ECE 571 – Advanced Microprocessor-Based Design Lecture 16

Vince Weaver

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

vincent.weaver@maine.edu

4 April 2017

Announcements

- HW8 was assigned, read about Newer Intel chips for Thursday
- Slow getting back to you about project ideas
- Sorry about voice, mild cold

How to save Energy in TLB?

- Turn off Virtual Memory completely (aside about ARM1176 manual and caches). Can you run Linux without VM? ucLinux

- TLB is similar to cache, can make similar optimizations (drowsy, sizing, etc)

- Assume in current page (i.e. 1-entry 0-level TLB)

  (Kadayif, Sivasubramaniam, Kandemir, Kandiraju, Chen. TODAES 2005).

(Kadayif,Sivasubramaniam, Kandemir, Kandiraju, Chen. Micro 2002)

- Use virtual cache (Ekman and Stenström, ISLPED 2002)
- Switch virtual to physical cache on fly (hybrid) (Basu, Hill, Swift. ISCA 2012)
- Dynamically resize the TLB (Delaluz, Kandemir, Sivasubramaniam, Irwin, Vijaykrishnan. ICCD 2013)
- Try to keep as much in one page as possible via compiler. (Jeyapaul, Marathe, Shrivastava, VLSI'09)

(Lee, Ballapuram. ISLPED'03)

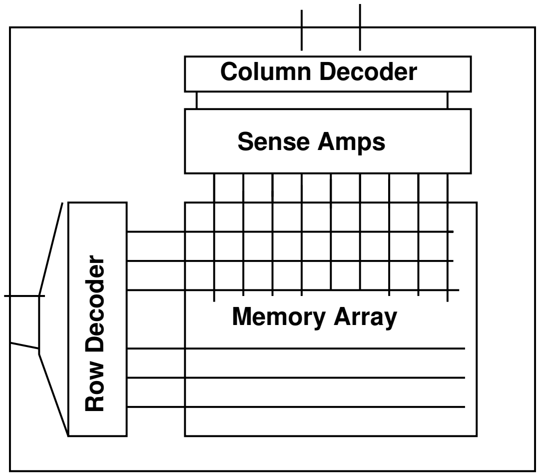
DRAM

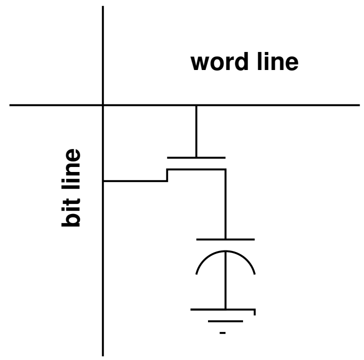
- Single transistor/capacitor pair. (can improve behavior with more transistors, but then less dense)
- Compare to SRAM that has 6 transistors (or 4 plus hard-to make resistors with high static power)
- In 90nm process, 30fF capacitor, leakage in transistor 1fA. Can hold charge from milliseconds to seconds.
- DRAMs gradually lose charge, need to be refreshed. Need to be conservative. Refresh every 32 or 64ms

