ECE 571 – Advanced Microprocessor-Based Design Lecture 27

Vince Weaver

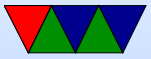
http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

6 November 2020

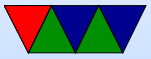
Announcements

- Homework 8 due (hopefully you read the papers)
- Homework 9 will be another reading, modern CPUs

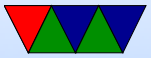


Reading

A Validation of DRAM RAPL Power Measurements
by Desrochers, Paradis and Weaver

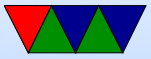


Digression on Academic Papers

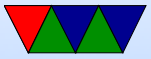


Page 1

- Work I did with some students, undergrad and grad
- MEMSYS'16. conference. Won an "award".
- RAPL, powercapping. What's that good for?
- RAPL
  - Package
  - Cores total
  - DRAM
  - GPU
  - SoC (skylake)

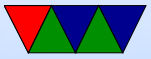


- Haswell-EP server with 80GB RAM is 13W of power that's not even with all slots full
  428GFLOPS incidentally (2.1 GFLOPS/w)
  130W CPU/16 cores, DRAM more than a core.

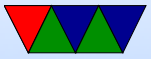


Page 2

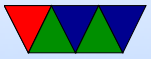
- Notes on the documentation. Intel tries, but their documentation can be a real pain sometimes, often conflicting and out of date. Also their terminology an be really confusing.
- We sort of noticed that Haswell desktop DRAM support was added accidentally, it was documented in some obscure sub-document (not the main documentation)
- PP0 (cores) does not seem to be supported on server-class machines, again, Intel does not really say why



- Lack of timestamp is an issue, it makes it hard to measure small intervals, and also makes it easy to double-count some intervals if trying to do phase charts. Aggregate is mostly OK.
- Haswell-EP with "RAPL Mode 1" (Real measurement due to integrated voltage regulator)
- Again with documentation, the DRAM energy quantum was different, only obscurely mentioned (and people noticed when you got really bizzarre readings)
- Three ways to read RAPL results. A pain. PAPI makes this worse.

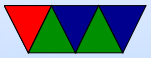


- RAPL measured using perf tool
- Related work: tried measuring DRAM on Sandybridge (the one Chad fried) but for whatever reason the HP server turned off support for some reason
- Related work: previous validations, including the original Intel authors, mostly had one fuzzy graph and that was it
- DRAM RAPL. Parametric model. Genetic algorithms. Calibrated at boot.



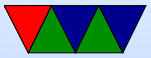
Page 3

- Instrumenting the hardware
  P4 power connector
  ATX power measurement and previous students
  Why a hall effect sensors vs sense resistor? Tens of amps of power. 10A * .1Ohm = 1V voltage drop, 10W of power.
- DIMM extender card
  Various voltages (how many? how many relevant?)
  DDR3 has 5 voltages

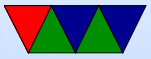


- VDD (main supply) 1.5V
- VDDQ (I/O driver, but tied to VDD)
- VREFDQ – reference
- VREFCA –reference
- VDDSPD – for the eeprom

- DDR4 Voltages
  - Vdd (main supply) 1.2V
  - Vtt termination
  - Vpp activation 2.5V
  - 12V – not used on our dimms
  - Vddspd – eepro

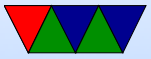


  - Vrefca – reference
- PCIe extender cards
  small resistance. Instrumentation amplifier
  Data acquisition board.
- Measure with perf.
- Synchronizing the measurements.
  - Hard at high frequencies.
  - RAPL measured locally (you have to)
  - Voltages logged on separate machine
  - Used serial port triggered by perf to click one of lines on DAQ board



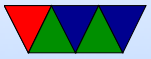
- Other ways to do it?
- On green500 list/wattsup just use NTP to make sure within a second.

