ECE 5730
Memory Systems
Spring 2009

More Overview
Defect Management
Drive Interfaces

Cornell University

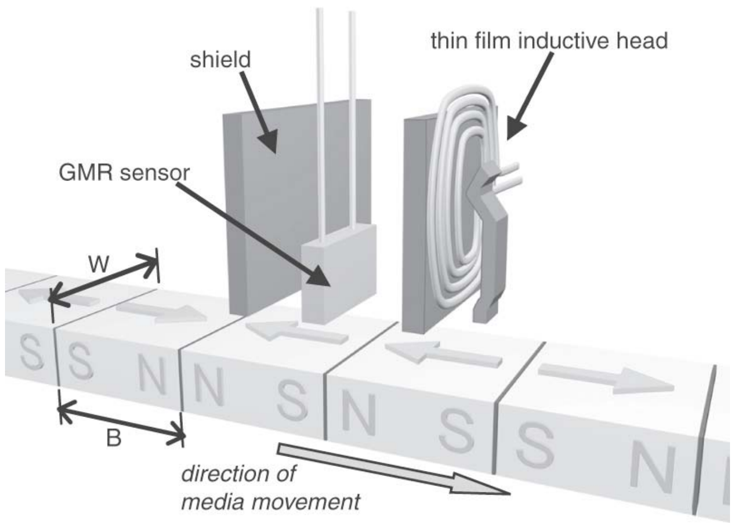
Announcements

- Class tomorrow
  - 6:00-7:15pm, PH 403
  - Pizza
  - No material will show up on a quiz or Exam II

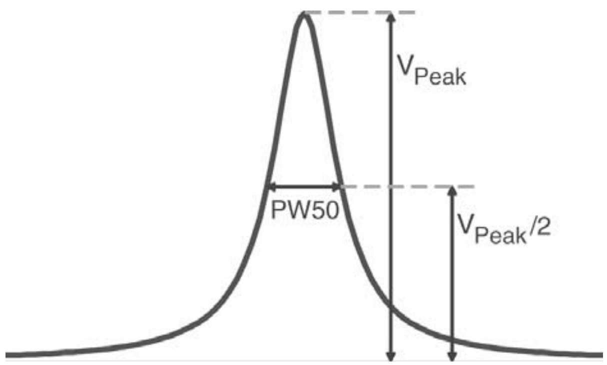
- Exam II
  - May 7, 7:00-10:00pm, Hollister 314
  - Covers material from 3/10-4/28 but excluding 4/22 (Lectures 14-21, 23-24)
  - Let me know immediately if you require a make-up

tpi and bpi Scaling Factors

- **tpi**
  - Width of the write head
  - Deviation in actuator staying in the center of the track
  - Wobble in the spindle motor
  - Mechanical vibration
  - Disk wobble



- **bpi**
  - Media properties
  - Reducing size of the heads
  - Reducing media-head gap

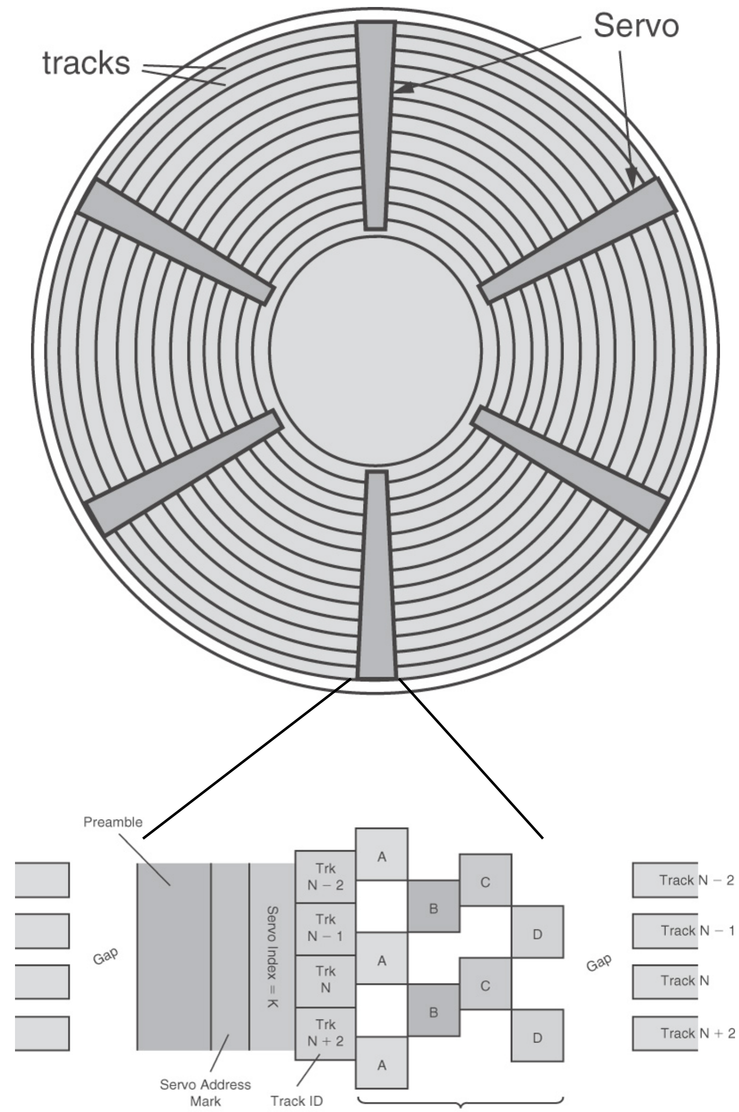


[17.19,17.27]

Simultaneously Read Multiple Heads?

