ECE 571 – Advanced Microprocessor-Based Design Lecture 6

Vince Weaver

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

vincent.weaver@maine.edu

4 February 2016

Announcements

- HW#3 will be posted
- HW#1 was graded

First Half of Class – Discuss Paper

Producing Wrong Data Without Doing Anything Obviously Wrong! by Mytkowicz, Diwan, Hauswirth and Sweeney, ASPLOS'09.

Measuring Power and Energy

Why?

- New, massive, HPC machines use impressive amounts of power
- When you have 100k+ cores, saving a few Joules per core quickly adds up
- To improve power/energy draw, you need some way of measuring it

Energy/Power Measurement is Already Possible

Three common ways of doing this:

- Hand-instrumenting a system by tapping all power inputs to CPU, memory, disk, etc., and using a data logger
- Using a pass-through power meter that you plug your server into. Often these will log over USB
- Estimating power/energy with a software model based on system behavior

Measuring Power and Energy

- Sense resistor or Hall Effect sensor gives you the current
- Sense resistor is small resistor. Measure voltage drop. Current V=IR Ohm's Law, so V/R=I
- Voltage drops are often small (why?) so you made need to amplify with instrumentation amplifier
- Then you need to measure with A/D converter
- $P = IV$ and you know the voltage
- How to get Energy from Power?

More on Hall Effect Sensors

Power and Energy Concerns

Table 1: ATLAS 300x300 DGEMM (Matrix Multiply)

Machine	Processor	Cores	Frequency	Idle	Load	Time	Total Energy
Raspberry Pi	ARM 1176	1	700MHz	3.0W	3.3W	23.5s	77.6J
Gumstix Overo	Cortex-A8	1	600Mhz	2.6W	2.9W	27.0s	78.3J
Beagleboard	Cortex-A8	1	800MHz	3.6W	4.5W	19.9s	89.5J
Pandaboard	Cortex-A9	2	900MHz	3.2W	4.2W	1.52s	6.38J
Chromebook	Cortex-A15	2	1.7GHz	5.4W	8.1W	1.39s	11.3J

Questions

- Which machine consumes the least amount of energy? (Pandaboard)
- Which machine computes the result fastest? (Chromebook)
- Chromebook is a laptop so also includes display and wi-fi
- Consider a use case with an embedded board taking a picture once every 20 seconds and then performing a

300x300 matrix multiply transform on it. Could all of the boards listed meet this deadline?
No, the Raspberry Pi and Gumstix Overo both take longer than 20s and the Beagleboard is dangerously close.

- Assume a workload where a device takes a picture once a minute then does a 300x300 matrix multiply (as seen in Table 1). The device is idle when not multiplying, but under full load when it is. Over an hour, what is the energy usage of the Chromebook? What is the energy usage of the Gumstix?

Chromebook per minute: $(1.39s \times 8.1W) + (58.61s \times 5.4W) = 327.75J$
Chromebook per hour: 327.75J * 60 $=$ 19.7kJ

Gumstix per minute: $(27s \times 2.9W) + (33s \times 2.6W) = 164.1J$
Gumstix per hour: 164.1J * 60 $=$ 9.8kJ

Pandaboard Power Stats

- Wattsuppro: 2.7W idle, seen up to 5W when busy
- `http://ssvb.github.com/2012/04/10/cpuburn-arm-cortex-a9.html`
- With Neon and CPU burn:

Idle system	550 mA	2.75W
cpuburn-neon	1130 mA	5.65W
cpuburn-1.4a (burnCortexA9.s)	1180 mA	5.90W
ssvb-cpuburn-a9.S	1640 mA	8.2W

Easy ways to reduce Power Usage

DVFS

- Voltage planes – on CMP might share voltage planes so have to scale multiple processors at a time
- DC to DC converter, programmable.
- Phase-Locked Loops. Orders of ms to change. Multiplier of some crystal frequency.
- Senger et al ISCAS 2006 lists some alternatives. Two phase locked loops? High frequency loop and have programmable divider?

