ECE 571 – Advanced Microprocessor-Based Design Lecture 17

Vince Weaver

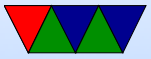
http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

1 November 2022

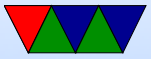
Announcements

- HW7 was assigned, TLB (some note on that)
- Midterm not graded yet
- Useful readings:
  - "A performance and power comparison of modern high-speed DRAM arch" from MEMSYS 2018
  - "DRAM Refresh Mechanisms, Penalties, and Trade-offs" Bhati, Chang, Chishti, Lu and Jacob. IEEE transactions on Computers, 2016.



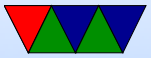
Static RAM (SRAM) Review

- Used on chip: caches, registers, etc. Made in same process as CPU
- 6 transistors (or 4 plus hard-to make resistors with high static power)
- Cross-coupled inverters
  - For read, precharge both bitlines. Raise wordline.
  - Lots of capacitance so hard to swing whole way, so sense amp which amplified the small voltage shift
  - For write, set bitline and not-bitline, set wordline.



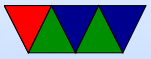
Overpowers inverters

- Clocked or no, clocked saves power
  (synchronous vs asynchronous. synchronous can be pipelined and only operate sense amp when needed)
- Bitlines might be braided to avoid noise



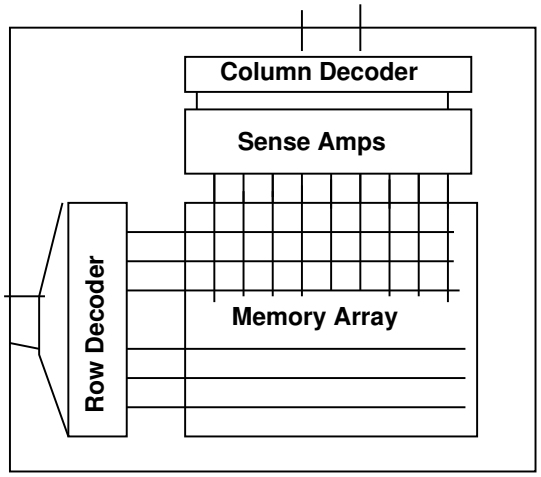
Why not have large SRAM

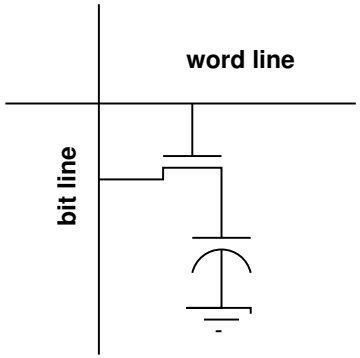
- Low power at low frequency, but more power at high freq
- It is harder to make large SRAMs with long wires
- It is a lot more expensive while less dense (Also DRAM benefits from the huge volume of chips made)
- Leakage for large data structures
- Price: (November 2022)
  - 16Mbit (2Mbyte) $18.70, $9000/GB
  - 8GB DDR4 DIMM, $54, $6/GB



