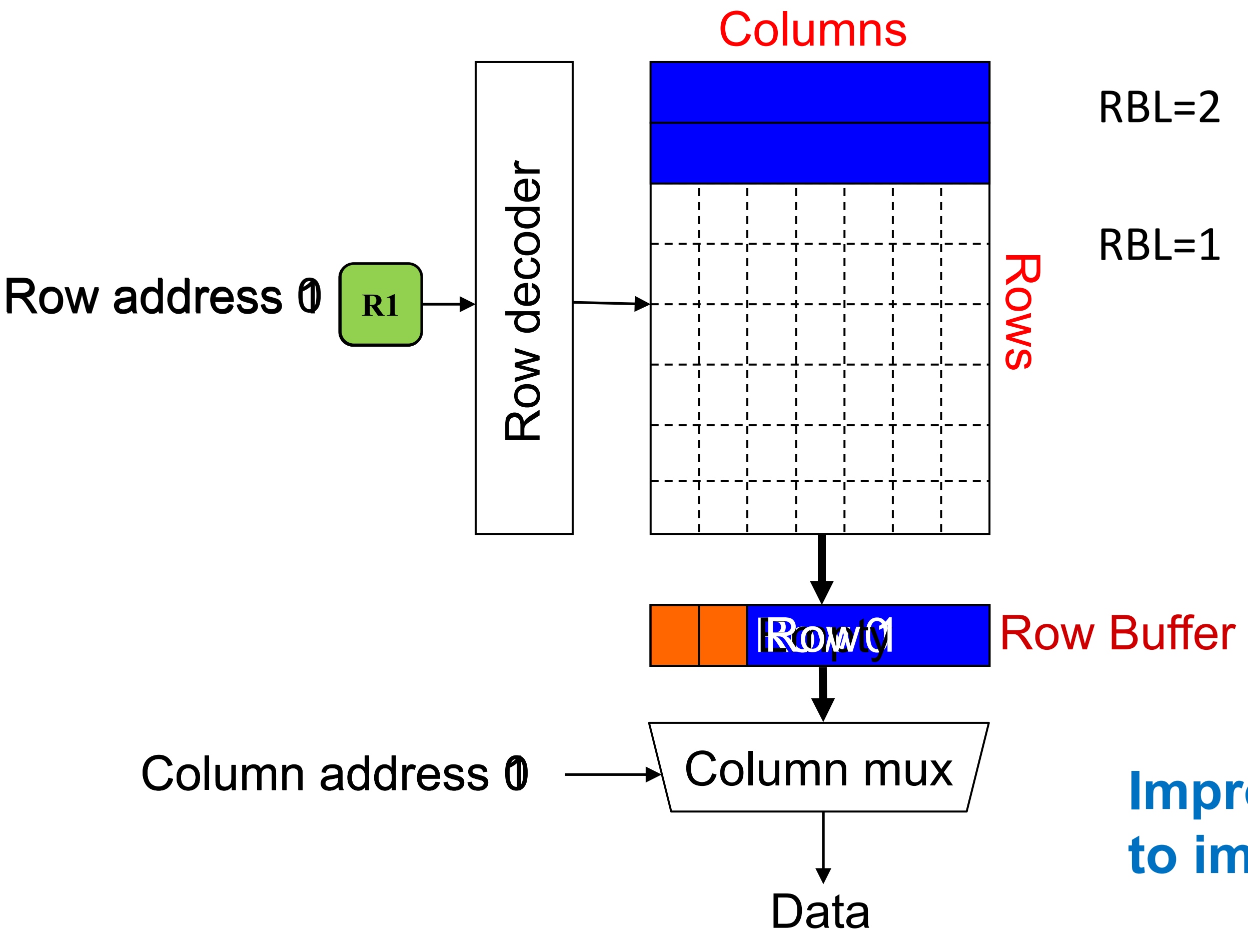
CMPE 200
Computer Architecture & Design

Lecture 4.
Memory Hierarchy (6)

Haonan Wang

SJSU
SAN JOSÉ STATE
UNIVERSITY

Row Operations & Row Buffer Locality



Access Address: **Row Operation:**

RBL=2
- (Row 0, Column 0) — Activation
- (Row 0, Column 1) — No operation

RBL=1
- (Row 1, Column 0) — Restore, Precharge, Activation

CONFLICT !

