ECE 571 – Advanced Microprocessor-Based Design Lecture 13

Vince Weaver

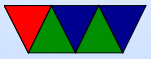
http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

30 September 2020

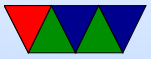
Announcements

- Don't forget HW#4. A little trickier than previous as you run on 3 different machines



HW#3 Review – The Benchmarks

- sleep – does nothing
- stream – stresses memory subsystem
- matrix – loads the CPU
- iozone – disk I/O



HW#3 Review – The System

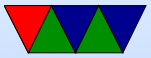
- Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz
  20MB cache, 22nm, 90W TDP

  `https://ark.intel.com/content/www/us/en/ark/products/83359/intel-xeon-processor-e5-2640-v3-20m-ca`

  `html`

- 80GB DDR4 RAM

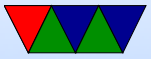
- Regular spinning hard drive



HW#3 Energy

Energy

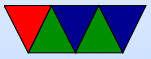
	sleep	stream	matrix	iozone
cores	0.00	0.00	0.0	0
sleep	427	465	305	49
ram	55	85	45	7.8
time	10.0	7.4	5.2s	1.0



HW#3 Power

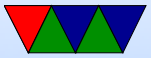
Power

	sleep	stream	matrix	iozone
cores	0.00	0.00	0.0	0
pkg	42.7	63	58	49
ram	5.5	11	9	7.8



HW#3 Notes

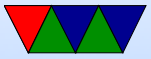
- highest power pkg: stream
- highest power dram: stream
- cores: always zero
  This is likely a bug with the Haswell-EP chips
  Intel slow to acknowledge this
- server class does not have integrated GPU
- guesses: old, virtual machine
- HPL (20k): pkg: 117W dram: 11W
- stream stresses memory. iozone, CPU is waiting?



HW#3 Energy-Delay

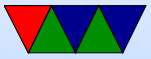
Energy-delay

	1	2	4	8	16	32	64
E	13.9k	8.3k	7k	4.4k	4.8k	5.2k	5.7k
time	220s	120s	79s	39s	35s	36s	47s
ED	3058k	996k	553k	171k	168k	187k	267k
ED2	672M	120M	44M	7M	5.8M	6.7M	12M
Power	63W	69W	89W	112W	137W	144W	121W

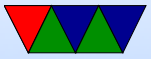


HW#3 Energy-Delay Discussion

- interesting, half class got results where 32 best
- a) fastest time = 16
- b) lowest energy = 8
- c) lowest E-D = 16
- d) lowest E-D-D = 16
- e) scaling? only can show strong (problem size same)
  - Poorly written benchmark? (possible)
  - Not enough memory? (unlikely, benchmark from 2001)
- f) 32 threads, but only 16 cores

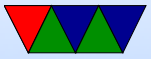


Oh No, More Caches!



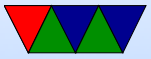
Cache Terms

- Line – which row of a cache being accessed
- Blocks – size of data chunk stored by a cache
- Tags – used to indicate high bits of address; used to detect cache hits
- Sets (or ways) – parts of an associative cache



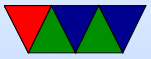
Replacement Policy

- FIFO
- LRU
- Round-robin
- Random
- Pseudo-LRU
- Spatial



Load Policy

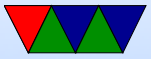
- Critical Word First – when loading a multiple-byte line, bring in the bytes of interest first



Consistency

Need to make sure Memory eventually matches what we have in cache.

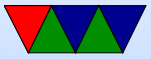
- write-back – keeps track of dirty blocks, only writes back at eviction time. poor interaction on multi-processor machines
- write-through – easiest for consistency, potentially more bandwidth needed, values written that are discarded
- write-allocate – Usually in conjunction with write-back Load cacheline from memory before writing.



Inclusiveness

- Inclusive – every item in L1 also in L2
  simple, but wastes cache space (multiple copies)

- Exclusive – item cannot be in multiple levels at a time



Other Cache Types

- Victim Cache – store last few evicted blocks in case brought back in again, mitigate smaller associativity
- Assist Cache – prefetch into small cache, avoid problem where prefetch kicks out good values
- Trace Cache – store predecoded program traces instead of (or in addition to) instruction cache