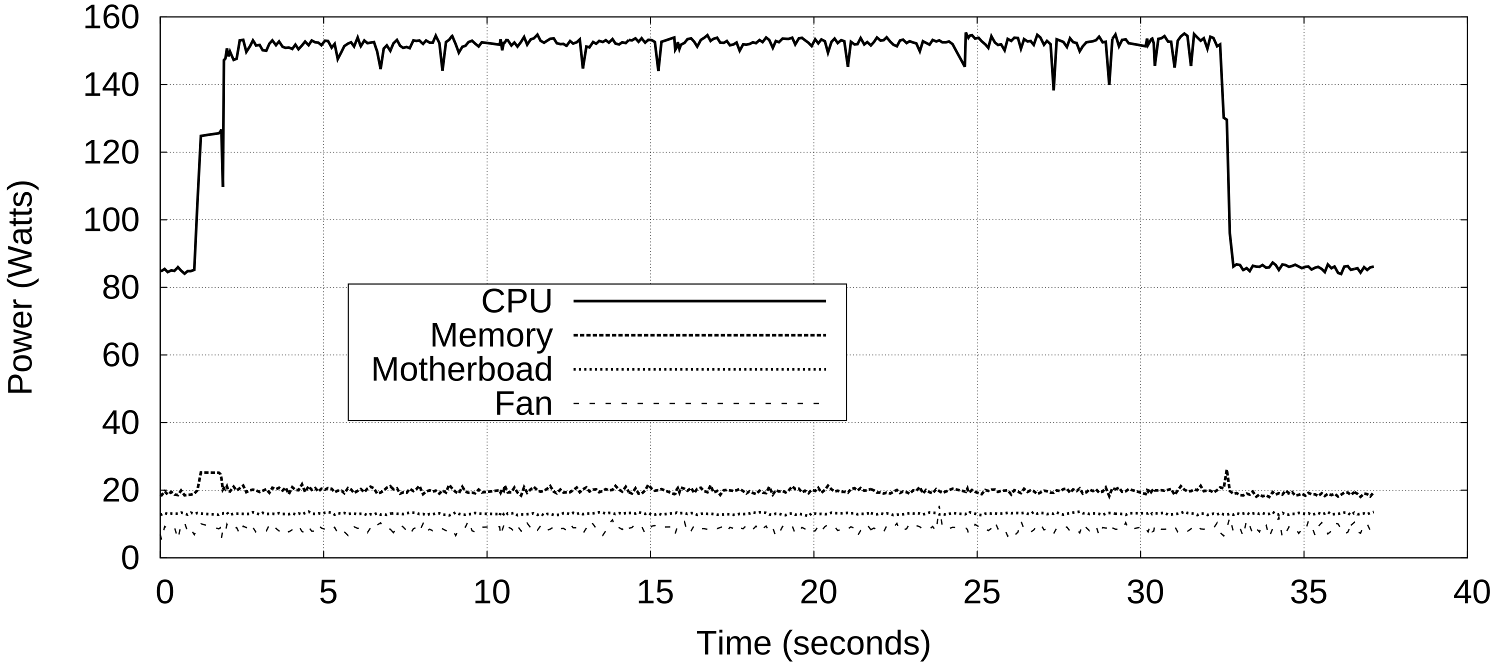
- Often takes time, on order of milliseconds, to switch frequency. Switching voltage can be done with less hassle.

When can we scale CPU down?

- System idle
- System memory or I/O bound
- Poor multi-threaded code (spinning in spin locks)
- Thermal emergency
- User preference (want fans to run less)

Existing Related Work

Plasma/dposv results with Virginia Tech's PowerPack



Powerpack

- Measure at Wall socket: WattsUp, ACPI-enabled power adapter, Data Acquisition System

- Measure all power pins to components (intercept ATX power connector?)

- CPU Power – CPU powered by four 12VDC pins.

- Disk power – measure 12 and 5VDC pins on disk power connecter

- Memory Power – DIMMs powered by four 5VDC pins

- Motherboard Power – 3.3V pins. Claim NIC contribution is minimal, checked by varying workload