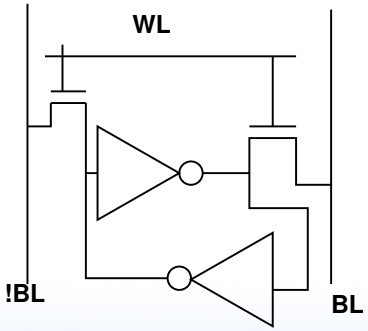
Diagrams

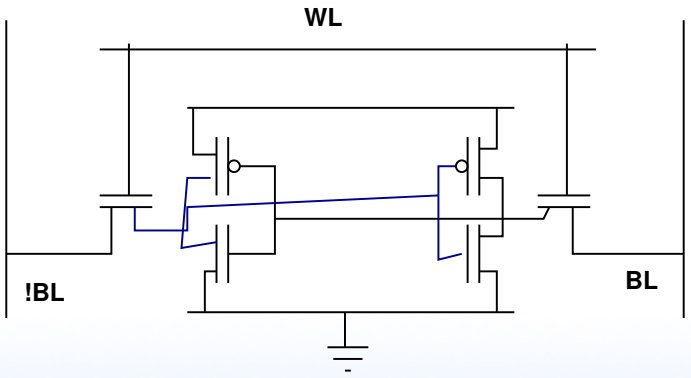
DRAM

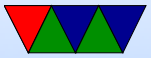




SRAM

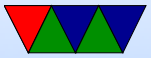




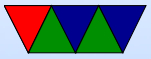


DRAM

- Single transistor/capacitor pair. (can improve behavior with more transistors, but then less dense)
- In 90nm process, 30fF capacitor, leakage in transistor 1fA. Can hold charge from milliseconds to seconds.
- DRAMs gradually lose charge, need to be refreshed.
- Need to be conservative.
  Refresh each row every 32 or 64ms
  (if 8192 rows, then 64ms/8192 is 7.8us)
- DRAM read is destructive, always have to write back

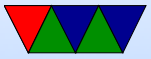


- Interesting article on history of 4116 16k RAM chip
  https://www.righto.com/2020/11/reverse-engineer
  html



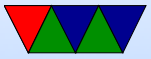
DRAM Actions

- *Precharging* the bitlines
- *Activating* entire row (discharge capactiors into the bitlines)
- *Sensing* the voltage change
- *Reading/Writing*



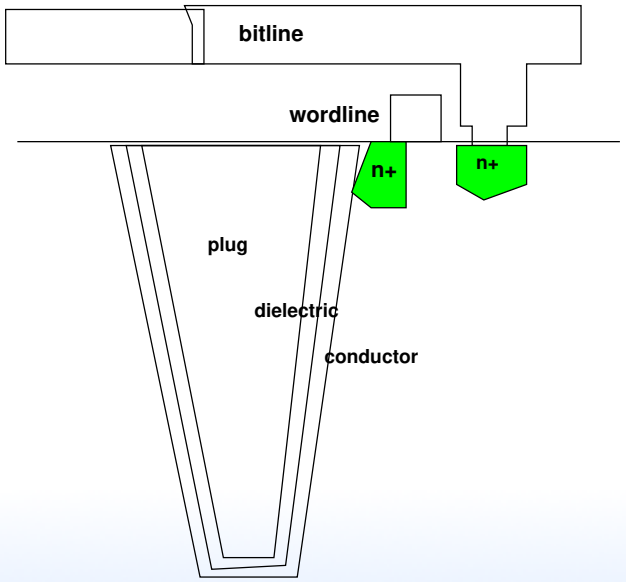
DRAM Notes

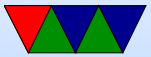
- Two parts of modern DRAM, core bit-arrays and data I/O
- Core bit-arrays have not really increased in speed in years, always in 10s of nanoseconds
  issues are physical: resistance, capacitance
- Improvements come by parallelizing things, reading out lots of bits at once, buffering them, then clocking out
- Summary: hard to improve latency (random-access), easier to improve bandwidth



Low Level

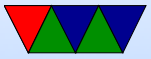
- Planar (old)
- Stacked Capacitors (transistors are below)
- Trench Capacitors (transistors are above)





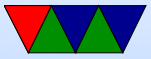
Memory Packaging

- DIMM – dual inline memory module
- Why dual? Replaced SIMMs
- SIMM had pins on both side but just duplicated signal
- SIMM also 32-bit, when modern systems moved to 64-bit bus (P5 pentium) you needed to have SIMMs in pairs
- DIMMs 64-bit memory bus and you only needed to add one module at a time



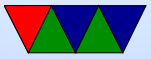
DIMMSs

- How many chips on DIMM? 8? 9?
  9 usually means ECC/parity
- Chips x1 x4 x8 bits, how many get output at a time. Grouped together called a "bank"
- Banks can mask latency, sort of like pipelining. If takes 10ns to respond, interleave the request.
- DIMM can have independent "ranks" (1 or 2 per DIMM), each with banks, each with arrays. (Rank is like a full



