ECE 571 – Advanced Microprocessor-Based Design Lecture 29

Vince Weaver

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

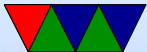vincent.weaver@maine.edu

15 November 2024

Announcements

- No in-person class Monday and Wednesday.
- There might be a video sent out, will send e-mail with details.
- HW#10 will be posted, another set of readings

We Read some Articles

But first some Intel history

Intel x86 Ancient History

- 8086 – 3 micron (μm) process – from 1970s
- 8088 (8 bit bus) thrust into history once IBM used it for PC
- 186 (only embedded) 286, 386 (first 32-bit)
- 486 (pipelined?)
- pentium (why not 586?) dual issue
- Pentium Pro (first out-of-order)
- Pentium II/III
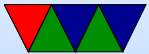- Pentium 4 – gamble on high frequencies, not work out

AMD64 happened around this time

- Core / Core2 was sort of separate issue to extend old PIII architecture

Intel "Modern" Generations

Generation	Name	tech		Year
-	Conroe/Merom	65nm	Tock	2006
-	Penryn	45nm	Tick	2007
1	Nehalem	45nm	Tock	2008
1	Westmere	32nm	Tick	2010
2	Sandy Bridge	32nm	Tock	2011
3	Ivy Bridge	22nm	Tick	2012
4	Haswell	22nm	Tock	2013
5	Broadwell	14nm	Tick	2014
6	Skylake	14nm	Tock	2015
7	Kaby Lake	14nm	Tock	2016
8/9	CoffeeLake	14nm	Tock	2017
9	CannonLake (mobile,rare)	10nm	Tick	2018
10	IceLake/SunnyCove	10nm+	?	2019
11	RocketLake/Tigerlake			2020
12	Alder Lake/(Golden Cove/Gracemont)	Intel 7?		2021
13	Raptor Lake	Intel 7		2022
14	Raptor Lake			2023
15	Arrow Lake			2024

Intel Confusion

- Codenames various places/resorts, often in Pacific Northwest or Israel
- They've had problems so recent generations are all mixed up
- Also P vs E cores, so there's a different set of codenames for the subcores
- Previously named things like "E5-2640v3" (the chip in Haswell-EP)
- These days things like i7-13700 (chip in raptor lake).

i3/i5/i7/i9 various levels of performance, the "13" means 13th generation chip
- Also have Xeon (server) and Pentium (low-end) brands that apply across generations
- The only way to know for sure is family/model/stepping from cpuinfo, but that falls apart on hybrid systems

Intel FAB Issues

- Intel early with 22nm FinFET (Intel 16)
- 14nm Skylake, stuck for a long time
- 14+
- 14++
- 10nm delayed for long time
  - Canon Lake but only very few shipped
- Intel 10SF (10nm Superfin)
- Intel 7 (third gen 10nm, 10ESF Enhanced Super Fin)
- Intel 4 (previously 7nm)

- ○ COAG – Contact Over Active Gate
- ○ Use a lot of Cobalt, reduce resistance 60% compared to tungsten, but difficult to work with
- ○ But in end back to copper, apparently cobalt problem
- ○ Tradeoff high-performance vs high density
- ○ Intel last company to introduce EUV (Extreme UV) which allows reducing number of masks, before that DUV (Deep UV) possibly immersion?
- ○ Hillsboro / Leixlip (Ireland)
- Intel 3
- Intel 20A

- Intel 18A – High-NA EUV (numerical aperture)

Alder Lake Reading

Intel Architecture Day 2021: Alder Lake, Golden Cove, and Gracemont Detailed by Dr. Ian Cutress and Andrei Frumusanu

Alder Lake – Hybrid

- High-end "performance" core
  Based on Golden Cove
- Low-end "efficiency" code Based on Gracemont
- 4 e-cores take up same space as 1 p-core
- Desktop hardware 32 EUs (execution units) Xe-LP graphics
- Gaussian Neural Accelerator (GNA 3.0)
- Only mobile gets Thunderbolt 4?
- Mobile also get IPU (image processing unit)

- Desktop 16 cores, 24 threads, up to 30MB non-inclusive L3 cache
- DDR5, DDR4, LPDDR5, LPDDR4X
- 20 lanes of PCIe, split 4.0 and 5.0 (should give a lecture on PCIe)
- Dual Bandwidth ring, 1000GB/s bandwidth, one of two rings can be disabled to save power (a ring takes as much power as an E-core?)

Alder Lake – Intel Thread Director

- Modern mobile (ARM) processors have 1+3+4 or 2+4+4 super high / high / efficiency cores
- Work with Windows 11 (what about Linux?)
- Three levels: One thread per core p-cores, only thread on e-cores, SMT threads on p-cores
- Embedded microcontroller on each core
- Give hints to the OS, can notice in 30us while the OS can take 100s of ms
- Monitors fancy AVX (from future, AVX-512 disabled on

p-cores because e-cores not support)
- All p-threads + all e-threads higher performance than all p-threads+hyperthreads
- Enhanced hardware frequency Interface (EHFI)
- Working on Linux support?

Alder Lake – Golden Cove uarch

- P-core
- Largest uarch change in a decade
- Move from 4-wide to 6-wide decode
- Decoded clock gated 80% of time, instead relying uop cache
- uop cache now 4k
- icache still 32k, l1 iTLB from 128 to 256 entries
- mispredict penalty up, so increased branch predictor performance. L2 BTB now 12k (we ran the experiment

in homework though, did we see improvement?)

