Hard Disk Drives (HDDs)

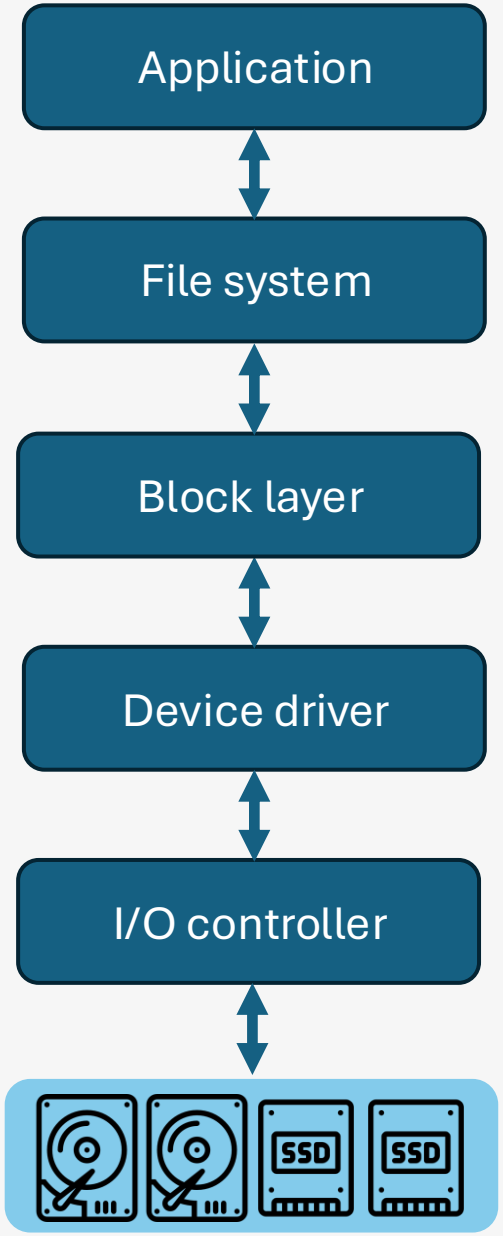
Juncheng Yang

Harvard John A. Paulson
School of Engineering
and Applied Sciences

HARVARD UNIVERSITY
MADSYS
Measurement and Design of Systems Lab

Agenda

- HDD internals
- HDD performance
- HDD reliability
- HDD density
- Future trend



Three Key Questions

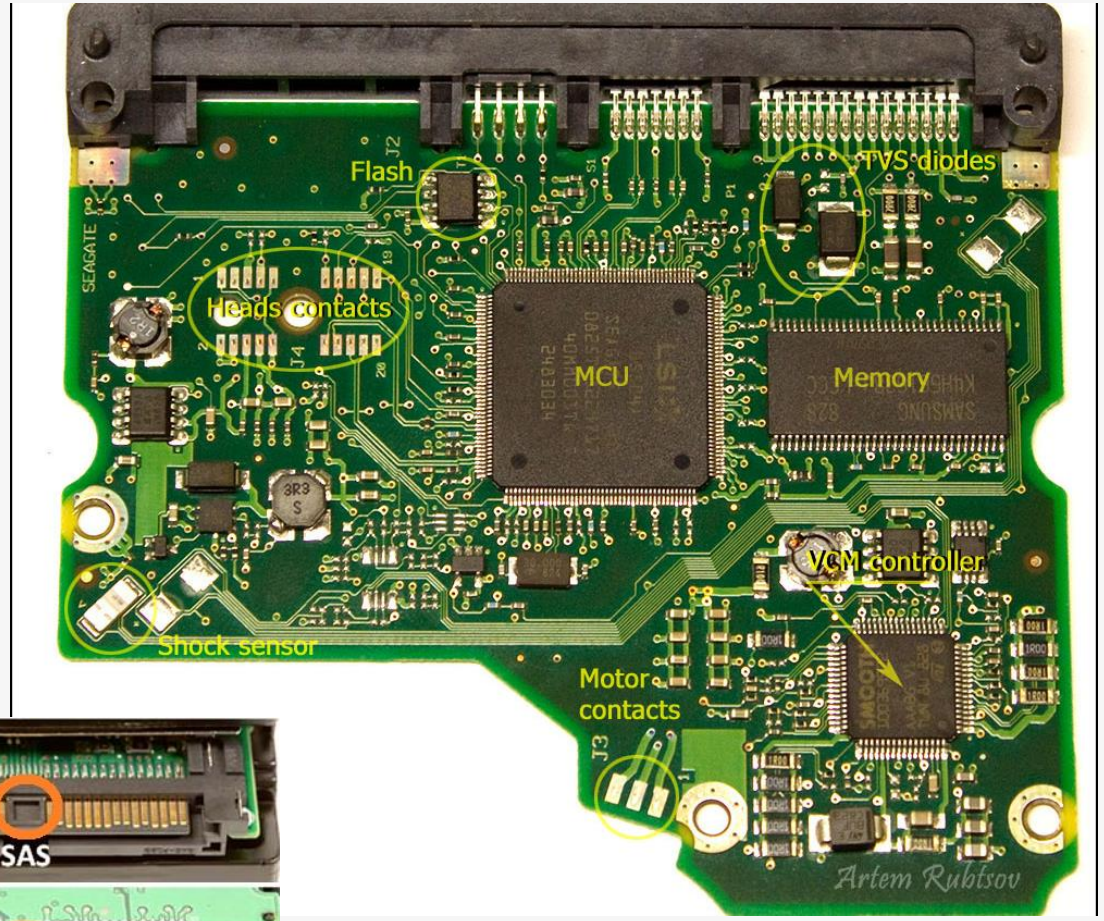
- What are the performance characteristics of a hard disk drive? Why?
- How do we measure disk reliability and how often do they fail?
- How has disk capacity and density improved?

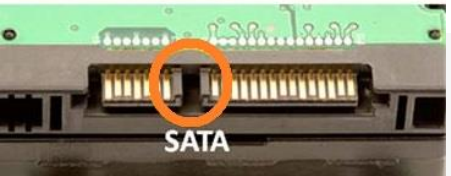
* Although this class we will discuss a bit hardware, they are not the focus

Basics

- Capacity: up to 30+TB
- Performance: 100-200 IOPS, 100-200MB/s, ~10ms
- Cost: $20/TB
- Physical size: 3.5” or 2.5”
- Rotation: 7200RPM
- Interface: SATA or SAS
- Lifetime: ~5 years