- System fans

Shortcomings of current methods

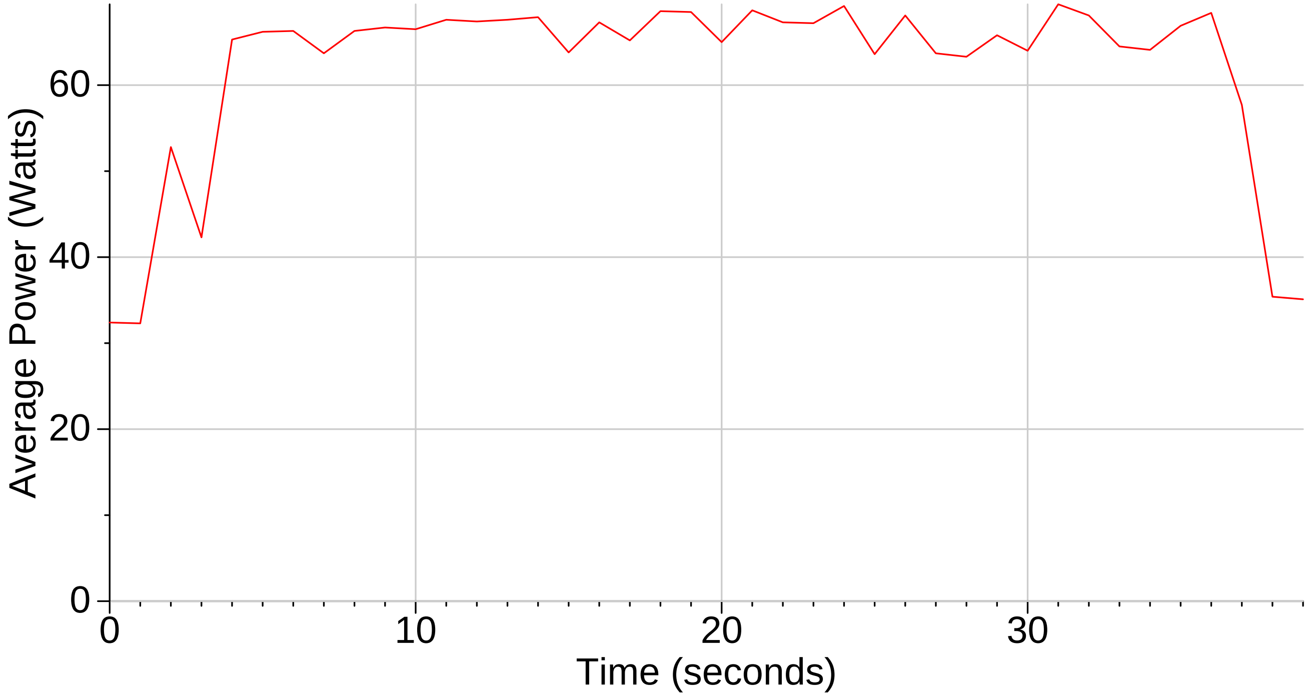
- Each measurement platform has a different interface
- Typically data can only be recorded off-line, to a separate logging machine, and analysis is done after the fact
- Correlating energy/power with other performance metrics can be difficult

Watt's Up Pro Meter

Watt's Up Pro Features

- Can measure 18 different values with 1 second resolution (Watts, Volts, Amps, Watt-hours, etc.)
- Values read over USB
- Joules can be derived from power and time
- Can only measure system-wide

Watt's Up Pro Graph



PLASMA Cholesky Factorization N=10,000 threads=2

Measured on Core2 Laptop

RAPL

- **R**unning **A**verage **P**ower **L**imit
- Part of an infrastructure to allow setting custom per-package hardware enforced power limits
- User Accessible Energy/Power readings are a bonus feature of the interface

How RAPL Works

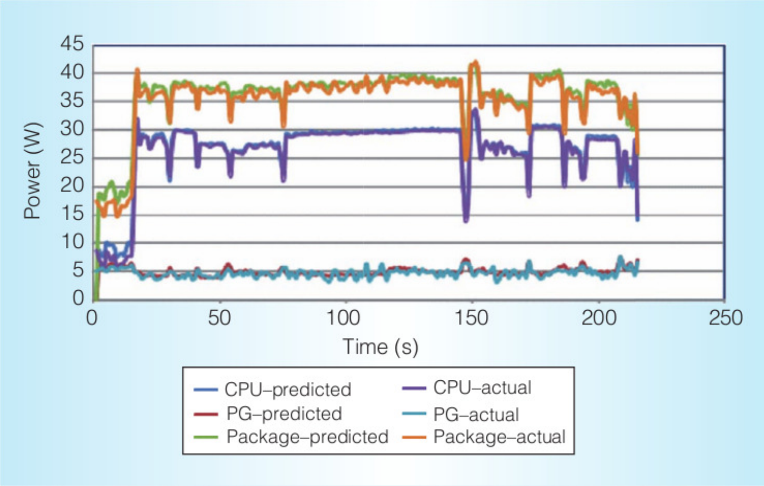
- RAPL is *not* an analog power meter
- RAPL uses a software power model, running on a helper controller on the main chip package
- Energy is estimated using various hardware performance counters, temperature, leakage models and I/O models
- The model is used for CPU throttling and turbo-boost, but the values are also exposed to users via a model-specific register (MSR)

