ECE 571 – Advanced Microprocessor-Based Design Lecture 19

Vince Weaver

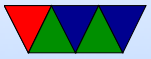
http://www.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

7 April 2016

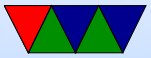
Announcements

- HW#9 will be another reading



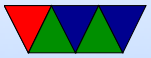
DRAM Further info

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



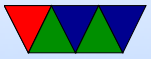
Memory Controller

- 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

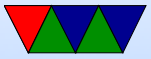
- Higher density, faster speed, lower voltage than DDR3
- 1.2V with 2.5V for “wordline boost” This might be why power measurement cards are harder to get (DDR3 was 1.5V)
- 16 internal banks, up to 8 ranks per DIMM
- Parity on command bus, CRC on data bus
- Data bus inversion? If more power/noise caused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



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

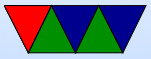
- pins are 0.85mm rather than 1.0mm Slightly curved edge connector so not trying to force all in at once

- 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



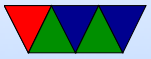
HBM2 RAM

- 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



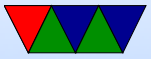
NVRAM

- 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



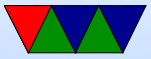
blocks)

- 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



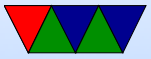
Why not have large SRAM

- 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



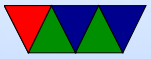
Reading 1

ARM Reveals Cortex-A72 Architecture Details by Andrei Frumusanu http://www.anandtech.com/show/9184/arm-reveals-cortex-a72-architecture-details

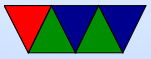


From website

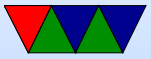
- Cortex-A72 Announced Feb 2015, this article from April 2015.
- A72 direct successor to A57
- How much performance from design, how much from shrink to 14nm/16nm process? FinFET
- Process size is getting more and more hand-wavy
- 3.5x performance of Cortex A15 in smartphone envelope



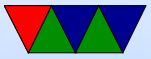
- 75% less energy for same workload
- 16nm FF POP? (finfet package-over-package?)
- Note they are comparing against Cortex-A15 not A57 directly. A15@28nm, A57@20nm 1.9x better, A57@16nm 2.6x better, A72@16nm=3.5x better. Normalize, 1.3 times better than A57?
- Performance, interesting mix of benchmarks. SPEC2006, Stream, LMBench, bunch of others.
- Performance, Power, Area



- Better branch predictor “sophisticated new, reduce energy from mispredict and specualtion” Bypasses completely if doing a bad job
- Fancier decode, ARM64 instruction fusion. Lots of power optimization
- 5-wide dispatch
- Advanced FP/SIMD unit: 3-cycle FMUL, 3-cycle FADD, 6-cycle FMAC, 2-cycle CVT ???, radix-16 FP divider
- L1/L2 bandwidth improved by 30%. Sophisticated

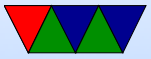


prefetcher? Lots of power optimization



Reading 2

A walk through of the Microarchitectural improvements in Cortex-A72 https://community.arm.com/groups/processors/blog/2015/05/04/a-walk-through-of-the-microarchitectural-improvements-in-cortex-a72



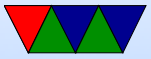
Blog Posting

Single core FP performance (For me, linpack/cores)

Type	Cores	ARM Reported	My own measurement (LINPACK)
Cortex-A8	1	1.0	1.0 (0.068)
Cortex-A9	2	?	7.0 (0.95/2)
Cortex-A9	4	?	??
Cortex-A15	8	11.0	11.0 (3.0/4)
Cortex-A57	?	12x	58.0 (16.0/4)
Cortex-A57		15x	
Cortex-A72			

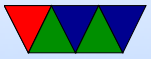
5x speedup Cortex-a15 to cortex-a57 (they only say 2?)

- Tradeoff – larger branch predictor can cost more power



but reduce speculation so save energy

- Cache– found way for 3-way cache to have similar power to direct mapped
- TLB – can turn off high bits when assuming locality?
- Branch predictor – optimize for close branches?
- Micro-op handling?
- Lots of power optimization
- Better prefetcher



- L1 dcache-hit predictor
- Power optimization in L2-idle case
- Big-little with Cortex-a53