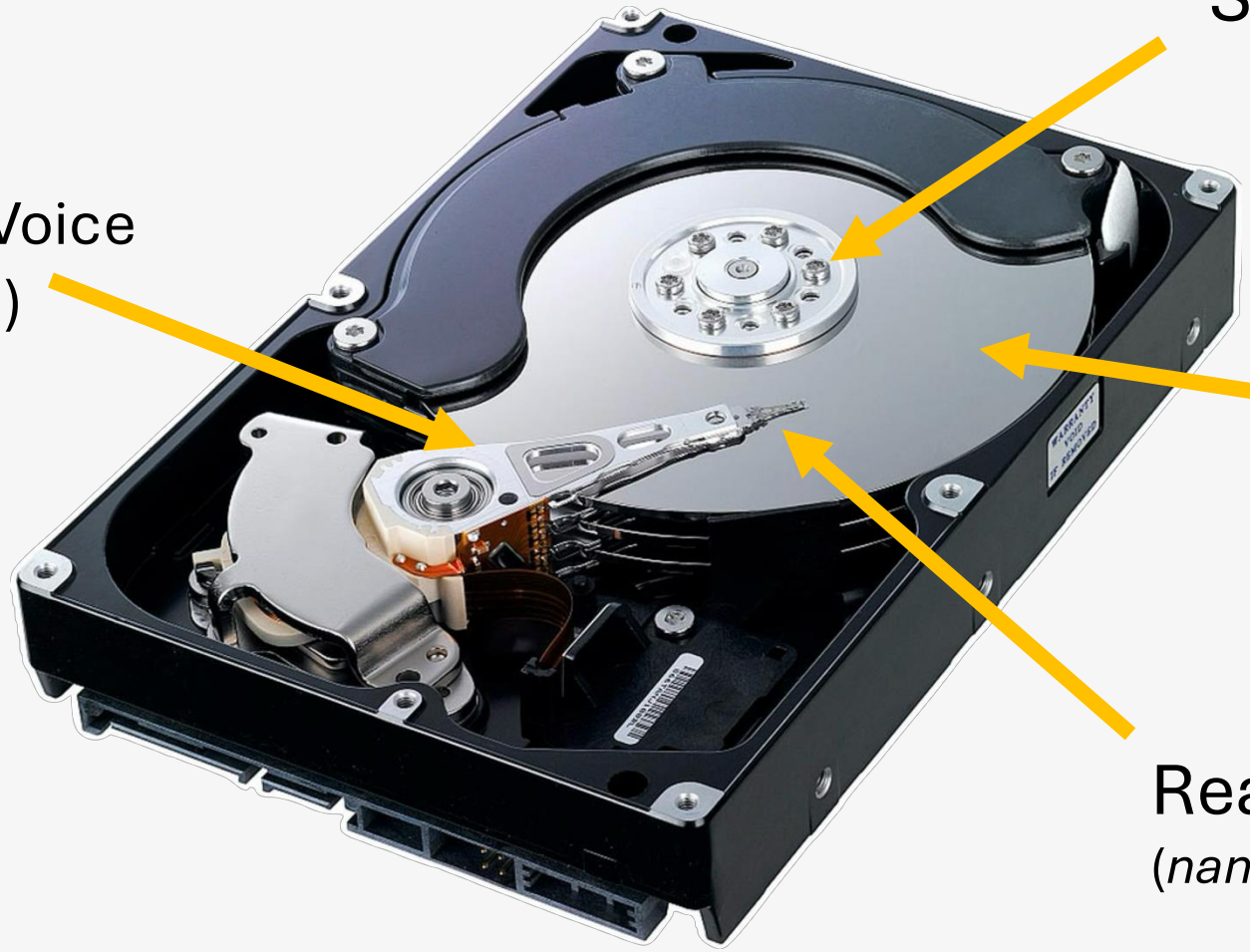
HDD internals: electronic components





- Just like a small computer
  - processor, memory and I/O interface
- ASIC
  - signal processing
  - error detection and correction
  - servo processing
  - motor/seek control
- Firmware running on the controller
  - request processing, queueing and scheduling
  - LBN-to-PBN mapping
  - defect management

HDD internals: mechanical components



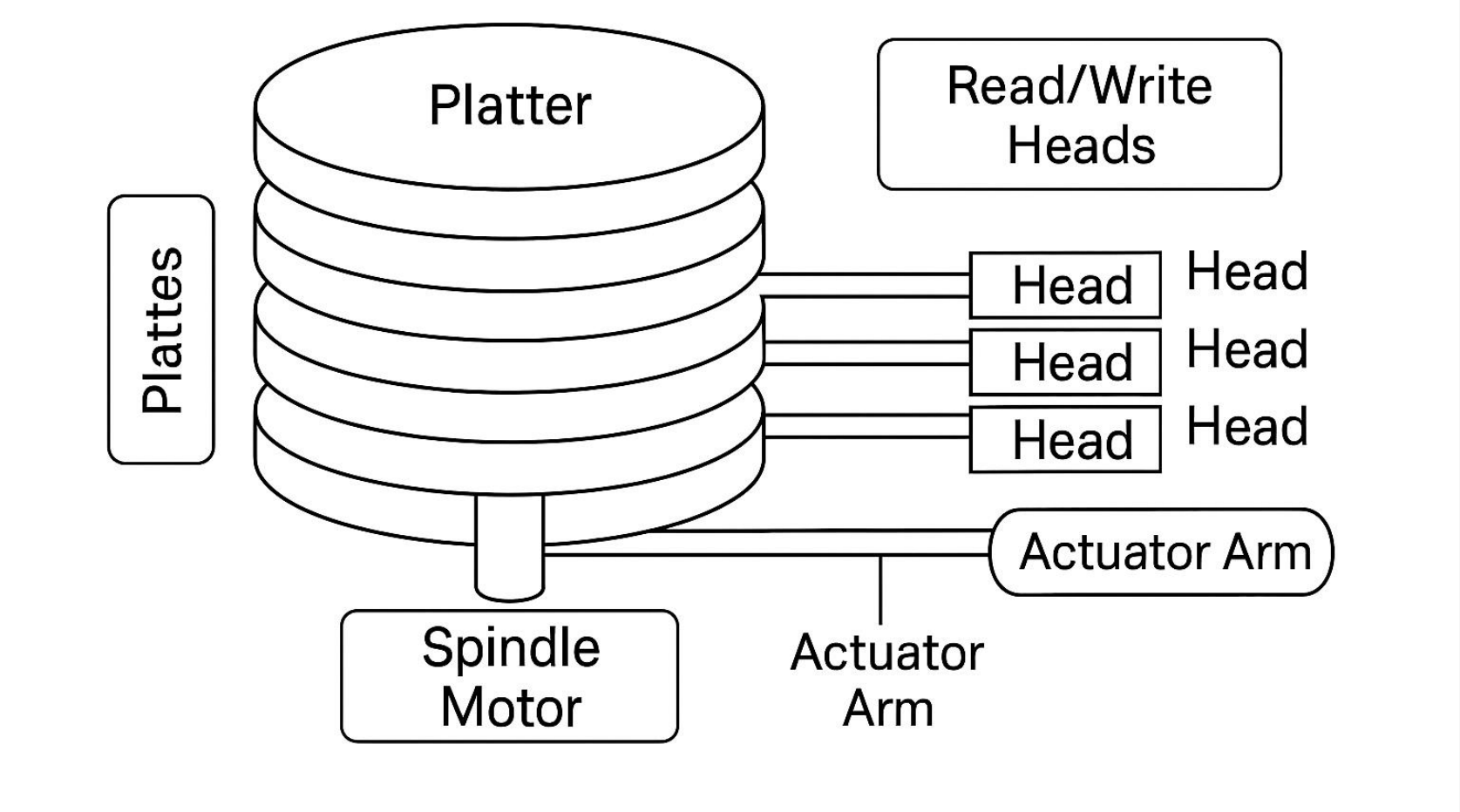
Spindle motor

Actuator Arm & Voice Coil Motor (VCM)

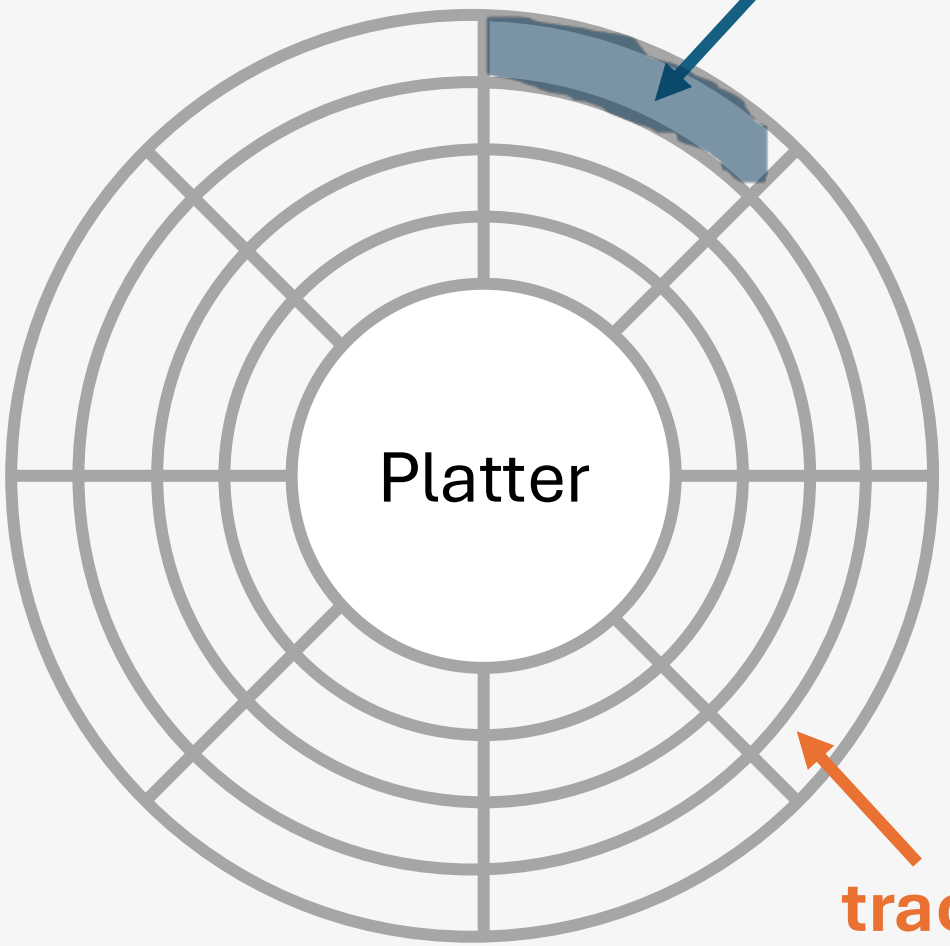
Platter (aluminum or glass substrates with **magnetic** coatings)

Read / write head
(*nanometers* away from platter)

HDD internals: mechanical components



HDD geometry



Does anyone see an efficiency problem?

