

ECE 571 – Advanced Microprocessor-Based Design Lecture 18

Vince Weaver

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

`vincent.weaver@maine.edu`

7 November 2019

Announcements

- Homework 8 was last minute (sorry about that)



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



Saving Power/Energy with RAM

- AVATAR: A Variable retention time aware refresh for DRAM systems by Qureshi et al.
 - JEDEC standard: cell must have 64ms retention time
 - Why refresh bad? Block memory, preventing read/write requests
 - Consume energy (6,28,35)
 - The bigger DRAM gets, more refresh needed
 - predict that in 64Gb chips 50% of Energy will be in refresh



- Multi-rate refresh possible – detect which cells need more and refresh them more often (can be a 4-8x difference)
- VRT (variable retention rate) a problem. Some cells switch back and forth between. So when you probe it might check fine, but then fail later.
- They find that addition of cells stabilized to one new cell/15 mins over time
- Use ECC to catch these errors, though relying on ECC in this case can lead to uncorrectable error every 6 months



- They propose using ECC to adjust the VRT at runtime based on errors that are found
- They find on a 64Gb chip improves perf by 35% and **Energy-Delay** by 55%
- “Refresh-wall”
- Memory controller keeps track of this info
- VRT first reported in 1987. Fluctuations in GIDL (gate-induced drain leakage) presence of “trap” near the gate region
- Intel and Samsung say VRT one of biggest challenge in scaling DRAM



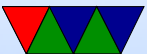
- VRT not necessarily bad – can cause retention to get better!
- Test – use FPGA to talk to 24 different DRAM chips, at controlled temperatures.

Why do they use an FPGA?

- Actually it's just 3 chips from different vendors, each with 8 chips (for 24)
- Look into ECC. Soft-error rate is 200-5000 FIT/Mbit. Every 3-75 hours for 8GB DIMM. Soft errors happen 54x-2700x lower rate than VRT
- Downside of ECC ... have to scrub memory to check



- for errors. Also has energy/perf overhead. Energy to refresh DIMM 1.1mJ, energy to scrub 161mJ (150x) but if you scrub every 15 minutes it's a win.
- Use memory system simulator USIMM



Cryogenic Memory

- Dip DIMMS in liquid nitrogen
- Low power? Faster? Interface with quantum circuits?



Rowhammer

- Been observed for years, adjacent rows discharging can affect nearby rows
- Particularly bad in DDR3 from 2012-2013
- Accessing same row over and over can make voltage fluctuations in nearby rows, causing faster leakage than normal
- Mitigations? Refresh more often? ECC? Refresh nearby lines if a lot of row hammering going on?



- Can cause exploit. Google NaCl disable “cflush” exploit (need to force access to row)
- Can also trigger just with lots of cache misses
- If you can flip bits of kernel/trusted pointers to point to something you control, then you win.



Notes from last time

- Can you buy phase change ram?
Micron sold from 2012-2014? No one wanted.
Amorphous if you heat and quench, crystal if cook a while
- Millipede memory, tiny bumps, MEMS devices to read
- Can you buy Optane?
April 24th? Special M.2 slot on Gen7 (Kaby lake? motherboards)



For now, 16GB and 32GB modules, using like a cache of your hard disk.

- Hybrid Memory Cube, Micron, 15x as fast as DDR3. Fujitsu Sparc64 2015 has some
- High Bandwidth Memory (AMD+Hynix) AMD Fiji, NVIDIA Pascal, Intel Knight's Landing Interposer (diagram)



DRAM – Mobile DRAM

- From Micron: “TN-46-12: Mobile DRAM Power-Saving Features”, 2009
- Temperature-Compensated Self Refresh (TCSR) – Auto adjust refresh timings based on temperature
- Partial Array Self Refresh (PASR) – only refresh parts of RAM that have data in them
- Deep Power Down (DPD) – enable turning off the voltage generators when maintaining DRAM not needed
- Has equations for estimating power usage



DRAM – Elsewhere

- Tom's Hardware. 2010. "How Much Power Does Low-Voltage DDR3 Memory Really Save?" Using low-voltage (1.25 or 1.35 rather than 1.5) DDR3 DRAM can reduce power by 0.5-1W. Slower performance settings, but not really noticeable.
- Linus Torvalds Rant from 2007: DRAM Energy not a prime concern. Just don't use FBDIMMs if you want low-power.



DRAM – Recent Academic

- “Rethinking DRAM Power Modes for Energy Proportionality”, Malladi et al, Micro 2012.
 - DRAM spends lots of time idle, but latency is so high for wakeup it cannot utilize powerdown modes
 - Reference 25% of data-center energy usage is DRAM?
 - Use LPDDR2 trades bandwidth for efficiency
 - Current modes involve turning off DLLs (Delay-locked loops?) which are slow to turn on again, 700ns+
 - some background on DRAM operation



- Low-power mode sounds good, but then it takes 512 memory cycles of power to re-start (a lot of energy)
- Propose MemBLAZE. Moves clock generation out of DIMM and into memory controller, allowing fast wakeup
- “Towards Energy-Proportional Datacenter Memory with Mobile DRAM”, Malladi et al, ISCA 2012.
 - Look at using LPDDR2 in servers rather than DDR3.
 - DDR3 often in “Active-idle” as many workloads do not allow sleep.



- “A Predictor-based Power-Saving Policy for DRAM Memories”, Thomas et al, EuroMicro 2012.
 - Use a history based predictor to pick when to powerdown.
 - Say up to 20% of mobile devices and 25% of data center is DRAM
- “Rethinking DRAM Design and Organization for Energy-Constrained Multi-Cores”, Udipi et al., ISCA 2010
 - DRAMs “overfetch” which hurts energy
- “A Comprehensive Approach to DRAM Power



Management” , Hur and Lin, HPCA2008.

- Throttling and Power Shifting – slowing down to fit in power budget
- Put DRAMs in low power mode – available commercially but no one seems to use this yet
- Simulate for Power5 and DDR2-533
- Modify the memory controller



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 can 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 bizarre 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. $10A * .1\Omega = 1V$ voltage drop, 10W of power.

- DIMM extender card

Various voltages (how many? how many relevant?)

DDR3 has 5 voltages



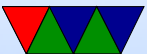
- VDD (main supply) 1.5V
- VDDQ (I/O driver, but tied to VDD)
- VREFDQ – reference
- VREFCA –reference
- VDDSPD – for the eeprom
- DDR4 Voltages
 - Vdd (main supply) 1.2V
 - Vtt termination
 - Vpp activation 2.5V
 - 12V – not used on our dimms
 - Vddspd – eeepro



- 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.

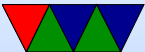


- Phase Plots. Do they, match well? Underestimate when idle, but spot on in a few cases.
- Haswell-EP results are better.
Notice that V_{pp} never amounted to much



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- μ 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-pong
 - 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 –

