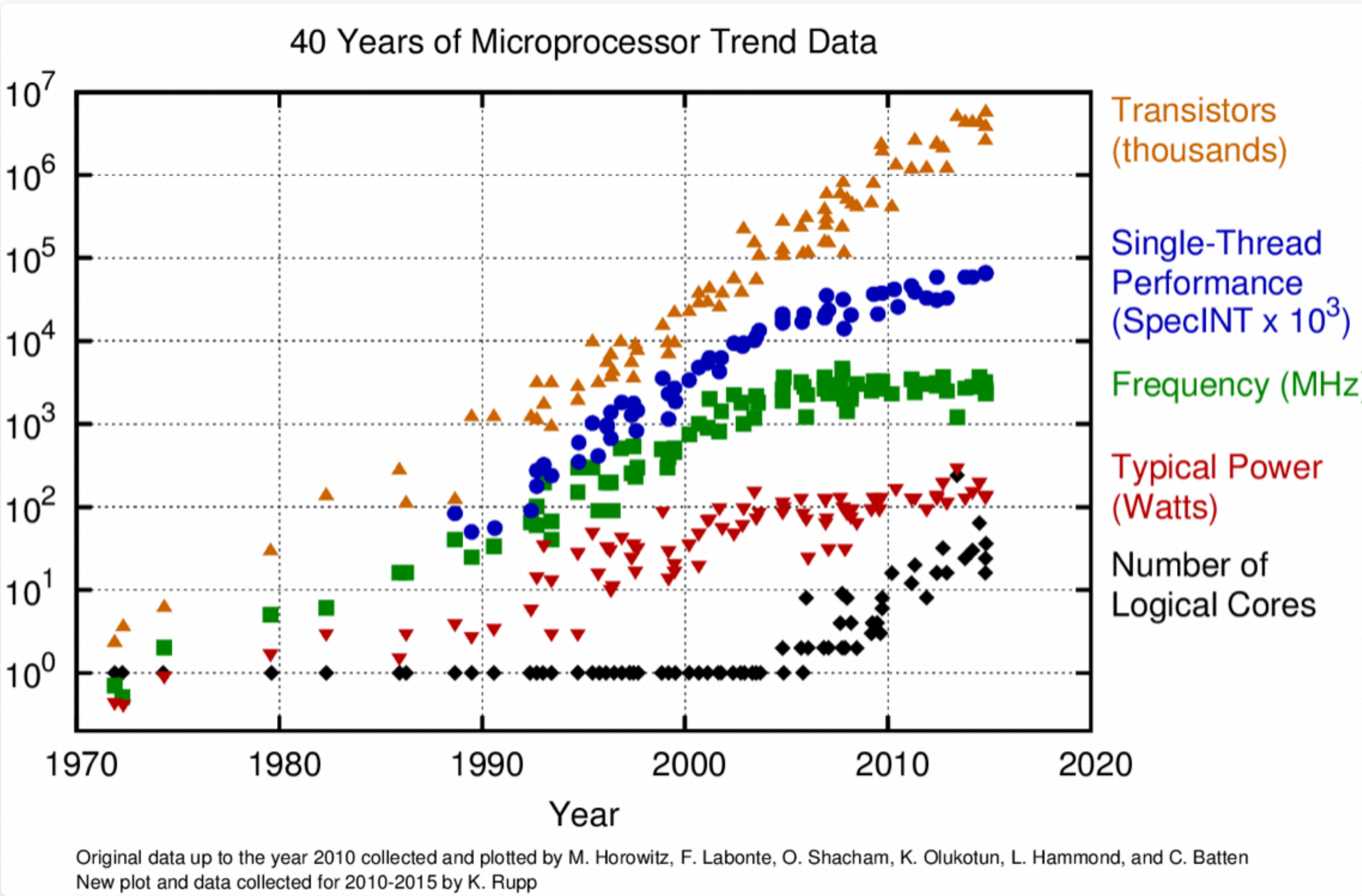
Data Structures and Architecture

Thomas Schwarz, SJ

Memory Hierarchy

- Stable for the last 50 years:
  - CPU with registers
  - Memory (DRAM)
  - Storage (Disk Drives)
  - Tape (Archival)
- Changes:
  - Increased size and number of on-chip caches between CPU and Memory
  - Change from disks to flash memory