Sector

smallest read/write unit

Track

consecutive sectors in the same concentric circle of a surface

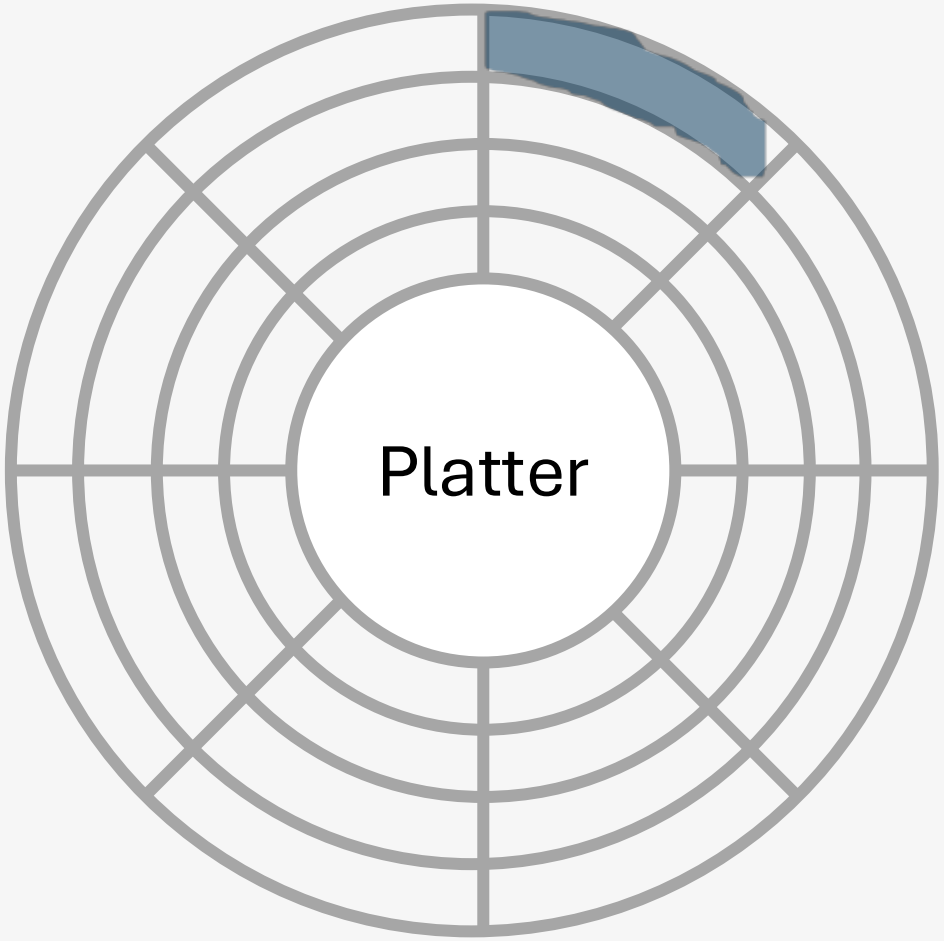
Cylinder (no longer used)

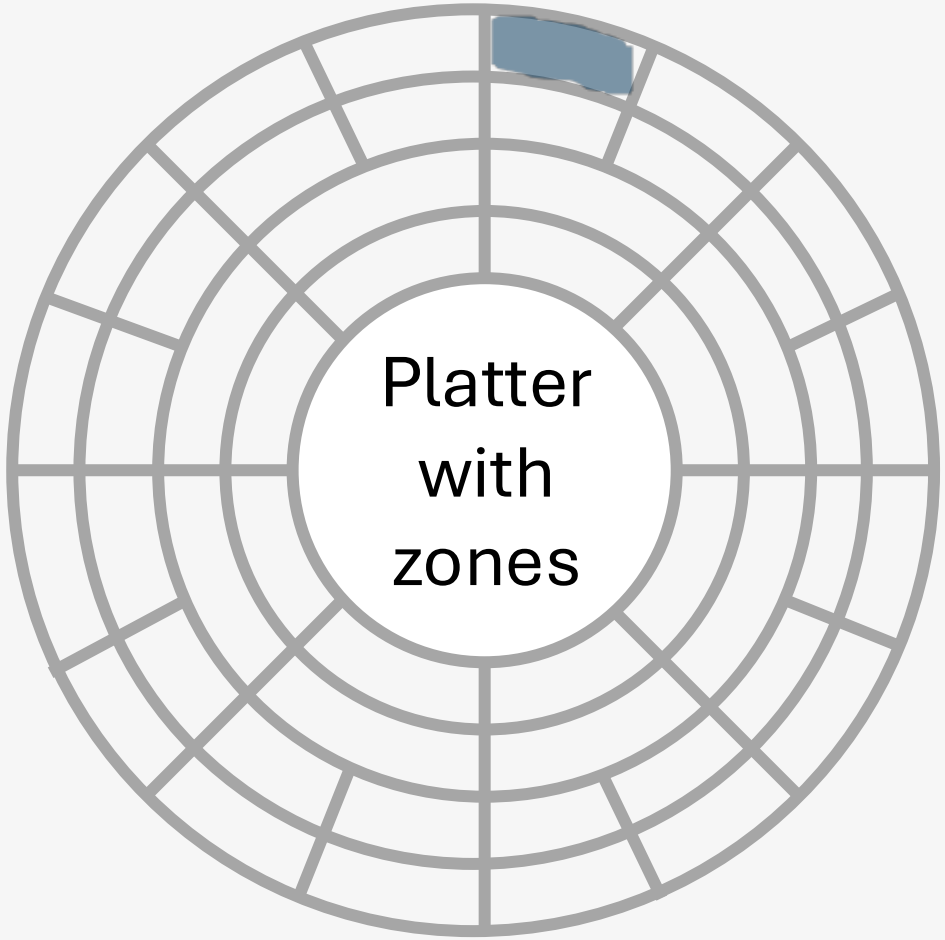
tracks on different surfaces at the same radius

HDD geometry

Zone

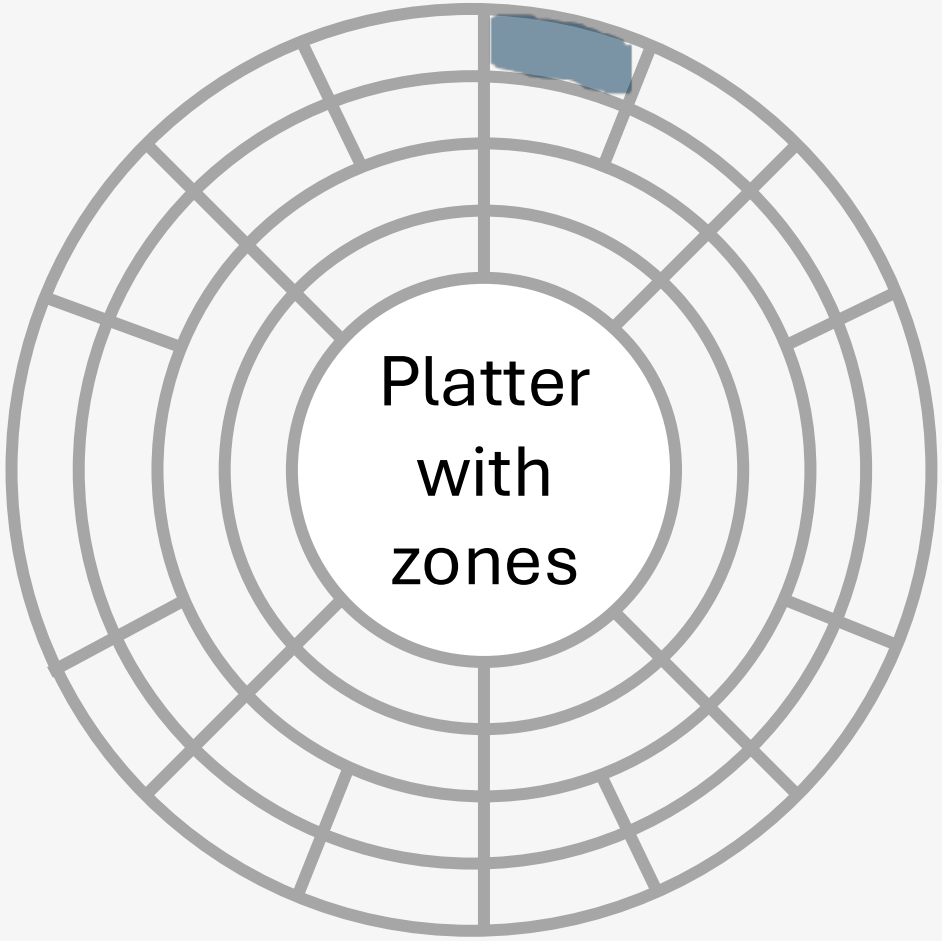
- allow outer tracks to have more sectors
- adjacent tracks with same #sectors