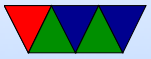
additional 64-bit memory dimm enabled with chip-select, but on same package)

- Layout, multiple mem controllers, each with multiple channels, each with ranks, banks, arrays
- Has SPD "serial presence detect" chip that holds RAM timings and info. Controlled by smbus (i2c)
- SODIMM – smaller form factor for laptops "small outline"



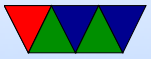
Refresh (more on this later)

- Need to read out each row, then write it back. every 32 to 64ms
- Old days; the CPU had to do this. Slow
  Digression: what the Apple II does
- Newer chips have “auto refresh”

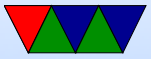


Low-Level Memory Bus

- JEDEC-style. address/command bus, data bus, chip select
- Row address sends to decoder, activates transistor
- Transistor turns on and is discharged down the column rows to the sense amplifier which amplifies
- The sense amplifier is first "pre-charged" to a value halfway between 0 and 1. When the transistors are enabled the very small voltage swing is amplified.

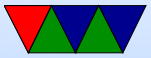


- This goes to column mux, where only the bits we care about are taken through



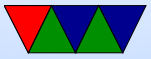
Memory Access

- CPU wants value at address (cache miss?)
- Passed to memory controller
- Memory controller breaks into rank, bank, and row/column
- Proper bitlines are pre-charged
- Row is activated, then $\overline{RAS}$, row address strobe, is signaled, which sends all the bits in a row to the sense