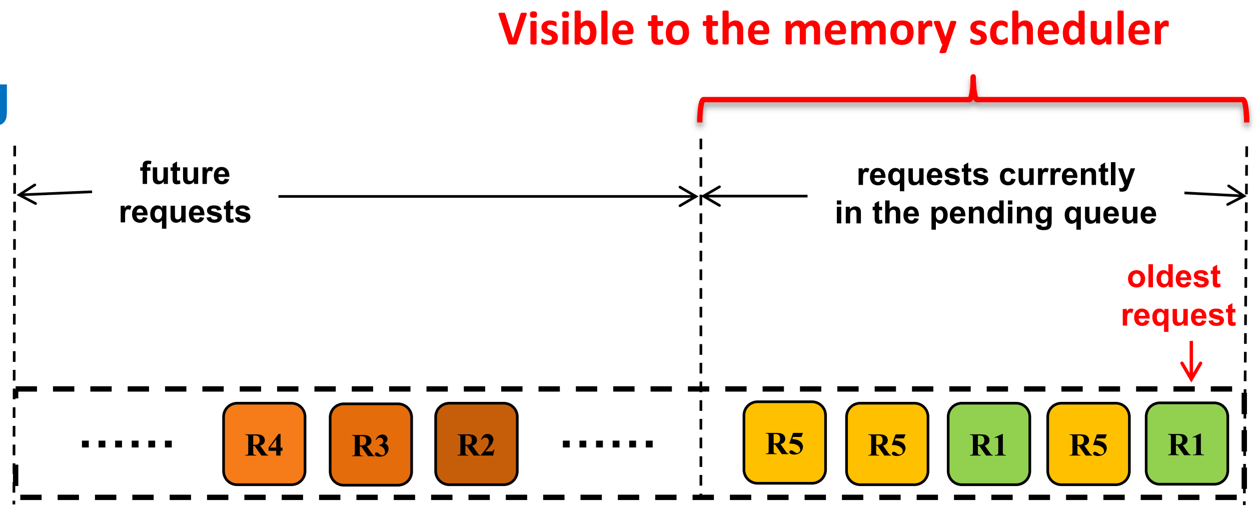
Improving Row Buffer Locality (RBL) is the key to improve DRAM efficiency

Credit: Onur Mutlu.

RBL & Memory Scheduling Schemes

In-order scheduling (FIFO):

Request Stream →

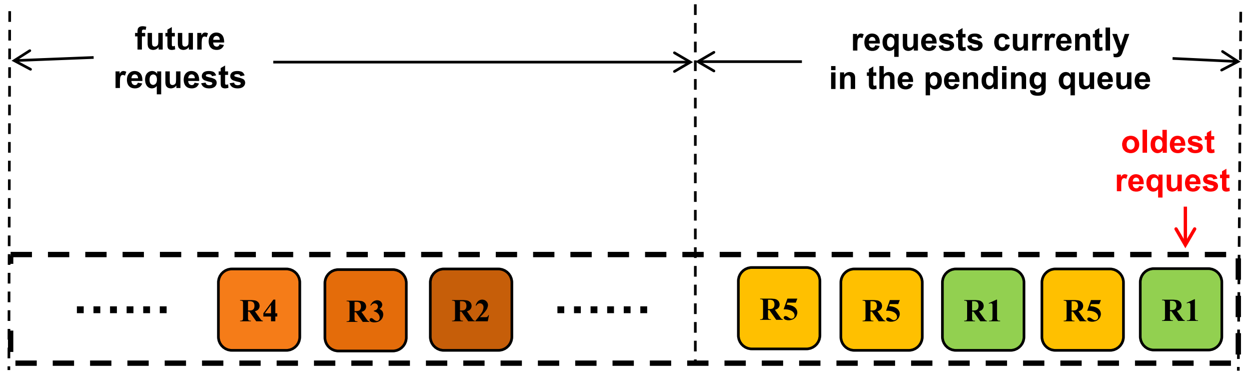


Activation Counter:

R1: Activation = 1
R5: Activation = 2
R1: Activation = 3
R5: Activation = 4
R5: Activation = 4
} Same activation
Avg RBL = 5 / 4 = 1.25

Out-of-order scheduling (FR-FCFS):

Request Stream →



Activation Counter:

R1: Activation = 1
R1: Activation = 1
} Same activation
R5: Activation = 2
R5: Activation = 2
R5: Activation = 2
} Same activation
Avg RBL = 5 / 2 = 2.5

DRAM Milestones

	DRAM	Page DRAM	Page DRAM	Page DRAM	SDRAM	DDR SDRAM
Module Width	16b	16b	32b	64b	64b	64b
Year	1980	1983	1986	1993	1997	2000
Mb/chip	0.06	0.25	1	16	64	256
Die size (mm^2)	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
BWidth (MB/s)	13	40	160	267	640	1600
Latency (nsec)	225	170	125	75	62	52