- **Main problem is temperature**
  - **Servo makes corrective adjustments to keep head centered on the track (*track following*)**
  - **Differences in temperature and in coefficients of expansion of arms and disks causes relative displacement of the heads**
  - **Resulting imprecision requires some sacrifice in tpi to achieve reliable operation**

limit on track density based on their imprecisions

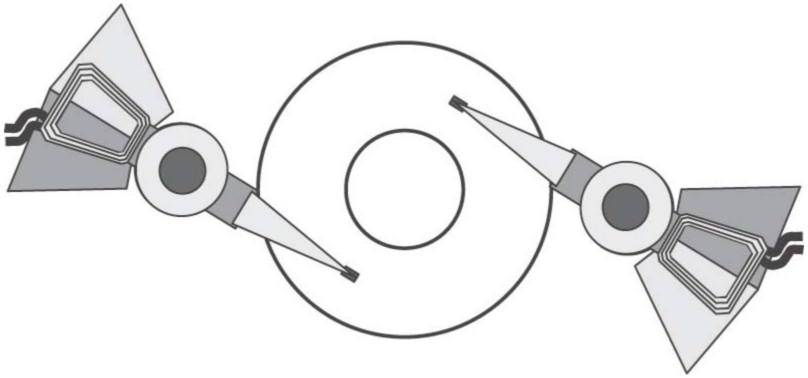


information for location

alignment pattern

[18.13,18.19]

Multiple Actuators?



- **Two independent head assemblies that simultaneously read two different locations**
  - performance advantage
- **Have been built but increases cost by 30-40%**
  - last known example ~1994
  - heads very expensive
  - added overhead of circuitry to do 2 streams of data

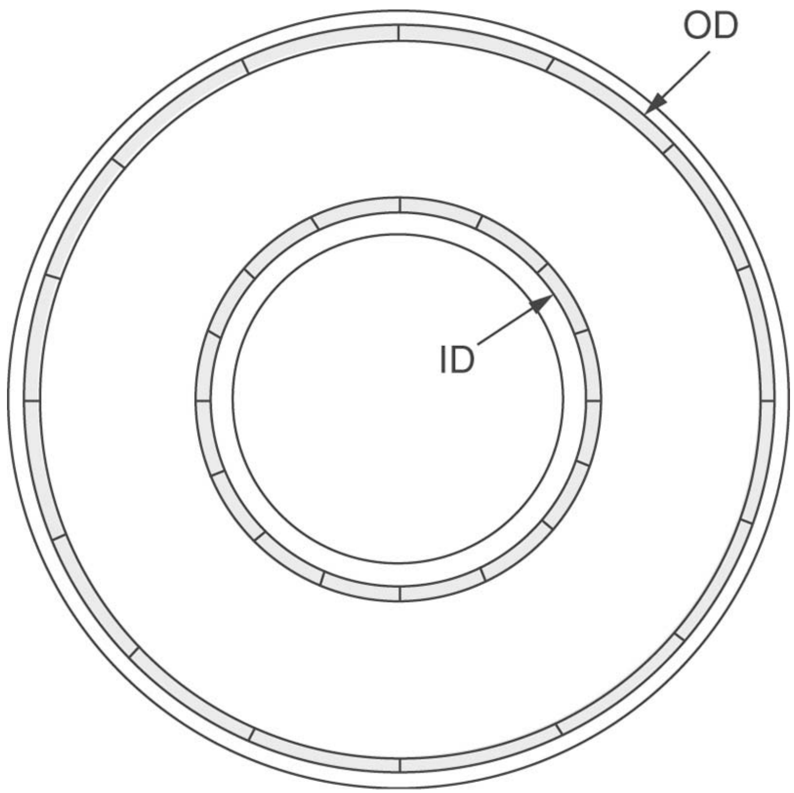
[21.1]

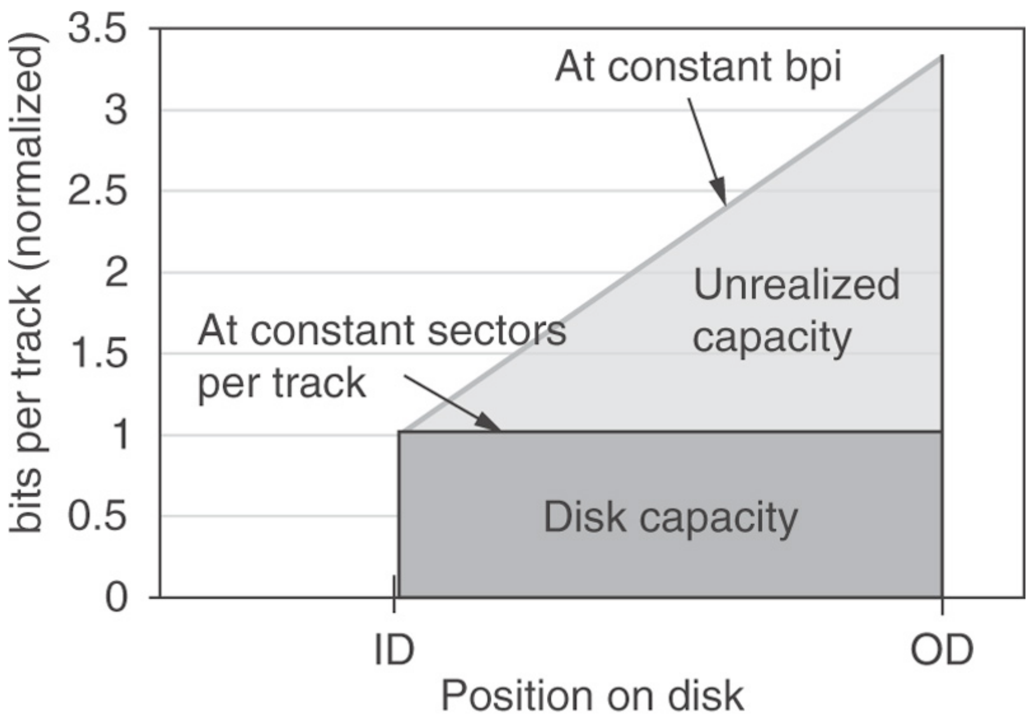
Recording Density

outer diameter — inner diameter

- **OD distance may be 3X ID distance**
- **At constant write rate, low bpi in outer sectors**