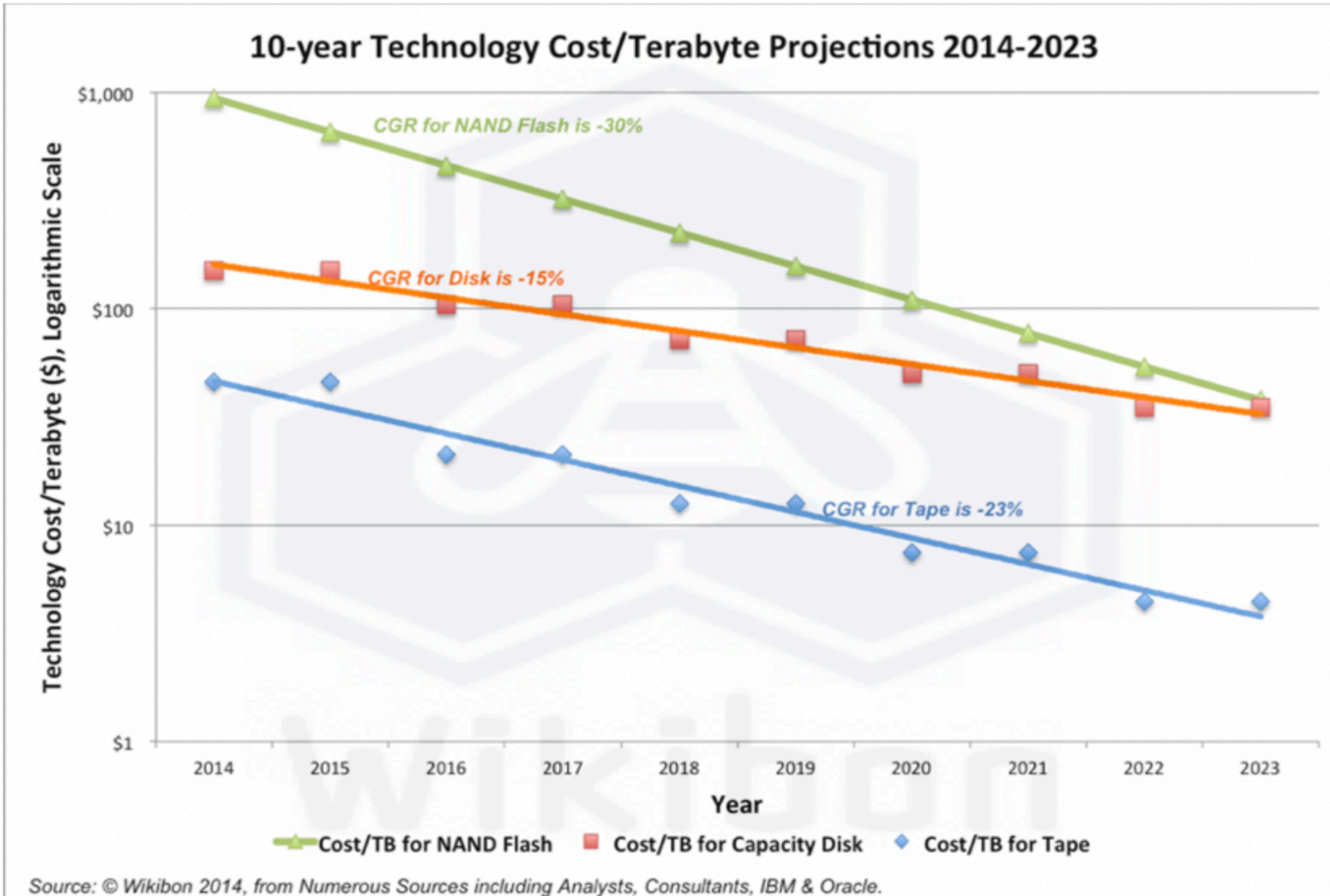
Memory Hierarchy



Memory Hierarchy

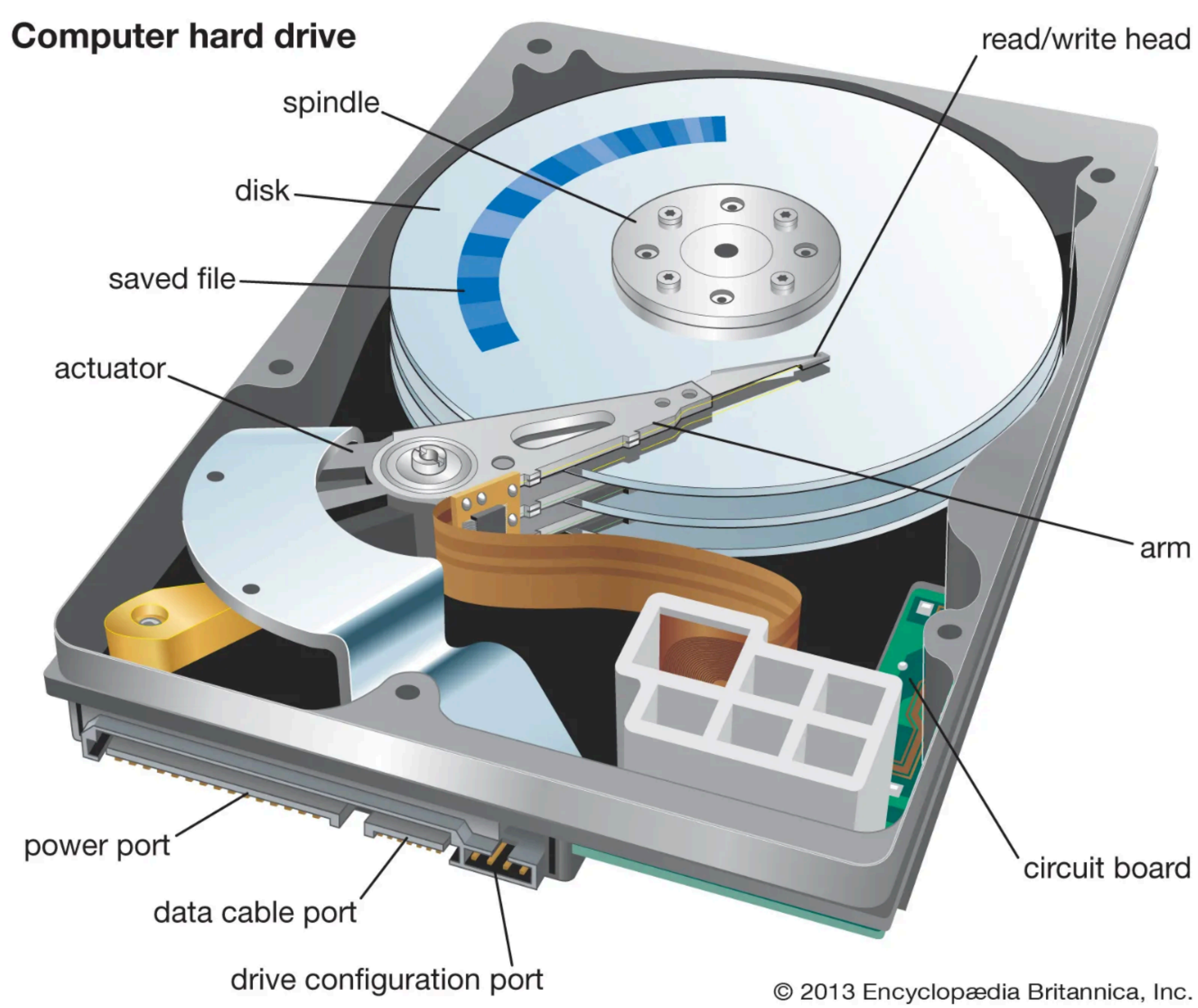
- Tape:
  - Low cost, high capacity, needs mounting
  - Go-to solution for archival data and backup

Storage Developments



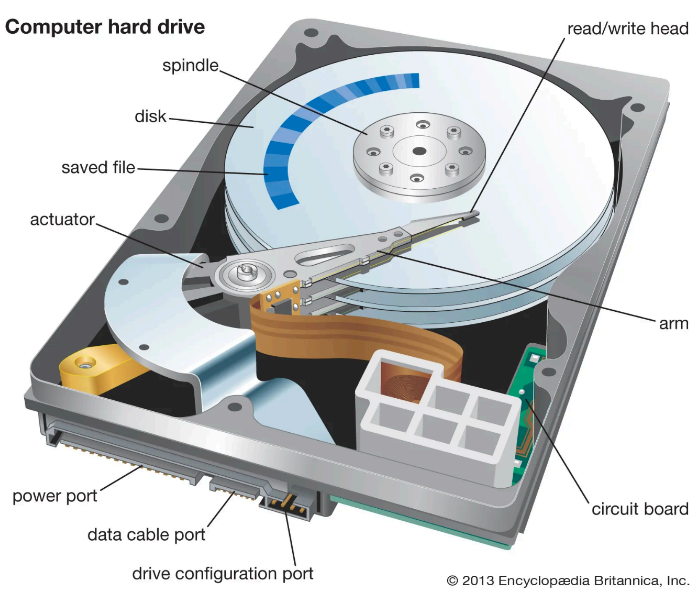
Hard Disk Facts

- Hard disk drives:
  - Electro-magneto-mechanical devices



Hard Disk Facts

- Hard disk drive access time:
  - Place actuators over track (**seek time**)
    - Use servo-information within a track for placement
    - Dependent on surface as size of the actuator arms differ because of different temperatures
  - Wait for disk sector to appear under selected actuator head (**rotational delay)**
    - Start transferring data (**transfer time**)



