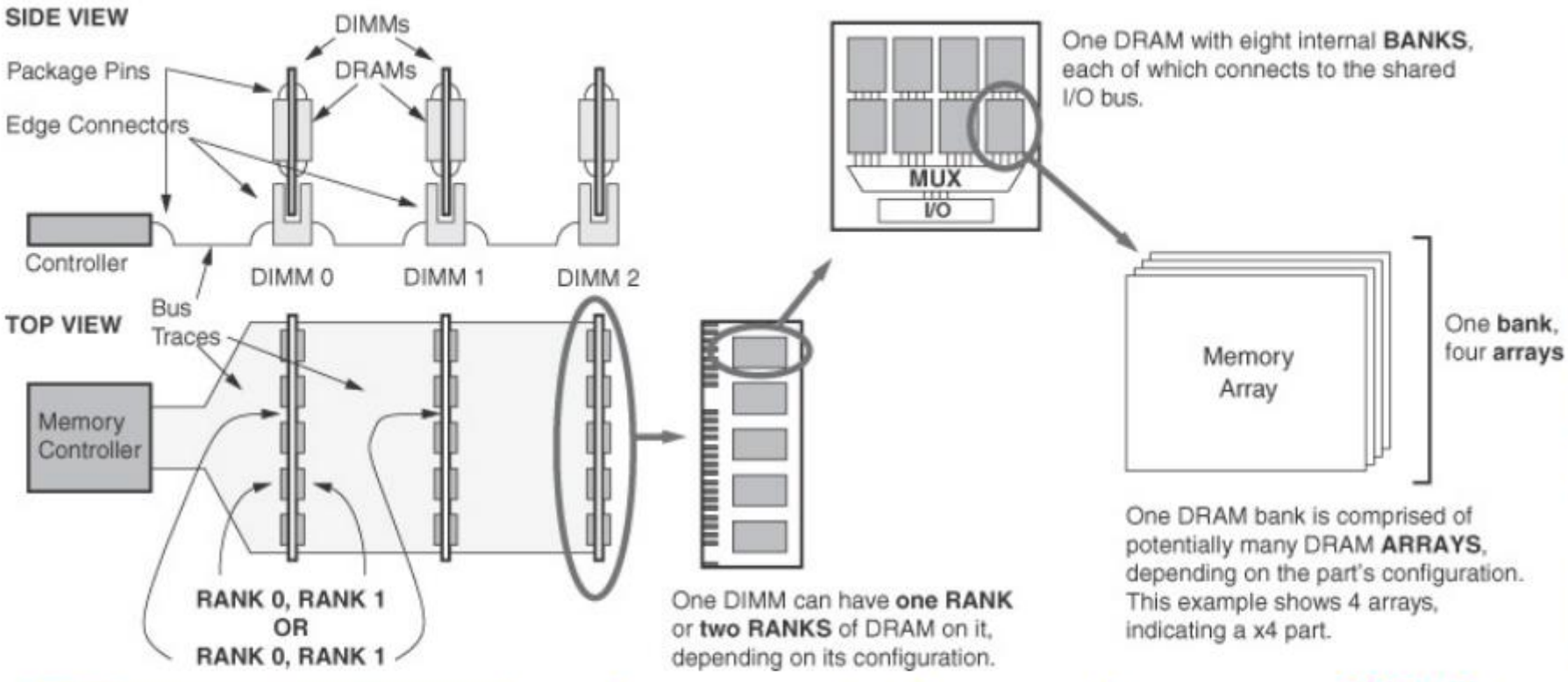
[Source: ITRS]

A 64M-bit DRAM: Logical Organization



- **Highest memory cell density**
  - **– only 1 transistor used to store 1 bit**
  - **– to prevent data loss, each bit must be refreshed periodically**
    - **DRAM access periodically all bits in every row (refresh)**
    - **about 5% of the time a DRAM is not available due to refreshing**
- **To limit package costs, address lines are multiplexed**
  - **e.g., first send 14-bit row address (Row Access Strobe), then 14-bit column address (Column Access Strobe)**

DIMMs, Ranks, Banks, and Arrays



Source: B. Jacob et al. "Memory Systems"

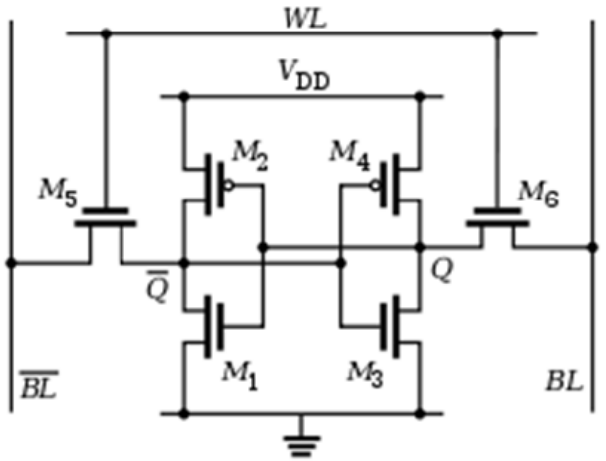
- A memory system may have many DIMMs, each of which may contain one or more *ranks*
- Each rank is a set of engaged DRAM devices, each of which may have many *banks*
- Each bank may have many constituent *arrays*, depending on the part's data width