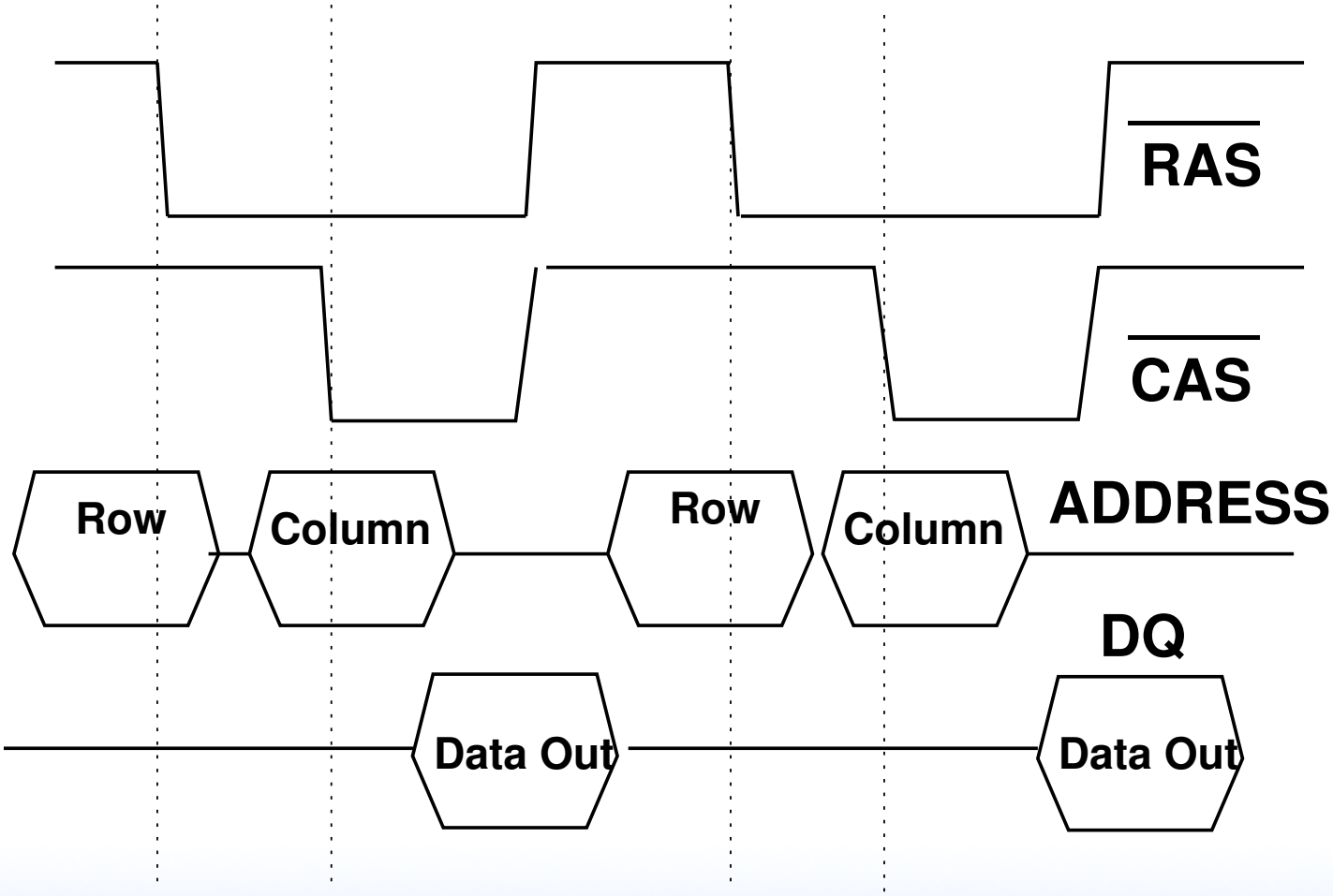


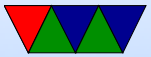
amp. can take tens of ns.

- Then the desired column bits are read. The $\overline{CAS}$ column address strobe sent.
- Again takes tens of ns, then passes back to memory controller.
- Unlike SRAM, have separate CAS and RAS? Why? Original DRAM had low pincount.
- Also a clock signal goes along. If it drives the device it's synchronous (SDRAM) otherwise asynchronous



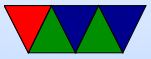
Async DRAM Timing Diagram





Memory Controller

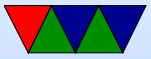
- Formerly on the northbridge
- Now usually on same die as CPU



Advances in Memory Technology

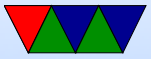
In general the actual bit array stays same, only interface changes up.

- Clocked
- Asynchronous
- Fast page mode (FPM) – row can remain active across multiple CAS.
- Extended Data Out (EDO) – like FPM, but buffer