- DRAM read is destructive, always have to write back

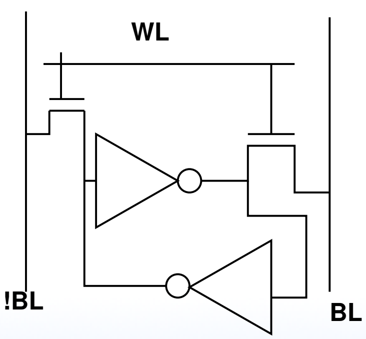
Diagram

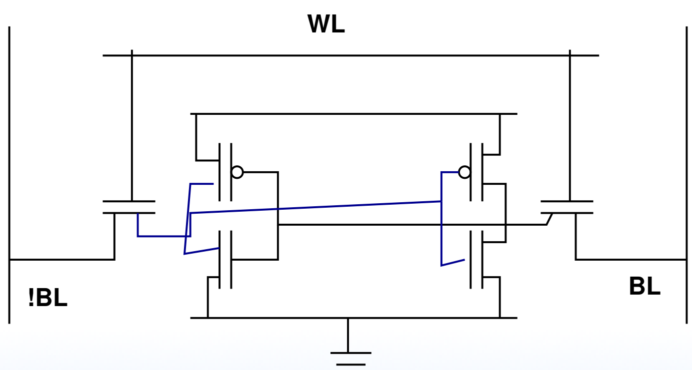
DRAM





SRAM





Low Level

- Trench Capacitors
- Stacked Capacitors

SIMMs/DIMMS

- 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

- 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

Refresh

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

