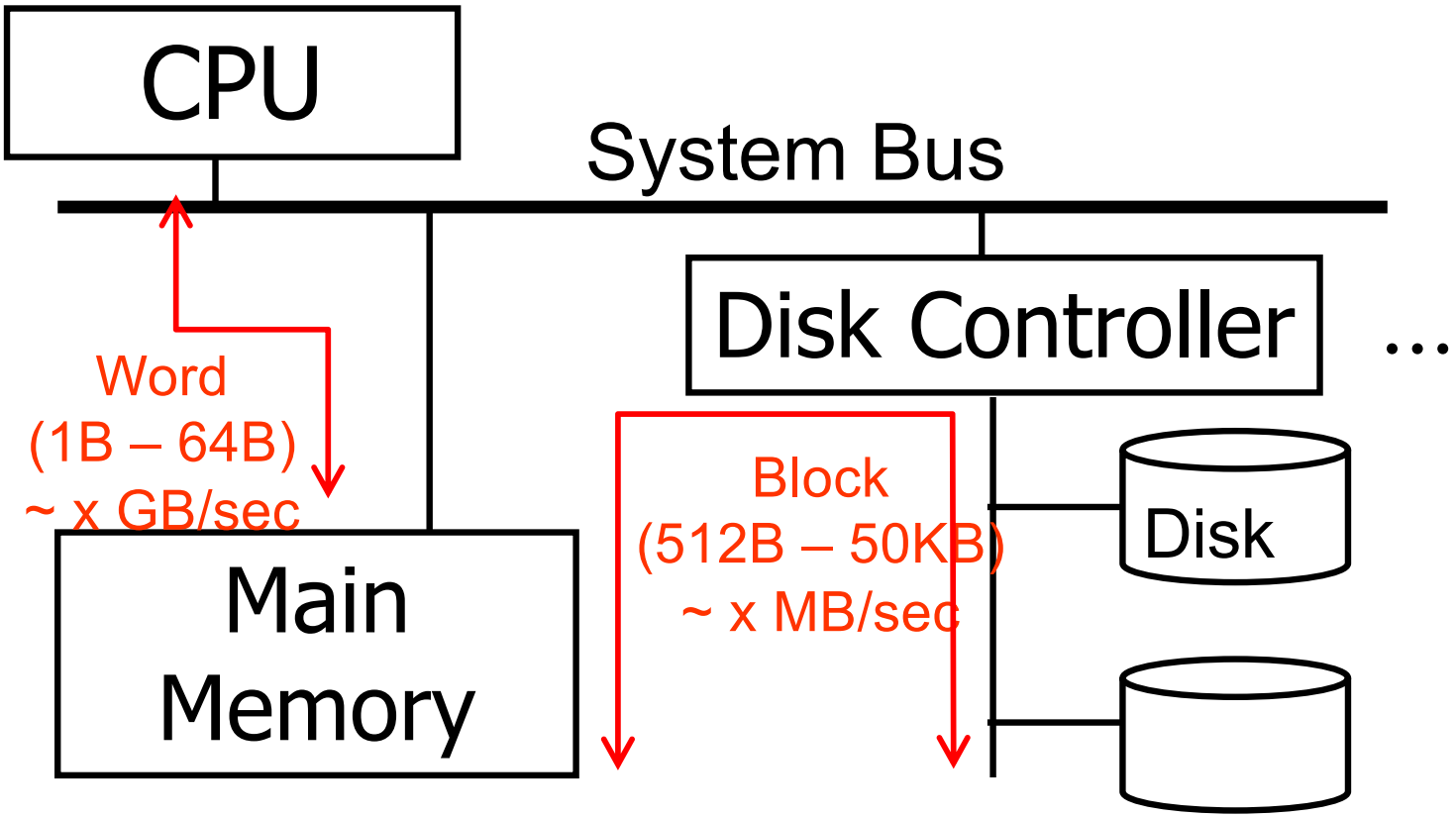
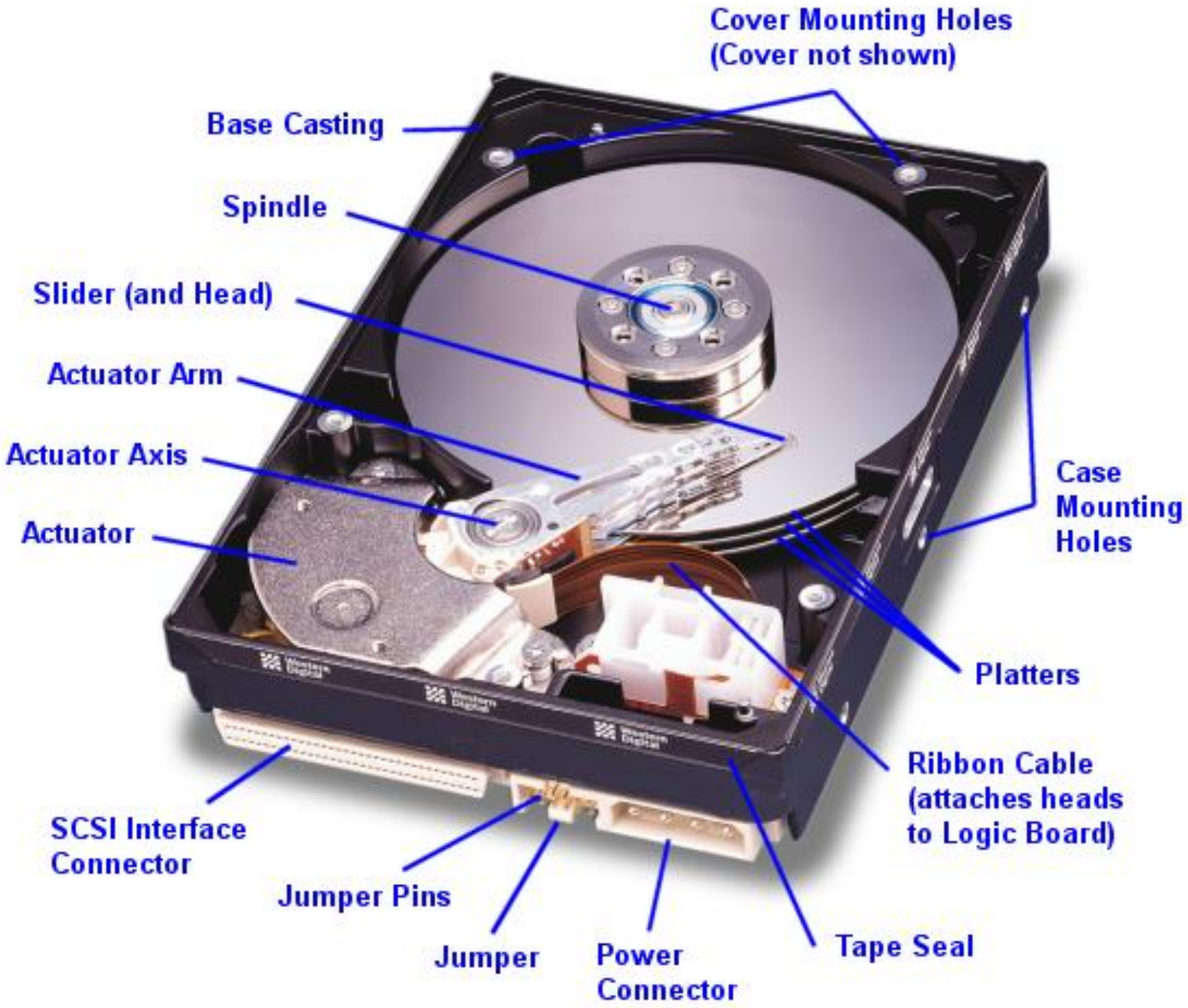
CS143: Disks and Files

