ECE 571 – Advanced Microprocessor-Based Design Lecture 21

Vince Weaver

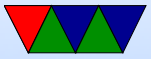
http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

15 November 2022

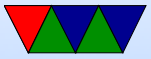
Announcements

- HW#9 will be ARM+Intel readings
- Will respond to topics, don't forget related work in Project update
- Busy week with Supercomputing and Demosplash
- Let me know if you want to borrow power measurement devices



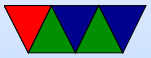
Disk Storage – History

- First disk, IBM, 1950s (size of two refrigerators, 3.75MB), oxide was similar to paint used on golden gate bridge, 1s access time
- At the time already had tape
- Disk vs Disc (usually magnetic vs optical)
- PATA/SATA/SCSI/USB



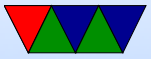
Magnetic Storage

- Magnetism to store onto drives
- Non-volatile
- Historical
  - Wire recording (early 1900s)
  - Magnetic tape (analog, 1928)
  - Magnetic drum – story of Mel
  - Core memory, Core rope
  - Twistor? Bubble memory?



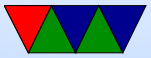
Disk Storage – Terminology

- Originally CHS (cylinder/head/sector)
- Now LBA (logical block access)
- Constant linear velocity (mostly CDs) – same speed no matter where track is. More data on outer tracks
- Constant angular velocity – HDD and FDD
- Audio/Video disks often one big spiral track, hard disks usually separate tracks



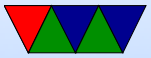
Magnetism

- Magnetic domains. Start out random, but can be arranged by external field to line up, making much stronger field.
- This is not a minimal energy config, but very stable
- Temperature can cause to disappear, at Curie point

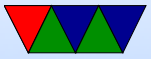


Hard Disk

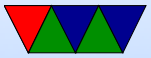
- Rigid disk, spinning
  Originally aluminum, now more exotic
- Two motors
  - spindle (spins)
  - 4200 to 15k rpm, consumer often 5400 or 7200 Which is better 5400 or 7200? 7200 faster, 5400 quieter and lower-power. Western digital causing issues 2020 by having "5400 class" drives that were actually 7200. why is that a problem?



- Want: low vibration, low noise
- Minimal wobble. Causes Non-Repeatable Runout (NRRO) (NRRO means not centered over track) causing track mis-registration (TMR)
- Used to have ball bearings, but hard to make them that good. Recently use fluid dynamic bearings.
- Constant angular velocity, though recent (since 90s) zone bit recording (store more on outer rim)
  To keep from being really complex, it's not per track but a bunch of different zones, each with different density

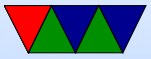


- actuator (move arm)
  - Linear or rotary
  - Arm moved using voice coil (stepper motor on old) NIB (niobium Iron Boron) high-flux magnet, coil (sort of like speaker coil) rapidly moves in magnetic field

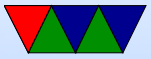


Reading/Writing

- Read/write head
- Close over surface of disk, often nanometers. Any dirt bad.
- N/S polarity, read as 0/1
- Hard to read 0 or 1, easier to measure transition, so encoding used. This means write a sector at a time, not possible to change an individual bit. NRZI encoding. FM/MFM Run-length limited (RLL) codes
- Magnetic "domains" of grains that can be aligned

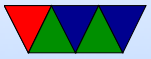


- Magnetic dipole forming magnetic field
- Prior to 2005 or so these were horizontal and parallel
- These days, perpendicular (see silly movie) `https://www.youtube.com/watch?v=xb_PyKuI7II`
- Originally iron(III) oxide, now cobalt-based
- Need to resist self-demagnetization
- Writing
  - Early used electromagnet to both read/write
  - Metal in Gap (MIG) and thin film heads later
  - Magnetoresistance
  - Spintronics, "Giant" magnetoresistance



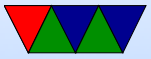
- Modern, separate read/write heads (close to each other) read is magneto-resistance, write thin-film inductive
- write-wide read-narrow
- Guard band on either side to keep from bleeding (affects directly TPI)

- "servo" or how to find the track
- A DSP takes the signal from the read head and converts to digital



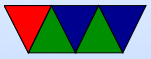
High Density

- Roughly follow Moore's Law?



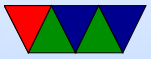
Problems with High Density

- Superparamagnetic trilemma involving grain size, grain magnetic strength and ability of the head to write
- Data can be lost for thermal reasons
- “superparamagnetic limit”
- Platters covered in two parallel magnetic layers, separated by 3 atoms of ruthenium. Magnetized in opposite directions.



