Initial Validation of DRAM and GPU RAPL Power Measurements

Spencer Desrochers, Chad Paradis, and Vince Weaver

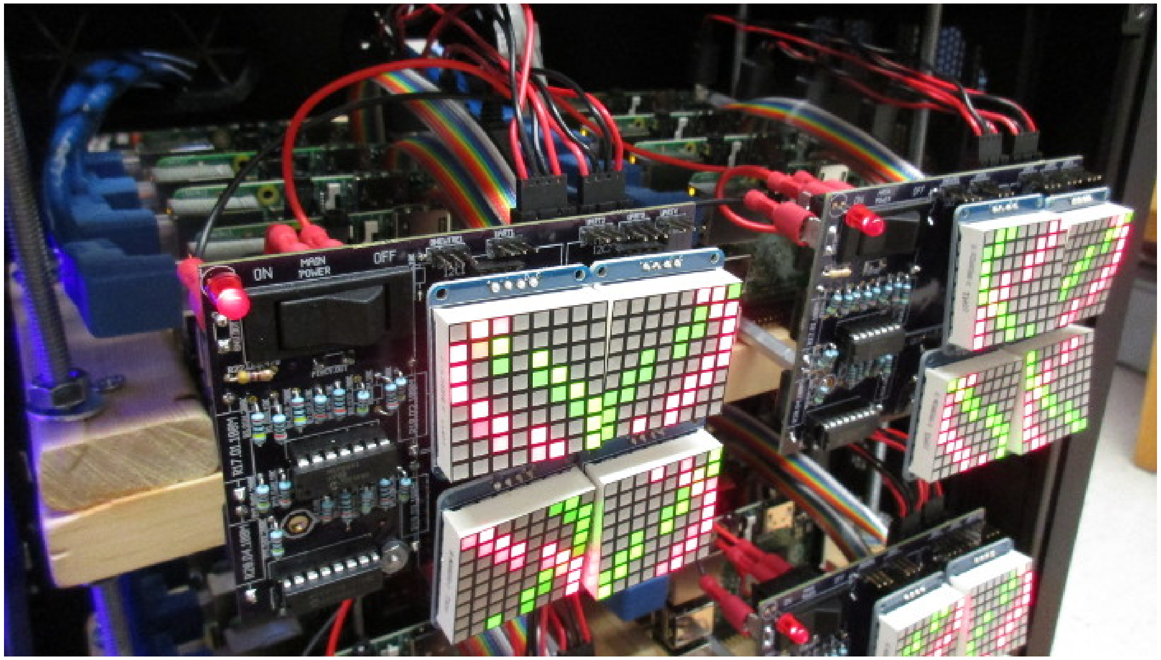
Research Group

University of Maine

{spencer.desrochers,chad.paradis,vincent.weaver}@maine.edu

Workshop on Extreme-scale Programming Tools
16 November 2015

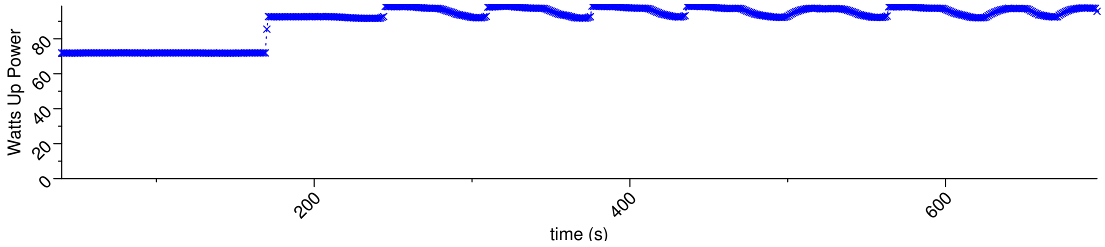
Power Measurement is Hard

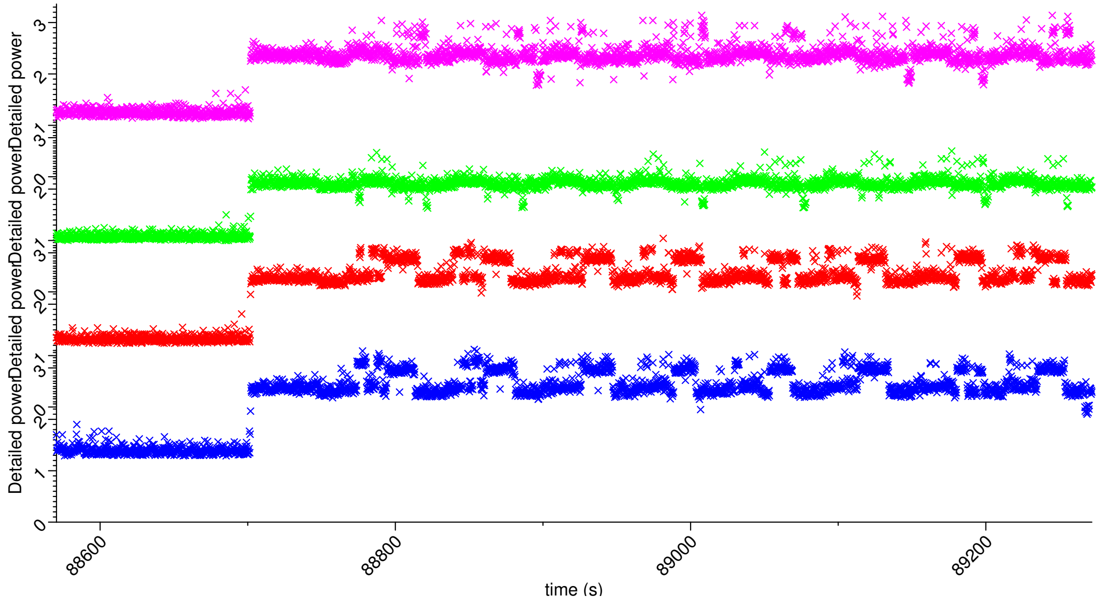


Raspberry Pi 2 Cluster, Instrumented for Detailed Power

Linpack 92W, 15.4GFLOPS = 0.17 GFLOPS/W

Detailed Cluster Power Graphs





Can we get Power Measurements w/o Hardware

- Estimated Power (based on performance counters, etc)
- Intel RAPL (Running Average Power Limit)

Intel RAPL

- Per-socket Energy Readings
- Produced by chip for power-capping, provided to user