Hard Disk Facts

- Rotational delay determined by rotational speed (rotations per minute)
  - Needs to be high enough such that air resistance lifts heads
  - Needs to be small enough such that head lift remains constant
    - Otherwise resulting in head crashes
- Smaller disks can have slower rotational delay

Hard Disk Facts

- Seek time:
  - Depends on actuator movement
  - Limited by actuator mass and range of movement

Hard Disk Facts

- Sectors are made up of blocks
  - Initially of size 512B, then 4KB
  - Blocks have a logical block number
  - Successive blocks on a track receive successive numbers
  - When we switch to the next track, next logical block is such that we can stream with only a track-to-next-track seek time in between

Hard Disk Facts

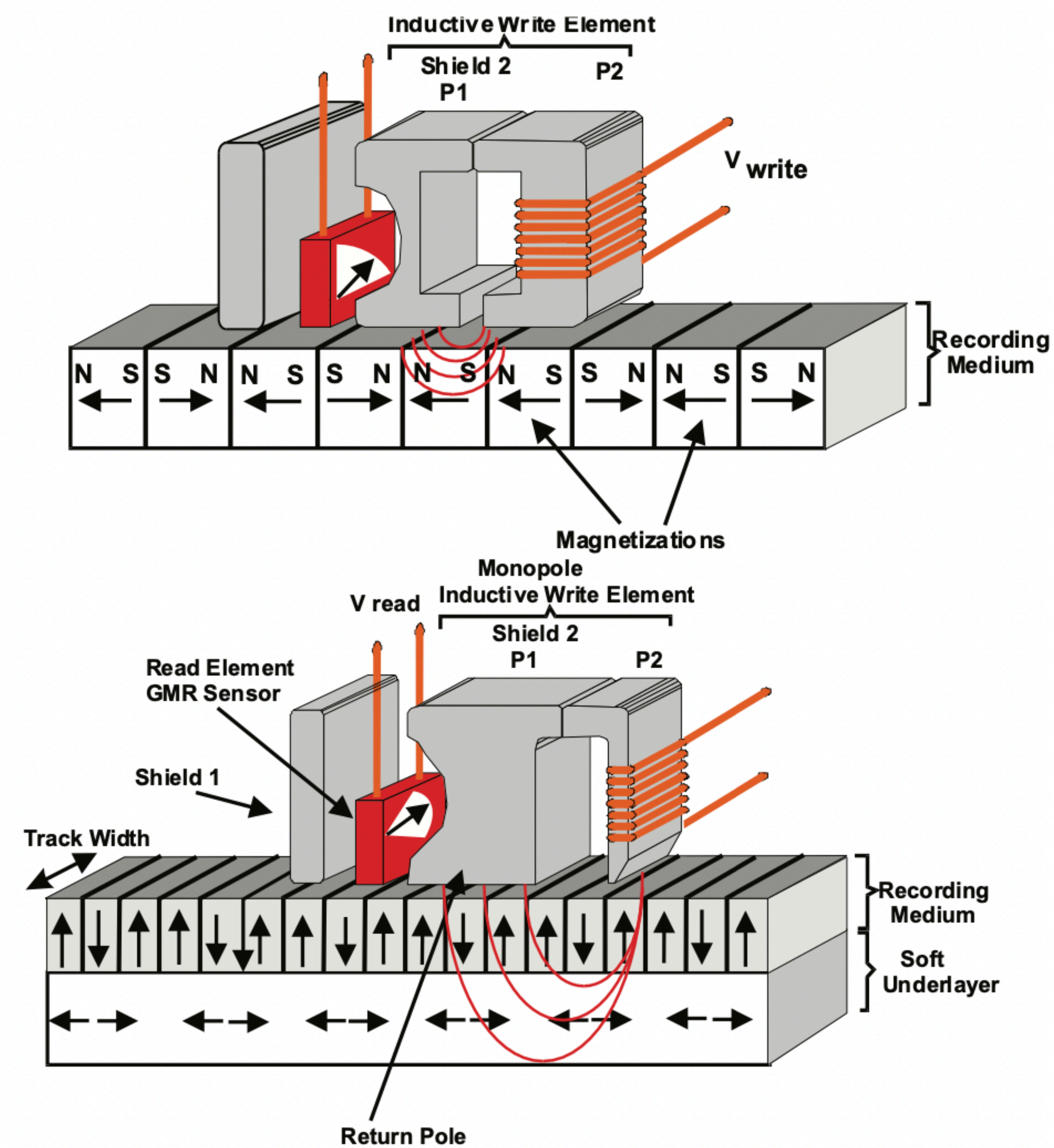
- Integrated error control and magnetic code
  - About 100 B parity data per 4KB block
- Use spare blocks and spare tracks to deal with material defects
  - Blocks cannot be used and are electronically replaced

Hard Disk Facts

- Streaming from Hard Disks
  - Continuous reads can reach 200 MB/sec
  - Rotation time is dominant factor
  - Outer tracks have more blocks and streaming is faster
- Random accesses from Hard Disk:
  - Performance is poor
    - Rotational delay (15000 rot/min): 2 msec
    - Seek time (optimistic): 2 msec
    - Gives $\frac{4 \text{ KB}}{4 \text{ msec}} = 1 \text{ MB/sec}$