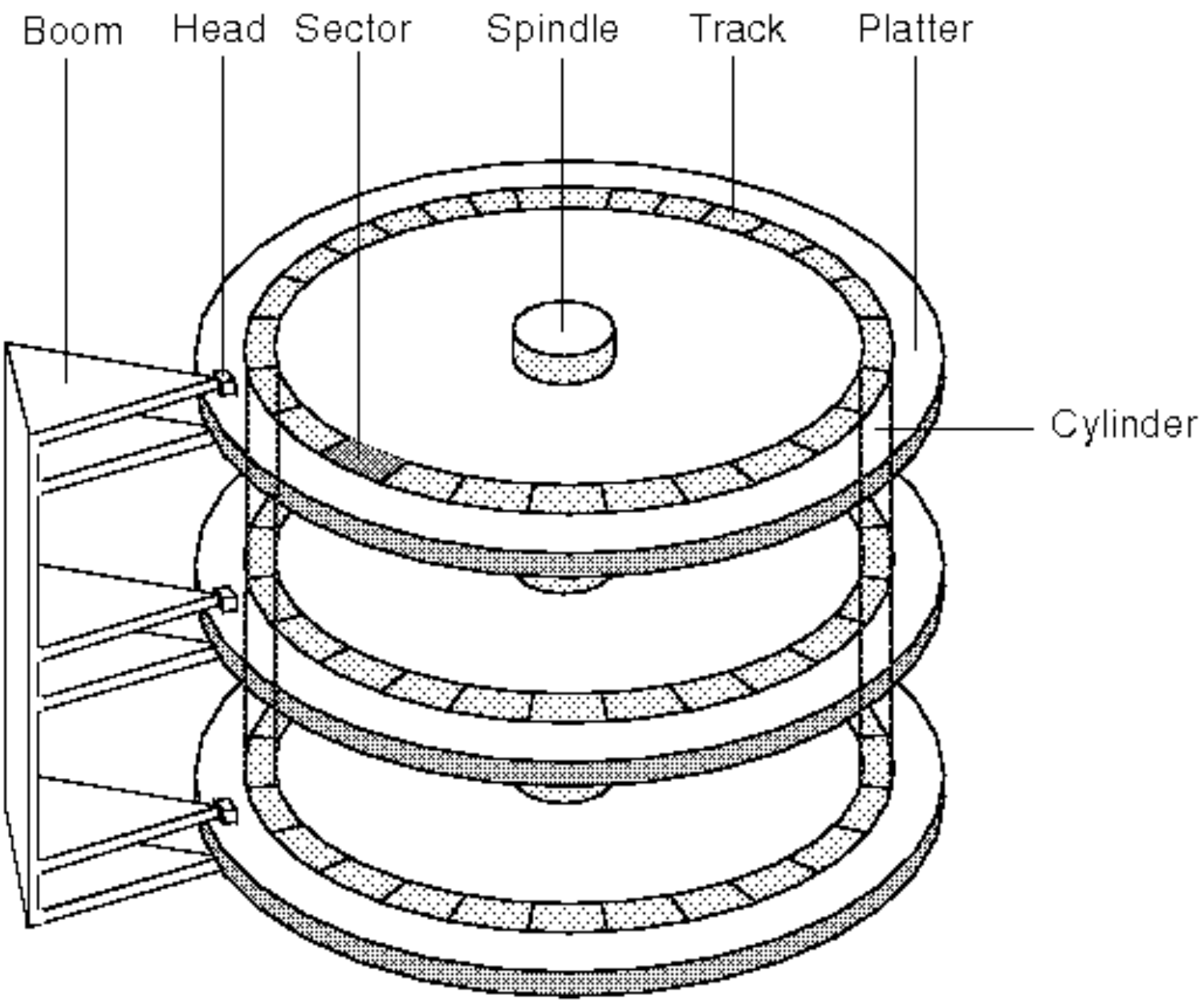
System Architecture



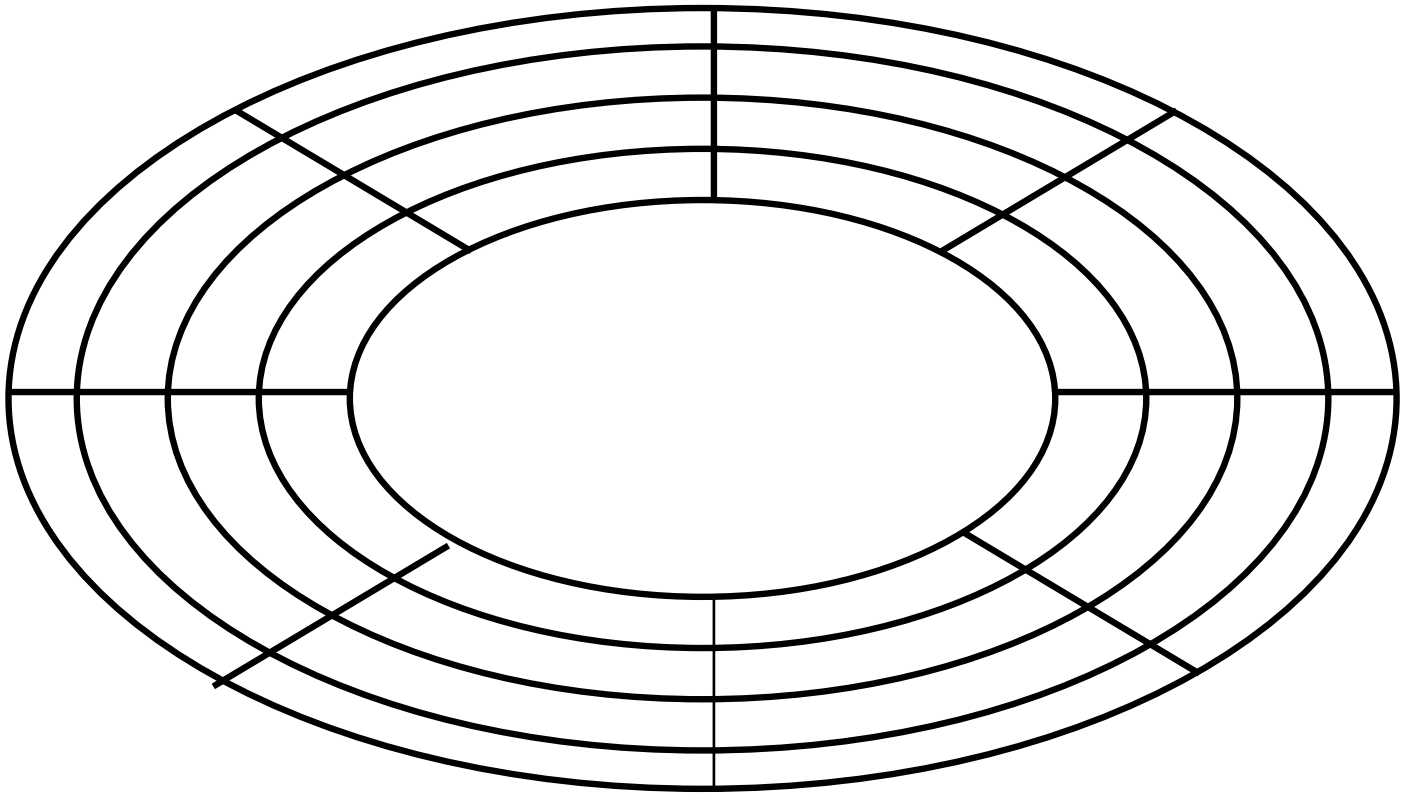
Magnetic disk vs SSD

- Magnetic Disk
  - Stores data on a magnetic disk
  - Typical capacity: 100GB – 10TB
- Solid State Drive
  - Stores data in NAND flash memory
  - Typical capacity: 100GB – 1TB
  - Much faster and more reliable than magnetic disk
  - But, x10 more expensive and limited write cycles (~2000)


Cover Mounting Holes
(Cover not shown)
Base Casting
Spindle
Slider (and Head)
Actuator Arm
Actuator Axis
Actuator
Case
Mounting
Holes
Platters
Ribbon Cable
(attaches heads
to Logic Board)
SCSI Interface
Connector
Jumper Pins
Jumper
Power
Connector
Tape Seal