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

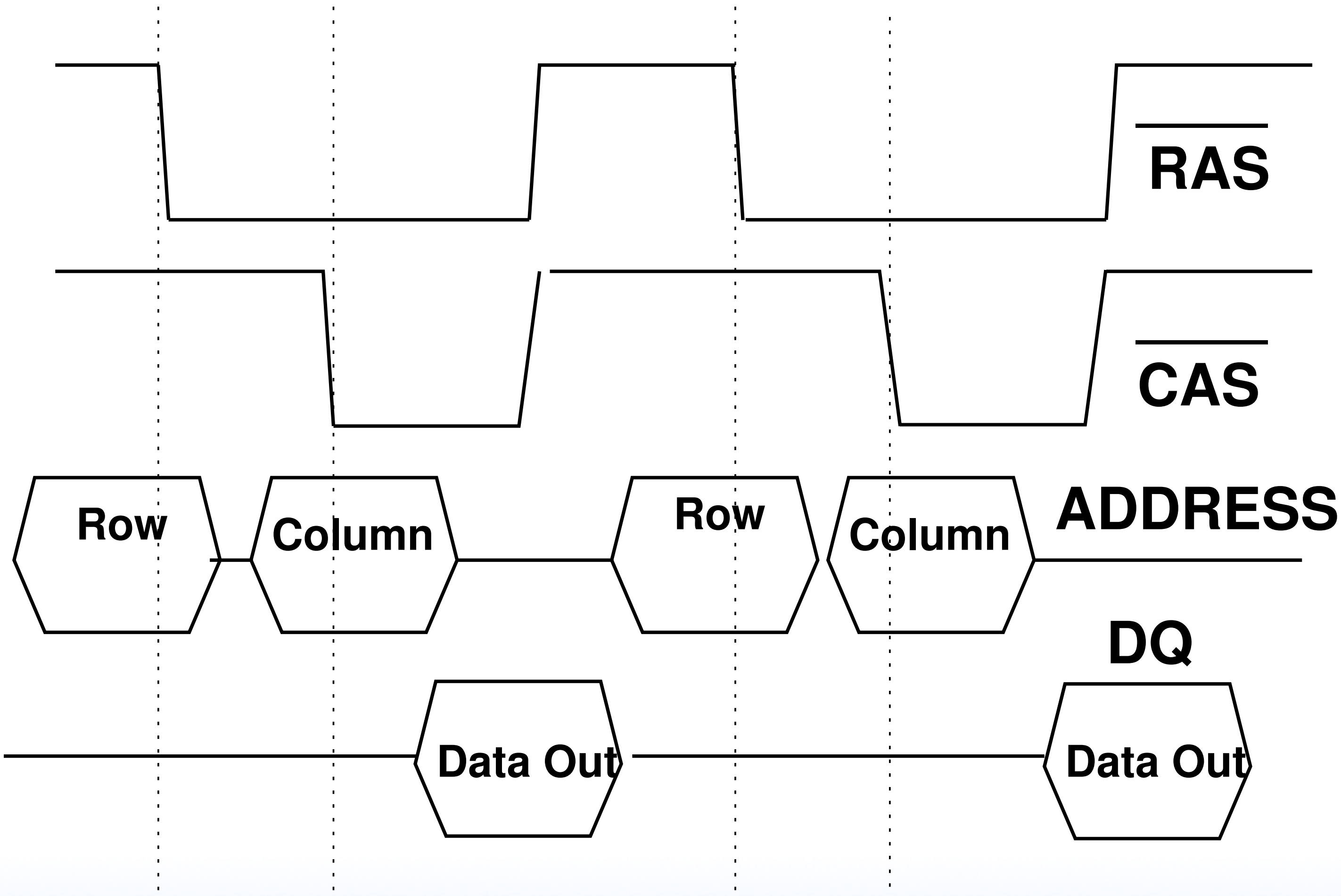
Memory Access

- CPU wants value at address
- 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 general the actual bit array stays same, only interface changes up.

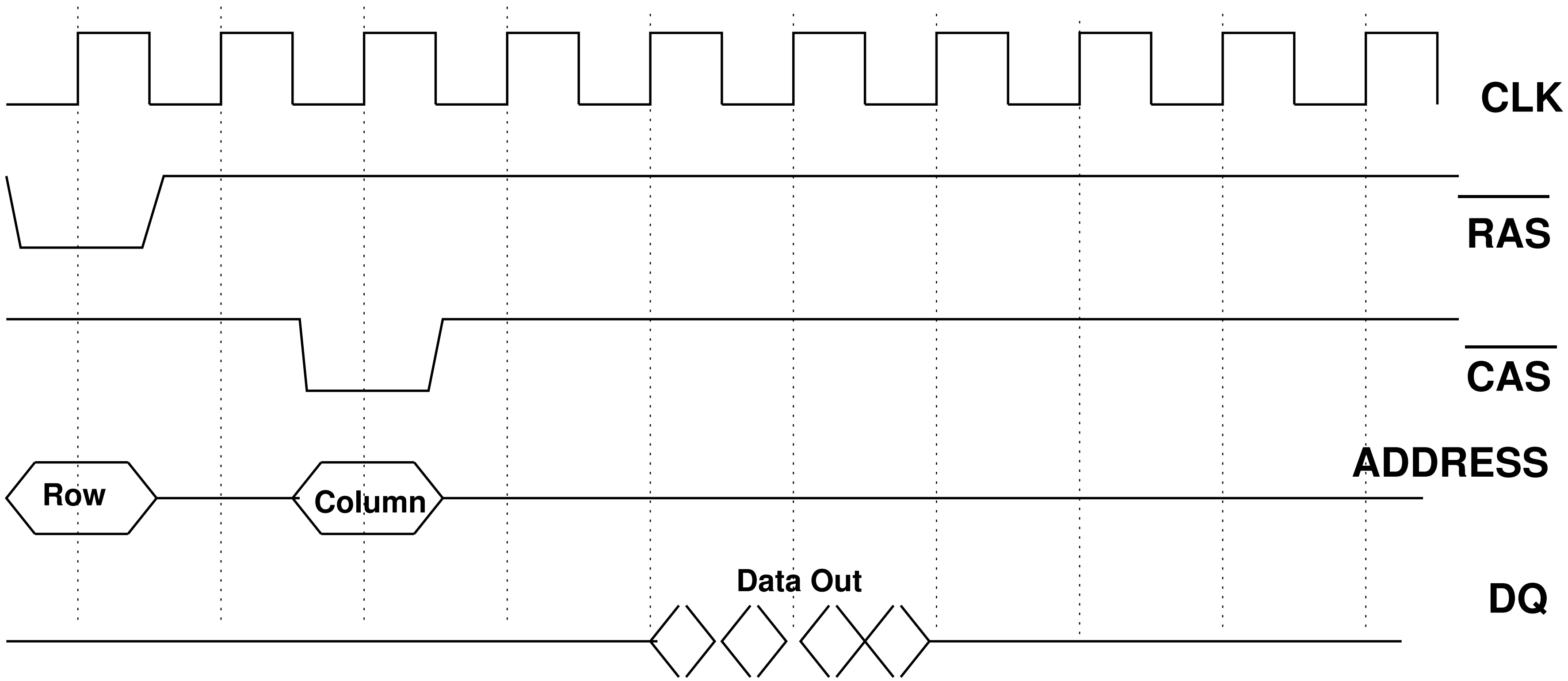
- Clocked
- Asynchronous
- Fast page mode (FPM) – row can remain active across multiple CAS.
- Extended Data Out (EDO) – like FPM, but buffer