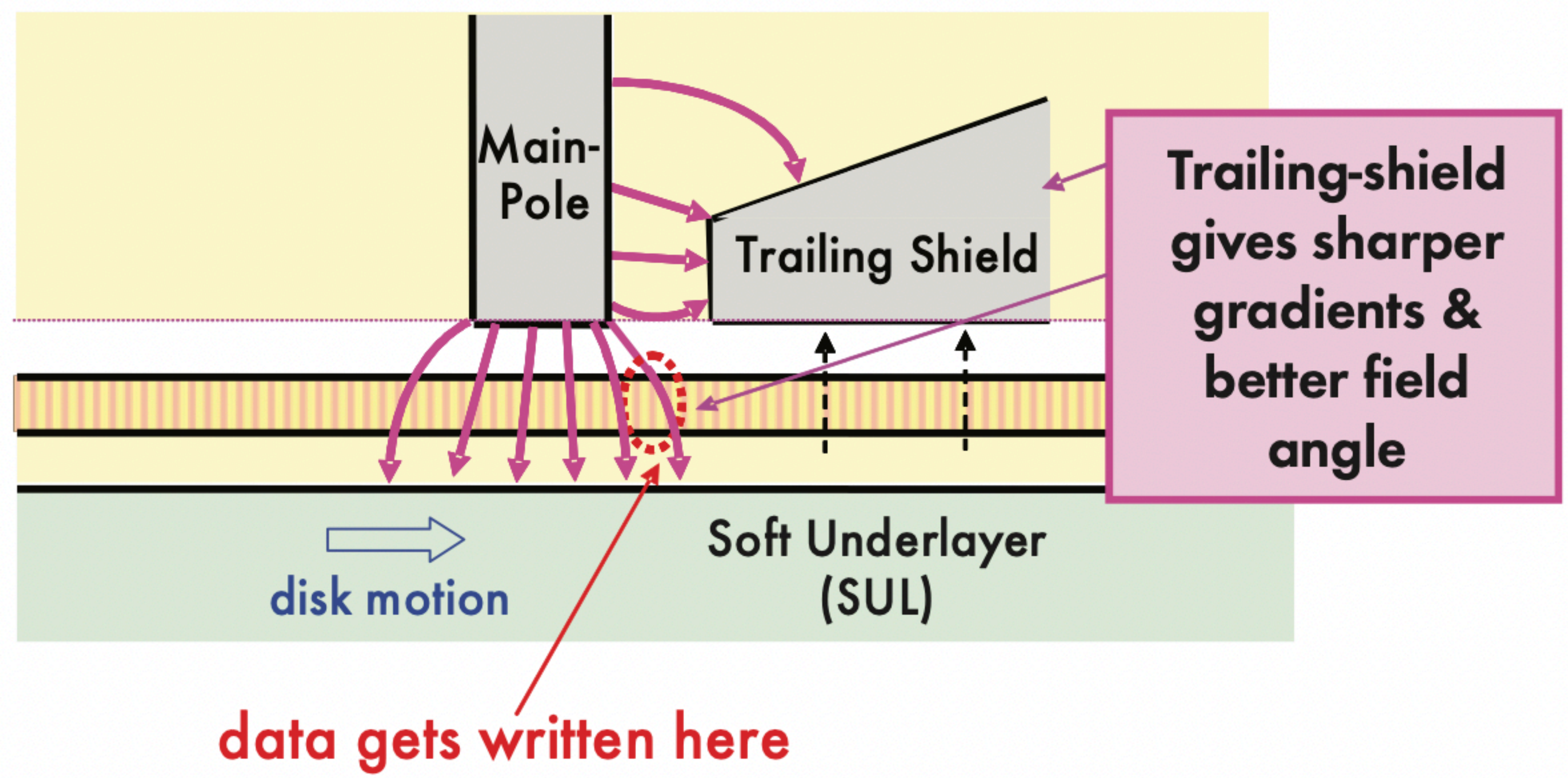
Hard Disk Facts

- Magnetic Recording Trends:
    - Longitudinal vs. Perpendicular



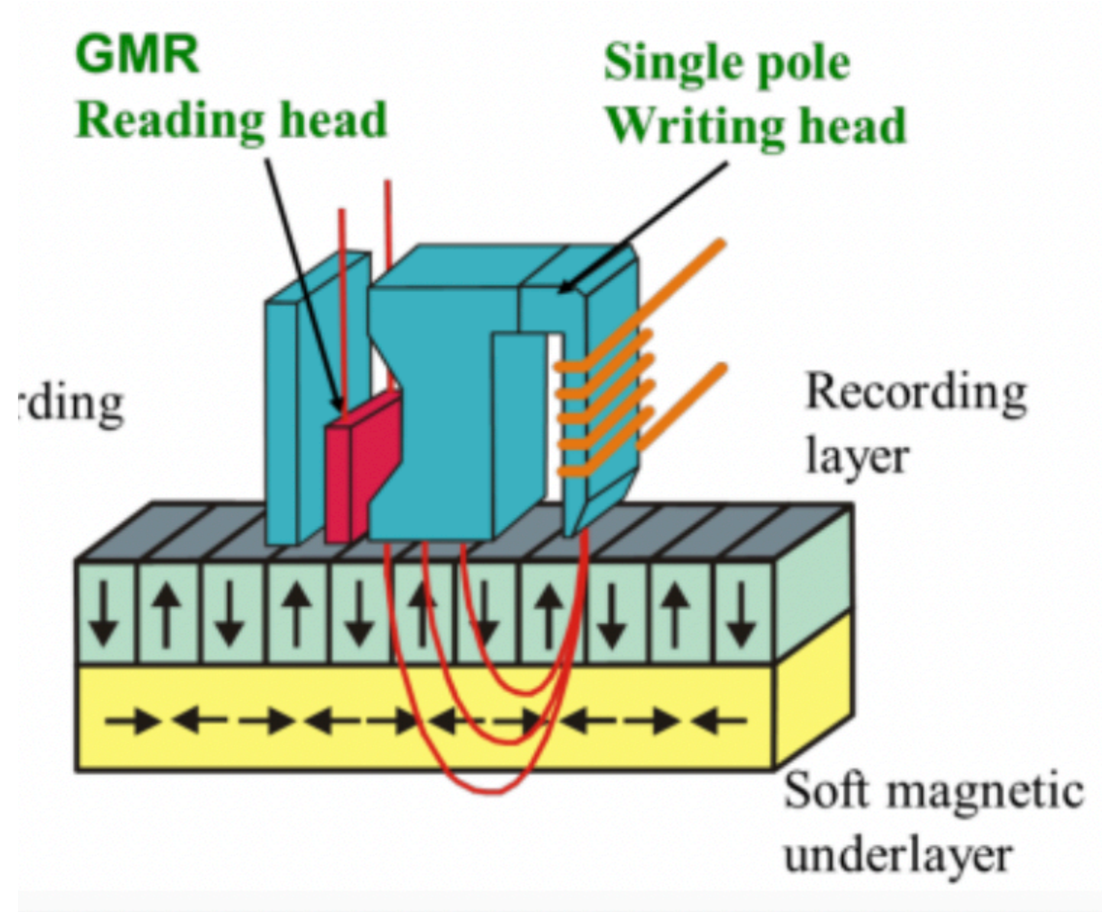
Hard Disk Facts

- Need to control stray magnetization
  - Prevent neighboring track from being disturbed