HDD geometry: some real numbers

- Sector
  - 1980-2011: 512 bytes, 2011-now: 4096 bytes
  - 1000s sectors per track (vary by zone)
- Track
  - 100,000s per surface
- Zone
  - 8-20 zones (not to be confused with ZNS)
- Platters
  - up to 10 platters (20 surfaces)



Addressing data

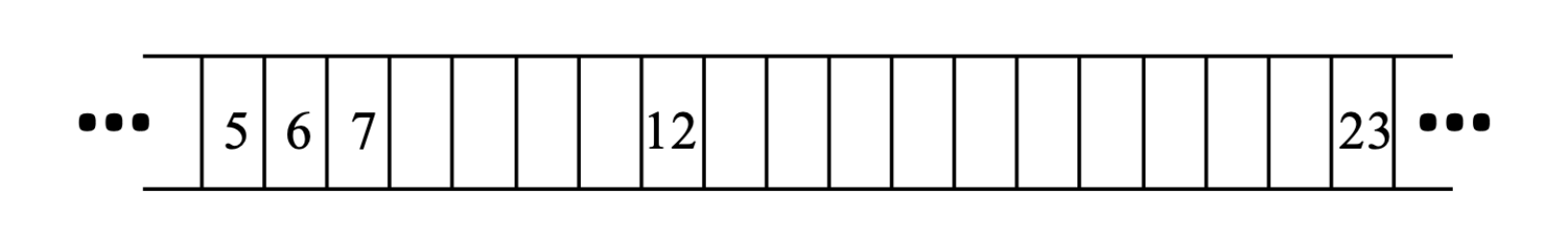
CHS (Cylinder-Head-Sector)

- legacy and no-longer used
- C: 0–1023 (10 bits), H: 0–254 (8 bits), S: 1–63 (6 bits)
- limited to 8.4 GB
- Problem?
  - OS needs to be aware of physical disk structures
  - OS needs to manage bad blocks

Addressing data

LBA (logical block addressing)

- expose the storage as a **linear logical** address space
  - block size: typically sector size or 4 KB
  - #block=device capacity / block size
- disk controller and firmware decide data placement
  - translation layer decouples software from hardware details
  - enable many disk internal management and optimizations



OS view of storage device

Implications of the HDD design

PERFORMANCE

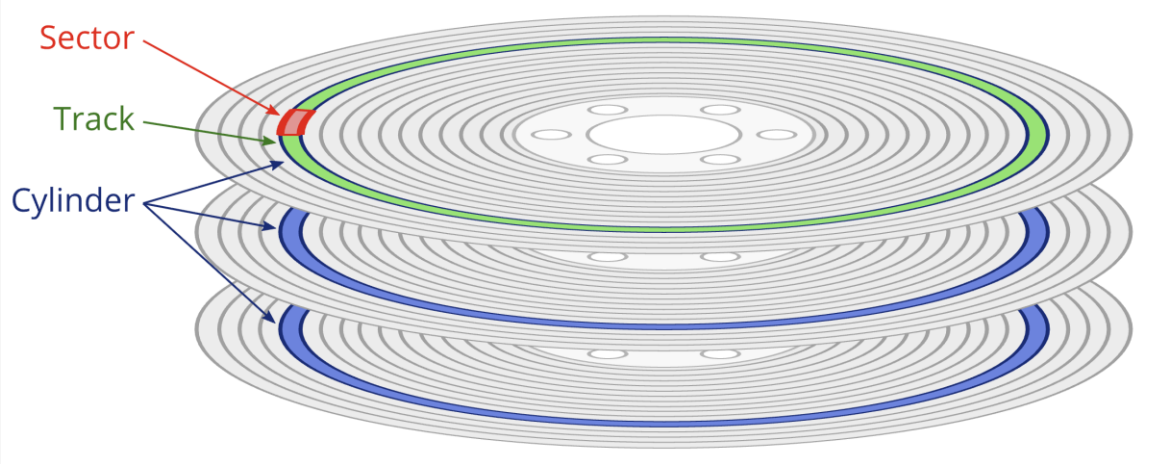
RELIABILITY

Access time

- Reading data
  - actuator arm ***seeks*** to the right track
  - platter is ***rotated*** to the right sector
  - activate the corresponding head

Slow physical movement

- Data access latency
  - seek
  - rotational
  - data transfer
  - command overhead (<1ms)



Access time

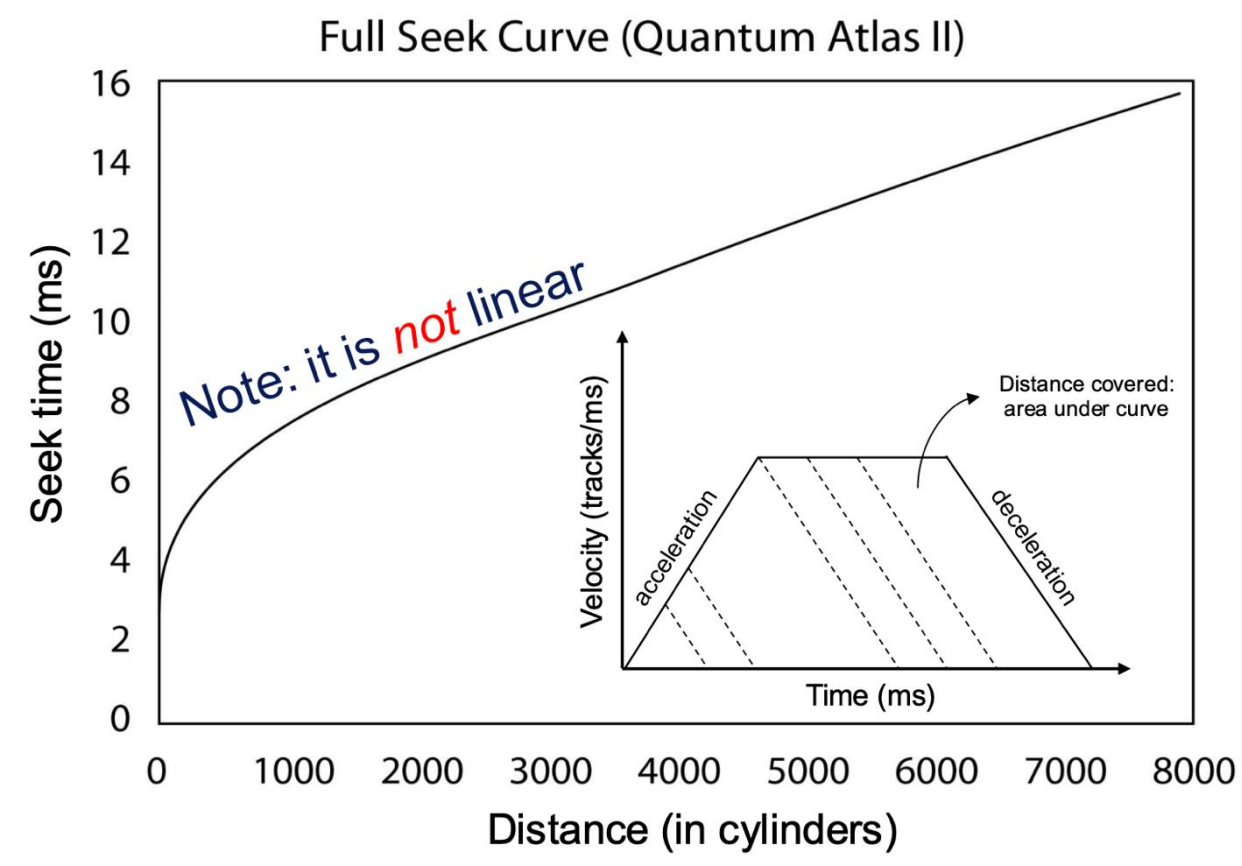
Seek time

Four components

- accelerate
- coast at max velocity (if go far enough)
- decelerate
- settle onto correct track (1-2ms)

Seek time (modern disk)

- average seek time: 2-10 ms
- track-to-track seek: 0.2-1 ms
- lower for high-end legacy drives



Trend: not improving much

Access time

Rotational latency

- Time required for head to reach the first desired sector
- Depends on rotation speed
  - measured in Rotations Per Minute (RPMs)
- Computing average rotational latency
  - assume an equal likelihood of landing on any sector of the track
  - this gives equal probability of each rotational latency from 0 sectors to N-1 sectors
  - thus, average rotational latency is time for 1/2 revolution

Example (7200 RPM)

- one rotation
  - 60s / 7200 = 8.33 ms
- average rotational latency
  - 8.33 / 2 = 4.16 ms

Trend: lower RPM, higher latency

Access time

Data transfer time

- Time for the sectors to rotate under head
- Depends on **data size** and **sustained transfer rate**
  - $T_{\text{tranfer}} = \frac{\text{data size}}{\text{sustained transfer rate}}$
- Sustained transfer rate in modern drives: ~200MB/s