Boom
Head
Sector
Spindle
Track
Platter
Cylinder

Structure of a Platter

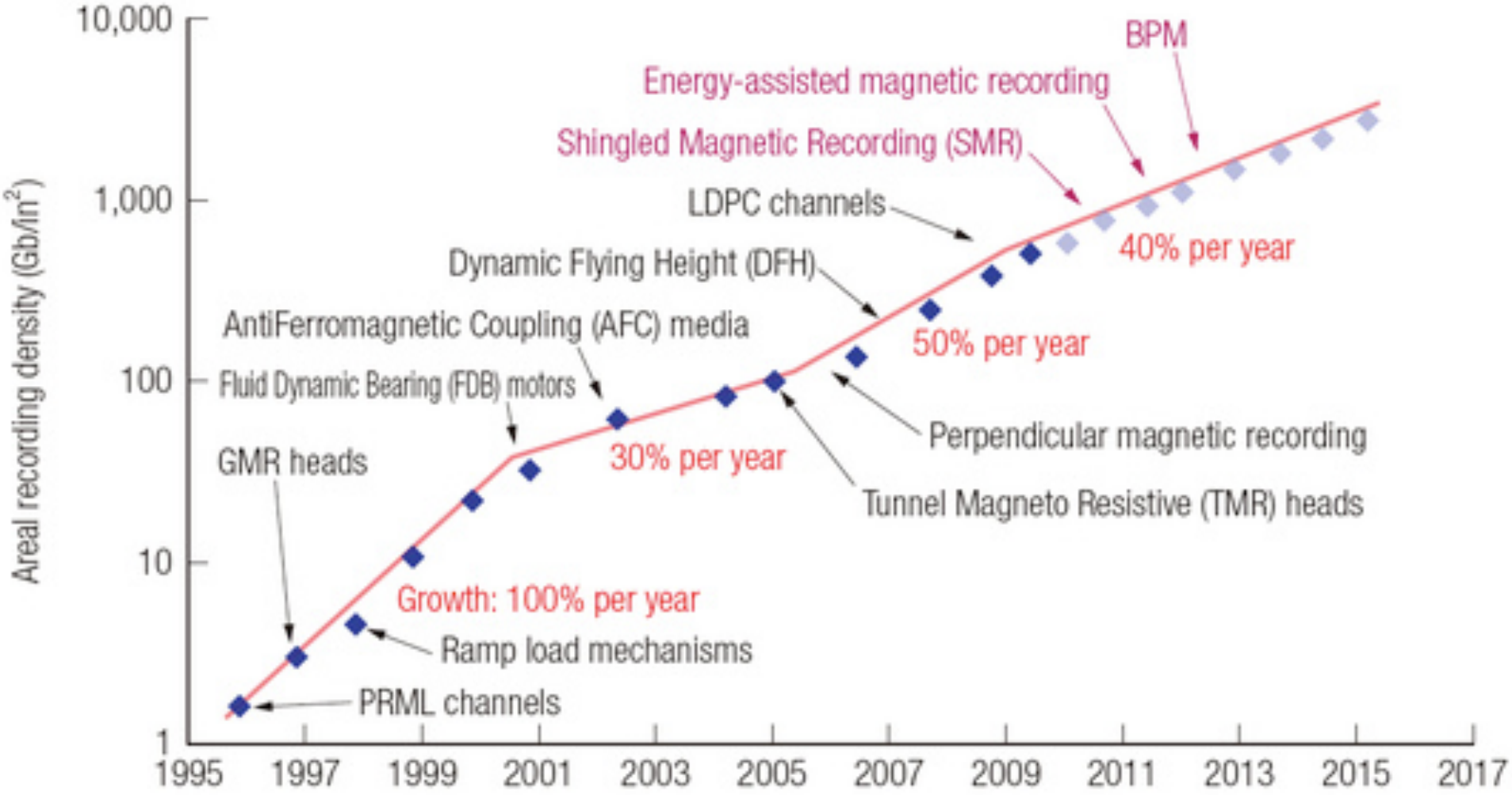


- Track, cylinder, sector (=block, page)

Typical Magnetic Disk

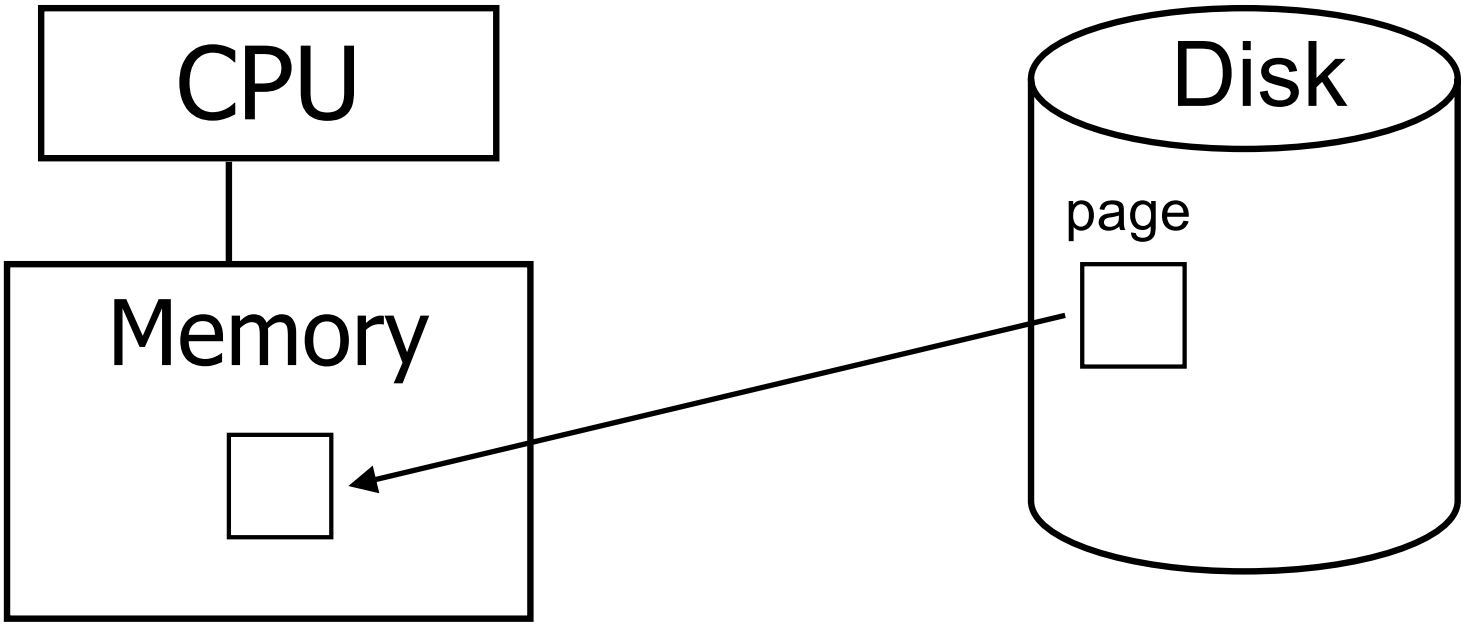
- Platter diameter: 1-5 in
- Platters: 1 – 20
- Tracks: 100 – 5000
- Sectors per track: 200 – 5000
- Sector size: 512 – 50K
- Rotation speed: 1000 – 15000 rpm
- Overall capacity: 100G – 10TB
- Q: 2 platters, 2 surfaces/platter,
  5000 tracks/surface,
  1000 sectors/track, 1KB/sector.
  What is the overall capacity?

Capacity of Magnetic Disk



- Capacity keeps increasing, but what about speed?

Access Time



- Q: How long does it take to read a page of a disk to memory?

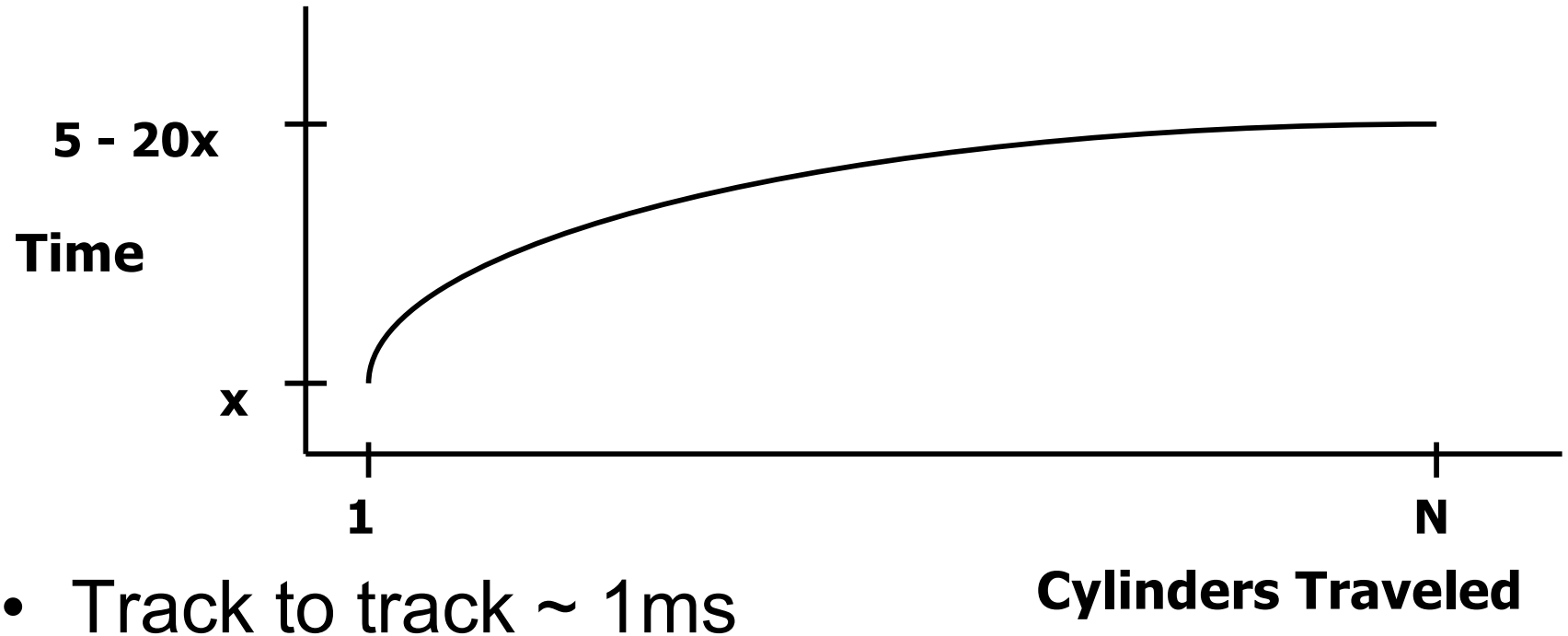
- Q: What needs to be done to read a page?