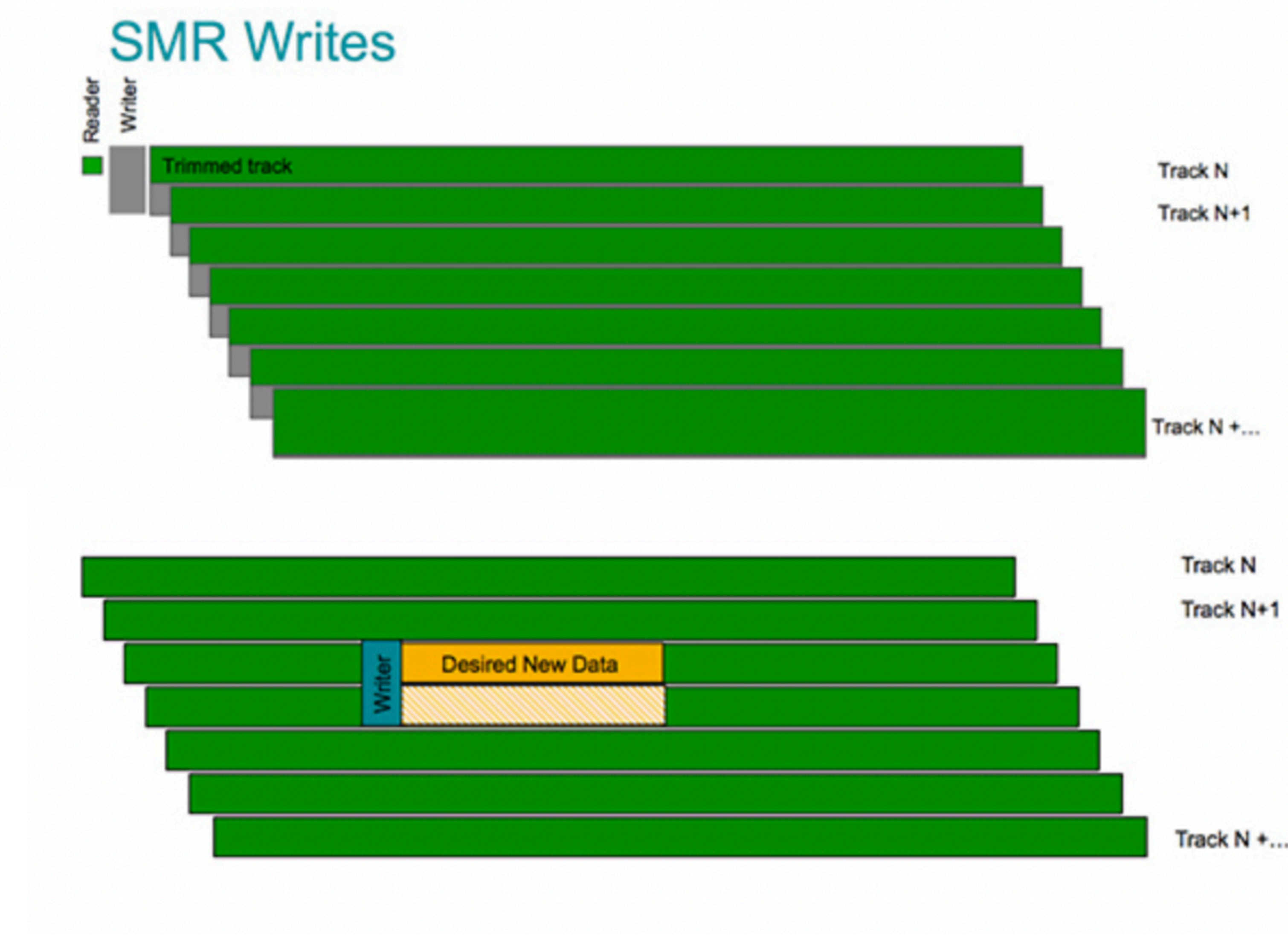
Hard Disk Facts

- Giant Magneto-Resistance
  - Quantum effect
    - Magnetic multilayer materials have a resistance dependent on a magnetic field
- Allows to identify much smaller magnetic areas
- Split head into read (GMR) and write head



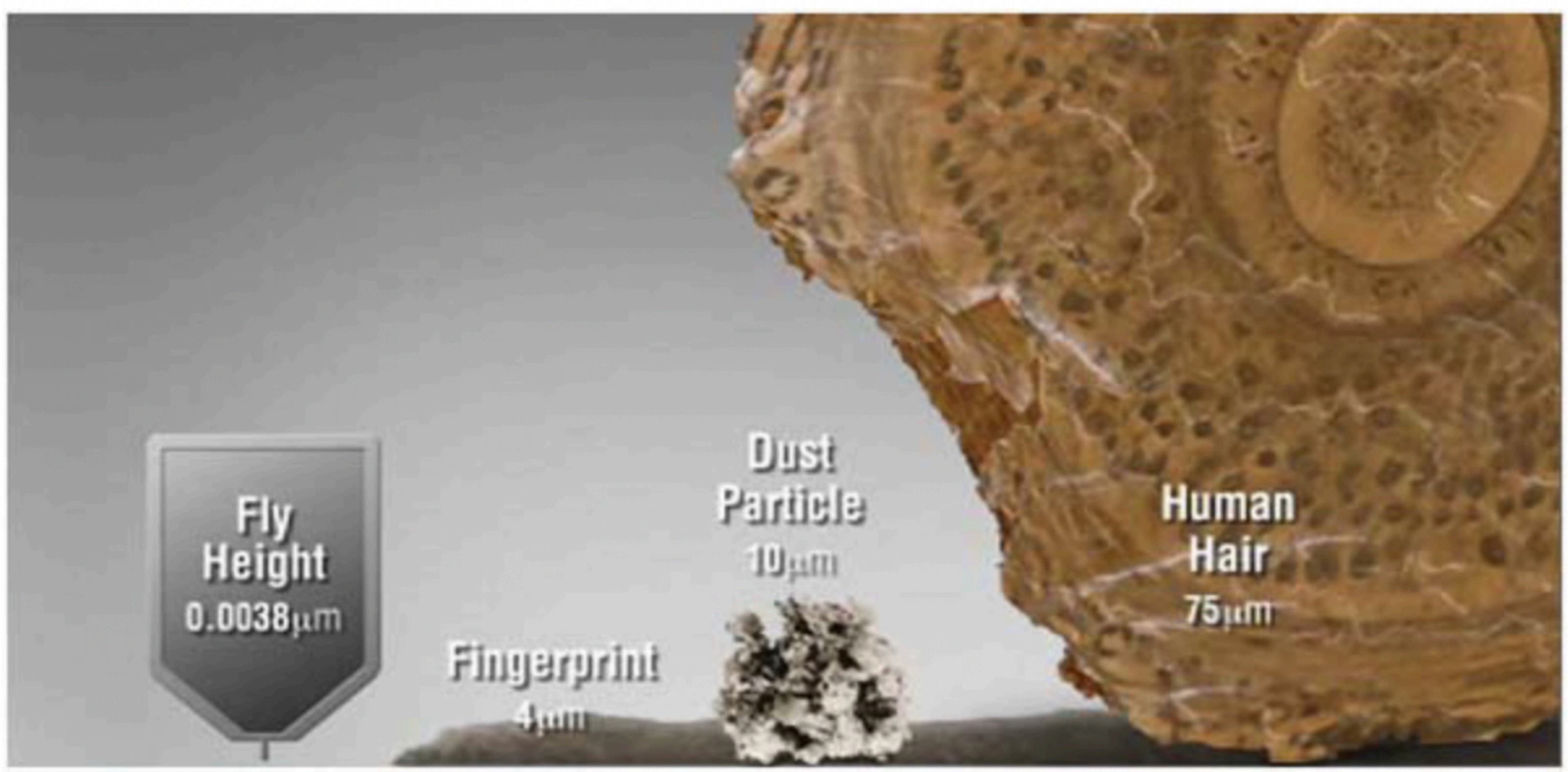
Hard Disk Facts

- Write field can be controlled only in one direction
- Read is much more localized
- Shingled Magnetic Recording
  - Overwrite part of the previous track
  - Higher density, but writes now can destroy previously written data