DRAM Generations

Year of Introd.	Chip Size (bit)	$ per GB	Total Access Time to a new row/column	Total Access Time to existing row
1980	64K	1,5M	250ns	150ns
1983	256K	500k	185ns	100ns
1985	1M	200k	135ns	40ns
1989	4M	50k	110ns	40ns
1992	16M	15k	90ns	30ns
1996	64M	10k	60ns	12ns
1998	128M	4k	60ns	10ns
2000	256M	1k	55ns	7ns
2004	512M	250	50ns	5ns
2007	1G	50	40ns	1.25ns

SRAMs

- **SRAM memory cell is bigger than DRAM cell**
  - **typically 6 transistors per bit**
- **Better for low-power applications thanks to stand-by mode**
  - **only minimal power is necessary to retain charge in stand-by mode**
- **Access Time = Cycle Time**
  - **Address lines are not multiplexed (for speed)**
- **In comparable technologies…**
  - **SRAM has a only 1/4 - 1/8 of DRAM capacity**
  - **SRAM cycle time is 8-16 times faster than DRAM**
  - **SRAM cost-per-bit is 8-16 times more expensive than DRAM**



ROM and Flash Memory

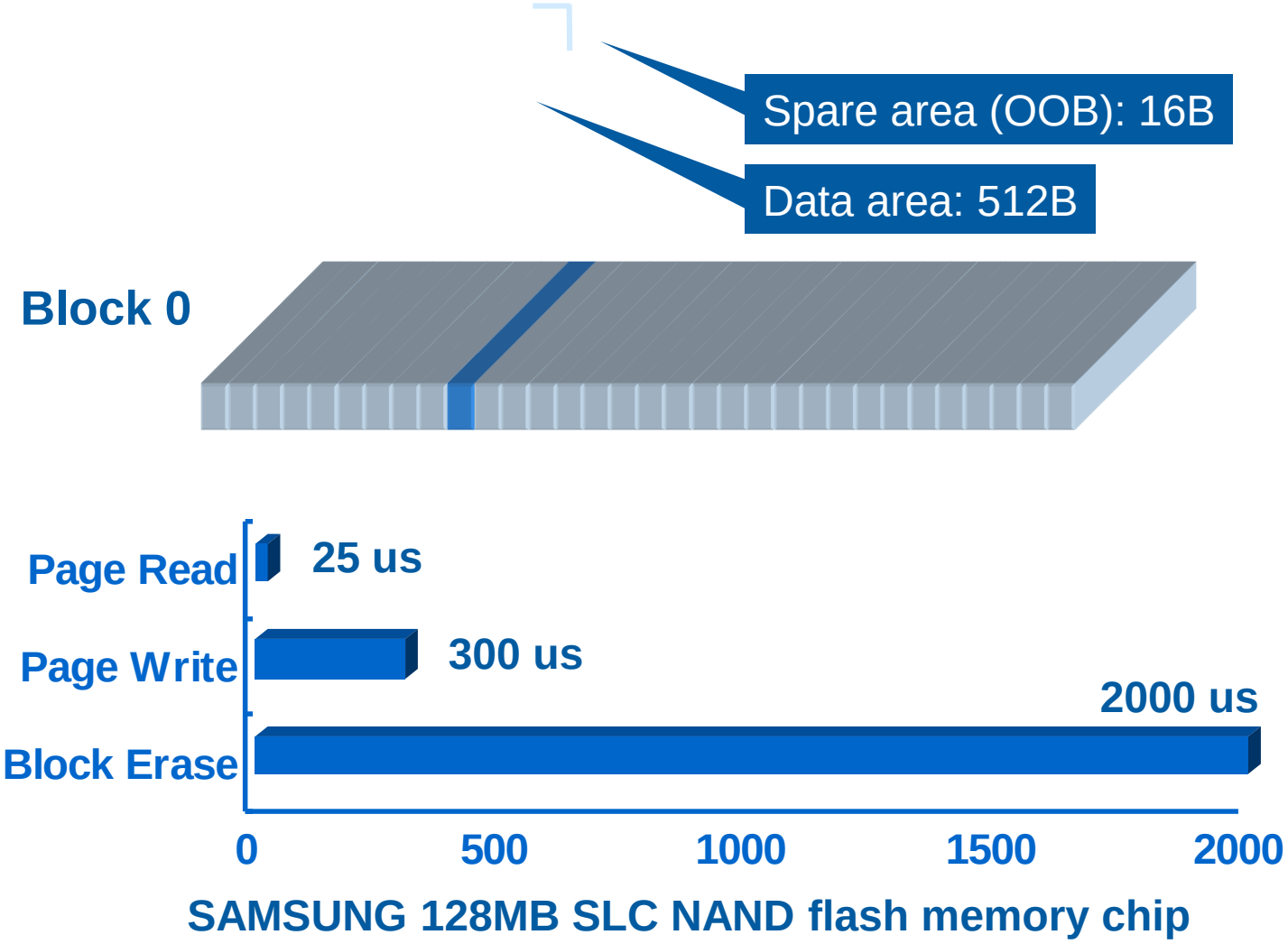
- **ROM**
  - **– programmed once and for all at manufacture time**
  - **– cannot be rewritten by microprocessor**
  - **– 1 transistor per bit**
  - **– good for storing code and data constants in embedded applications**
  - **• replace magnetic disks in providing nonvolatile storage**
  - **• add level of protection for embedded software**
- **Flash Memories**
  - **– floating-gate technology**
  - **– read access time comparable to DRAMs**
  - **• 50-100us depending on size (16M-128M)**
  - **– write is 10-100 slower than DRAMs (plus erasing time 1-2ms)**
  - **– price is cheaper than DRAM but more expensive than magnetic disks**
  - **• Flash: $2/GB , DRAM: $40/GB; disk = $0.09/GB**
  - **– Initially, mostly used for low power/embedded applications**
  - **• but now also as solid-state replacements for disks**
  - **– or efficient intermediate storage between DRAM and disks**