Access Time

- Access time =
  (seek time) + (rotational delay) +
  (transfer time)

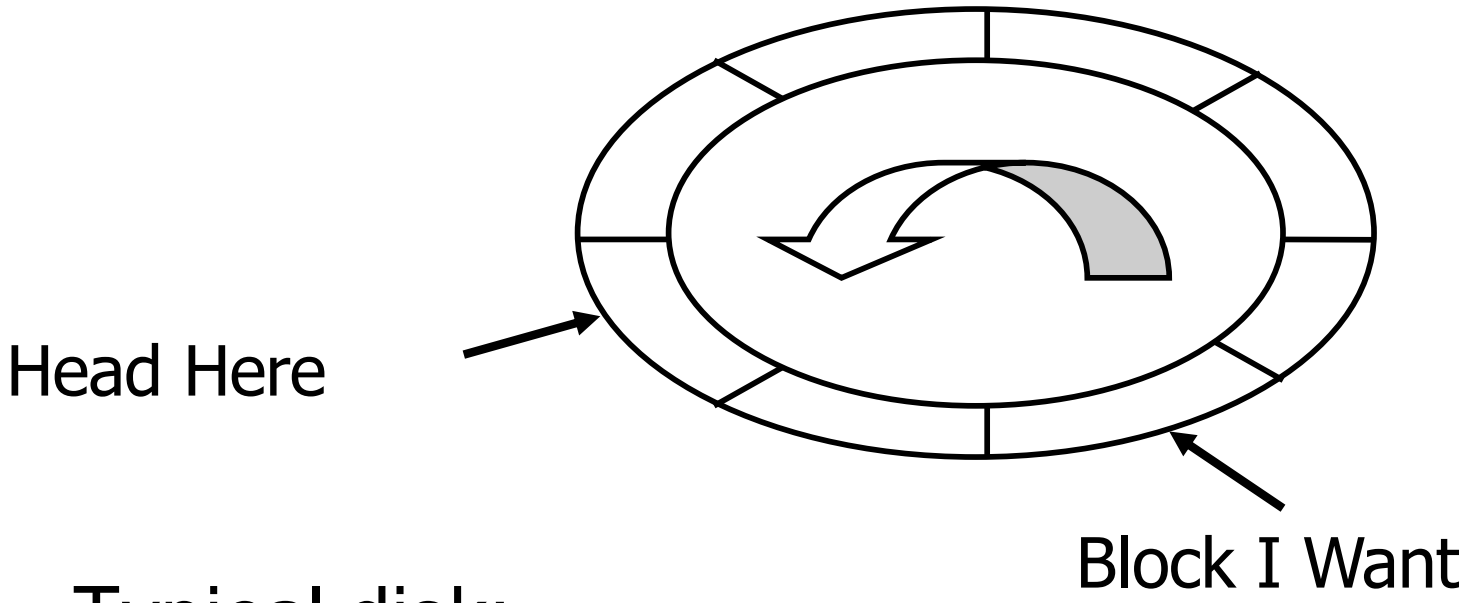
Seek Time

- Time to move a disk head between tracks



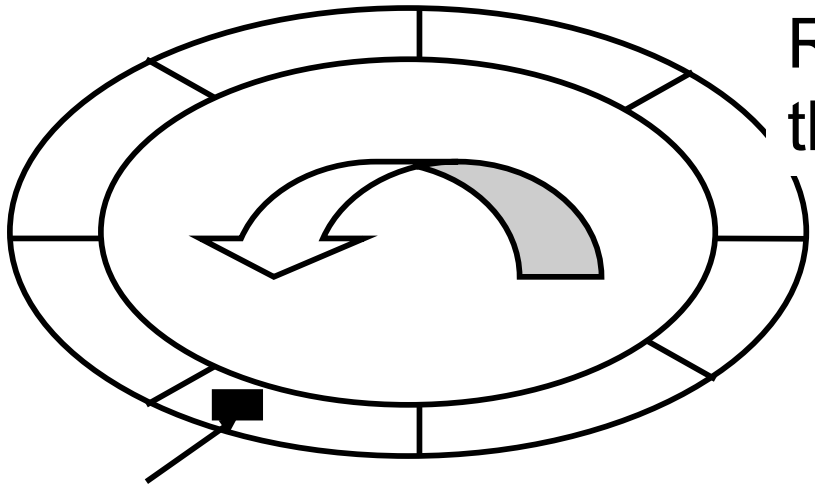
- Track to track ~ 1ms
- Average ~ 10 ms
- Full stroke ~ 20 ms

Rotational Delay



- Typical disk:
  - 1000 rpm – 15000 rpm
- Q: For 6000 RPM, average rotational delay?
- 100 rots per sec: 10 ms per a full rotation

Transfer Rate



6000 RPM, 1000 sectors/track, 1KB/sector

- Q: How long to read one sector?
- Q: What is the transfer rate (bytes/sec)?

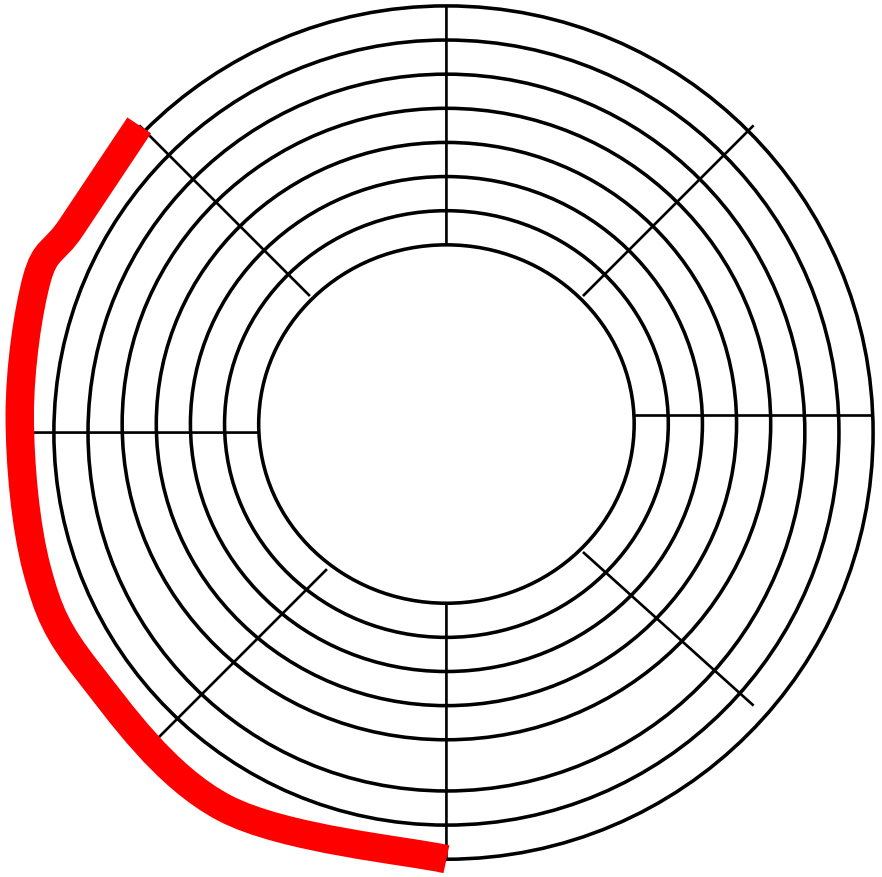
(Burst) Transfer Rate

- (Burst) Transfer rate =
  (RPM / 60) * (sectors/track) * (bytes/sector)

- *(A sector holds 1000 bytes, a track with 1000 sectors is read in 10 ms:*
  *Bytes per ms: 1000*1000/10*
- *So, 100 KB per ms = 100 MB per sec.*

Sequential vs. Random I/O

- Q: How long to read 3 sequential sectors?