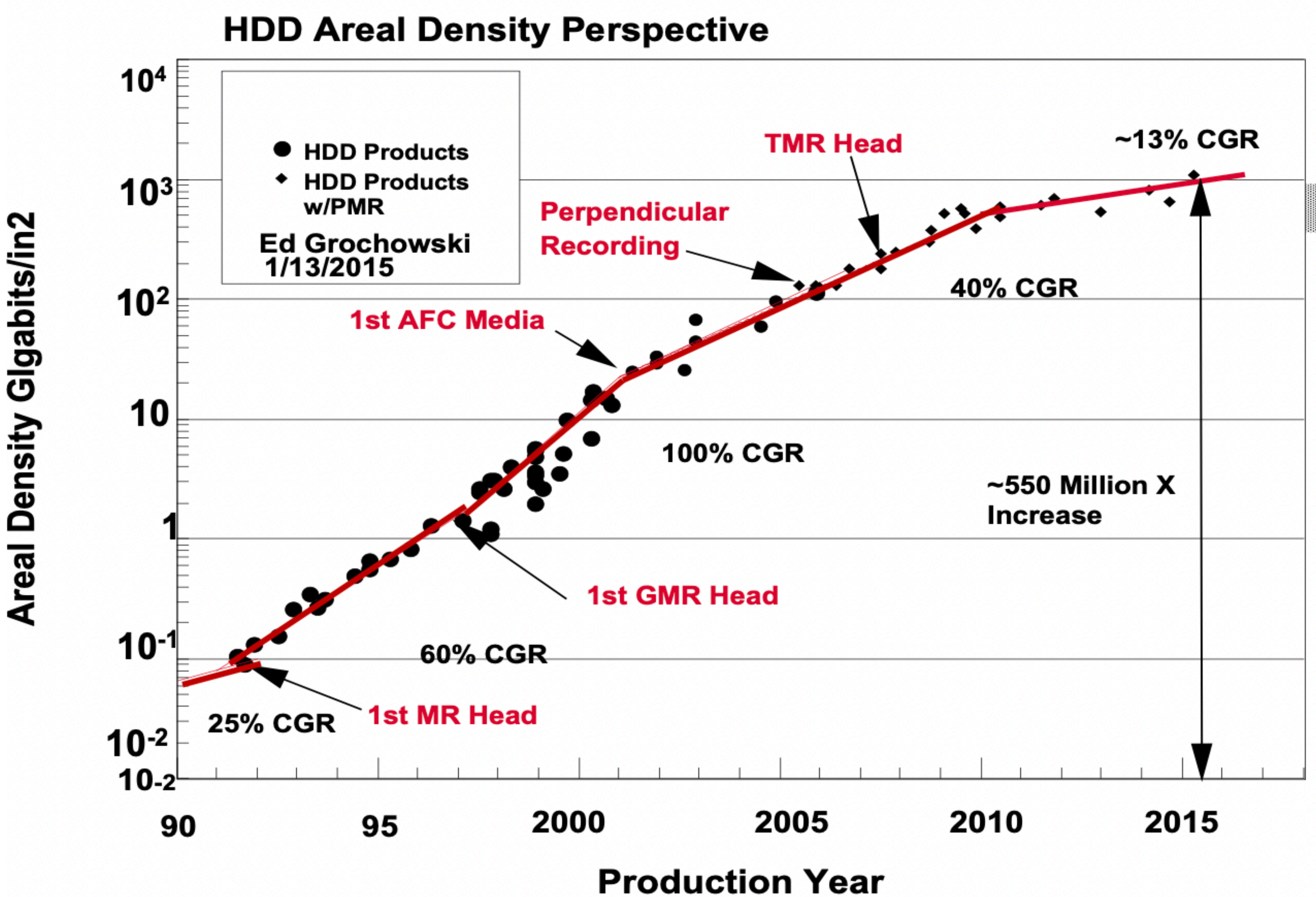


Hard Disk Facts

- Reliability:
  - Disks can fail
  - Failure can sometimes be predicted

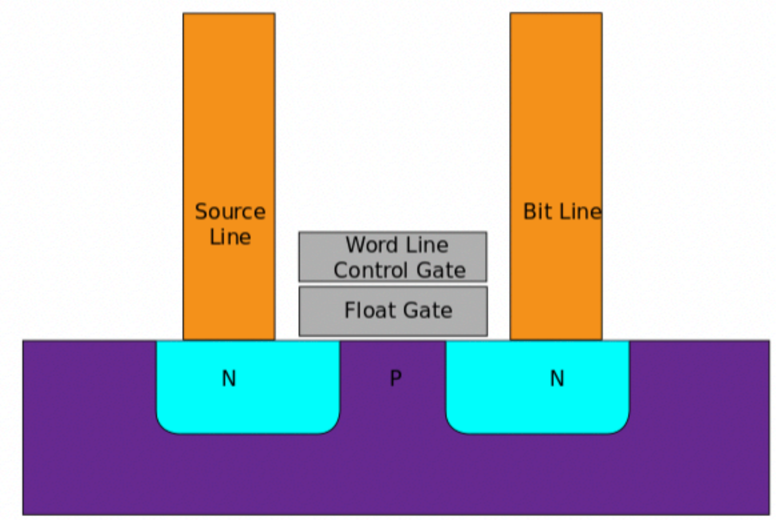


Hard Disk Drive Development



Flash Memory

- Floating Gate MOSFET
  - Float gate charge controls resistance between source and bit line
  - To move / remove electrons from / to Control gate to Float gate, use high voltage (Fowler Nordheim tunneling)
  - High electric fields when writing slowly destroy the surrounding tunnel oxide