Trend: ~10% improvement every year

Access time

Seek time

Rotational latency

Data transfer time

- Reading **4 KB** of data
  - seek time: 6 ms
  - rotational latency: 4 ms
  - data transfer time: 4 KB / 100 MB/s = 0.04 ms
  - effective bandwidth: 4 KB /10.4 ms = 0.4 MB/s
- Reading **8 MB** of data
  - seek time: 6 ms
  - rotational latency: 4 ms
  - data transfer time: 8 MB / 100 MB/s = 80 ms
  - effective bandwidth: 8 MB / 90ms = 89 MB/s

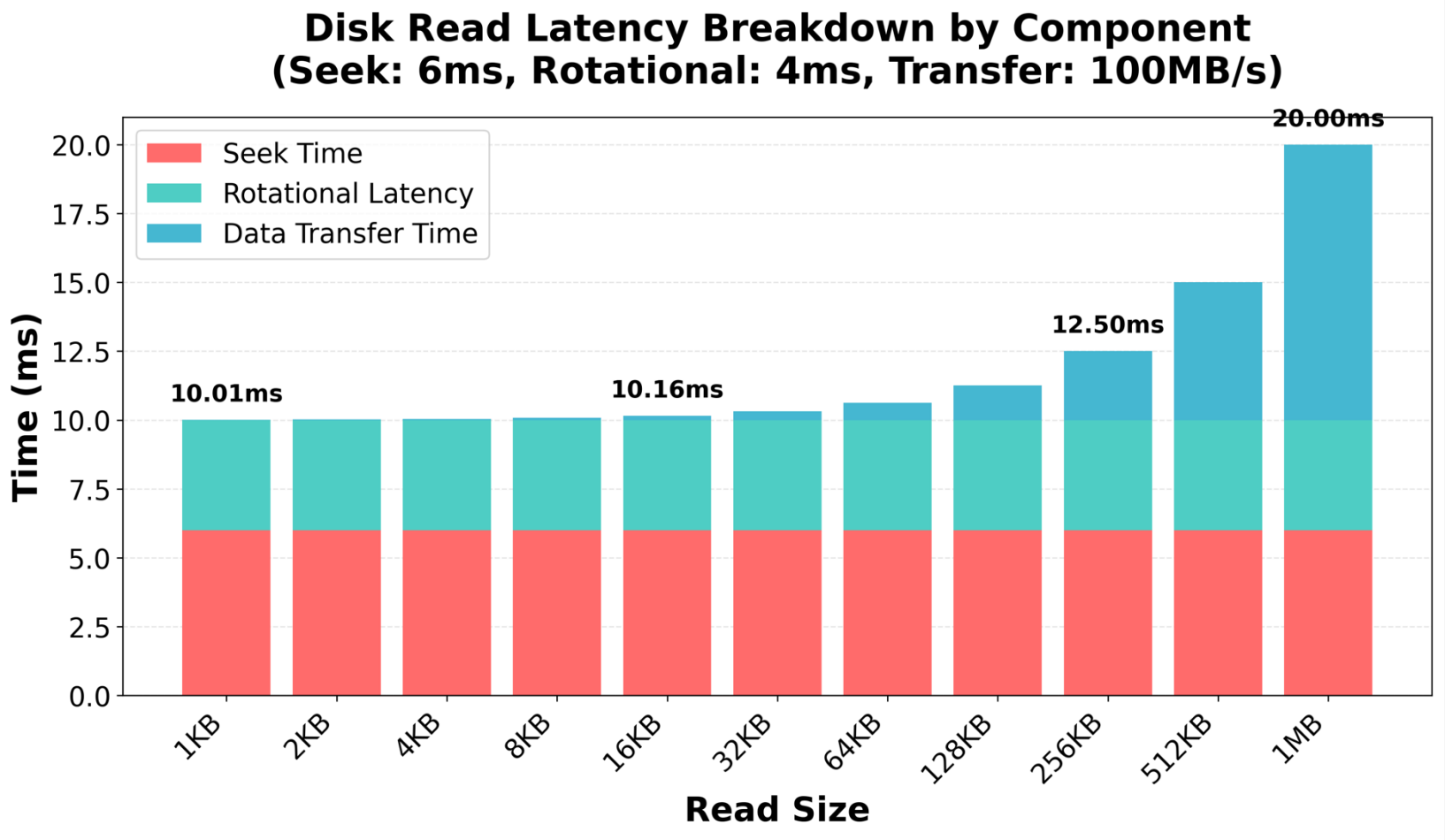
HDDs are not good at serving small random I/O operations.

Access time

Seek time

Rotational latency

Data transfer time

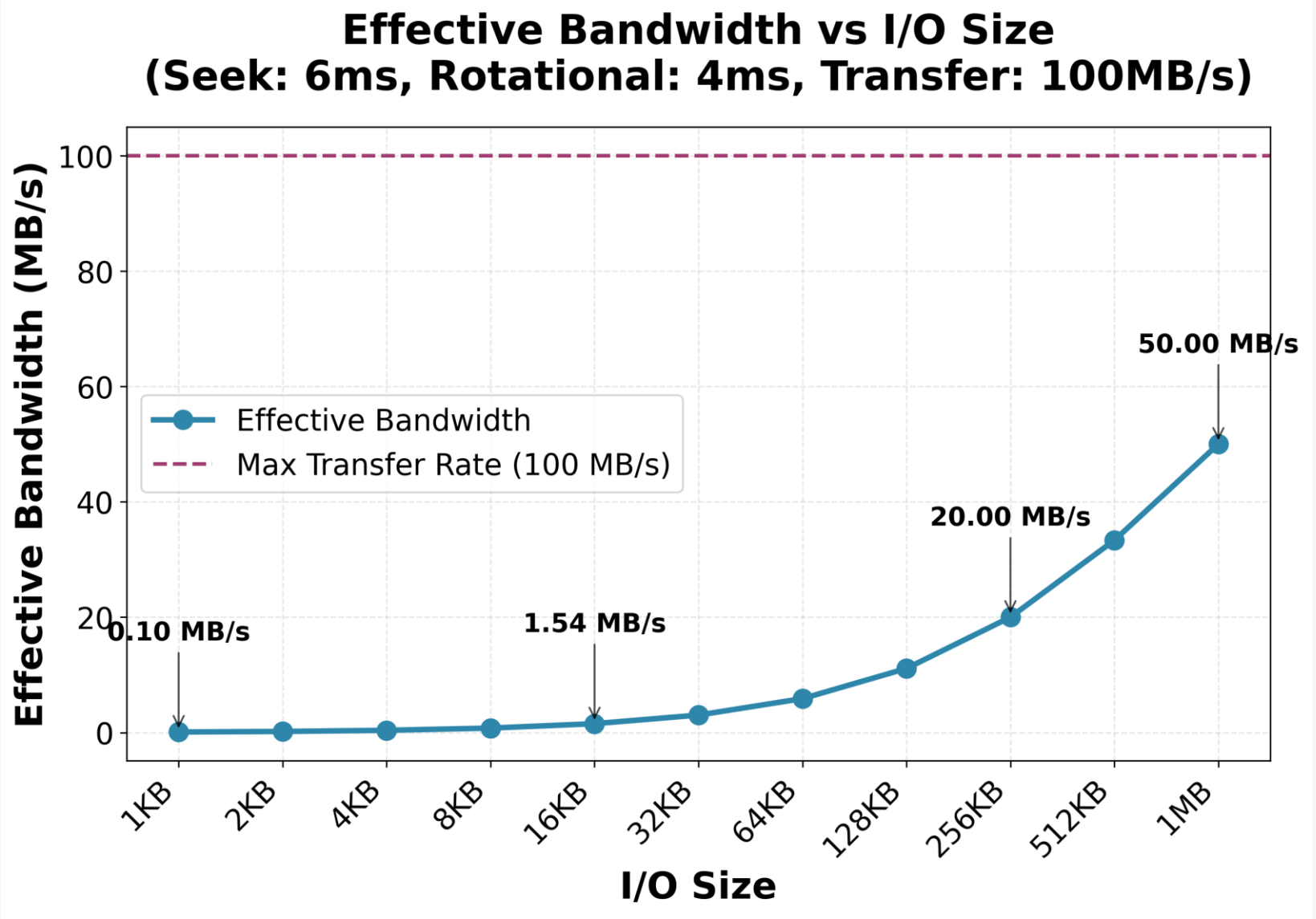


Access time

Seek time

Rotational latency

Data transfer time



HDDs are not good at serving small random I/O operations.

How do we achieve a high performance?

- Disk (device side)
  - Larger and tiered cache
  - Better scheduling algorithm with request reordering
  - Higher RPM (no longer used)
- Systems (host side)
  - Larger caches and prefetch
  - Defragmentation
  - Write combining
  - Careful data placement, e.g., create locality
- Software
  - Larger and sequential reads

How reliable is HDD?

Q: What type of vibration is allowed?