- Updated every millisecond, but no timestamp so not sure where in window
- On most chips “estimated” based on an internal model and performance readings.
- Newer server Haswell might actually measure data too.

Intel RAPL

- Hard to get documentation
- Available measures:
  - Total Package
  - PP0 “cores”
  - PP1 “uncore” usually GPU
  - DRAM
- Support Varies, GPU not available on servers, until Haswell DRAM not available on desktop.
- Validated? Can we trust them?

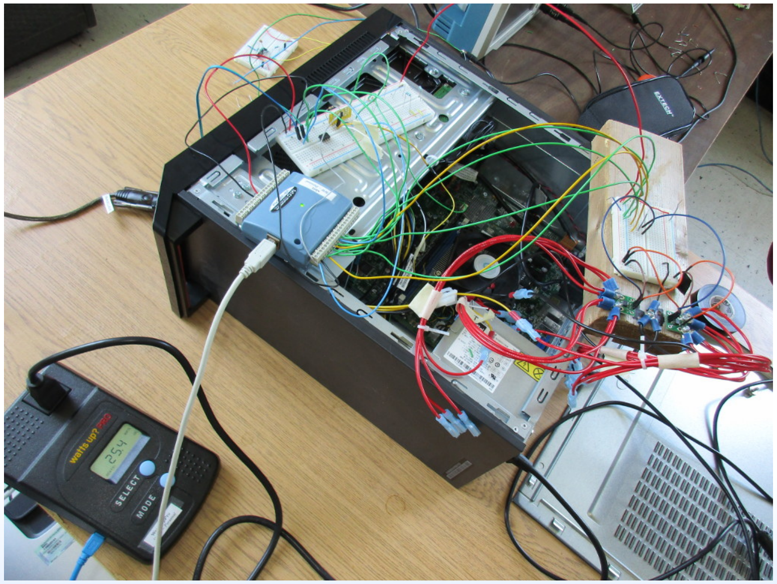
RAPL Validation

- CPU has been looked at various times, although results are often for microbenchmarks, or are small plots tucked away in long articles, etc.
  Hähnel et al, Rotem et al, Demmel and Gearhart, Hackenberg et al, Mazous et al.