- RAPL overhead, only measure at 10Hz.
  Overhead of too many interrupts, writing to disk. Also power overhead.

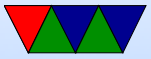


Page 4-5-6

- Measurement accuracy concerns
  - Power conversion from 12V down (we measure after conversion)
  - Synchronization
  - Long wires, breadboards
  - Non-linearity in instrumentation amplifier
  - BIOS firmware variation
  - Temperature dependencies
- Does putting the DIMM in make things better/worse?

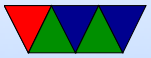


- Overhead of using perf. 0.5W more power gathering at 100Hz. at 1kHz perf interrupts taking more than 25% of CPU time



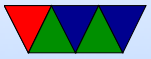
Page 6-7

- Benchmark choice
  - idle: sleep
  - dram: stream OpenMP
  - CPU/FP: Linpack, with BLAS: ATLAS, OpenBLAS, MLK
  - CPU/Int: gcc compiling PAPI
  - GPU: OpenCL ray-tracer, KSP

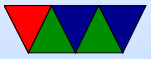


Page 8

- Results
- Benefit of sharing all raw data
- Do Tables tell full story?
- Figure 8 can see on i5 under-report, plus really bad on Samsung
- Intel-MKL matches well
- Same DIMMs are being used
- CPU power rises above total power? Artifact of sample rates.

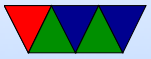


- Phase Plots. Do they, match well? Underestimate when idle, but spot on in a few cases.
- Haswell-EP results are better.
  Notice that Vpp never amounted to much



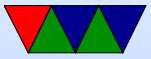
Easy Future Experiments

- Conduct same measurements on other machines SODIMMs? Skylake?
- Get another memory extender and see how it works with two DIMMs
- Measure RAPL overhead, can we run at 1kHz if we read MSR directly too a buffer w/o any other overhead? Still need a timer of some sort.



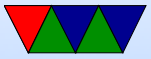
Another Reading

- Power Measurement Techniques on Standard Compute Nodes: A Quantitative Comparison
- Hackenberg, Ilsche, Schoene, Molka, Schmidt, Nagel, TU-Dresden
- ISPASS 2013 (Austin, TX)
- Tell bat story.

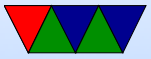


Page 1 + 2 + 3

- IPMI interface – for server machines
  I had Chad look at this but he got weird results
- PDUs
- AC Instrumentation
  - ZES ZIMMER LMG450 (how much does it cost?)
  - IPMI/PDU
- DC Instrumentation
  - p8 connector – found it powers CPU and DRAM but not refresh?

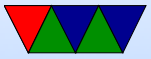


  - Hall effect sensor
  - National instruments PCI-6255 DAQ
  - PCIe by using a 12V-¿ATX converter, measure 12V
- RAPL
- APM – AMD Application Power Management – have had problems with that. Only measure last 10ms?

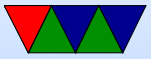


Page 4

- Synthetic Workloads
  - sleep
  - dgemm
  - memory
  - sin
  - sqrt
  - mult-add
  - OpenMP ping-poing
  - syscall (gettimeofday)

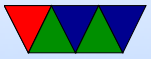


- Vampir – from Dresden
- RAPL MSR 0.46us. Full scan 8.6us
- APM with libpci, 70us
- Synchronization: NTP, also “defined workload signal”



Page 5

- PDUs have trouble, but the LMG450 did not
- Mainboard (ATX?) power consumption 33-35W
- p8 connector – 1W to 100W
- Small enough sample rate, can see interrupts
- RAPL does not account for hyperthreading?
- APM results not as good
- Filtering
- SpecOMP



Results

- Measuring total energy of compute job – all methods OK except maybe APM
- Coarse grained – OK. Some people told them more than 1 sample/second won't work on AC due to filtering caps, but that's not what they saw. Don't use PDU/IPMI for this
- High resolution –