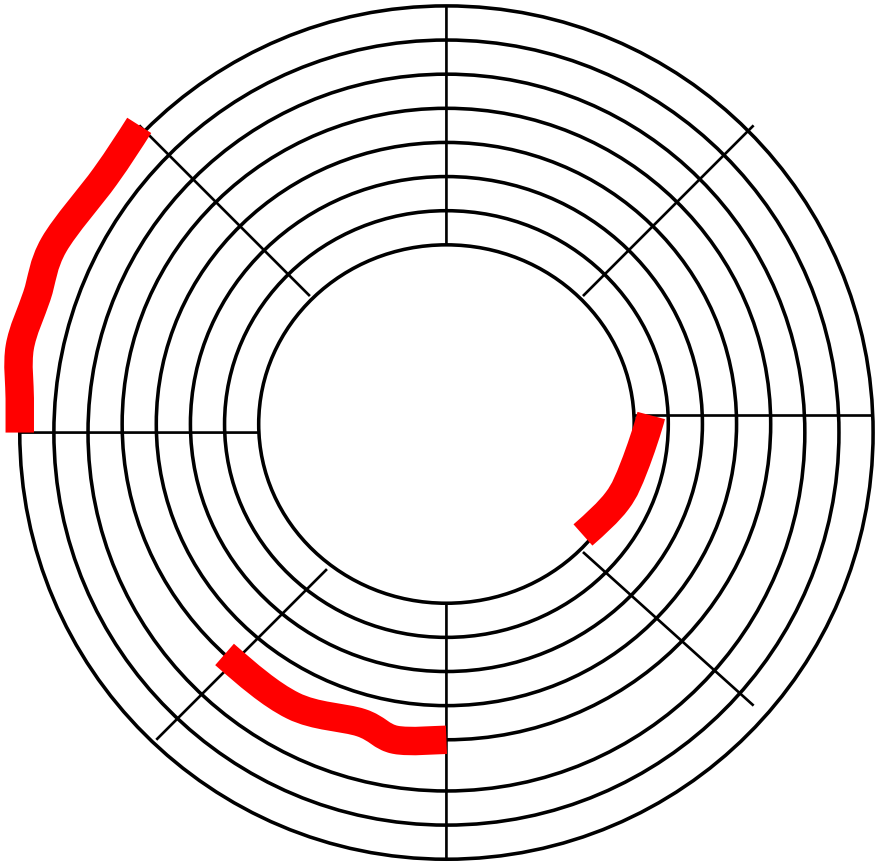


- 6000 RPM
- 1000 sectors/track
- Assume the head is above the first sector

Answer: 0.03ms

Sequential vs. Random I/O

- Q: How long to read 3 random sectors?



- 6000 RPM
- 1000 sectors/track
- 10ms seek time
- Assume the head is above the first sector

Answer: 30 ms

Random I/O

- For magnetic disks:
  - Random I/O is VERY expensive compared to sequential I/O
  - Avoid random I/O as much as we can

Magnetic Disk vs SSD

	Magnetic	SSD
Random IO	~100 IOs/sec	~100K IOs/sec
Transfer rate	~ 100MB/sec	~500MB/sec
Capacity/$	~1TB/$100 (in 2014)	~100GB/$100 (in 2014)

SSD speed gain is mainly from high random IO rate

RAID

- Redundant Array of Independent Disks
  - Create a large-capacity “disk volumes” from an array of many disks
- Q: Possible advantages and disadvantages?

RAID Pros and Cons

- Potentially high throughput
  - Read from multiple disks concurrently
- Potential reliability issues
  - One disk failure may lead to the entire disk volume failure
  - How should we store data into disks?
- Q: How should we organize the disks and store data to maximize benefit and minimize risks?

RAID Levels

- RAID 0: striping* only (no redundancy)
- RAID 1: striping + mirroring
- RAID 5: striping + parity block

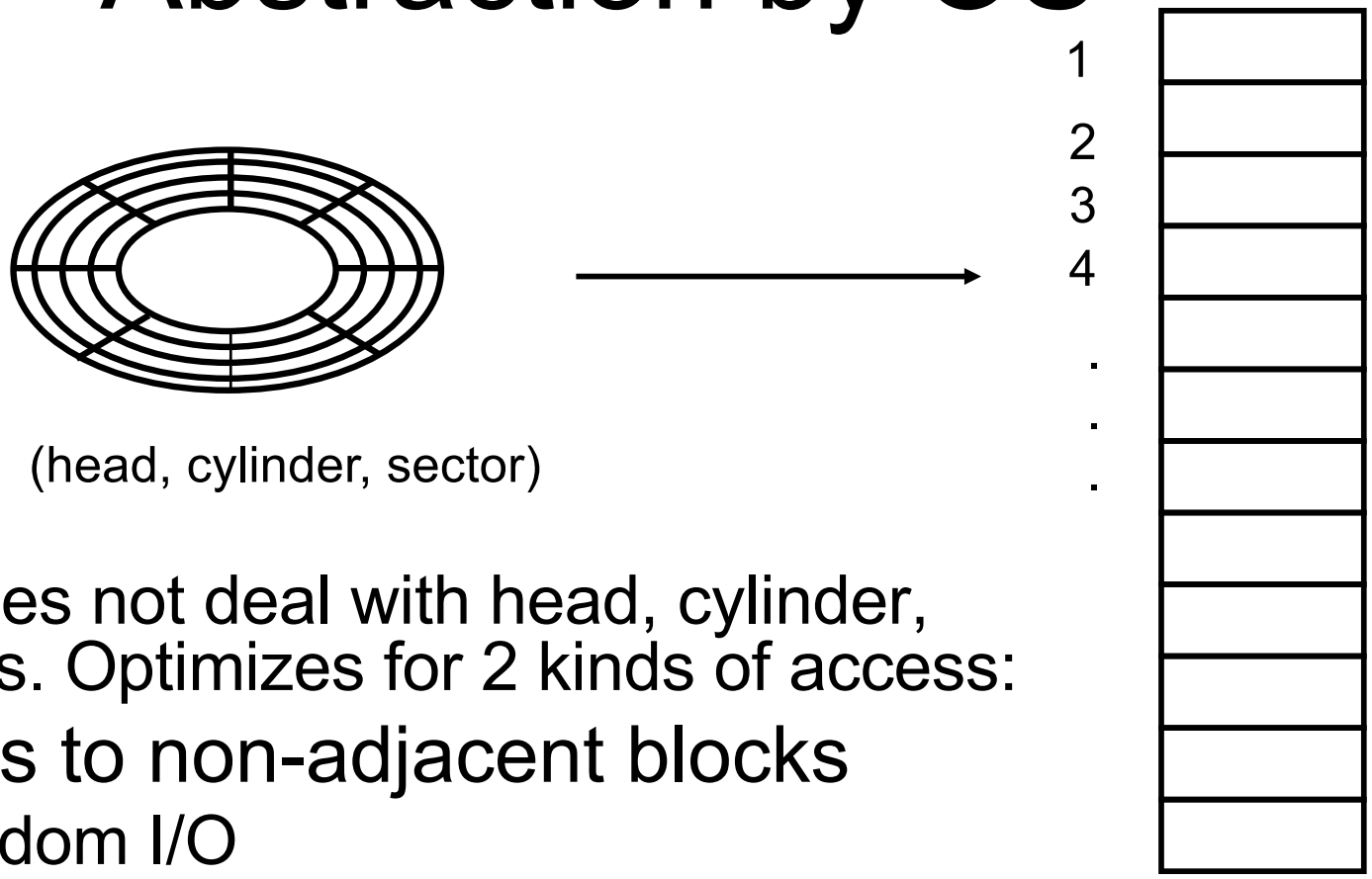
Striping: dividing a file into multiple data blocks and spread them across multiple disks

Data Modification for Files

- Byte-level modification not allowed
  - Can be modified by blocks

- Q: How can we modify only a part of a block?

Abstraction by OS



OS does not deal with head, cylinder, sectors. Optimizes for 2 kinds of access:

- Access to non-adjacent blocks
  - Random I/O
- Access to adjacent blocks
  - Sequential I/O

Buffers, Buffer pool

- Temporary main-memory “cache” for disk blocks
  - Avoid future read
  - Hide disk latency
  - Most DBMS let users change buffer pool size

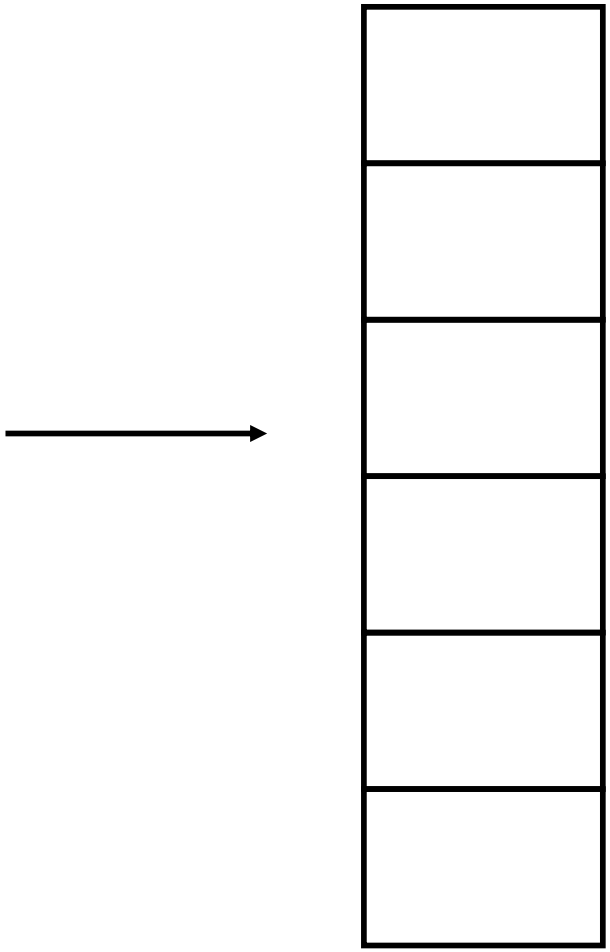
Reference

- Storage review disk guide
  - http://www.storagereview.com/guide2000/ref/hdd/index.html

Files: Main Problem

- How to store tables into disks?