- GPU and DRAM have not been validated at all.

So it’s back to hardware instrumentation

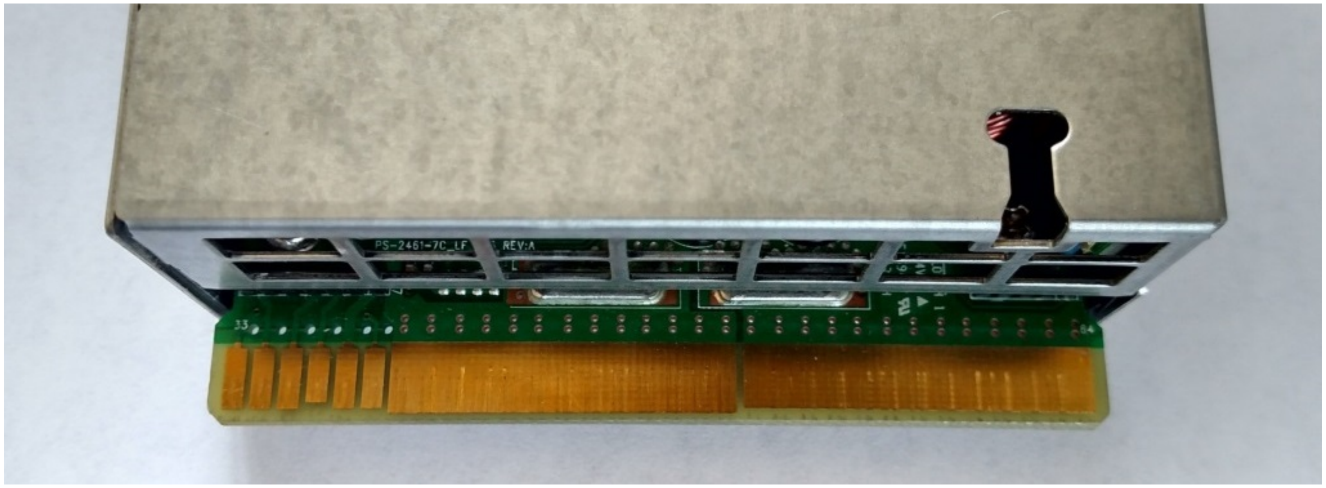
Goal is a small, cheap, drop-in power measurement device to put in all nodes of server. In practice this ends up being much harder than you could hope.