(due to angular vs linear speed)





[18.7,18.8]

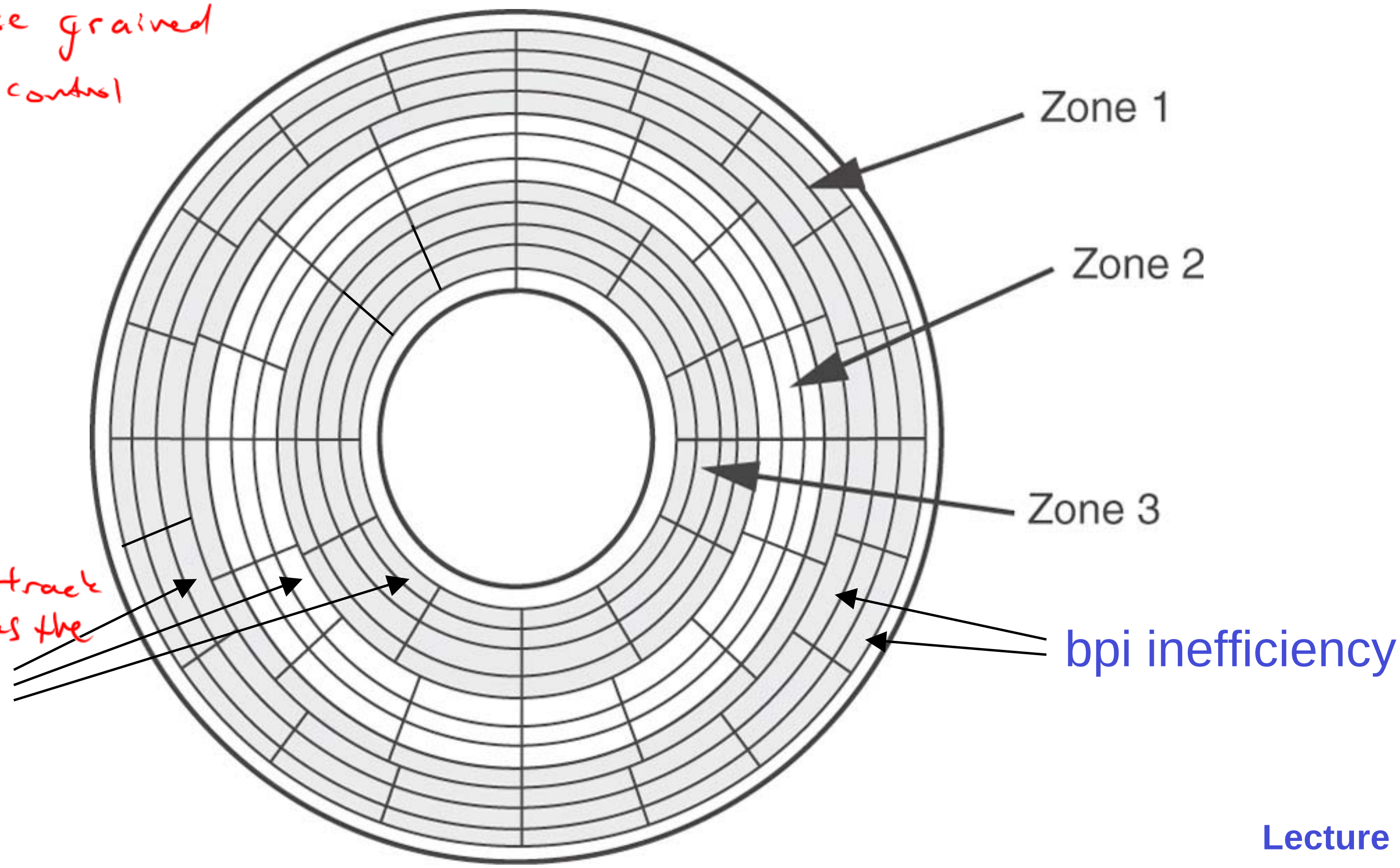
Zoned-Bit Recording

- **To get max density, use max bpi for all tracks**
  - **Different recording rates for 100's of 1000's of tracks**
  - hard.
- **Simplification: Zones with same sectors/track**

- coarse grained

bpi control

inner most track of zone has the



[18.11]