Available RAPL Readings

- `PACKAGE_ENERGY`: total energy used by entire package
- `PP0_ENERGY`: energy used by “power plane 0” which includes all cores and caches
- `PP1_ENERGY`: on original Sandybridge this includes the on-chip Intel GPU
- `DRAM_ENERGY`: on Sandybridge EP this measures DRAM energy usage. It is unclear whether this is just the interface or if it includes all power used by all the DIMMs too

RAPL Measurement Accuracy

- Intel Documentation indicates Energy readings are updated roughly every millisecond (1kHz)
- Rotem at al. show results match actual hardware

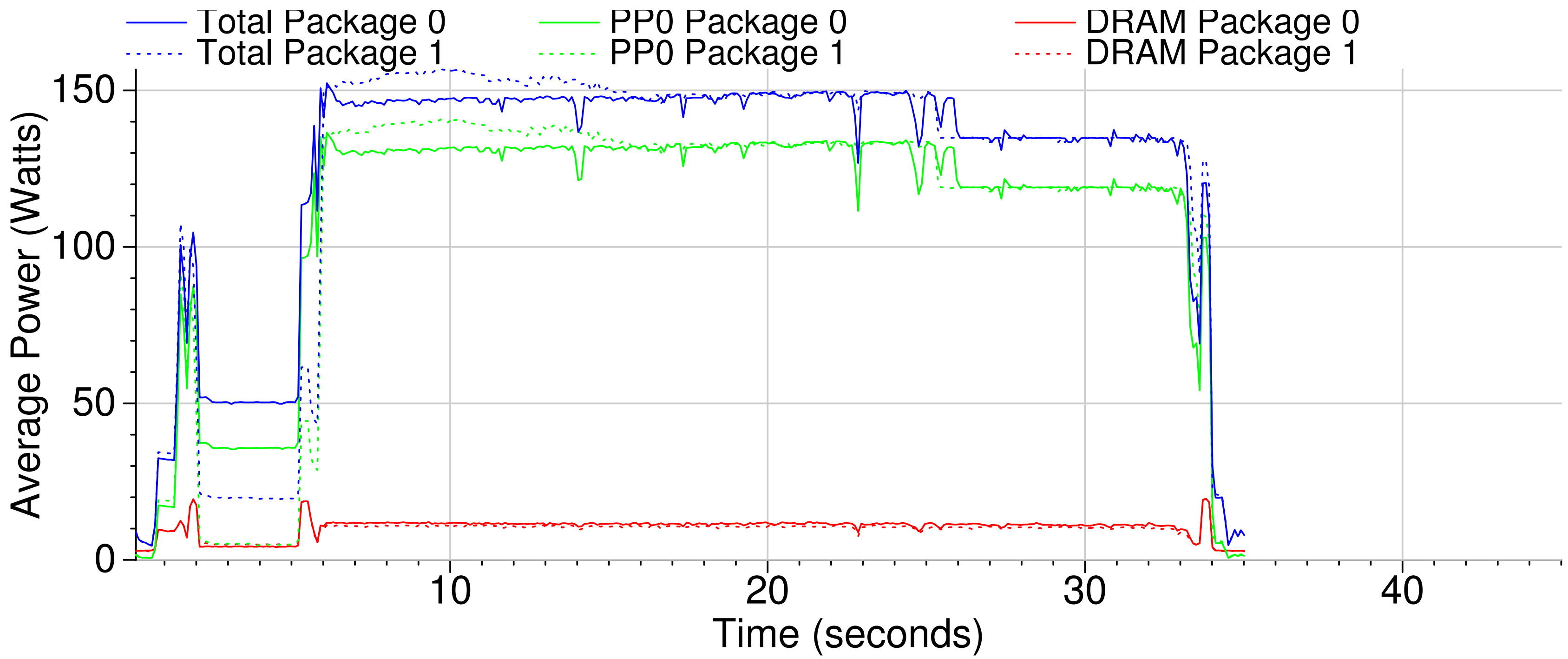


Rotem et al. (IEEE Micro, Mar/Apr 2012)

RAPL Accuracy, Continued

- The hardware also reports minimum measurement quanta. This can vary among processor releases. On our Sandybridge EP machine all Energy measurements are in multiples of 15.2nJ
- Power and Energy can vary between identical packages on a system, even when running identical workloads. It is unclear whether this is due to process variation during manufacturing or else a calibration issue.

