# ECE 571 – Advanced Microprocessor-Based Design Lecture 22

Vince Weaver

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

vincent.weaver@maine.edu

19 April 2018

# Announcements

- HW#11 will be posted

# Reading 1

Exploring DynamIQ and ARMs New CPUs: Cortex-A75, Cortex-A55
by Matt Humrick

https://www.anandtech.com/show/11441/dynamiq-and-arms-new-cpus-cortex-a75-a55

# Intro

- Multiple design teams in multiple locations? Where? Austin, Cambridge, Taiwan
- Bifrost GPU
- Cortex A73 – new big CPU
- Successor to A53 (A53 is in the Raspberry Pi 3)
- 1.7 billion in 3 years
- Big went from A57 to A72 to A73
- Octane performance? Javascript benchmark

# Cortex-A53

- Dual-issue in-order
- Good throughput, improve memory system
- New prefetcher, integrated L2 (50% less latency)
- “Extra level” of L3 cache
- 18% Specint2006, 38% Specfp2006
- LMBench memory test?
- Power consumption (mW/GHz) up 3% but “power efficiency” up 15% on SPECint
- Virtual Host Extensions (VHE) automotive virtualization

- "ASIL D" – a risk classification scheme ISO 26262 - Functional Safety for Road Vehicles standard
  ASIL A-D, D is the highest
  SIL 1 is dangerous failure limit of $10^{-5}$ per hour while SIL 4 is associated with a probability of dangerous failure rate limit of $10^{-9}$. ASIL D has been shown aligned to SIL 3
- RAS support aarch64 spec – Reliability, availability, serviceability
  Propagating soft errors

# Cortex-A73/A75

- A73 was thermal/power efficiency
- A75 increase performance staying in same envelope
- Single thread performance over Cortex-A73
  - 22% more perf
  - 33% higher fp/NEON
  - 16% memory throughput
  - 48% more Octane, 34% more geekbench
- A73 wasn't necessarily faster than A72 on floating point
- A75 at 3GHz/10nm better perf and same efficiency (so

more power)

- Thermally limited with 4 cores, but common for mobile chips
- Mobile limited 750mW, ARM wants to go beyond
- At 1W/core and 2W/core outperforms
- A73 specifically went for power efficiency removing features (ECC on L1, 256-bit AMBA 5 CHI)

# DynamIQ

- Extension of big.Little
- Cortex-A75 with Cortex-A55
- Private L2, Shared L3
- Snoop Control Unit (SCU)
- Place up to 8 CPUs (rather than 4) in a cluster scale up to 256 total
- Each core in own power domain (can be turned off) but each cluster share DVFS frequency
- Up to 8 volt/frequency domains so in theory could be at

core level (but requires own voltage regulator)

- Instead of 4+4 pairings, can have 1+7, 2+6, 3+5, 4+4
- 1+7 with big A75 good performance
- DynamIQ Shared Unit (DSU)
- Move L2 closer to core, instead shared L3. 16-way, 0, 1,2,4MB, mostly exclusive
- L3 ECC parity, SECDEC (for dirty data) SED (on only clean). ASIL-D
- “Cache-stashing”: GPU or accelerator can write directly into L2/L3

# Power Management

- Faster power down?
- Leakage reduction
- Finer-grained CPU power management

# Cortex-A75 uarch

- ARM Sophia class, A73, A17, A12
- Move from ARMv8.0 to ARMv8.2
- 11-13 stage OoO
- 3-wide decode
- can dispatch 6uop/cycle
- Note, A75 vulnerable to Meltdown?
- Slot based?
- 64k L1-icache 4-way VIPT
- 0-cycle branch prediction? Upstream of main predictor?

- L1 Dcache, 64kb, 4-way, VIPT (PIPT programmer view?) 4-way VIPT looks like 8-way 32k or 16way 64k
- improved prefetcher
- Integrate (rather than shared) L2. 256k or 512k. Larger only helps with DynamIQ
- L2 hit rate biased for instructions
- Non-blocking TLB, no need for everyone to wait while pages being walked
- NEON. 16-bit FP, Int8 dot product (neural nets)
- Atomic operations?

# Cortex-A55 uarch

- Dual-issue in-order 8-stage pipeline
- No real speedup going from 16/14nm to 10nm to 7nm (scaling gains mostly in power/leakage saving)
- Configurable cache. 16,32,64k, 4-way, optional parity (SED). VIPT
- 15-entry L1 TLB
- neural-network based branch prediction
  loop termination prediction, 0-cycle uop predictors
- Improved data prefetcher

- Fully exclusive (vs pseudo-exclusive) cache
- Moves from PIPT to VIPT cache (but handles aliases in hardware?)
- Integrated L2, 0, 64, 128,256k
- L2 is PIPT, simpler, less power
- Improved L2 TLB
- Improved AGU
- NEON optional
- FP16 support. A53 could fetch but converted to fp32 first
- Fused-multiply-add

# Reading 2

Cortex-M7 Launches: Embedded IoT and Wearables by Stephen Barrett

https://www.anandtech.com/show/8542/cortexm7-launches-embedded-iot-and-wearables

# Background

- Big difference between A and M is lack of MMU
- Can you run general purpose OS w/o a MMU?
- Lots of A and R class cores come with M cores attached
- Combating 8 and 16-bit mcus
- M series more popular than all other ARM, 1.7 billion in 1st half 2014

# Cortex-M7

- Why M7?
- IoT
- six-stage in-order dual issue
- integer and float
- Unlike previous M: branch predictor, caches, TCM (tightly coupled mem)
- brpred help on DSP-like code
- Thumb/Thumb2/hw multiply,divide,DSP extensions, saturate, FP, TCP, ARMv7-M

# Hybrid Systems

- low power M core might be on a lot of time time, only waking A core occasionally
- Wearables?