ZBR Implementation Options

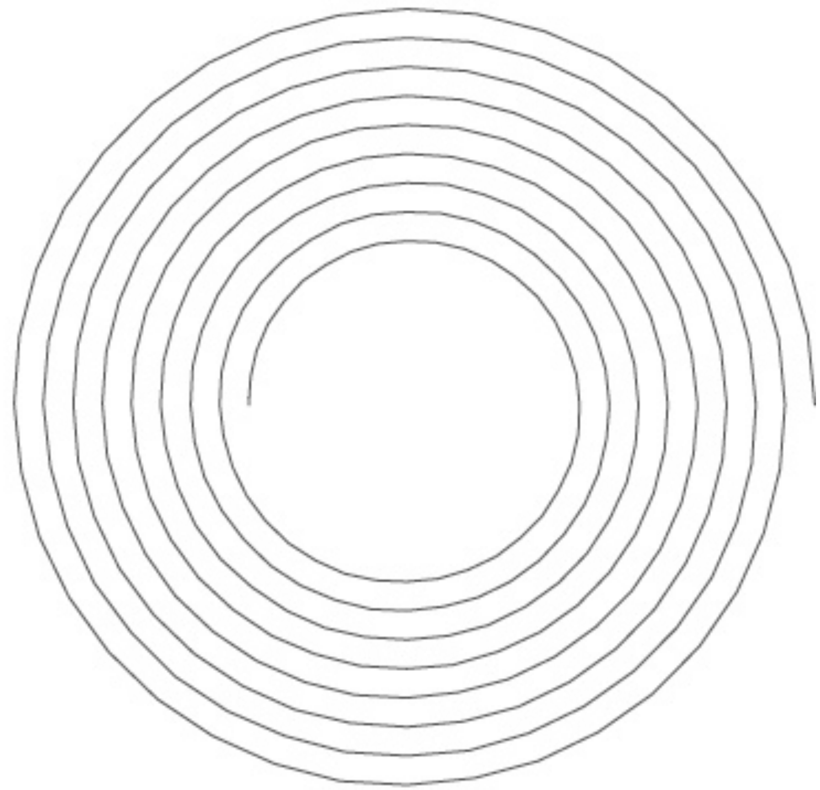
- **Variable rotation speed**
  - **Rotational speed varied from zone to zone**
  - **Takes 100's of ms to change drive speed**
  - **Since 3X data is stored at OD than ID, rotational speed must be 3X slower**
  - **Not a viable option for hard disk drives**

KK sez: no good!

- **Variable data rate**
  - **Data rate varied from zone to zone**
  - **Data rate change is very fast**
  - **3X data is stored at OD and data rate is 3X faster**
  - **Used in hard disk drives**

Optical Disks (CDs and DVDs)

- **Spiral track rather than concentric tracks**

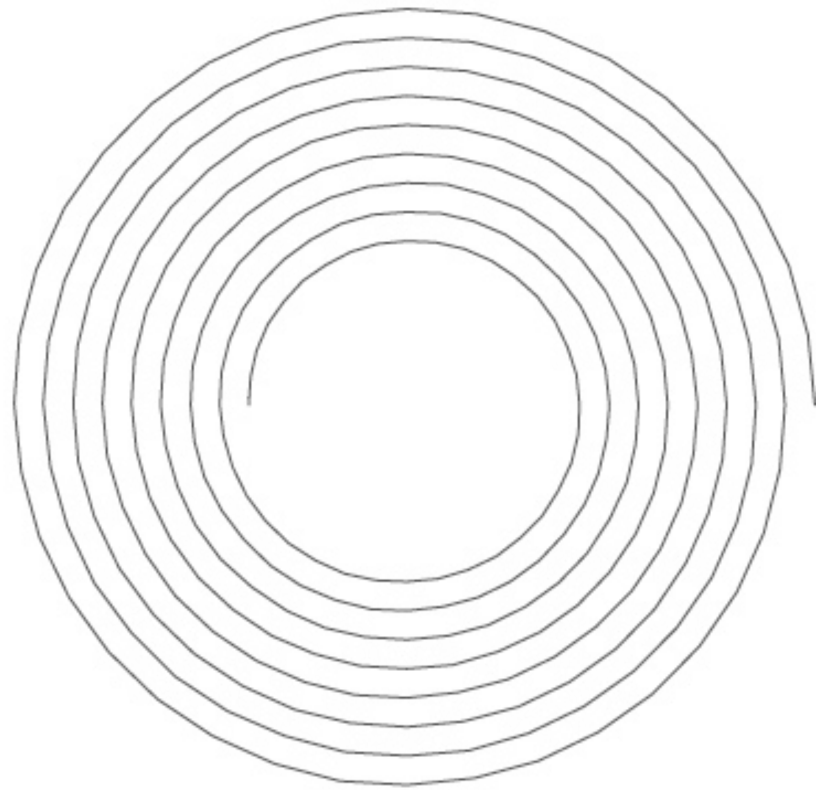


- **+ Density advantage**
  - **No lost space at end of tracks**
  - **No lost space due to intra-zone bpi inefficiency**
- **+ Higher average data rate on sequential accesses**
  - **No head track switching time** → gradual moving of head to follow spiral
  - **Constant linear velocity (CLV):** → easy to implement
    Rotational speed gradually decreases as move from ID to OD (1400-580rpm for DVD)

[18.2]

Optical Disks (CDs and DVDs)

- Spiral track rather than concentric tracks
- + Simple to implement
  - Constant data rate
- - Lower average data rate on random accesses
  - Large, abrupt rotational speed changes take 100's of ms
  - this is really sad. :(



[18.2]

Masking Hard Drive Errors

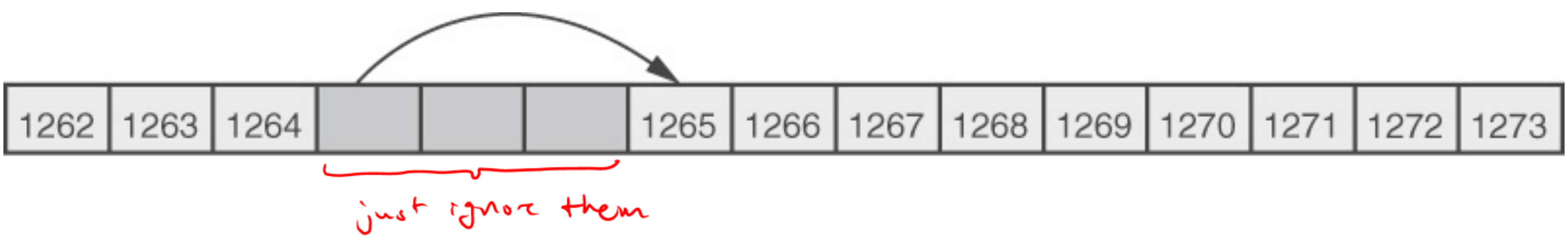
- ***Manufacturing defects*** **create areas on the media where bits cannot be reliably written or read**
  - **Many of these may be correctable by sector ECC**
  - **Those that cannot must be masked by the controller**