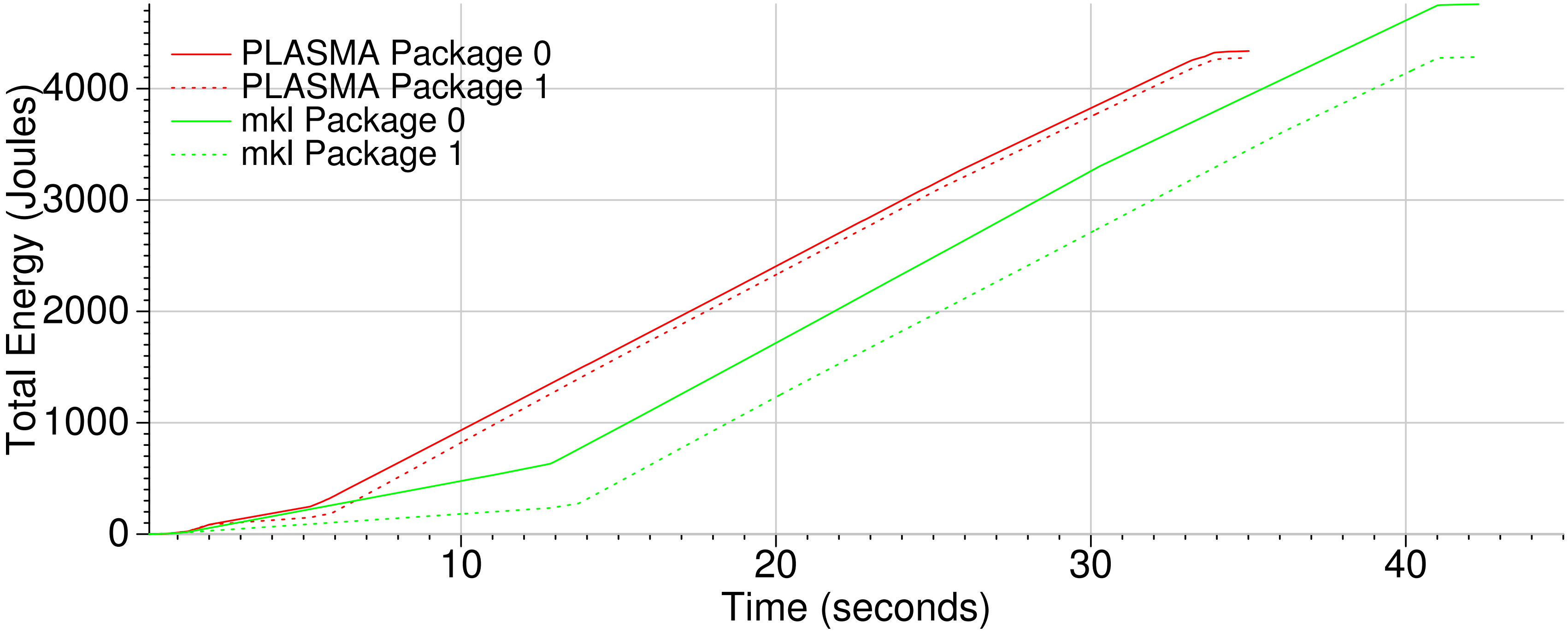
RAPL Power Plot



PLASMA Cholesky Factorization N=30,000 threads=16

Measured on SandyBridge EP

RAPL Energy Plot



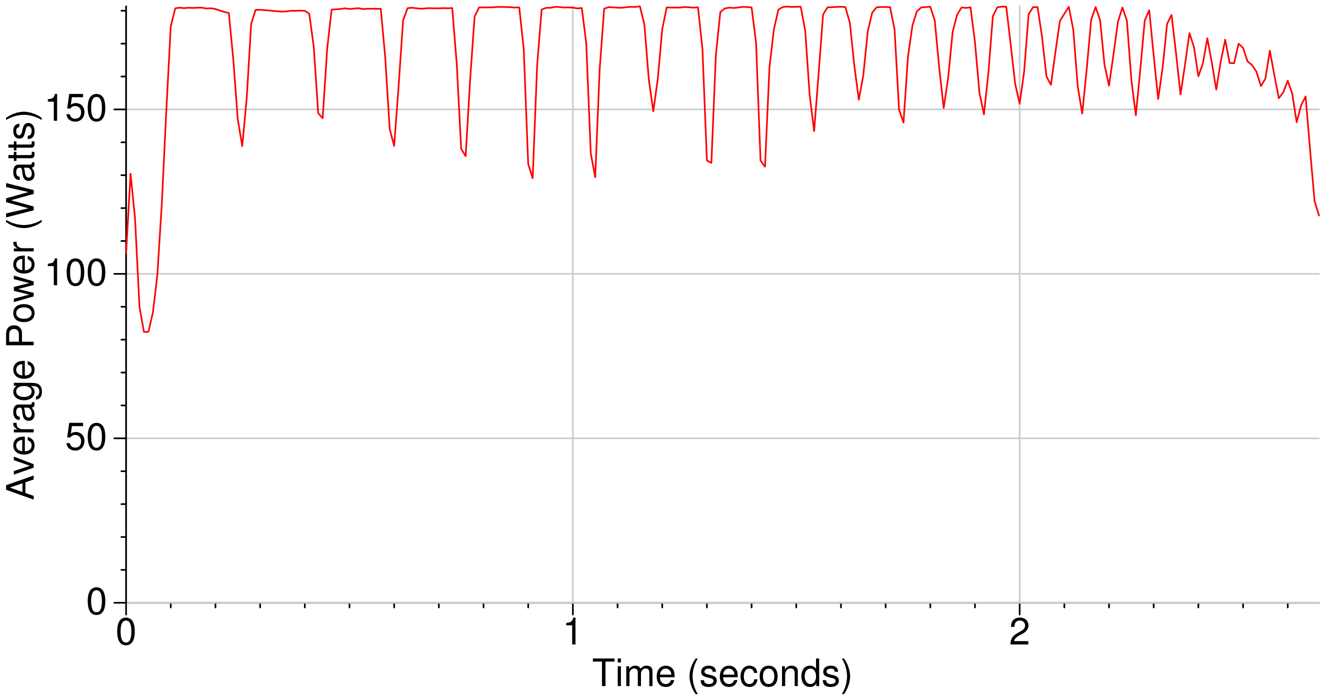
Cholesky Factorization N=30,000 threads=16

Measured on SandyBridge EP

NVML

- Recent NVIDIA GPUs support reading power via the NVIDIA Management Library (**NVML**)
- On Fermi C2075 GPUs it has milliwatt resolution within ±5W and is updated at roughly 60Hz
- The power reported is that for the entire board, including GPU and memory

NVML Power Graph



MAGMA LU 10,000, Nvidia Fermi C2075

AMD Application Power Management

- Recent AMD Family 15h processors also can report "Current Power In Watts" via the Processor Power in the TDP MSR
- Support for this can be provided similar to RAPL
- We just need an Interlagos system where someone gives us the proper read permissions to `/dev/cpu/*/msr`

PowerMon 2

- PowerMon 2 is a custom board from RENCI
- Plugs in-line with ATX power supply.
- Reports results over USB
- 8 channels, 1kHz sample rate
- We have hardware; currently not working

Using RAPL

Listing Events

```
$ perf list
...
  power/energy-cores/                                 [Kernel P
  power/energy-gpu/                                   [Kernel P
  power/energy-pkg/                                   [Kernel P
  power/energy-ram/                                   [Kernel P
...
```

Measuring

```
$ perf stat -a -e power/energy-cores/,power/energy-ram/,instru

 Performance counter stats for 'system wide':

             63.79 Joules power/energy-cores/
              2.34 Joules power/energy-ram/
       21038123875        instructions              #    1.06
       19782762541        cycles

       3.407427702 seconds time elapsed
```

Measuring

- The key is -a which enables system-wide mode (needs root too if not configured as such)
- Why do you need system-wide?
- What does that do to the other metrics?