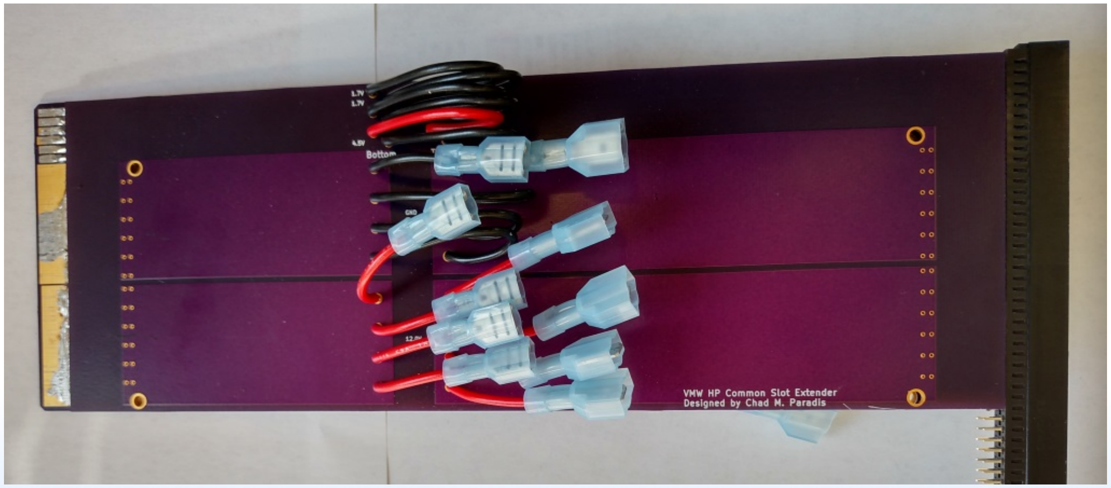


Full System Power Measurement

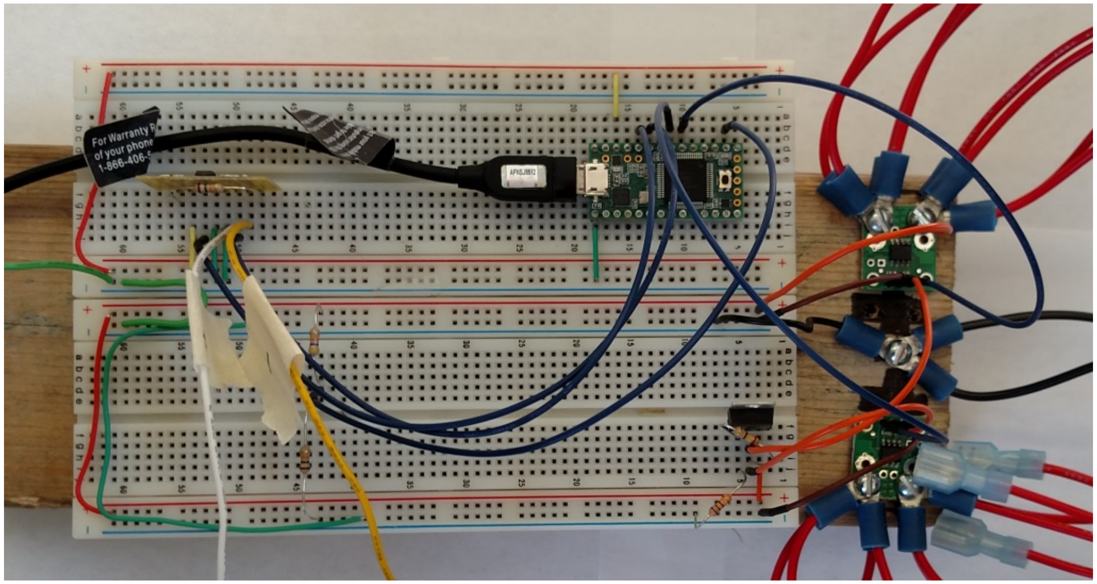
- WattsUpPro?
  Wall outlet measurement (low frequency, 1Hz)
- IPMI (also low frequency)
- Custom
  - Measure *every* ATX pin. There are 24+
  - Server hardware has custom power connectors

HP Connector in our SNB-EP system





Measurement Setup

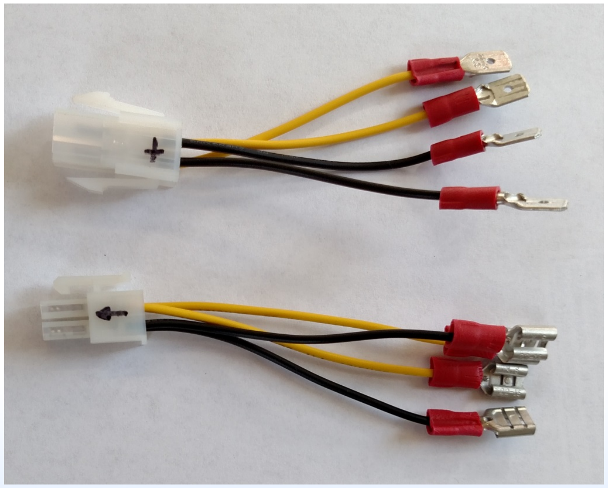


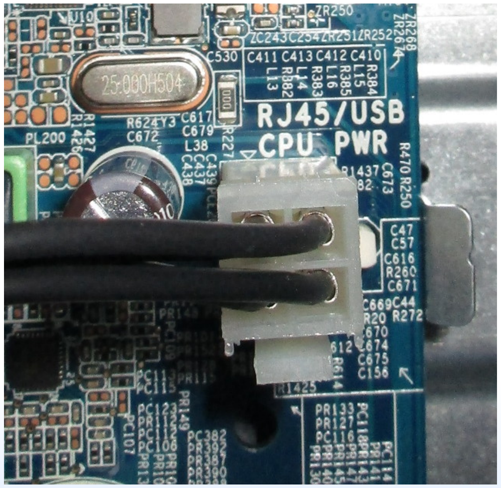
- Sense Resistor or Hall Effect Sensor
- Instrumentation Amplifier
- A/D Converter