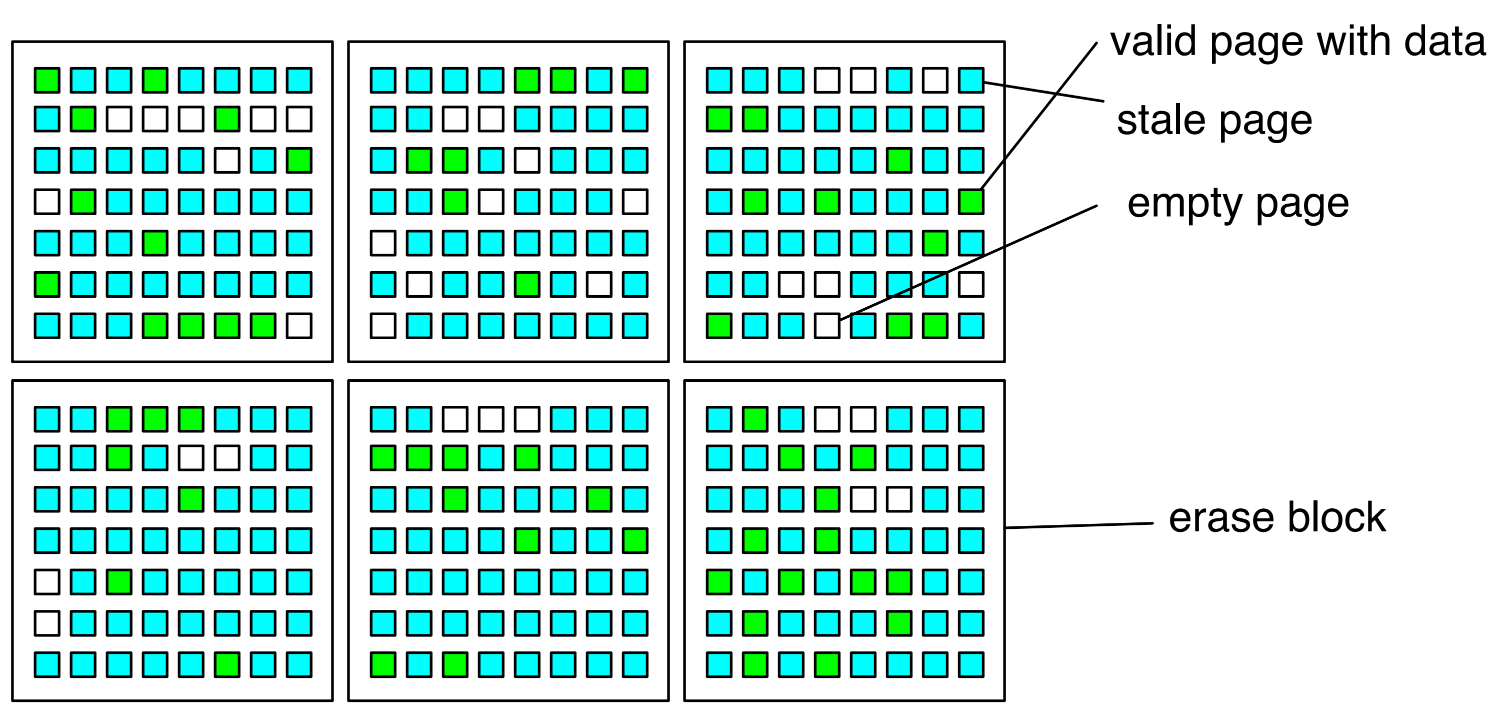


Flash Memory

- Programming / Erasure
- Write complete pages (from 512B to 4KB)
  - Program selected bits in the page
- Read complete pages
- Erase blocks of pages

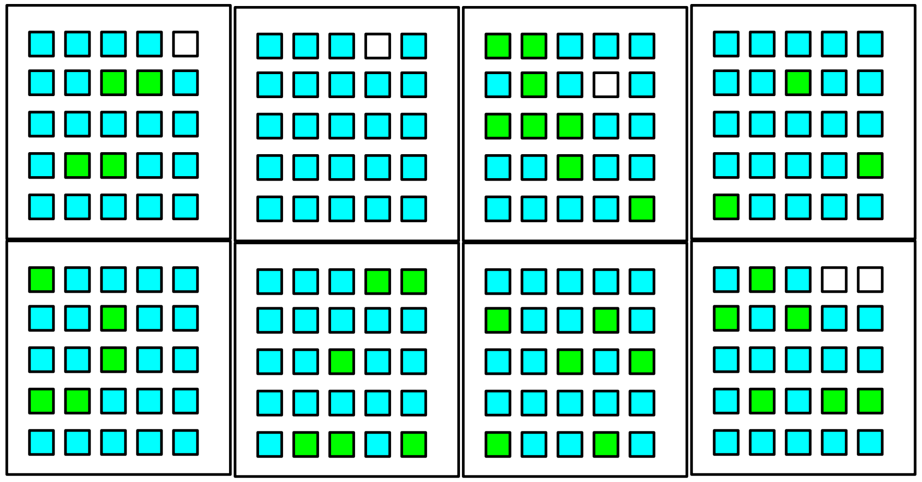
Flash Memory

- Write amplification:
    - Assume a somewhat loaded Solid State Drive (SSD)
        - Pages are empty (erased), in use, or stale