- ***Hard errors*** **arise in the field due to media wearout, shock, contaminants, etc**
  - **Those that cannot be corrected by sector ECC must be masked by the controller**

also can be fixed by ECC, but you might have to mask it

Defect Management

- Manufacturing defects (*primary defects*) are found through testing before the drive ships
- *P-List* of uncorrectable sectors is maintained
- Fixed through *Sector Slipping*



  – **Negligible performance impact**

[18.23]

Hard Error Management

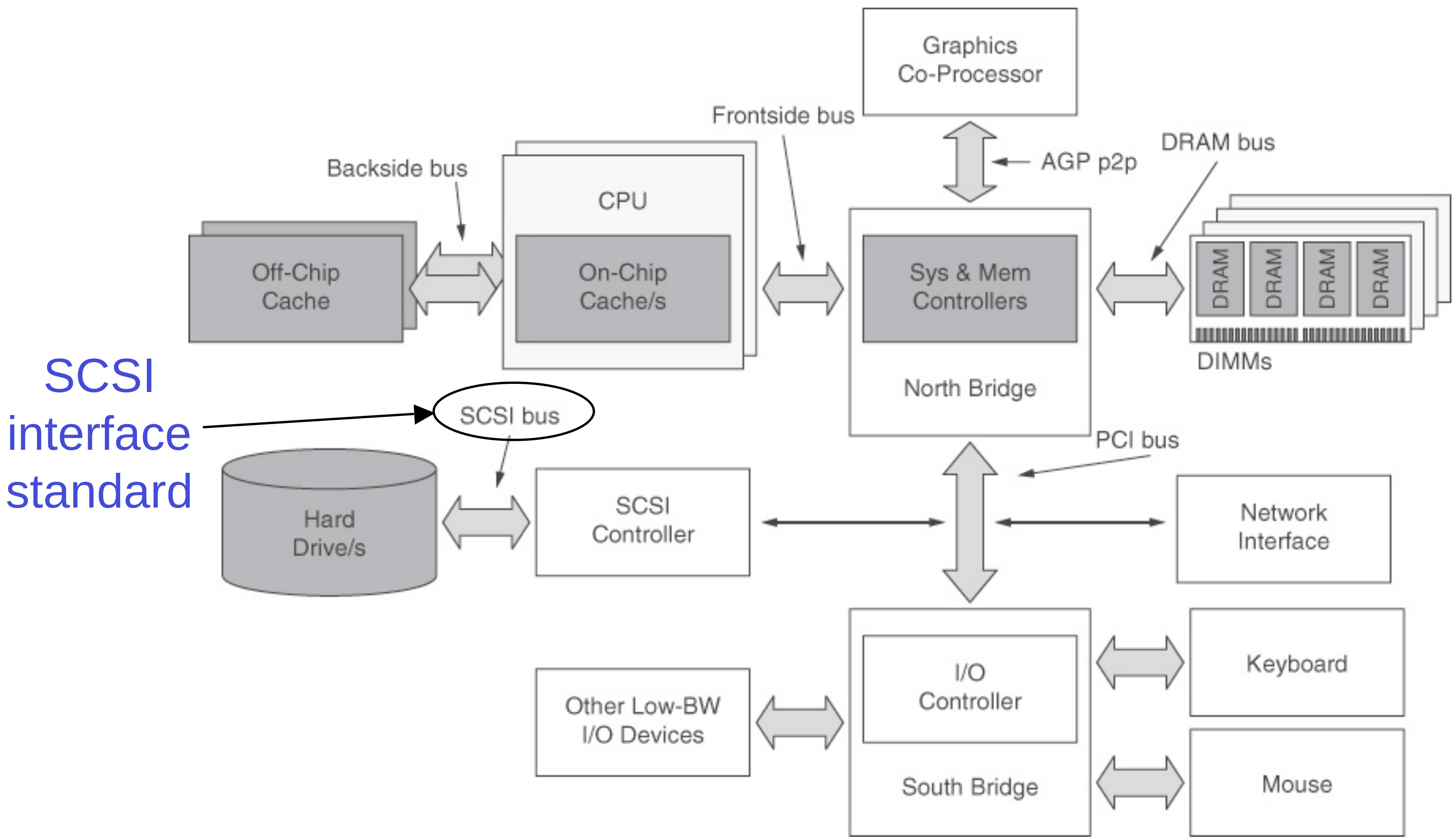
- **Hard errors (*grown defects*) develop after product shipment**

- **Rereading with different head positions, and ECC, are first tried**

- **Then *sector sparing* is used**
  - **Spare sectors are kept throughout the disk drive**
  - ***G-List* of mappings of defective sectors to spare sectors is maintained** → replace bad ones with spare ones
  - **Higher performance cost than sector slipping but avoids remapping the drive** → re-mapping would be a huge pain

Drive Interface

- **Drive interface provides a standardized means for communication between the host and drive**



[ov.3]

Drive Interface Standards

- Detailed specifications of the different required layers of the standard

- Physical, link, and transport layers
  - Voltages, timings, termination, interconnections, clocking, etc
  - all the super annoying EE stuff

- Logical layer
  - Communication protocol, command formats and actions, etc

Desirable Interface Characteristics

- **Simple protocol**
  - Little handshaking required between host and drive to reduce command overhead and bandwidth usage
- **Host autonomy**
  - Little host involvement so it can perform other tasks
- **Sufficient data rate**
  - Prevent the interface from becoming the bottleneck
- **Overlapped I/O operations**
  - Concurrent commands to multiple drives → parallelism
- **Command queuing**
  - Permit drive to reorder operations for performance