- **In the time that the memory to processor bandwidth has more than doubled the memory latency has improved by a factor of only 1.2 to 1.4**

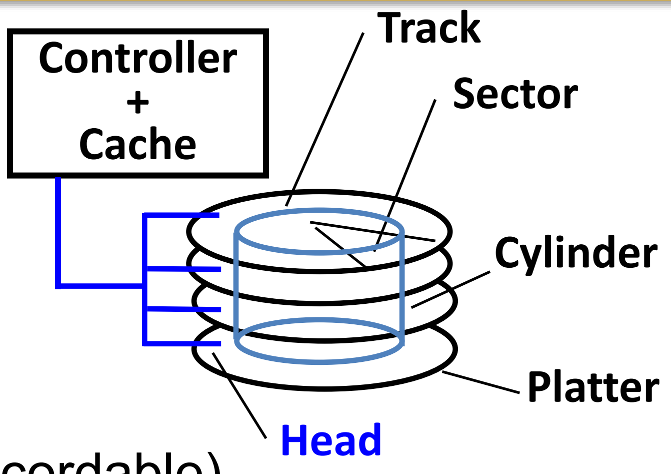
Patterson, CACM Vol 47, #10, 2004

Review: DRAM vs. SRAM

- **DRAM**
  - Slower access (capacitor)
  - Higher density (1T 1C cell)
  - Lower cost
  - Requires refresh (power, performance, circuitry)
  - Manufacturing requires putting capacitor and logic together

- **SRAM**
  - Faster access (no capacitor)
  - Lower density (6T cell)
  - Higher cost
  - No need for refresh
  - Manufacturing compatible with logic process (no capacitor)

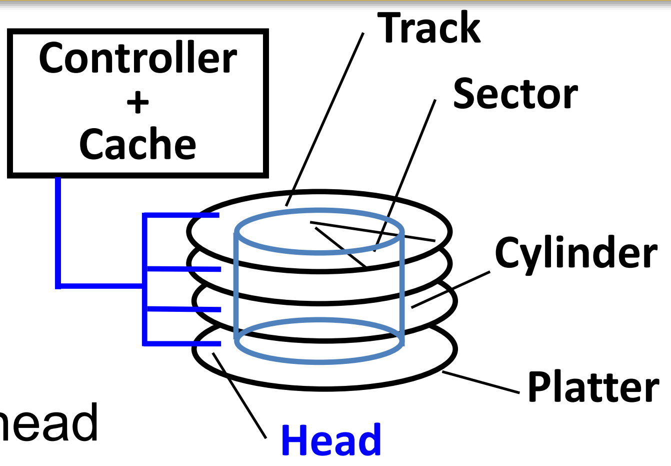
Magnetic Disk



- **Purpose:** Long term, nonvolatile storage
  - Lowest level in the memory hierarchy: large, cheap, slow

- **General structure**
  - 1 to 4 rotating platter coated with a magnetic surface (2 sides recordable)
    - Rotational speeds of 5,400 to 15,000 RPM
  - Moveable read/write head to access the information for each platter
  - 10,000 to 50,000 tracks per surface
    - Cylinder - all the tracks under the head at a given point on all surfaces
  - 100 to 500 sectors per track
    - The smallest unit that can be read/written (typically 512B)
    - Outer tracks can hold more sectors than the inner tracks

Magnetic Disk Characteristic



1. Seek time: position the head over the proper track
   - 3 to 12/15 ms on average
   - Due to locality of disk references, the actual average seek time may be only 25% to 33% of the advertised number

2. Rotational latency: wait for the desired sector to rotate under the head
   - ½ of 1/RPM converted to ms: 0.5R/5400RPM = 5.6ms to 0.5R/15000RPM = 2.0ms

3. Transfer time: transfer a block of bits (one or more sectors) under the head to the disk controller's cache (70 to 125 MB/s are typical disk transfer rates)
   - the disk controller's "cache" takes advantage of spatial locality in disk accesses
   - cache transfer rates are much faster (e.g., 375 MB/s)

4. Controller time: the overhead the disk controller imposes in performing a disk I/O access (typically < 0.2 ms)

Typical Disk Access Time

The average time to read or write a 512B sector for a disk rotating at 15,000 RPM with average seek time of 4 ms, a 100MB/sec transfer rate, and a 0.2 ms controller overhead