"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



Memory Types

- 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 – 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 – internal doubles again. Up to 6400MB/s, up to 8gigabit dimms. 1.5V or 1.35V
- DDR3L - low voltage, 1.35V
- DDR4 – just released. 1.2V , 1600 to 3200MHz. 2133MT/s, parity on command/address busses, crc on data busses.
- DDR4L – 1.05V

- 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

More obscure 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

Memory Latencies, Labeling

- DDR400 $=$ 3200MB/s max, so PC3200
- DDR2-800 $=$ 6400MB/s max, so PC2-6400
- DDR2 5-5-5-15
- CAS latency – T_{RCD} row address to column address delay – T_{RP} row precharge time – T_{RAS} row active 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

- Scrubbing

Issues

- Truly random access? No, burst speed fast, random speed not Is that a problem? Mostly filling cache lines?

Future

- Phase-change RAM
- Non-volatile memristor RAM

Announcements

- HW#9 will be another reading

DRAM Further info

- How do you configure DRAM/initialize/find timings?
- With DIMMs there's an i2c bus with an EPROM with the info

Memory Controller

- Can we have full random access to memory? Why not just pass on CPU mem requests unchanged?
- What might have higher priority?
- Why might re-ordering the accesses help performance (back and forth between two pages)

DDR4 Speed and Timing

- Higher density, faster speed, lower voltage than DDR3
- 1.2V with 2.5V for "wordline boost" This might be why power measurement cards are harder to get (DDR3 was 1.5V)
- 16 internal banks, up to 8 ranks per DIMM
- Parity on command bus, CRC on data bus
- Data bus inversion? If more power/noise caused by

sendings lots of 0s, you can set bit and then send them as 1s instead. New package, 288pins vs 240pins,

- pins are 0.85mm rather than 1.0mm Slightly curved edge connector so not trying to force all in at once

- 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

HBM2 RAM

- High bandwidth memorhy
- 3d-stacked RAM, stacked right on top of CPU
- In newer GPUs, AMD and NVIDIA. HBM2 in new Nvidia Pascal Tesla P100

NVRAM

- Phase Change or Memristors
- Phase change
  - bit of material can be crystaline or amorphous
  - resistance is different based on which
  - need a heater to change shape
  - Faster write performance than flash (slower than DRAM)
  - Can change individual bits (flash need to erase in

blocks)

– Flash wears out after 5000 writes, PCM millions
– Flash fades over time. Phase change lasts longer as long as it doesn't get too hot.
– needs a lot of current to change phase
– can potentially store more than one bit per cell

Stuff from Last Class

- Phase change RAM.
  - chalcogenide glass – used in CD-Rs
  - 100ns (compared to 2ns of DRAM) latency
  - heating element change from amorphous (high resistance, 0) to crystaline (low resistance, 1)
  - temp sensitive, values lost when soldering to board (unlike flash)
  - better than flash (takes .1ms to write, write whole blocks at once

– Newer methods might involve lasers and no phase change?
– Mapping into memory? No need to copy from disk?
– But also, unlike DRAM, a limit on how many times can be written.

- Memristors
  – resistors, relationship between voltage and current
  – capacitors, relationship between voltage and charge
  – inductors, relationship between current and magnetic flux

– memrister, relationship between charge and magnetic flux; “remembers” the current that last flowed through it

– Lot of debate about whether possible. HP working on memristor based NVRAM

Why not have large SRAM

- SRAM is low power at low frequencies but takes more at high frequencies
- 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