Jane	CS	3.7
Susan	ME	1.8
June	EE	2.6
Tony	CS	3.1

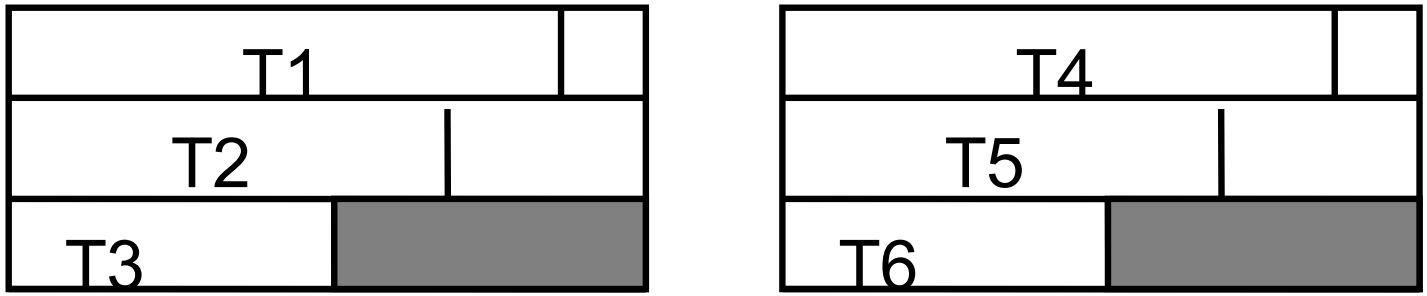


Spanned vs Unspanned

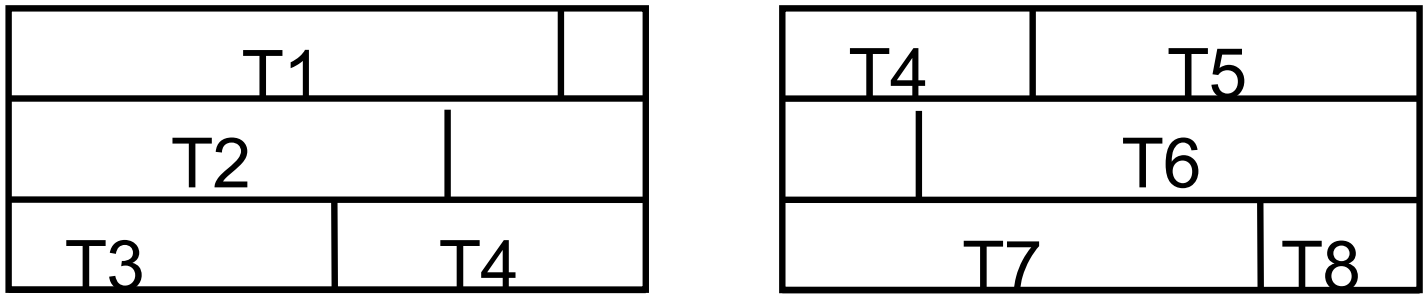
- Q: 512Byte block.
- 160 Byte tuples.
- How to store?

Spanned vs Unspanned

- Unspanned



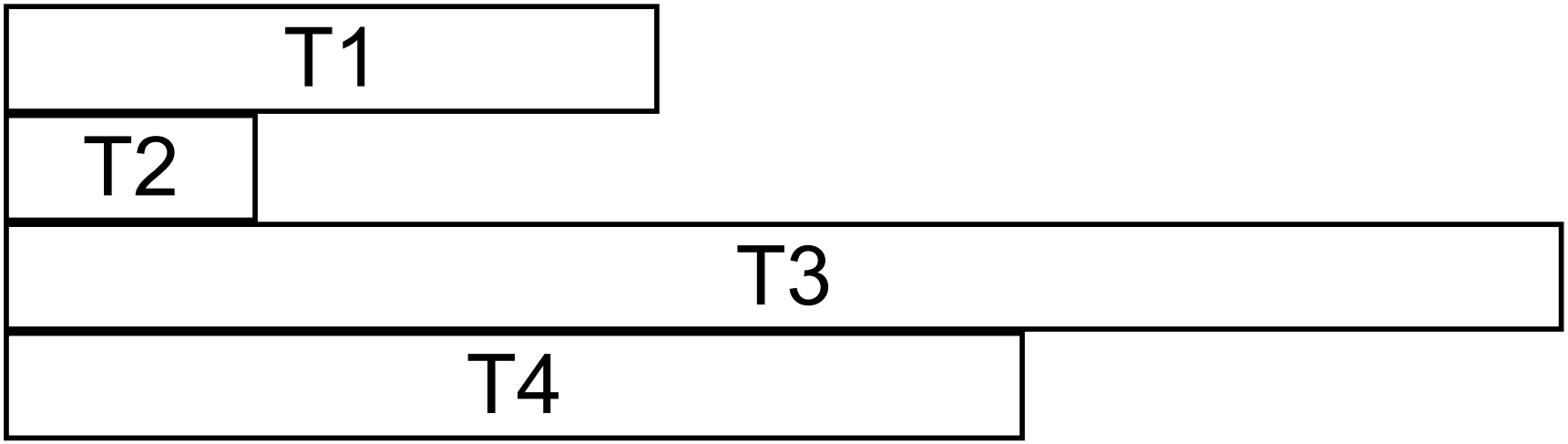
- Spanned



- Q: Maximum space waste for unspanned?

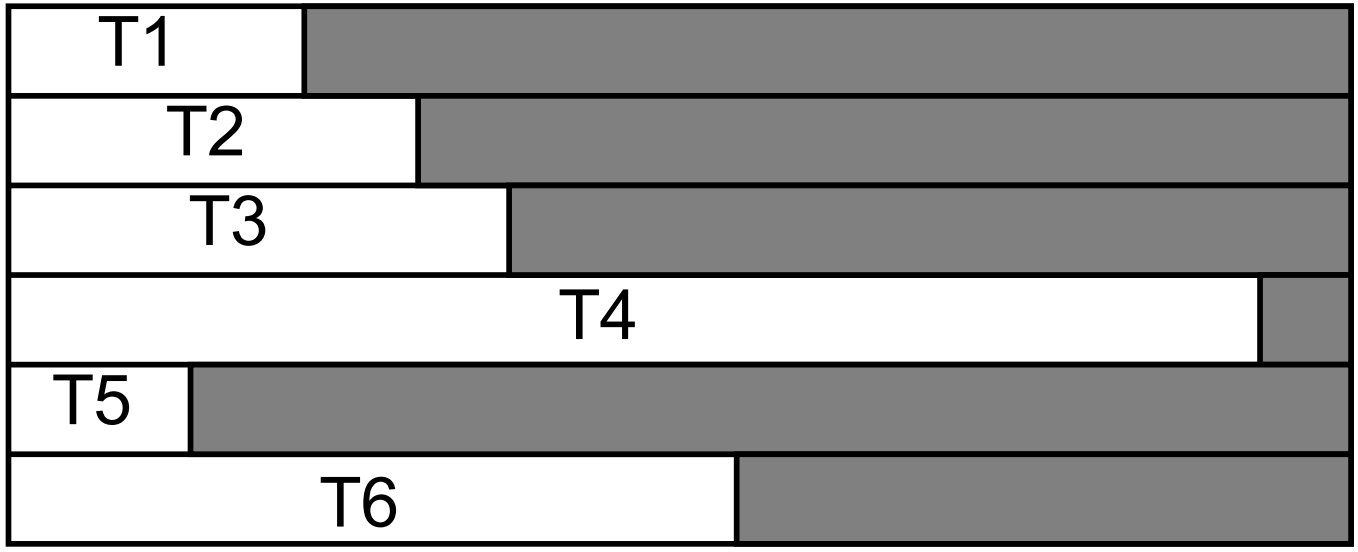
Variable-Length Tuples

- How do we store them?



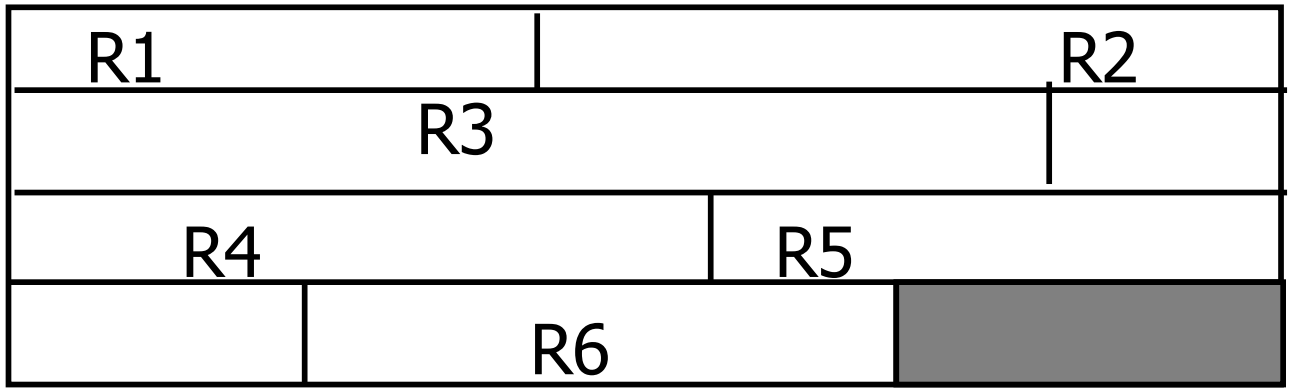
Reserved Space

- Reserve the maximum space for each tuple



- Q: Any problem?