- simple is to just buffer commands
- complex is to reorder

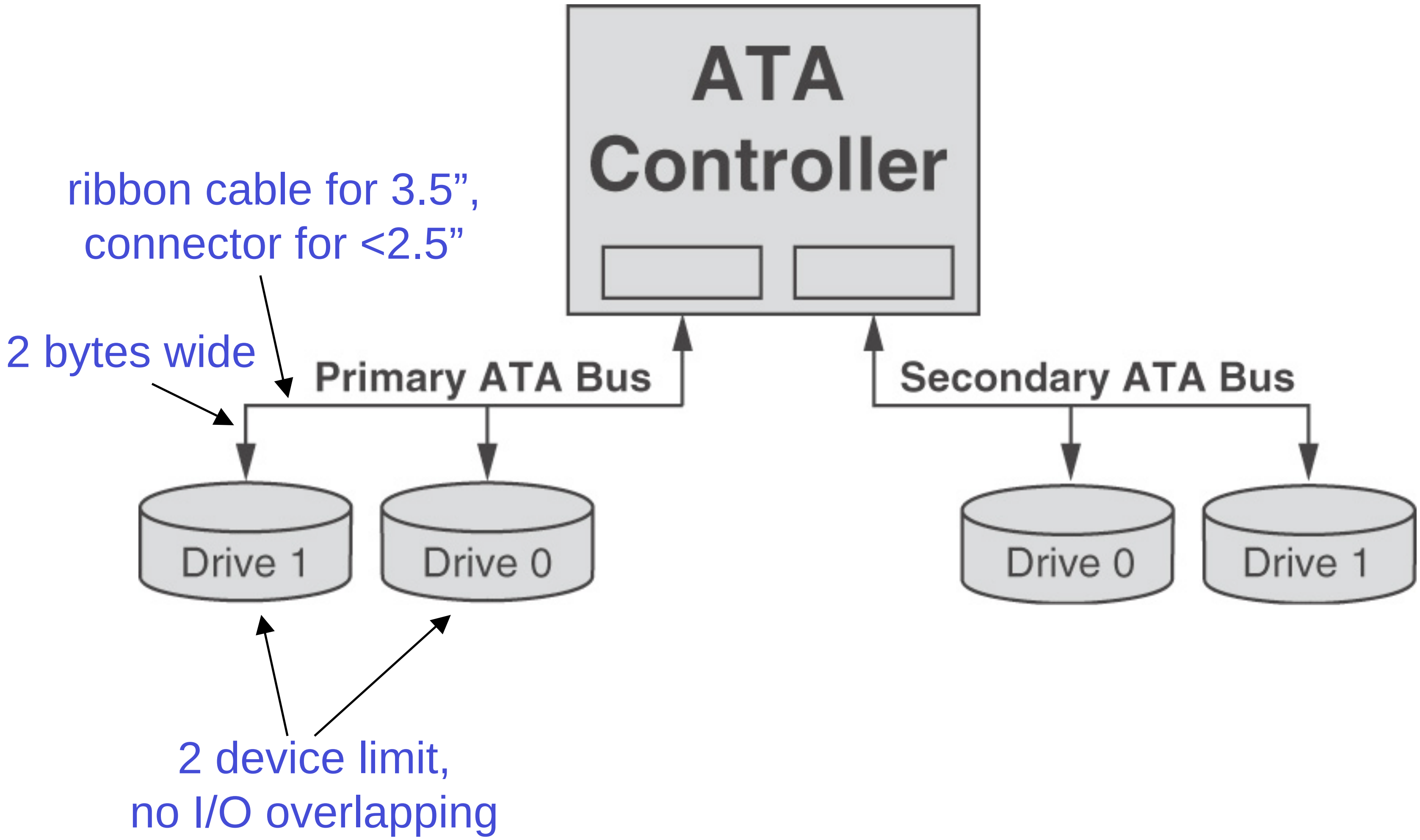
AT Attachment (ATA)

- Designed in the early 80's for the IBM PC-AT
  - Many revisions since standardization

richer command set

- ATA Packet Interface (ATAPI) permits SCSI commands to be sent over ATA
  - Use SCSI devices like CD-ROMs without a SCSI card

ATA Topology



ATA Operation Summary

logical block address — how much data to get

- **Host sends LBA and block size to registers on the selected drive**
  - **Drive is selected through the device address**
- **Host sends command code to trigger action**
  - **Command code indicates operation and mode**
- **Status registers read to determine completion**
- **CRC on data only**
- **Little command queuing support**

ATA Data Transfer Modes

- PIO (Programmed I/O)
  - Processor coordinates transfer of every 2B of data
  - No longer used for disk drives
  - super old school
- DMA
  - 16.7MB/s @11MHz
- Ultra DMA
  - DDR transfers (both clock edges)
  - 133MB/s @ 33MHz