Avg disk read/write

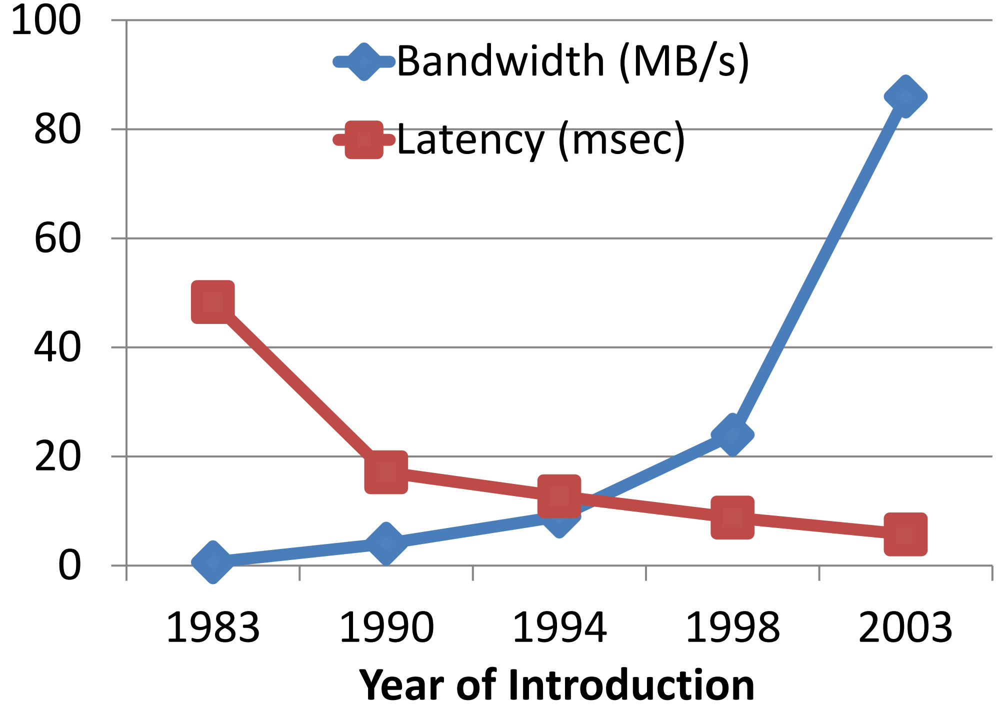
= 4.0 ms + 0.5/(15,000RPM/(60sec/min)) + 0.5KB/(100MB/sec) + 0.2 ms

= 4.0 + 2.0 + 0.005 + 0.2 = 6.2 ms

If the measured average seek time is 25% of the advertised average seek time, then

Avg disk read/write = 1.0 + 2.0 + 0.005 + 0.2 = 3.2 ms

Disk Latency & Bandwidth Improvement



- Disk latency = average seek time + rotational latency.
- Disk bandwidth is the peak transfer speed of formatted data from the media (not from the cache).

Patterson, CACM Vol 47, #10, 2004

Flash Storage

- **Flash memory is semiconductor memory that is nonvolatile like disks but has latency 100 to 1000 times lower and is smaller, more power efficient, and more shock resistant.**
  - Flash memory bits wear out. But with wear leveling it is unlikely that the write limits of the flash will be exceeded.
  - Example:

Storage	Write throughput	Read throughput	Latency
SSD	2,500 MB/S	3,500 MB/S	0.025 ms
DISK	250 MB/S	250 MB/S	5 ms

Flash Storage Characteristics

- **Flash memory is organized into blocks of pages**
  - page size 512B to 4KB, block size 32 to 128 pages (typical)
  - "read" is page read
  - "write" is block erase (~ 1 ms) followed by page write (and additional copying of other pages in that block)
- **Comes in two flavors: NOR Flash and NAND Flash**
  - NOR Flash is randomly addressable (Minimum access size 512 bytes)
  - NAND Flash is less expensive (greater storage density) and is not randomly addressable (minimum access size 2048 bytes); so more popular

Dependability, Reliability, Availability

- **Reliability – measured by the mean time to failure (MTTF). Service interruption is measured by mean time to repair (MTTR)**

- **Availability – a measure of service accomplishment**

  Availability = MTTF / (MTTF + MTTR)