Variable-Length Space

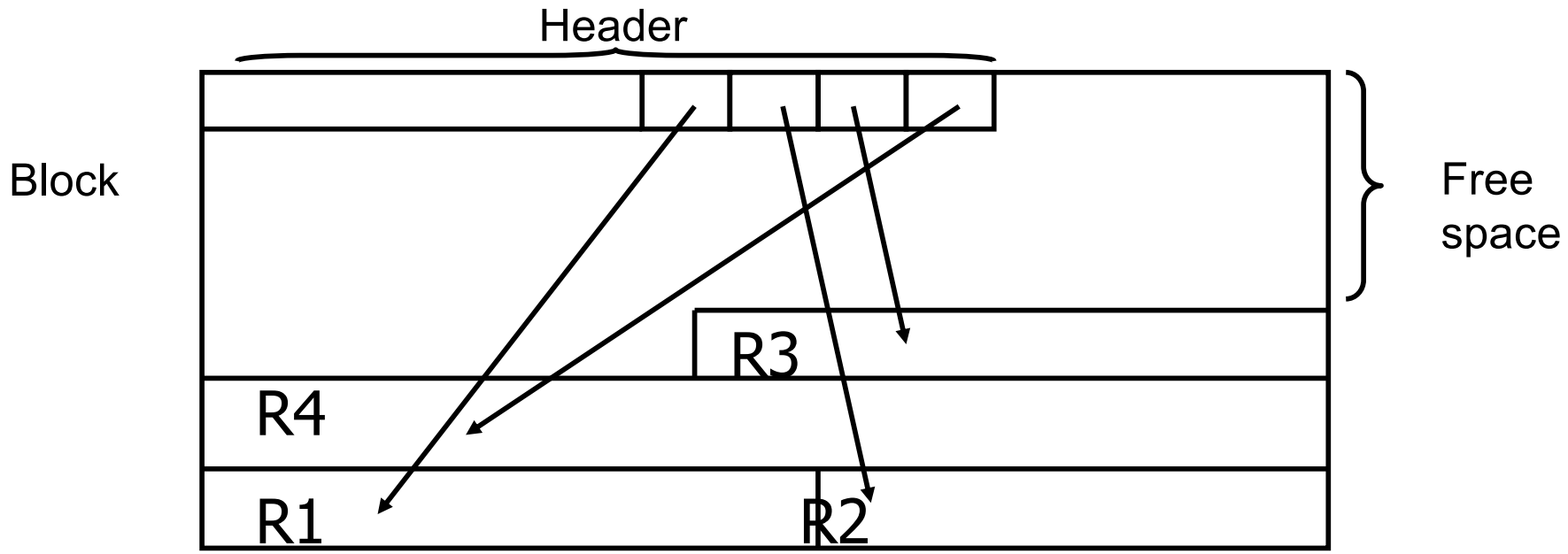


- Pack tuples tightly
- Q: How do we know the end of a record?

- Q: What to do for delete/update?

- Q: How can we “point to” to a tuple?

Slotted Page



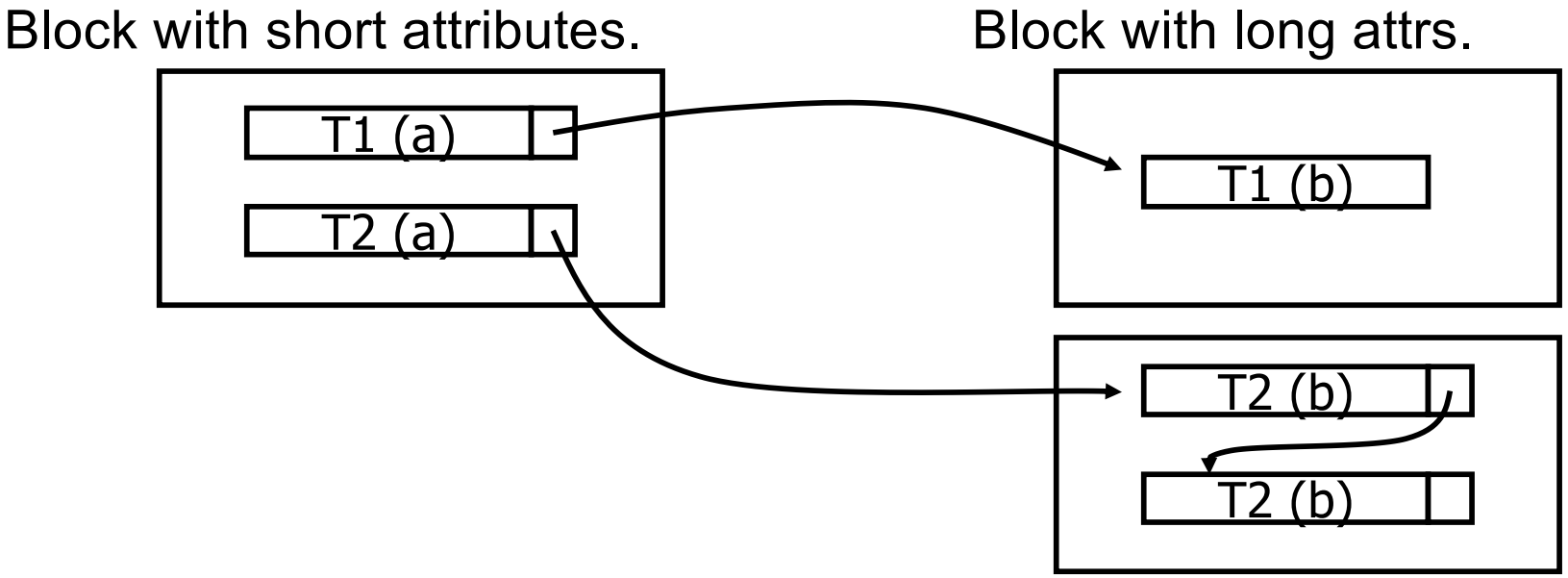
Q: How can we point to a tuple?

Long Tuples

- ProductReview(
  pid INT,
  reviewer VARCHAR(50),
  date DATE,
  rating INT,
  comments VARCHAR(1000))
- Block size 512B
- How should we store it?

Long Tuples

- Spanning
- Splitting tuples



Sequential File

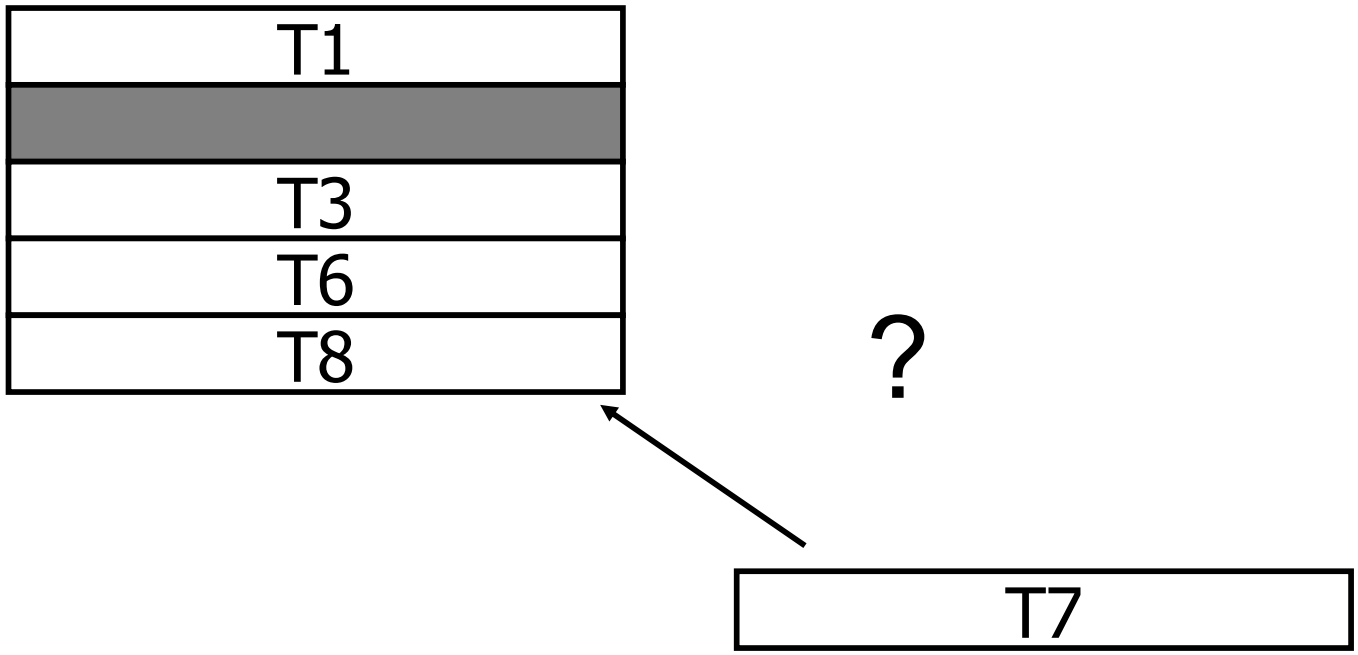
- Tuples are ordered by certain attribute(s) (search key)