Complicated. Also easy to ruin your system.

CPU Power Measurement

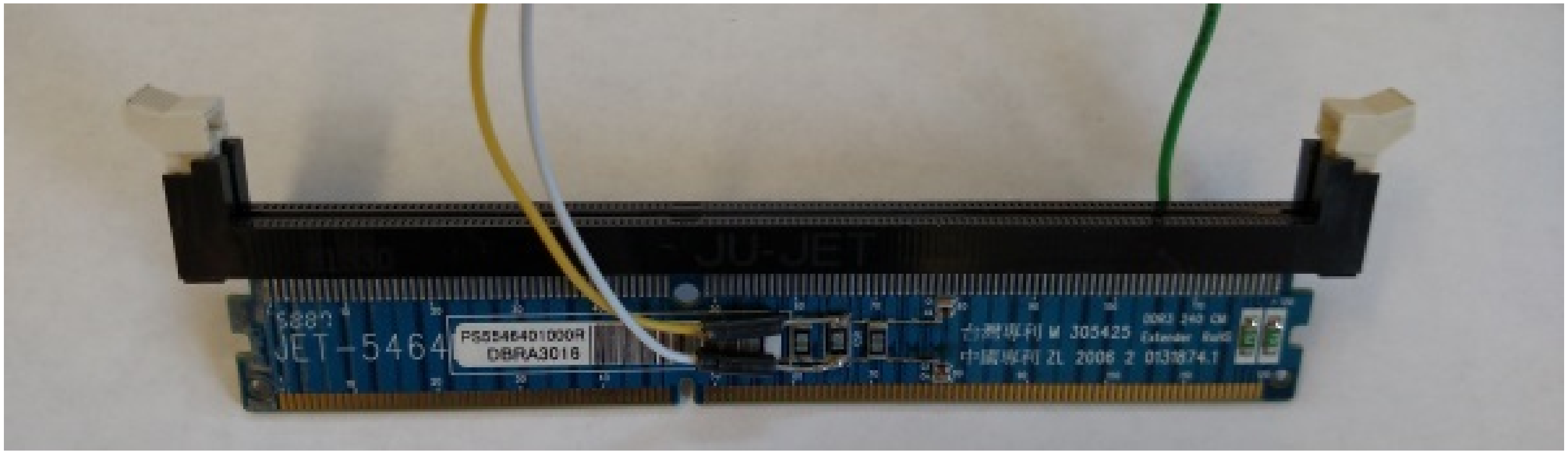
- Hard to do when power converters on motherboard, not way to intercept.
- Desktop systems have the P4 connector. Most papers look at this. Is it CPU only?





