ECE 571 – Advanced Microprocessor-Based Design Lecture 24

Vince Weaver

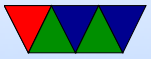
http://www.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

25 April 2013

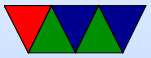
Project/HW Reminder

- Project
  Presentations. 15-20 minutes. Can use projector.
  2 on Tuesday the 30th, 3 on Thursday the 2nd.
  Papers due next Friday, can extend to Wednesday if necessary.



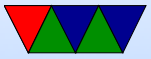
ISPASS Report

- A *lot* of interest in power measurement especially RAPL
- ARM / Samsung really interested in performance counters
- 64-bit multiply, 1ns, 50pJ
  cache access 10ns, 1,000nJ
  DRAM access 100ns, 10,000nJ

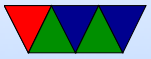


Power Measurement Techniques on Standard Compute Nodes: A Quantitative Comparison

- Hackenberg et al, TU Dresden
- Look at power consumption on servers and various tools for measurement
- ZES ZIMMER LMG450, 250V power since in Europe
- Do not use auto-range as it takes 1s to adapt



- Some claim you can't get much better than 1% while measuring A/C due to large power supply filter caps
- IPMI – built into Dell power supply, accessible over TCP/IP, Dell claims 1% accuracy
- PDU – power distribution unit for AMD system, collected from a separate machine with python script, manufacturer claims 2%
- on inside, monitor the 12V voltage lane, also a hall-effect sensor for even better resolution. Resolution limited by sensor and by capacitors on mainboard

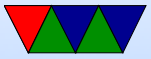


- RAPL provide only Energy and no timestamps so derived Power values not as accurate as could be

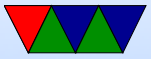
- AMD APM (application power measurement). Unlike RAPL only holds value for last 10ms capture, so can get power values

- Reading values like AMD APM through kernel (rather than direct libpci) adds overhead

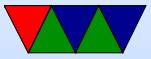
- Hard to synchronize measurements, especially when things happening on sub-ms time scales



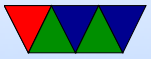
- Irregular workloads below the sample threshold hard to measure. Some tools properly average them, some do not
- RAPL not always accurate on all workloads. Does not correctly account for hyperthreading. Does handle Turboboost / DVFS.
- AMD APM works best with TurboCore disabled and C-state C6 enabled
- If trying to optimize small-length code sequences, RAPL and other measurements not fine-grained enough



- Energy Measurement of Full compute job – using power-supply measurements fine. RAPL is good but only covers CPU. APM has trouble with sleep modes
- Low-resolution power consumption measurement – 1 - 20 Hz. Show AC measurement can do well despite filter capacitors.
- High-resolution power consumption measurement – 100Hz or more. Can measure down to 1ms. Measure up to 10kHz, may need to filter.

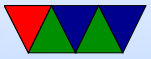


HotCHIPS 24

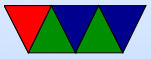


Centip3de: A 64-Core, 3D Stacked Near-Threshold System

- 3d chip, seven layers (2 core, 2 cache, 3 DRAM)
- 28 ARM Cortex-M3, 256MB DRAM
- 3930 DMIPS/W
- Voltage hotspots
- Clusters of 4-cores share 8kB L1-data (4-way), 1KB L1 insn(4-way)

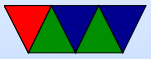


- Cache run at higher frequency and voltage than cores (can't scale memory as efficiently as logic)
- If want turbo boost, shut off 3 cores and scale core to speed of cache
- If cache read in single-core mode takes more energy as have to read cache tags in parallel
- Cortex A9, 40nm, 8000DMIPS/W
  Centip3de: 130nm, 3930DMIPS/W

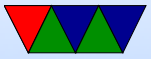


Reducing Transistor Variability for Higher-Performance Lower-Power Chips

- Threshold voltage has plateaued due to transistor variability at the fab
- VDD scaled from 5V at 0.5um to 1.2V at 130nm to 1V since then
- Doping is down to a few 100 atoms per channel. Thus just a few atoms either way can drastically change transistor properties

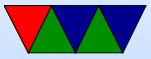


- Ways around this: Full Depleted Silicon on Insulator Finfets, Deeply Depleted Channel
- SRAM memory use smallest transistors, limit Vdd reduction
- Design tends to be conservative. Some places "bin" chips but SOC designs typically don't
- Using DDC down to 0.9V, 50% power reduction

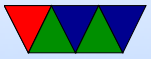


IA-32 Processor with a Wide-Voltage-Operating Range in 32nm CMOS

- 280mV - 1.2V original Pentium P54c
- 8kB I/D cache
- Memory on a separate voltage plane from logic
- Added Power PMU
- Can power-gate FPU

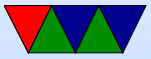


- Not to ramp up and down FPU too much, as would waste power
- At 500mV, 5 cycles break even point for power gating. 12 cycles at 1.2v
- Custom interposer and classic 15-year old Pentium motherboard
  Old hardware putting out 4.1V on 3.3V bus
- 3 to 920MHz, trouble for 66/133MHz of original bus
- Only power gate FPU as other parts were too active,



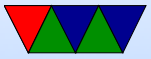
didn't idle enough

- 380mV at 10MHz, 1.5mW
- 1.1V at 780MHz, 444mW
- Boot into OS using a solar cell the size of processor die
- Leakage power 50% at low voltage

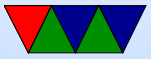


IBM zEC12: The Third-Generation High-Frequency Mainframe Microprocessor

- Power no object. Runs at 5.5GHz
- 6 Cores
- EDRAM – 48MB L3 Cache
- dedicated data compression/encryption engine
- Power features: some clock gating, save 25% of chip power

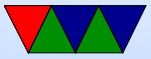


- high VT, non-pipeline adjusted
- No DVFS as expected to always run at 90% load or higher
- 2-level BTB, acts as instruction prefetcher
- L2 split 1MB cache, 1MB data
- Transactional Memory
- Advanced profiling available

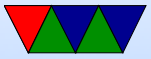


The Oracle SPARC T5 16-core Processor Scales to Eight Sockets

- 16-cores, 3,6GHz. L1 I/D = 16kB, L2=128KB, 8MB 16-way L3, banked, bank selected by bits 8:10
- DDR3 up to 512GB
- 2x8 PCIe gen 3
- Power: scale power with workload
- DVFS



- coherence links powered up and down to save energy
- PCI power management
- ECC
- DVFS, part of lights-out monitor
- Performance p-states: only scale for thermal reasons
- Elastic p-states: OS control
- Thermal and current capping



- Monitors I/O, fan, etc and throttles
- During production p-state threshold set in EFUSE rom
- Cycle skipping, can skip 1-8 of 8 cycles
- Gradually ramp up frequency one core at a time to avoid stressing voltage controller
- 25W power savings by shutting down coherence link
- Error in caches found then kick out