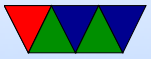
Solutions for High Density

- Perpendicular recording
- exchange spring media – hard/soft layers better for writing. “Exchange” is a quantum physics term
- Heat-assisted magnetic recording (HAMR) (lasers)



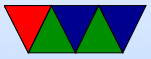
Error Handling – Errors

- Without error checking, error of 1 in 10^5 which is pretty high
- Primary vs Grown defects
  - Primary – permanent, there when drive is made
  - Grown/Growth – happen over time when using disk



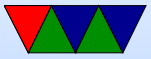
Error Handling – Solutions

- ECC, Reed-Solomon Coding
- With ECC drops to 1 in 10^{14} which is better, but more like one per petabyte or so
- Takes up space, 1TB disk might have 93GB of ECC
- Newer, LDPC (low-density parity check)
- Reserve pool, remap bad sectors. SMART
- For Primary
  - Sector slipping – if bad sectors found when formatting, just slip the numbers so sector numbers of good sectors



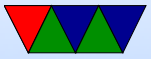
still contiguous

- For Growth
  - Sector Sparing – replace bad sectors with good ones when found



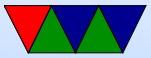
Other Disasters

- Head crash (park first)
  "Landing zone" to park head at power off, needs to be able to spin up with it parked
- Air density – has a filter letting outside air in, stops working at high altitudes
- Physical shock – accelerometer to auto-park heads
- Just noise or vibration, playing loud music can slow speed you get
  - News from 2022, apparently specific Janet Jackson



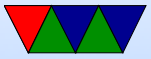
video “Rythym Nation” could crash certain laptop hard drives

- Entire disk crash – in old days large glass or metal platters spinning near speed of sound, shrapnel if break



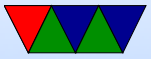
Security

- can you recover a drive that is over-written?
- Overwritten with zeros? Alternating zeros and ones? random?
- Do you need to shred/melt the drive?
- Problem with wear leveling/error where might leave behind old data in places that won't get erased



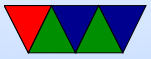
Ways to Increase Performance

- Helium – less turbulence and friction, can pack tighter
- Perpendicular
- Shingling
- Heat-assisted magnetic recording
- Microwave-assisted magnetic recording
- Two-dimensional magnetic recording
- Bit-pattern recording
- Giant-magnetoresistance
- Exchange-spring media



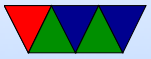
Form Factor

- Capacity – 1GB (10**9) vs 1GiB (2**30)
- Desktop – 60GB - 4TB, 5400 - 10k rpm
- Mobile – smaller in size and space, usually spin slower 5400 or 7200
- Enterprise – fast, often 10k or 15k. Sometimes smaller platters (2.5) for faster seek time
- Consumer – often slower, more shock resistant



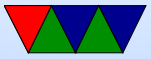
Performance

- Response time vs Throughput
- Seek time – time to get head to block of interest
  Fast drives today, 4ms
- Rotational latency – head seeks to right track but has to wait for block to spin by
  15k – 2ms, 7200 – 4ms, 5400 – 5ms
- Bits per second – 2010, 7200rpm drive 1Gb/s, depends on which track, rpm,.
  SATA can send about 3Gb/s with 10-bit encoding

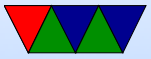


Formatting and 4k Transition

- Low-level. Logical blocks. Delimited by markers start, end, ECC, space to allow for timing
- Traditionally 512B but by moving to 4k can reduce percentage dedicated to the delimiters (though have to be backwards compatible)
- High-level format is just putting filesystem onto the blocks
- Interesting backwards compatibility issues and Linux drama when this first came out

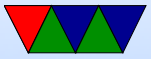


SSD



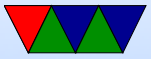
FLASH

- ROM, EPROM, EEPROM
- FLASH (NAND/NOR) can have parts erased, not whole thing at once
- Invented at Toshiba in the 1980s
- NAND vs NOR



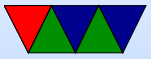
NOR Flash

- Faster, more expensive, slower write/erase, but can individually address bytes and can directly execute code
- Long erase
- Random access
- 100x - 1000000x erase cycles
- CF was originally NOR (but NAND cheaper)
- Like a NOR gate, one end to ground
- low read latency, can be used bit-by-bit ROM
- Program by writing at high voltage, Channel turned on,



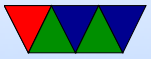
quantum tunneling via "hot electron injection"

- Resetting (actually to all 1 state). Same process, opposite direction, large voltage. Can in theory be individually reset bits but in practice in blocks

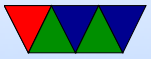


NAND Flash

- Bits are a bunch of nand gates connected serially, to read out have to read out whole row (sort of like a shift register). OK if streaming but tough if just want single bytes
- Read whole pages (4k-16k) but erase areas much larger?
- Reduced erase/write times
- Less chip area (higher density)
- 10x endurance of NOR
- Must read out in large blocks (not random access)