HDD is fragile

Shouting in the datacenter

https://www.youtube.com/watch?v=tDacjrSCeq4

What is reliability?

- **Reliability** (will it work?)
  - probability a device function as designed
  - measured by MTBF, AFR
- **Availability** (can I access it now?)
  - percentage of time a device is operational and accessible
  - measured by uptime percentage
- **Durability** (will data survive long-term?)
  - long-term persistence and integrity of the data
  - measured by data retention time and unrecoverable bit-error rate (UBER)

HDD failure modes

- Media degradation (sector-level)
  - surface defect and contamination
- Head failure
  - often catastrophic and lead to data loss
  - e.g., head crash from vibration or mechanical wear
- Firmware issues
  - software bug triggered in certain conditions
- Spindle / motor / actuator problems
- Electronics / PCB failure

Common HDD failure reasons

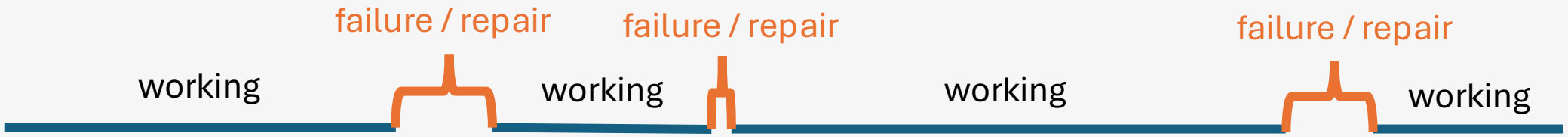
- Vibration, shock, dust
- Thermal issues
- Wear: mechanical and magnetic decay

Reliability metrics

- **MTBF (mean time between failure)**
  - **average time to failure of a large population of drives**
  - also called MTBR (mean time between repair/replacement)

$$\mathrm{MTBF} = \frac{\sum(t_{\mathrm{down}} - t_{\mathrm{up}})}{\#failures}$$

  - common MTBF: 1 million hours (114 years)



Reliability metrics

- **AFR (annual failure rate)**
  - approximate fraction of drives that fail in one year
  - calculating AFR from MTBF
    - Poisson process with exponential time-to-failure and a constant failure rate
    - failure rate: $\lambda = 1/\mathrm{MTBF}$
    - device survive for time t: $P(T > t) = e^{-\lambda t}$
    - AFR: $1 - P(T > \text{one year}) = 1 - e^{-t/\mathrm{MTBF}} \approx t/\mathrm{MTBF}$
- A MTBF of 1 million hours (114 years)
  - => each year there is a 0.9% chance failing, 0.9% of all disks fail each year

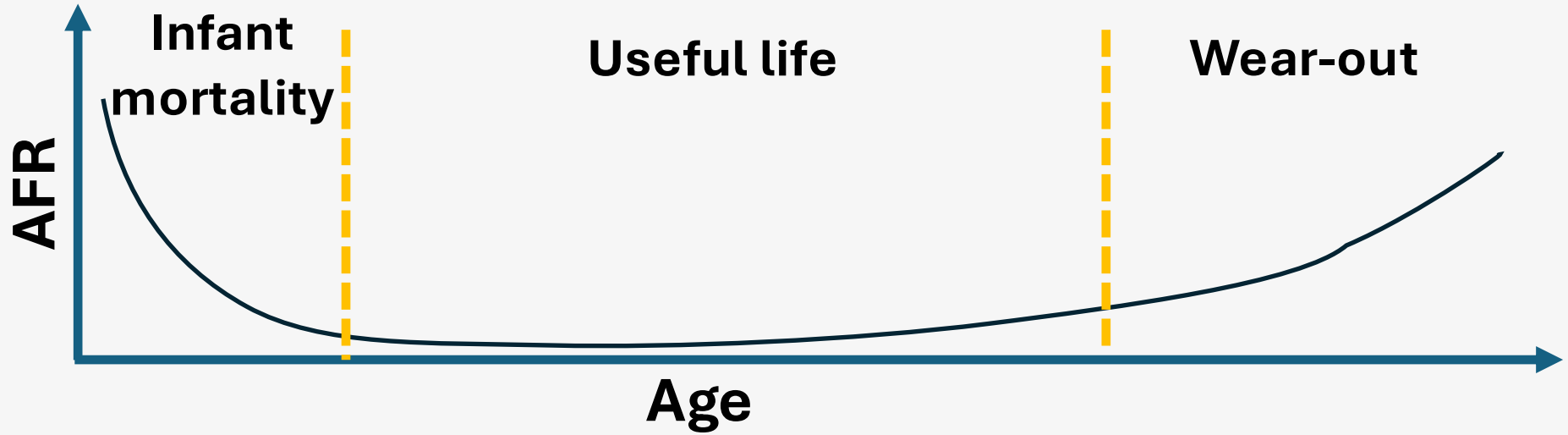
Reliability metrics

- **Unrecoverable Bit Error Rate (UBER)**:
  - Raw Bit Error Rate is much higher (10^{-1} - 10^{-2}), but most can be corrected
  - typical spec: 10^{-14} to 10^{-16} errors/bit
  - => an uncorrectable read while reading many TB of data
  - disk is still heathy when UBER happens
  - make high-capacity disk rebuild challenging

- **Workload ratings**:
  - TB/year, duty cycle (e.g., 24×7 vs 8×5)
  - typical value: 55 TB/year (consumer) 550 TB/year (enterprise)

Bathtub curve

- Failure rate over time
  - **Infant mortality:** early-life failures, manufacturing defects
  - **Useful life:** low, steady failure rate
  - **Wear-out:** aging, increasing failures
- Implication
  - burn-in new drives
  - proactively replace old drives



Gray failure

- So far, we have been focusing on fail-stop
- Gray failures are common
  - latent sector error (LSE)
  - intermittent read/write error
  - performance degradation (fail-slow)
- Gray failures are dangerous
  - hard to discover (non-deterministic)
  - increase risk during rebuild
  - silent data corruption

Production
gray failures are often masked, but cause more outages

Correlated failure

- Assumption: disk failures are **independent**
- Reality: no
  - same batch with manufacturing defects
  - environmental correlation: temperature, humidity, vibration, power surge
  - operational correlation: heavy and imbalanced workload
  - wear-out correlation: disks in the same server have similar ages
- Solution
  - diversity in sourcing
  - place data on different failure domains
- Not much impact when we consider a single disk, more important later (will revisit)

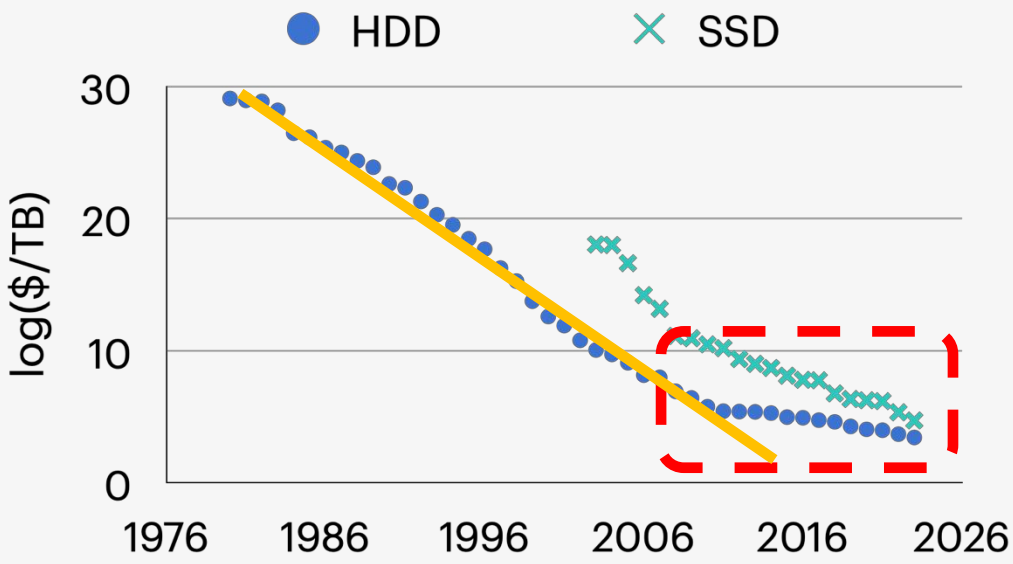
How do we detect and mitigate failure?