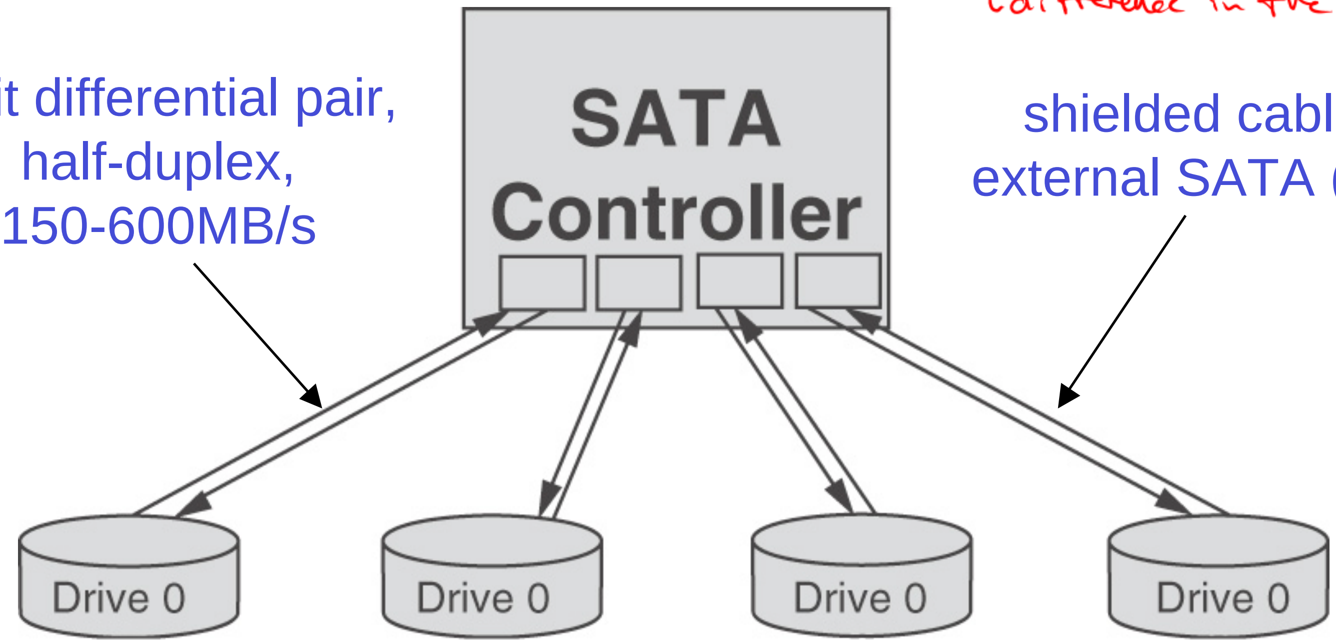
Serial ATA (SATA)

- **Uses serial point-to-point links to send packets**

(difference in the physical layer)

1 bit differential pair,
half-duplex,
150-600MB/s

shielded cables for
external SATA (eSATA)



- **Commands are compatible with ATA**
- **CRC on both data and commands**
- **Up to 32 entry command queue**

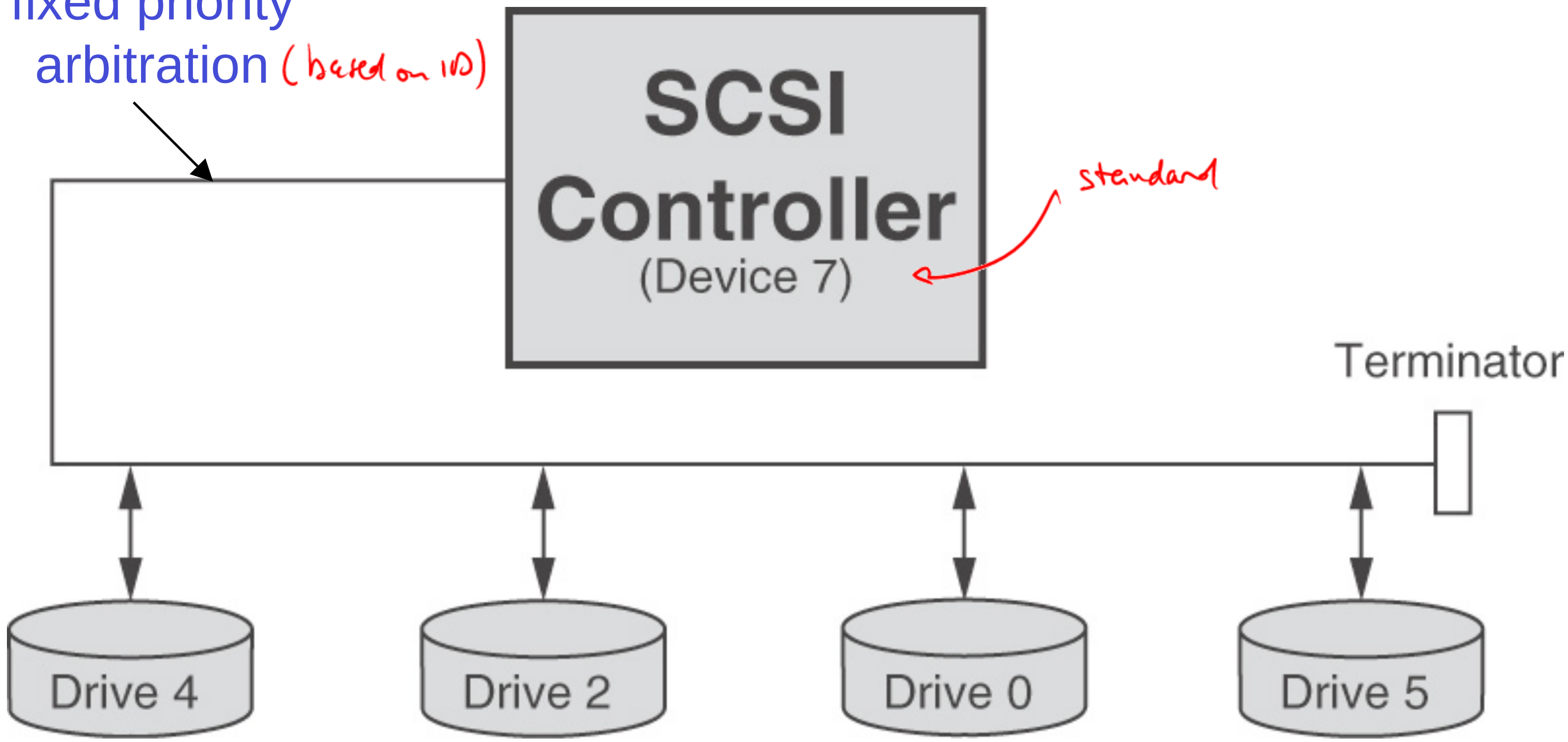
[20.3]