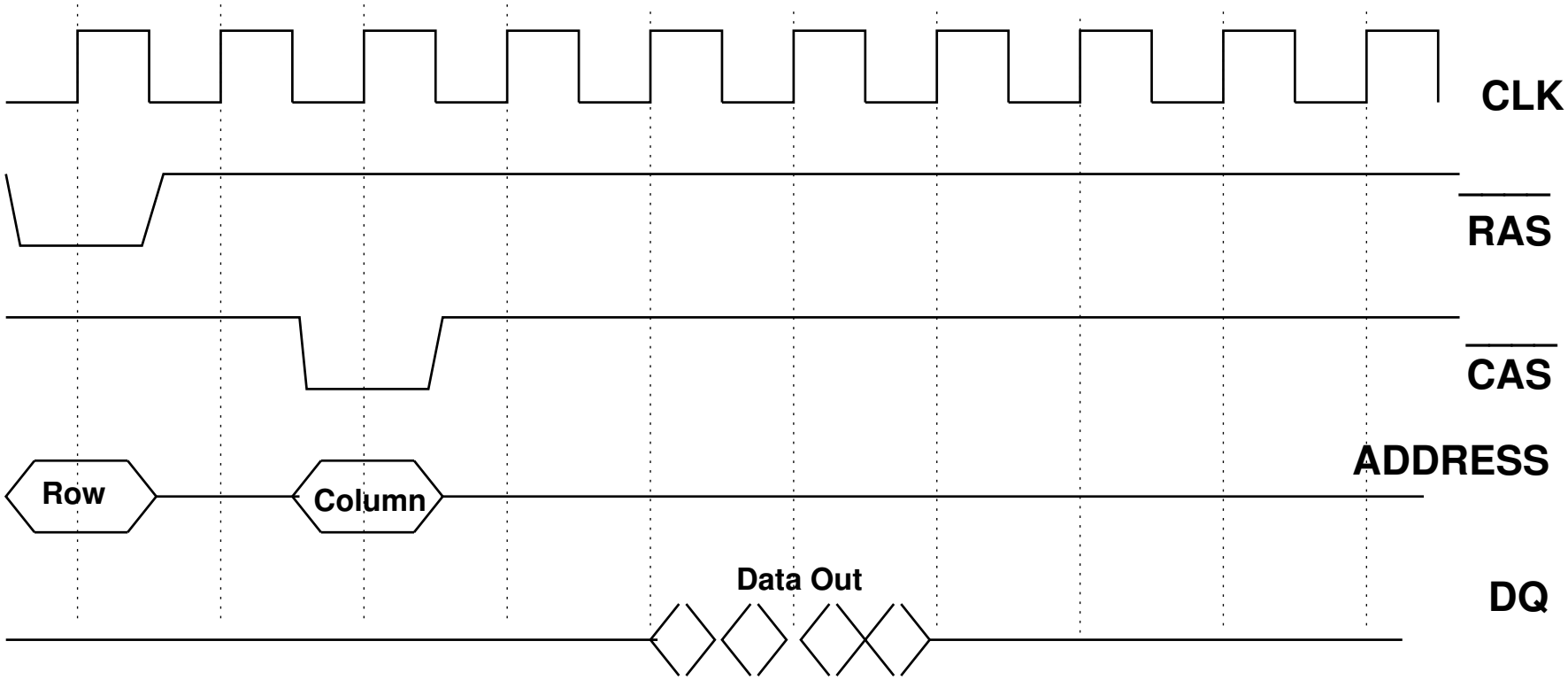


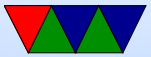
"caches" a whole page of output if the CAS value the same.

- Burst Extended Data Out (BEDO) – has a counter and automatically will return consecutive values from a page
- Synchronous (SDRAM) – drives internal circuitry from clock rather than outside RAS and CAS lines. Previously the commands could happen at any time. Less skew.

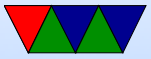


DDR Timing Diagram



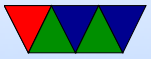


Historical Memory Type Rundown



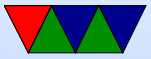
SDRAM – Synchronous DRAM

- SDRAM – 3.3V



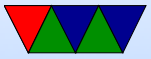
DDR

- transfer and both rising and falling edge of clock
- 2.5V
- Adds DLL to keep clocks in sync (but burns power)



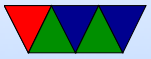
DDR2

- 2003
- runs internal bus half the speed of data bus
- 4 data transfers per external clock
- memory clock rate * 2 (for bus clock multiplier) * 2 (for dual rate) * 64 (number of bits transferred) / 8 (number of bits/byte) so at 100MHz, gives transfer rate of 3200MB/s.
- not pin compatible with DDR3.
- 1.8 or 2.5V



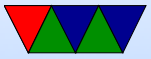
DDR3

- 2007
- internal doubles again
- Up to 6400MB/s, up to 16GB DIMMs.
- 1.5V or 1.35V



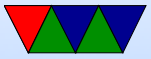
DDR4

- 2014
- Higher density, faster speed, lowe voltage than DDR3
- 1.2V, 2.5V auxiliary wordline boost
- 16 internal banks, up to 8 ranks per DIMM
- 1600 to 3200MHz. 2133MT/s
- Up to 64GB DIMMs
- parity on command/address busses, crc on data busses.
- Data bus inversion (if more power/noise causes by sending lots of 0s you can set this bit and then send



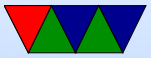
things as 1s instead)