- S.M.A.R.T. (**S**elf-**M**onitoring, **A**nalysis, and **R**eporting **T**echnology)
  - key attributes:
    - Reallocated Sector Count, Current Pending Sector Count
    - Uncorrectable Sector Count, Spin Retry Count
    - Power-On Hours, Power Cycle Count, Temperature (Current, Max)
  - limitation: noisy, not predictive of disk failure (disks may fail with good SMART attributes)
- Error correction & mapping
  - ECC on the fly
  - bad sector management
- Redundancy
- Background scrubbing

HDD density

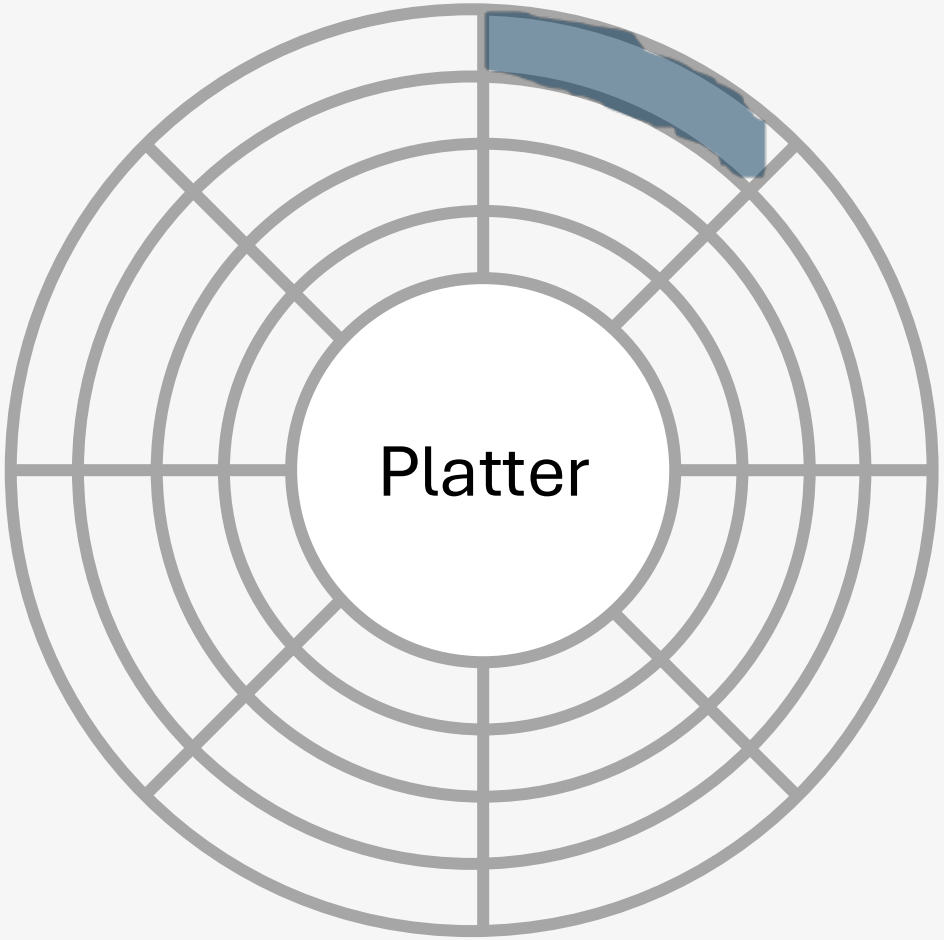
HDD areal density

- Areal density: bits per unit area
- Kryder's Law (~2000):
  - "*The density of information on hard drives doubles every 13 months*."
  - Stopped around 2011
- Density not expected to grow forever due to superparamagnetic limit
  - thermal forces bits to swap on their own
  - => data cannot be retained



Aspect ratio for bits

- **Linear density** (density within track)
  - 10-20X track density
- **Track density**
  - offer more room for improvement
  - doesn't help data transfer and can hurt track-following



How do we get around?

MR: Magnetic Recording

- **Reducing magnetic grain size**
  - **LMR** (Longitudinal MR):
    - bits are magnetized *parallel* to the disk surface (along the track)
  - **PMR** (Perpendicular MR)
    - change bit direction for better stability at small sizes
    - bits are magnetized *perpendicular* to the disk surface (up/down through the thickness of the magnetic layer)
- **Improving track density**
  - **SMR** (Shingled MR): overlapping tracks (write head is inherently wider than read head)
  - **TDMR** (Two-Dimensional MR): better readback technology using a second sensor
- **Making stable bits easier to write**
  - **HAMR** (Heat-Assisted MR), **MAMR** (Microwave-Assisted MR)

HDD Power Consumption

- bonus materials

HDD power consumption

- **Consumer 3.5" drives (7200 RPM):**
  - Idle: 3-5 watts, active (random I/O): 6-10 watts, sequential I/O: 8-12 watts
- **Enterprise 3.5" drives (7200 RPM):**
  - Idle: 4-6 watts, active: 8-12 watts, sequential I/O: 10-15 watts
- **High-performance enterprise (10K-15K RPM):**
  - Idle: 6-10 watts, active: 12-18 watts
- **2.5" laptop drives (5400 RPM):**
  - Idle: 0.5-1 watt, active: 1.5-3 watts
  - Much lower due to smaller platters and slower rotation

HDD power consumption

- Spindle motor: 40%-60% of active power
  - Must overcome aerodynamic drag—this increases with platter count, diameter, and RPM
  - Helium drives reduce drag, lowering spindle power by ~20–30%
- Actuator assembly: 20–30% of active power (strongly workload dependent)
  - Rapid seek movements require substantial current
  - High-IOPS workloads (random reads/writes) drive VCM power higher
- Others: 0.5-2W

Sequential workload is not only faster, but also saves energy

Cost of an HDD

- Platters (glass/aluminum + magnetic media): 30-40%
- Read/Write heads: 20-25%
- Motor assemblies (spindle + VCM): 10-15%
- PCB and electronics: 10-15%
- Mechanical enclosure: 5-10%
- Assembly and testing: 5-10%
- R&D and firmware development (amortized): 5-15%

Market

- Three manufacturers and oligopoly
  - barrier to entry (fab cost, HAMR/MAMR patent) is too high
  - **Seagate**: ~43% market share, betting on HAMR
  - **WD**: ~37% market share, more conservative using ePMR and MAMR, diversified with flash business
  - **Toshiba**: ~20% market share, lag behind 1-2 years on density
- Market Segmentation
  - client market: dying, shrinking by 10-15% annually
  - enterprise nearline: 90% of the market, cloud demands grow at 30% rate, a lot of demand from AI indirectly

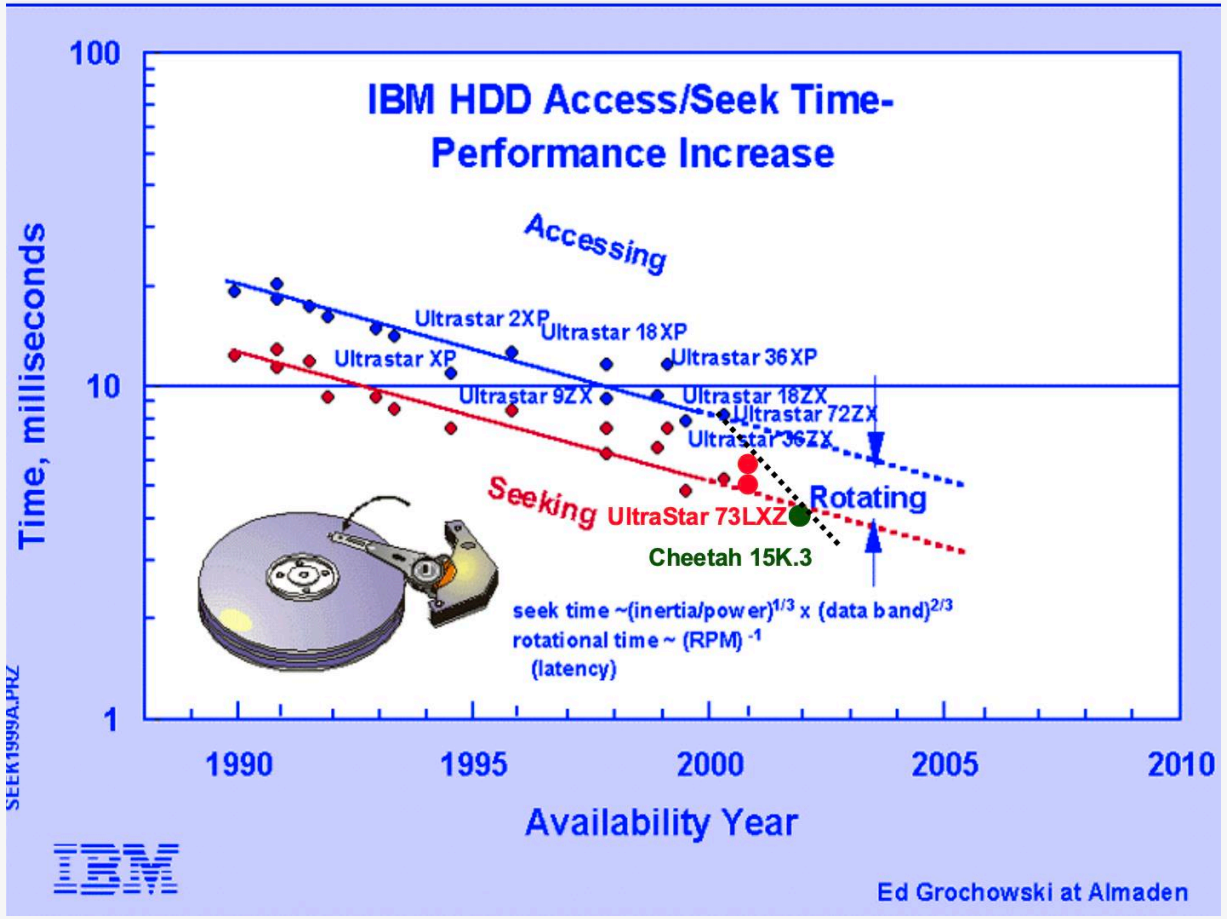
Why we discuss trend and why should you care?