Elaine	CS	3.7
James	ME	2.8
John	EE	1.8
Peter	EE	3.9
Susan	CS	1.0
Tony	EE	2.4

  – Search key: Name

Sequencing Tuples

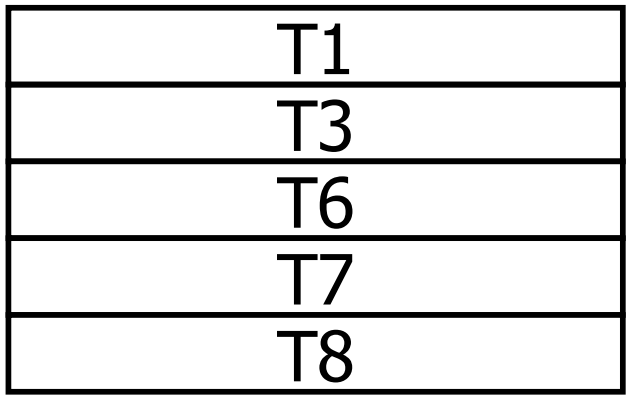
- Inserting a new tuple
  - Easy case



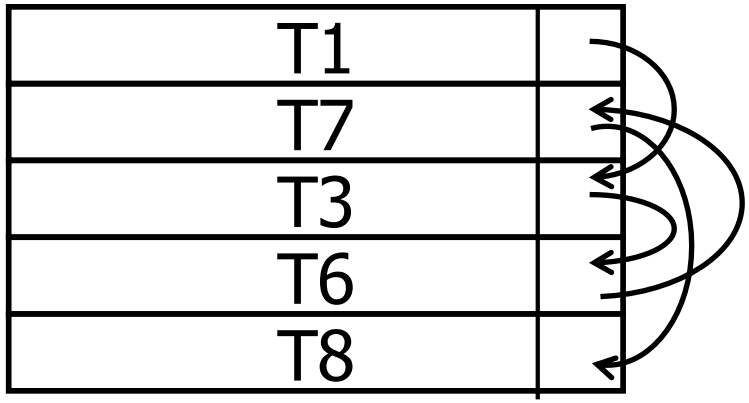
Sequencing Tuples

Two options

1) Rearrange

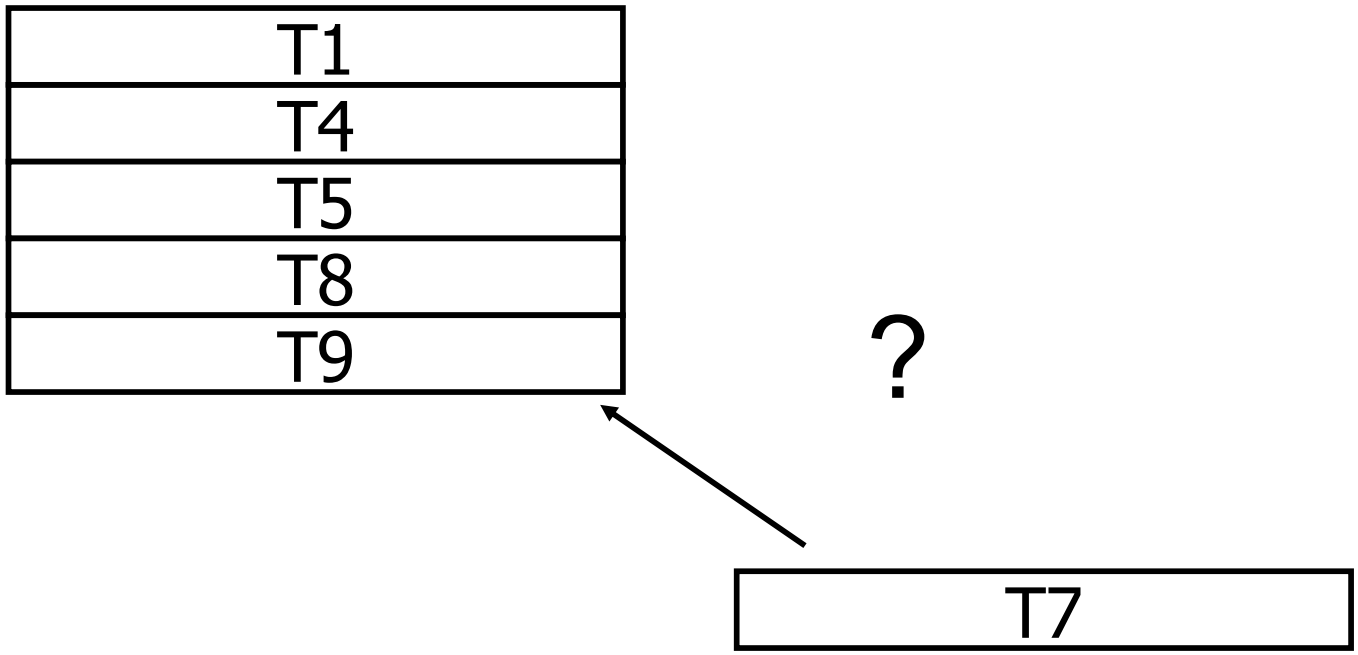


2) Linked list



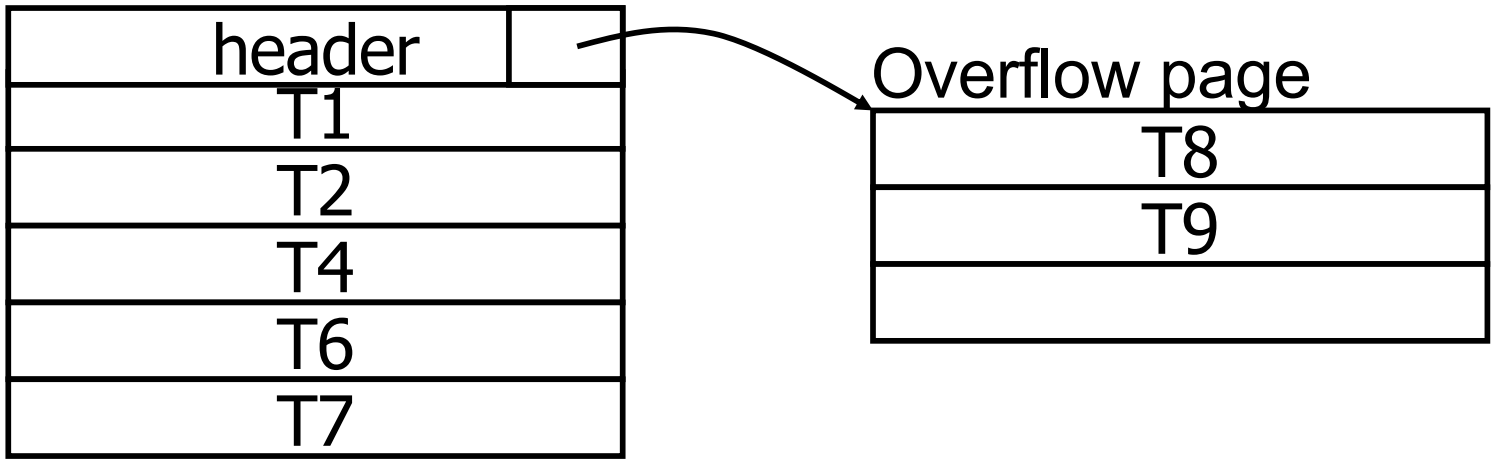
Sequencing Tuples

- Inserting a new tuple
  - Difficult case



Growth & Resequencing: 2 Options

1) Use Overflow pages



2) Reserve free space to avoid overflow

– PCTFREE in DBMS

CREATE TABLE R(a int) PCTFREE 20

(20% space is kept free in all data blocks used to store R. This allows for future growth & reordering)

Things to Remember

- Disk
  - Platter, track, cylinder, sector, block
  - Seek time, rotational delay, transfer time
  - Random I/O vs Sequential I/O
- Files
  - Spanned/unspanned tuples
  - Variable-length tuples (slotted page)
  - Long tuples
  - Sequential file and search key
    - Problems with insertion (overflow page)
    - PCTFREE