- Pins closer together (0.85mm vs 1.0mm)
- 288 pins vs 240 pins, new package. Slightly curved edge connector so not forcing all pins at once when pushing in
- Example: DDR4-2400R, memory clock 300MHz, I/O bus clock 1200MHz, Data rate 2400MT/S, PC4-2400 19200MB/s (8B or 64 bits per transaction) CAS latency around 13ns

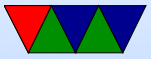


DDR5

- 2020
- Reduce power and Doubled bandwidth over DDR4
- 1.1V, on board voltage regulation (5V or 12V from board?)
- Up to 512GB DIMM
- On-chip ECC to avoid errors on chip, but not required to traditional report-to-CPU if something is wrong ECC
- Decision Feedback Equalizer (DFE)?
- LRDIMM and UDIMM

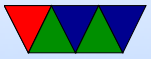


- 288 pins like DDR4 but different layout
- C0-C2 column bits removed to speed protocol, only get multiples of 8
- Built in temp sensor?

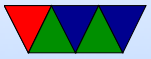


GDDR – Graphics Card RAM

- Despite similar name, not related to same DDR version usually optimized for high-bandwidth
- GDDR2 – graphics card, but actually halfway between DDR and DDR2 technology wise
- GDDR3 – like DDR2 with a few other features. lower voltage, higher frequency, signal to clear bits
- GDDR4 – based on DDR3, replaced quickly by GDDR5
- GDDR5 – based on DDR3. addresses sent at double rate

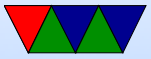


- GDDR5X – 2016
- GDDR6 – quad data rate
- GDDR6X



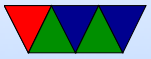
LPDDR

- Despite similar name, not related to same DDR version
- LP-DDR
- LP-DDR2 – low power states, 1.2V, different bus
- LP-DDR3 – higher data rate
- LP-DDR4 – change from 10-bit DDR to 6-bit SDR bus
- LP-DDR4X – I/o voltage 0.6V
- LP-DDR5 – (2019) 6.4Gbit/s/pin, differential clocks

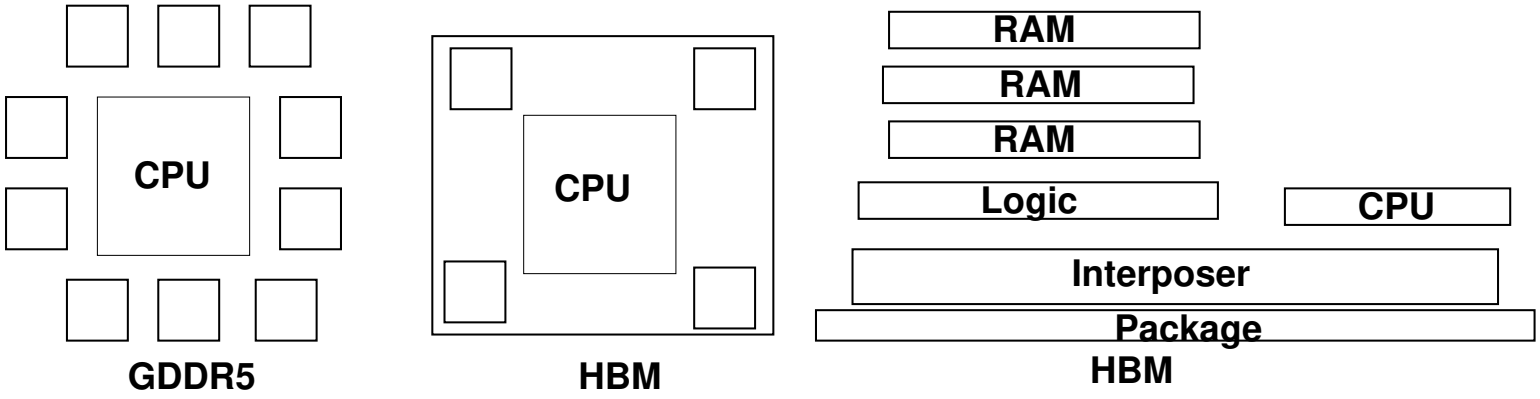


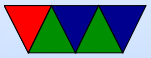
DDRL

- DDR3L - low voltage, 1.35V (not same as LPDDR3)
  DDR3U (ultra-low voltage) 1.25V
- DDR4L – does not exist?