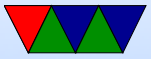


- Wired in series like a bunch of NAND gates
- Certain amount of errors allowed (unlike NOR)
- Tunnel injection for erasing
- ECC error correction
- Programming
  - Starts as all 1s
  - In general can change any 1 to 0 at any time, but if want to switch from 0 back to 1 have to erase whole block
- Floating-gate transistors
  - Each cell like a MOSFET, but with two gates



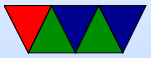
- Floating gate and control gate. Control to switch, float can trap electrons
- Floating gate raises the Vt by acting as a screen, so detect if 1 or 0 by putting an intermediate voltage

- Charge pumps
  - Need high voltage to write. But usually this is done from single voltage supply
  - In space applications usually the charge pump that fails (so chips can still be read, but no longer write)



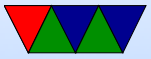
Flash Issues

- Memory wear – can only write so many times before wears out. 100k?
- Memory disturb – a bit like rowhammer, write too many times can change nearby
- Xray can reset bits (problem when trying to see if BGA solder went well)
- Data retention – trapped electrons steadily leak away, especially at warm temperatures



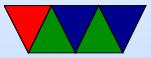
SSD

- Solid-state disk
- No moving parts
- Faster, lower-latency, more resistant to shock
- Still more expensive
- Most 3D TLC NAND-based (can store 2 to 3 bits per cell) Vertical (3D) Nand
  SLC=single level (bit) per core, MLC=multiple(2), TLC=triple level, QLC=quad level cell
- SSDs not permanent, will gradually leak and lose data



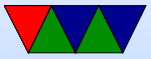
after 2-3 years (faster if worn? trapped electrons leak away)

- Originally was DRAM (battery backed) but these days NAND flash
- Controller that handles things
  - Bad-block remapping
  - Read scrubbing
  - Wear leveling



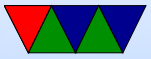
SSD Performance

- DRAM-based fastest
- Single NAND relatively slow
- Having lots in parallel helps

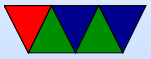


SSD Form Factor

- Can be SATA, but SATA while fast enough for magnetic disk cannot keep up with flash
- M.2 (formerly NGFF) – Intel
  Can provide PCIe, SATA 3.0, or USB 3, different keying to keep from plugging in wrong
- NVMe – non volatile memory express – hook up via PCIe
- U.2 (formerly) SSD Form Factor Working Group, provides SSD connector for enterprise. Hot-swap. mechanically indentical to SATA-express. 3.3V or 12V

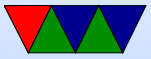


(M.2 only 3.3V)

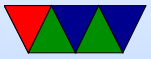


Trim Operation

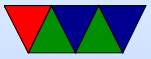
- On filesystem, erase file, usually just mark as deleted and blocks unused, even though never want the bytes again
- TRIM on flash lets you tell disk you don't want them anymore, and the drive can then reclaim them
- Also, when OS then re-uses freed block, flash sees this as an over-write of the block (expensive) rather than a fresh write to a new block
- Expensive because typically erases in big chunks (512kB)



so over-writing you have to erase a whole big chunk, then do a write back of existing values

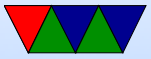


Hybrid Drive

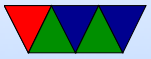


SSD vs HDD comparison

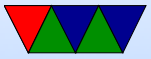
- Data Durability – SSD loses in a few years – HDD lasts longer, but motors/mechanical might fail
- Startup – HDD has to spin up, takes a while
- Random Access – HDD bad, has to spin to location
- Read latency – SSD better
- Bandwidth – SSD often higher
- Read perf – SSD fast but goes down with use
- Noise – SSDs silent
- Heating – both don't like high temps



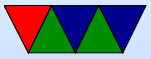
- Cooling – SSDs can operate at lower temps
- Air – SSDs don't require air
- Price – SSD cheaper
- Power –SSD usually better
- Storage size – HDD usually better



Compact Flash



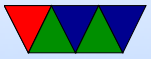
SD-card



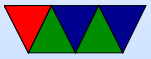
RAID arrays

Note: Becoming less and less common these days as disk capacities increase, easier just to have backups. Speed improvements have negated I/O benefits. Also some filesystems (ZFS, btrfs) might do similar

- Redundant Array of Independent/Inexpensive Disks
- RAID0 – Striping – just makes drive bigger (spreads across disks which might be faster)
- RAID1 – Mirroring, two copies



- RAID2 – bit-level striping not really used anymore
- RAID3 – byte-level striping
- RAID4 – block-level striping
- RAID5 – distributed parity, at least 3 disks, can survive on 2/3 disks
- RAID6 – double parity, any two disks can fail



Tape