Small Computer Systems Interface (SCSI)

- Geared towards higher end devices than ATA
  - fancy server stuffs
- Variety of cabling/signaling/data width options
  - Narrow (1 bytes) or wide (2 bytes)
  - Internal unshielded or external shielded cabling
  - Single ended, high voltage differential (HVD), or low voltage differential (LVD)
  - Typically wide with LVD

SCSI Topology

fixed priority arbitration (based on ID)



up to 8 devices for narrow, 16 for wide (includes controller)

[20.4]

SCSI Operation Summary

highest ID #

- *Initiators* arbitrate for bus, highest one wins
- Command descriptor block (CDB) sent to *target*
  - Command data structure (>60 commands)
  - Operation, LBA, other bits
- Data transferred between initiator and target
- Target returns a status code

checks command bits integrity as well

- CRC on commands and data
- Up to 256 entry command queue with reordering

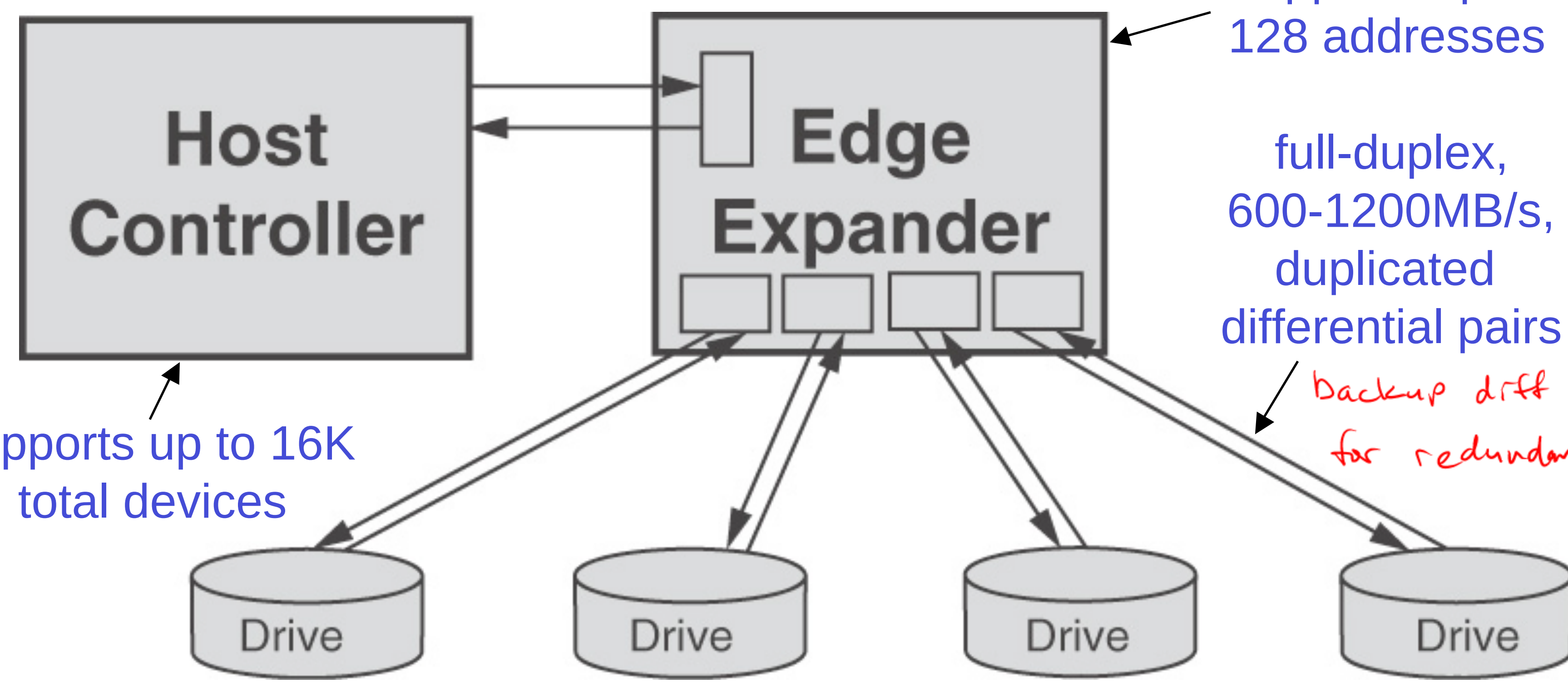
SCSI Transfer Modes

- Numerous modes added over time

- SCSI-1 (1986)
  - 1 byte wide
  - 5MB/s @ 5MHz

- Ultra-5 SCSI (2003)
  - 2 bytes wide
  - QDR transfers (both edges of 2 phase shifted clocks)
  - 640MB/s @ 160MHz

Serial Attached SCSI (SAS)



supports up to 128 addresses

full-duplex, 600-1200MB/s, duplicated differential pairs

backup diff pair for redundancy

supports up to 16K total devices

[20.5]

USB Mass Storage

- Device class for USB storage devices
- 60MB/s for USB 2.0; 625MB/s expected for 3.0

  pew pew pew!

- Includes read/write SCSI commands
- No command queuing support

Which Interface?

- SCSI has had a richer command set and more performance and reliability features (high end)
- ATA has been lower cost (PCs, laptops)
- SATA competes favorably with SAS in performance and reliability
- USB lagging in performance, but popular in lower cost applications (e.g., backup)

Next Time

Disk Drive Performance Issues and Design Tradeoffs