- More about research
- If you work on hardware, these are where future innovations are needed
- If you work on software systems, knowing the hardware capability is critical
- If a problem *will be* solved in a lower layer soon, innovations in upper layer is not meaningful—*think, plan and research for the future*
  - example 1: if disk will be 100% reliable in 5 years, how should we handle the redundancy in today's systems
  - example 2: improvements in pre-trained model makes many fine-tuning less meaningful
- If a problem *cannot* be solved in hardware, can we solve it in software systems?
- What new problems may appear in 5 years, 10 years, 20 years given the trend?
  - think ahead: this is where impactful research direction happens
  - identifying problems before they show up

HDD historical trend

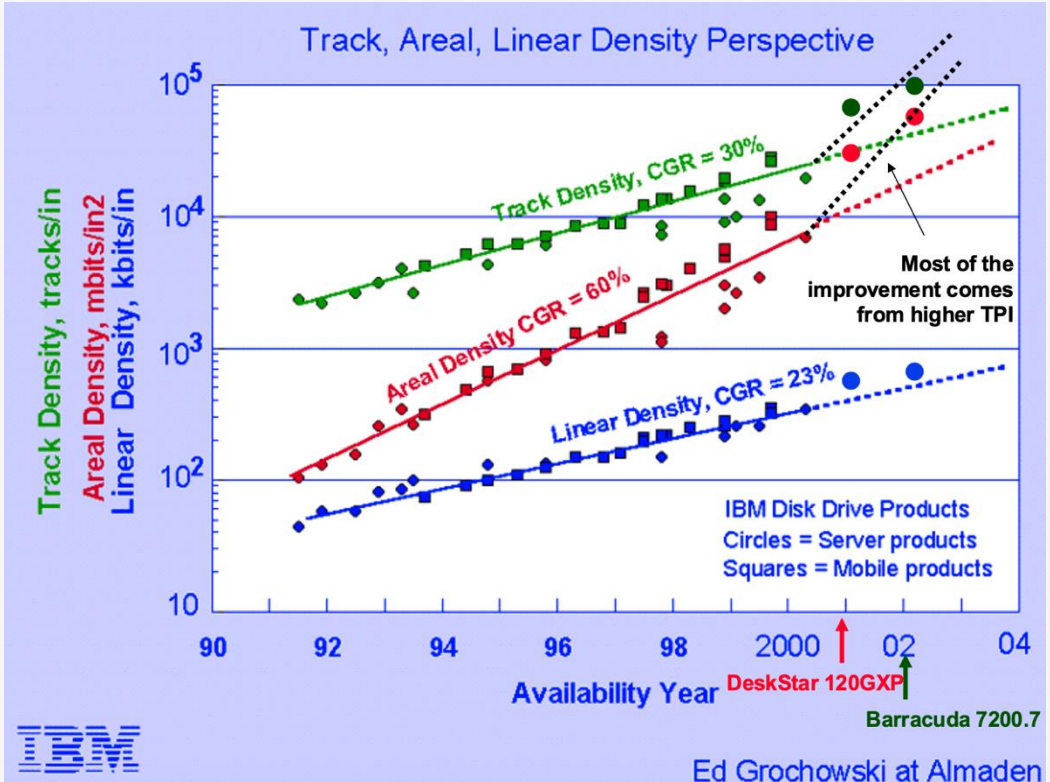
Performance

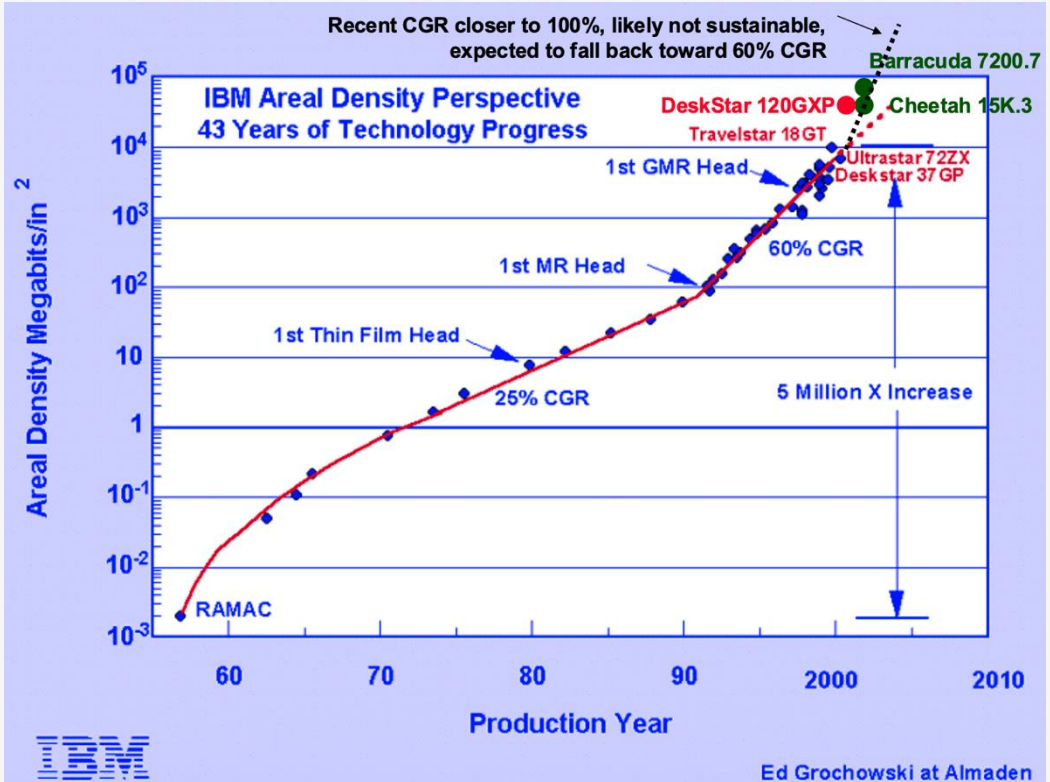
- **seek and access time**: exponential improvement till ~2000, then very slow
- **rotation latency**: barely improved
- **bandwidth**: slow improvement
  - come from density improvement
  - 1990s: ~10 MB/s
  - 2000s: ~60 MB/s (improved recording)
  - 2010s: ~120 MB/s consumer / ~220 MB/s enterprise (PMR maturity)
  - 2020s: ~150 MB/s / ~260 MB/s (helium, more platters)



HDD historical trend

- Capacity and density
  - exponential improvement (Kryder's law)
  - but has slow down since 2011





HDD historical trend: summary

- Capacity: **~7 orders of magnitude** increase
  - ~5 MB (ST-506, 1980) → 20–36 TB in 2025
- Bandwidth: **two to three orders of magnitude** improvement
  - ~0.6 MB/s (early PC) → 10–40 MB/s (1990s) → 100 MB/s (2000s) → 200 MB/s (2020s)
- Latency: only **a small multiple improvement**
  - seek: ~80–170 ms seeks -> ~8–12 ms on mainstream 7200 RPM drives
  - rotational latency ~4 ms hasn't changed much in 20+ years
- Random IOPS: similar to latency
  - still in low hundreds

Future trends

- Innovations have slowed down
- Performance
  - latency and IOPS: stagnant
  - bandwidth: improved slowly with density
- Density and capacity: continue slow improvement
  - improvement will show step function
- Other
  - SMR drive is becoming mainstream (cloud adoption)
  - disaggregated storage: better scaling, amortize compute, networking costs
  - market: less consumer demand, mostly industrial cloud infrastructure

Future trends: implications

- IOPS and bandwidth per TB will drop quickly
- High density brings challenges to reliability: slow data recovery

Summary

- HDD internals
- HDD performance
- HDD reliability
- HDD density
- Future trend

- What are the performance characteristics of a hard disk drive? Why?
- How do we measure disk reliability and how often do they fail?
- How has disk capacity and density improved?

Next time

- Solid-state drive