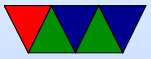
HBM RAM


CPU
GDDR5
CPU
HBM
RAM
RAM
RAM
Logic
CPU
Interposer
Package
HBM

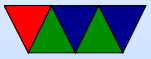


HBM/HBM2 RAM

- HBM
  - High bandwidth memory
  - 3d-stacked RAM, stacked right on top of CPU
  - Silicon through VIA
  - Higher bandwidth, two 128-bit channels per die
  - 4096 bit wide bus compared to GDDR5 where you might have 32-bit channel times 16 chips for 512 bit
- HBM2
  - Eight dies per stack, up to 2GT/s

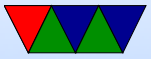


- HBM3/HBM4
  - Specified, doesn't exist yet? HPC?
- In newer GPUs, AMD and NVIDIA. HBM2 in new Nvidia Pascal Tesla P100



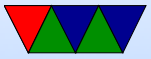
More obscure/obsolete Memory Types

- RAMBUS RDRAM – narrow bus, fewer pins, could in theory drive faster. Almost like network packets. Only one byte at time, 9 pins?
- FB-DIMM – from intel. Mem controller chip on each dimm. High power. Requires heat sink? Point to point. If multiple DIMMs, have to be routed through each controller in a row.
- VCDRAM/ESDRAM – adds SRAM cache to the DRAM
- 1T-SRAM – DRAM in an SRAM-compatible package,



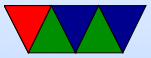
optimized for speed

- Hybrid Memory Cube – similar to High Bandwidth RAM (discontinued 2018)



Memory Latencies, Labeling

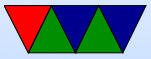
- DDR400 = 3200MB/s max, so PC3200
- DDR2-800 = 6400MB/s max, so PC2-6400
- DDR2 5-5-5-15
  - C_L CAS latency
  - T_{RCD} row address to column address delay
  - T_{RP} row precharge time
  - T_{RAS} row active time
  - CMD (optional), command time
- DDR3 7-7-7-20 (DDR3-1066) and 8-8-8-24 (DDR3-1333).



Memory Parameters

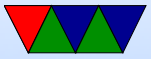
You might be able to set this in BIOS

- Burst mode – select row, column, then send consecutive addresses. Same initial latency to setup but lower average latency.
- CAS latency – how many cycles it takes to return data after CAS.
- Dual channel – two channels (two 64-bit channels to memory). Require having DIMMs in pairs



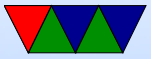
ECC Memory

- There's debate about how many errors can happen, anywhere from 10^{-10} error/bit*h (roughly one bit error per hour per gigabyte of memory) to 10^{-17} error/bit*h (roughly one bit error per millennium per gigabyte of memory
- Google did a study and they found more toward the high end
- Would you notice if you had a bit flipped?
- Scrubbing – only notice a flip once you read out a value



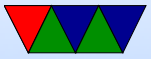
Registered Memory

- Registered vs Unregistered
- Registered has a buffer on board. More expensive but can have more DIMMs on a channel
- Registered may be slower (if it buffers for a cycle)
- Registered uses more power, and might be faster (even with the extra latency)
- RDIMM/UDIMM



Bandwidth/Latency Issues

- Is RAM truly random access these days?
- No, burst speed fast, random speed not.
- Is that a problem? How is RAM accessed these days? Mostly filling cache lines?



Memory Controller

- Do we even want full random access to memory?
  Should we just pass on CPU mem requests unchanged?
- What might have higher priority?
  Are some accesses more important? (regular vs prefetches)
- Why might re-ordering the accesses help performance (it's bad to switch back and forth between two pages)
- Can you improve mem performance with a better mem controller?