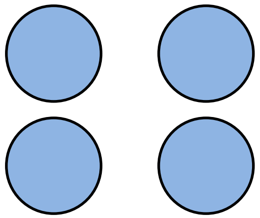
- **To increase MTTF, either improve the quality of components or design the system to continue operating in the presence of faulty components**
  1. Fault avoidance: preventing fault occurrence by construction
  2. Fault tolerance: using redundancy to correct or bypass faulty components (hardware)

RAID: Redundant Array of Independent Disks

- **Arrays of independent physical disks working together as one logical disk**
  - Increase potential throughput by having many disk drives
    - Data can be spread across multiple disk
    - Multiple disk accesses can be made simultaneously for higher throughput

- **Reliability for the array is lower than for a single disk**

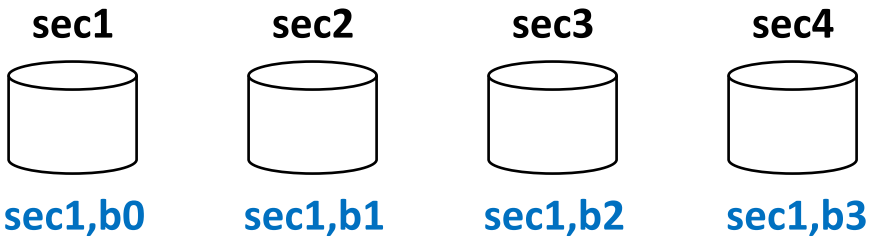
- **Availability can be improved by adding redundant disks**
  - Lost information can be reconstructed from redundant information
  - MTTR: mean time to repair is in the order of hours
  - MTTF: mean time to failure of disks is tens of years



RAID Disk Array

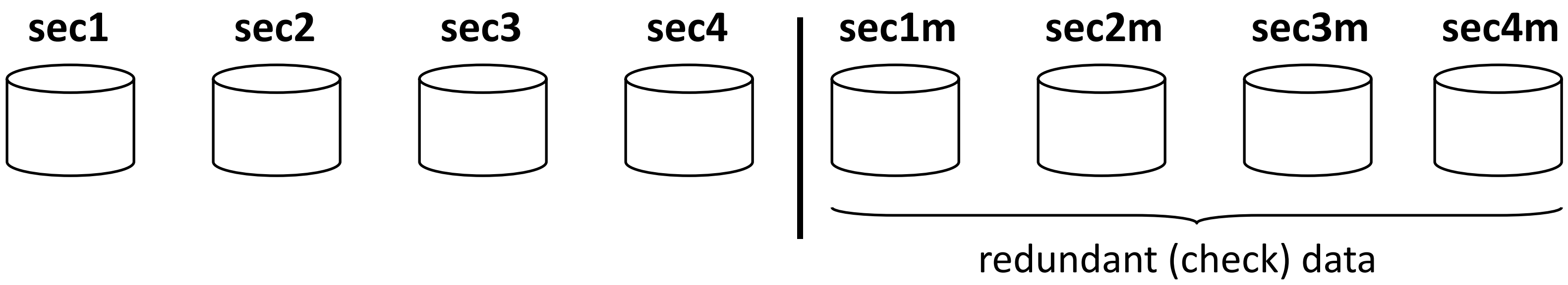
RAID: Level 0 (Striping, No Redundancy)

Assumes one stripe = four blocks



- **Multiple smaller disks as opposed to one big disk**
  - Multiple blocks can be accessed in parallel to increase the performance
  - Works well for large data requests
- **E.g., a 4-disk system gives four times the throughput (R/W) of a 1-disk system**
  - Same cost as one big disk – assuming 4 small disks cost the same as one big disk
- **What if one disk fails?**
  - No redundancy, data is lost
  - More likely to fail as the number of disks increases

RAID: Level 1 (Redundancy via Mirroring)



- **# redundant disks = # of data disks, so always two copies of the data**
  - Twice the cost of one big disk
  - Writes are made to both sets of disks and have no speed improvement
  - Reads can be 2 times faster

- **If a disk fails, the system just goes to the "mirror" for the data**

RAID: Design Choices

Category	Level	Disks	Features
Striping	0	N	No fault tolerance
Mirroring, no striping	1	2N	Expensive
Striping Rarely used	2 3	N + m N + 1	Hamming code Parity disk
Block level Striping Parity	4 5 6	N + 1 N + 1 N + 2	Parity disk Distributed parity Dual distributed parity

- **Parity: used for error detection**
  - E.g., data: 00001111, parity bit: 0
- **Hamming code: limited error correction**
  - Using multiple parity bits to locate 1 error bit

SAN JOSÉ STATE UNIVERSITY *powering* SILICON VALLEY