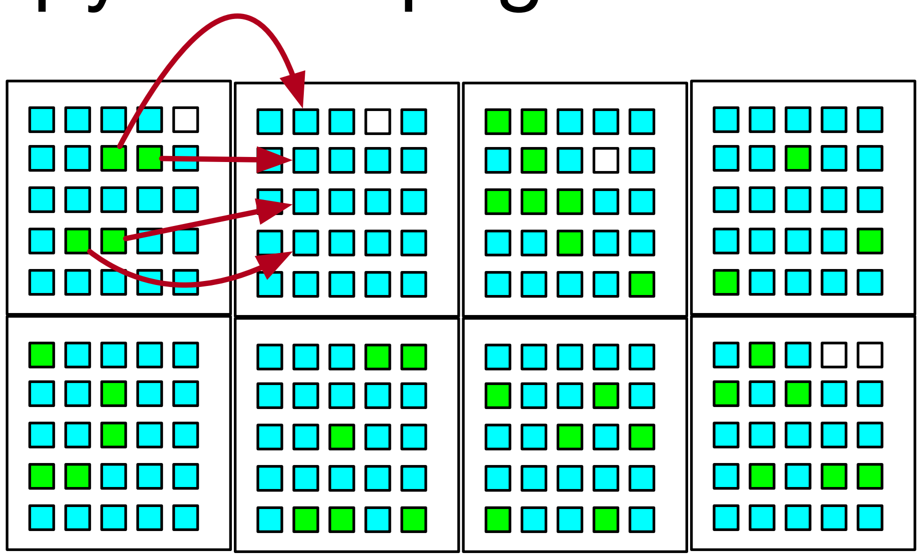


Flash Memory

- When a device runs out of programmable pages

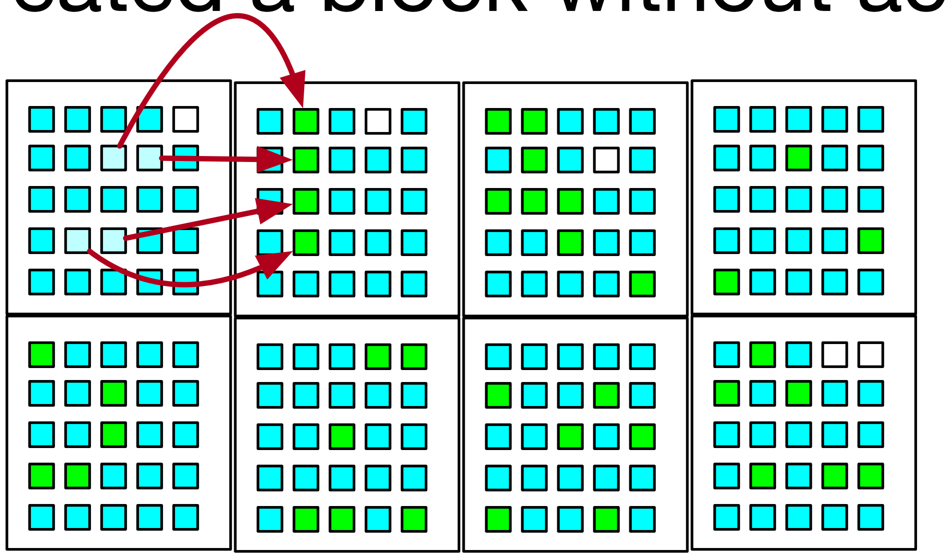


- Needs to copy active pages somewhere else

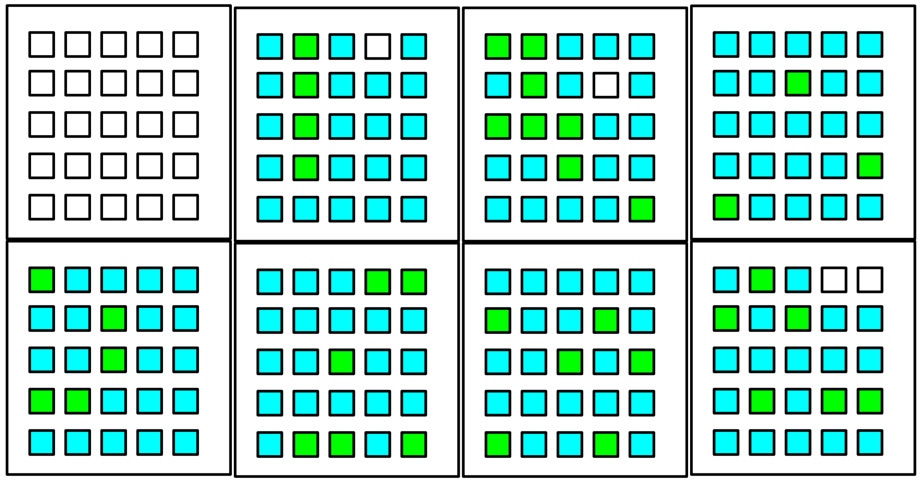


Flash Memory

- Now we have created a block without active content



- Which is erased and now can be used to store new data



Flash Memory

- To write one page, we had to write five
    - Write amplification
    - Reason while write performance for SSD goes down with increased storage utilization
- Research question:
    - Design data structures that work well with flash

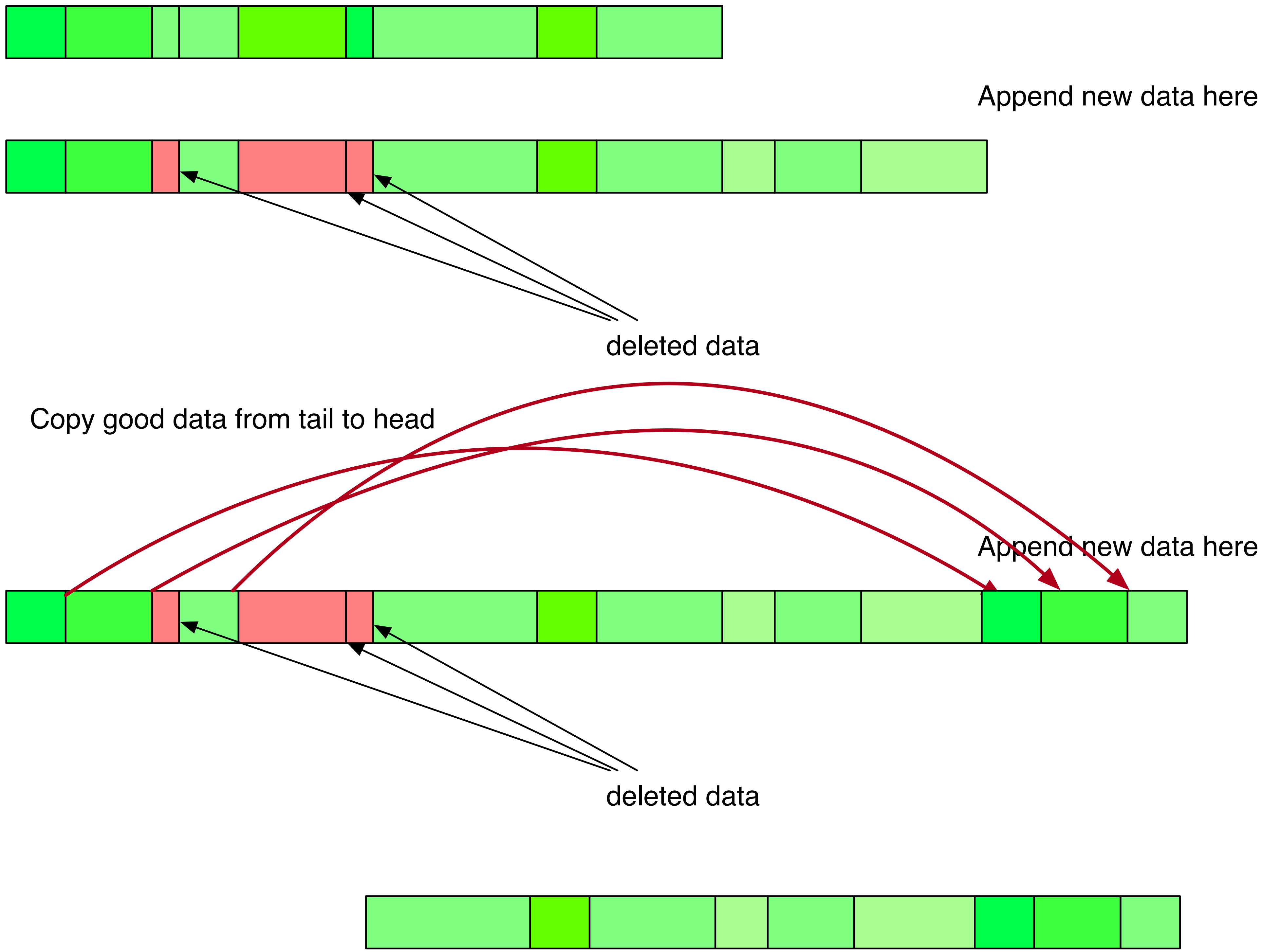
Flash Memory

- Endurance limitations
  - Pages can be written as little as 10^5 times
- Need to do (age-based) wear leveling
- Flash Translation Layer (FTL):
  - Present a virtual block view to the user (OS)
  - Maintain a virtual to physical mapping
    - That can be updated when pages are moved for erasure
    - That minimize erase cycles for frequently erased blocks

Flash Memory

- Log-structured file system
  - Data and meta-data are written to a cyclic buffer
  - From time to time, compaction and garbage collection occurs:
    - Move valid regions from tail to head
    - Reset tail

Flash Memory



Flash Memory

- Log-structured data structures
  - Example: Microsoft Flash File System