Why Flash Memory?

- Non-volatility
- Short read/write latency
- Low power consumption
- Small size and light weight
- Solid state reliability

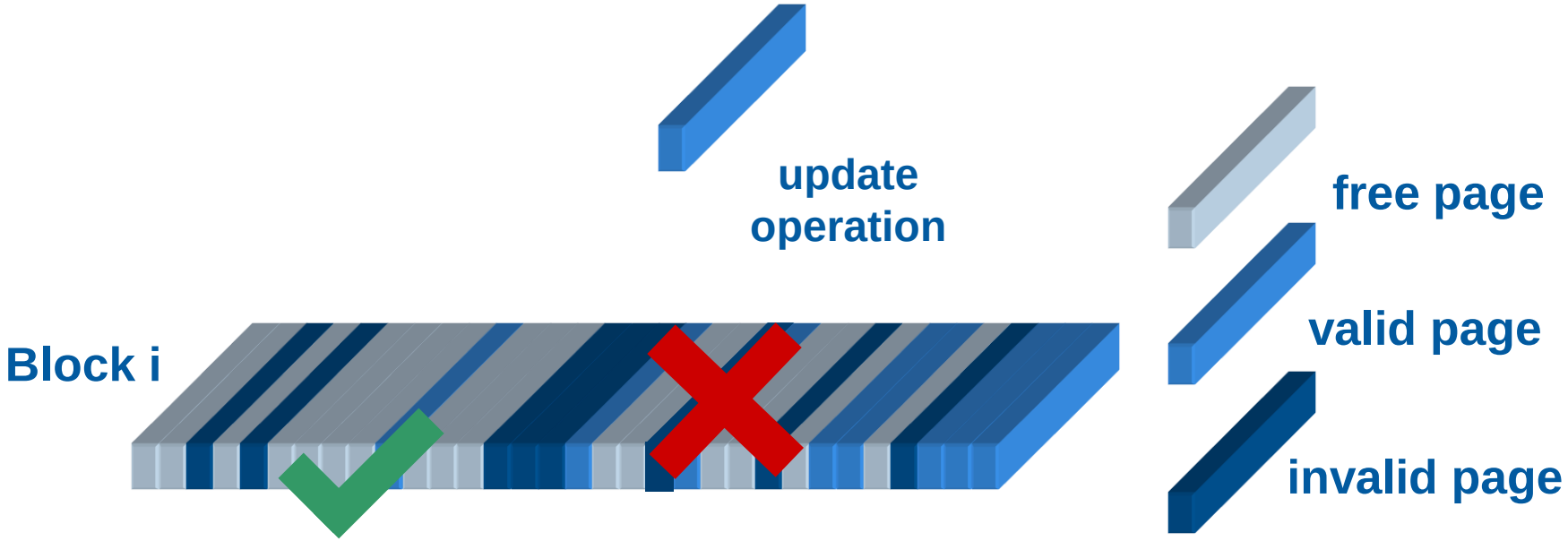
NAND Flash Memory Organization

- **Chip → Block → Page**
  - Block = 32 / 64 pages; Page = 512B + 16B



NAND Flash Memory Constraints

- **Out-of-place update**



- **Limited endurance: 10^4 (MLC) ~ 10^5 (SLC)**
- **Wear leveling**
- **Solution: Flash Translation Layer (FTL)**