- In theory can hit IPC of 6
- ROB from 352 to 512 entries (double Zen3, behind Apple)
- 5th ALU port, ALU and LEA capabilities
- AGU port, loads per cycle from 2 to 3
- AVX512 can load 1k-bit/cycle from l1 cache
- L1 DTLB from 64 to 96
- Double number of "page-walkers" from 2 to 4 to handle TLB misses
- prefetcher: better stride prefetch into L1

- prefetching improvements in the L2. What the company calls "full-line-write predictive bandwidth optimization"
- 19% IPC compared to Cypress cove, but older than Willow Cove/Tiger Lake (Intel often has subtly odd comparisons in their slides, have to watch out for it)

Alder Lake – Gracemont

- 4th Gen Out-of-Order Atom
- e-core
- Based on atom processor
- Claim better than skylake at 1C1T with less power
- Dual 3-wide decode paths (two streams, 3 per cycle?)
- 64k L1 icache
- 17 execution ports
- 32-k L1 dcache, 3-cycle pointer chasing latency?
- Shared 4MB l2-cache

- First atom with AVX2 support
- Control-flow enhancement tech (CET) (security)
- Tuned for voltage (power)

Academic Paper

- Intel Alder Lake CPU architectures by Rotem at al. IEEE Micro, May-June 2022, p13-19, vol42.
- Rotem wrote the RAPL paper
- Really weird typos at points, "4,000 pages in the TLB" instead of 4k
- Not really well written
- Pcores
  - Simultaneous 128 bytes of read/write bandwidth per cycle

- ○ L1 5-cycle load-to-use latency
- ○ 2048 entry second level TLB (STLB)
- ○ STLB misses sent to page miss handler (PMH) can do 4 page walks in parallel
- ○ L2 is 1.25MB in size? 15 cycle latency
- ○ New micro-controller in each core can track power in microseconds and update power budget
- Ecores
- ○ frontend 32-byte prediction. First is NLP (next line predictor) taken branch can be predicted each cycle. Second level is 5k entry target array (?) three-cycle

predictor
- Instruction boundaries stored in cache for decode
- on-demand instruction length decoder block
  two extra cycles?
- 5-instruction wide
- move and nop detection
- retirement is 8 instructions wide, out of order window
  256 entries
- Reservation stations
- peak execution 16 single-precision FP per cycle
- two 16-byte loads / two 16-byte stores per cycle

○ 3-4 cycles for load

○ 2MB L2 shared 4 cores

○ Memory disambiguation? (tries to figure out if memory accesses overlap or not)

○ L2 64 bytes/cycle, 17 cycles

○ PMH can also perform 4 walks in parallel

○ Power: each core can be individually clock gated (C1) or power gated (C6). L2 can be kept coherent independent of cores power.

○ L2 flushing supported in MC6 deep sleep state

○ "each core operates at the same frequency, which is

independent of the fabric and other cores"

- ○ Performance: E core is .8x that of P core for SPEC2017 int, .62x for FP

  Oddly a 2-core P int workload is 13% slower than 4-core E, but for FP is 17 lower. 4 e-cores roughly half area two p-cores. P core can run at higher freq at same voltage.

- SoC
  - ○ Desktop, Mobile, Ultramobile
  - ○ DDR4/DDR5/LPDDR4/LPDDR5
  - ○ I/O: PCIe Gen5, Wi-fi 6

- **Thread Director**
  - ○ Presented by hardware as all equivalent threads
  - ○ Thread Director tracks behavior of threads at ns granularity
  - ○ Categorizes workloads

    Class 0: Most, arithmetic. E-core good

    Class 1-2: Emerging AI and Vector workloads. P-core 2.75x better

    Class 3: spin loops, best on E-core
  - ○ Info stored at context switch
  - ○ E-core isn't always most energy efficient, e.g. if a job

finishes faster on P-core

- ○ Up to 65% performance improvement with thread director enabled
- ○ Windows 11
- ○ Despite some confusing news online, possibly support is not in Linux yet and unclear if it will be merged. Ricardo Neri patches

Sapphire Rapids (server version of Alder Lake)

- Server version
- Jim Keller design? Had famous career
  Alpha 2164, then AMD K7/K8, Broadcom, then PA who Apple Bought, Apple A4/A5, AMD Zen, Tesla, Intel, Left Intel
- Ill-fated Aurora supercomputer waiting on Intel Sapphire rapids and Ponte Vecchio GPU (Xe)
- Multi-die with 4 tiles, each 15 cores?

- Server (-SP) 4 to 8 sockets?
- Workstation with 56 cores, compete with ThreadRipper Xeon W5/W7/W9
- Some models HBMe L4 cache
- Built-in PCIe Accelerators
  - Intel Dynamic Load Balancer (DLB)
  - Data Streaming Accelerator (DSA) data copies and crc32
  - In-Memory Analytics Accelerator (IAA) compression
  - QuickAssist Technology (QAT) encryption
- Advanced Matrix Extension (AMX) – similar to tensor

cores

Sapphire Rapids Tech Paper (ISCC 2022) solid-state circuits

- Xeon Processor
- 100MB L3, 8 DDR5 channels
- 4 dies Intel 7
- Super-Fin transistors
- Asynchronous buffers
- Redundant circuitry in case of defect
- FIVR and motherboard VR (voltage regulators)
- Lots of low-level details on I/O

Sapphire Rapids writeup

- CXL support
- Golden Cove processors
- FP16 support
- Faster UPI
- HBM
- Based on Alder Lake
- AVX-512
- AMX – 2d register file, 8 registers 1k each T0-T7, TMUL instruction

- Tiles
- HBM, UPI, Optane

Raptor Lake

- Intel snuck out a small update to Alder Lake to better compete with AMD
- 5.8GHz, 2MB L2, new dynamic prefetch algorithm
- Can beat AMD but takes more power?
- Has lots of E cores, Ryzen only has P cores
- i9 moved from 8 to 16 cores
- i7/i5 from 4 to 8 cores