DRAM Measurement

- Sense resistor on DIMM extender

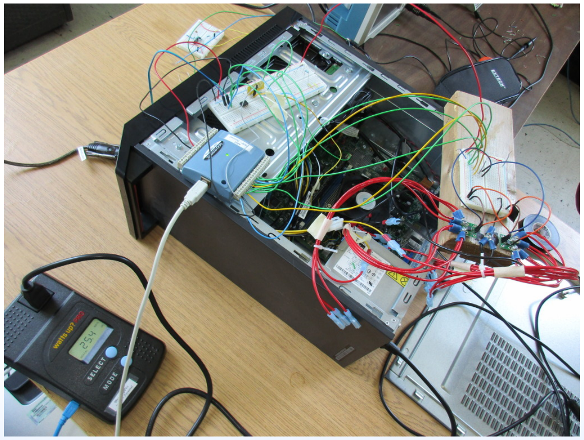


GPU Power Measurement

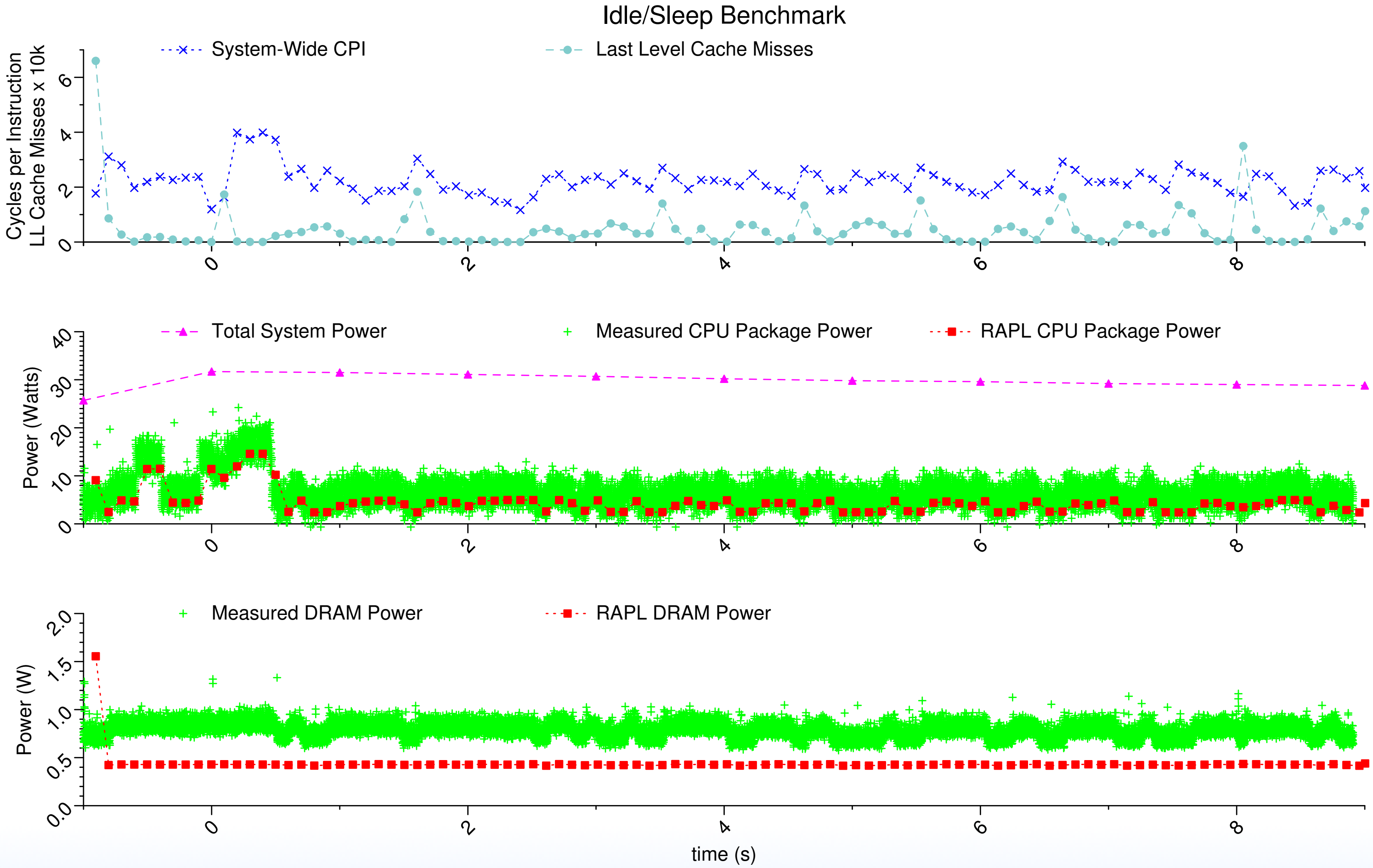
- Discrete GPUs can use a PCIe extender
- Some GPUs (NVidia) provide this info.
- To our knowledge Intel GPUs do not, although there is preliminary support for other hardware counters.

Our Validation Setup

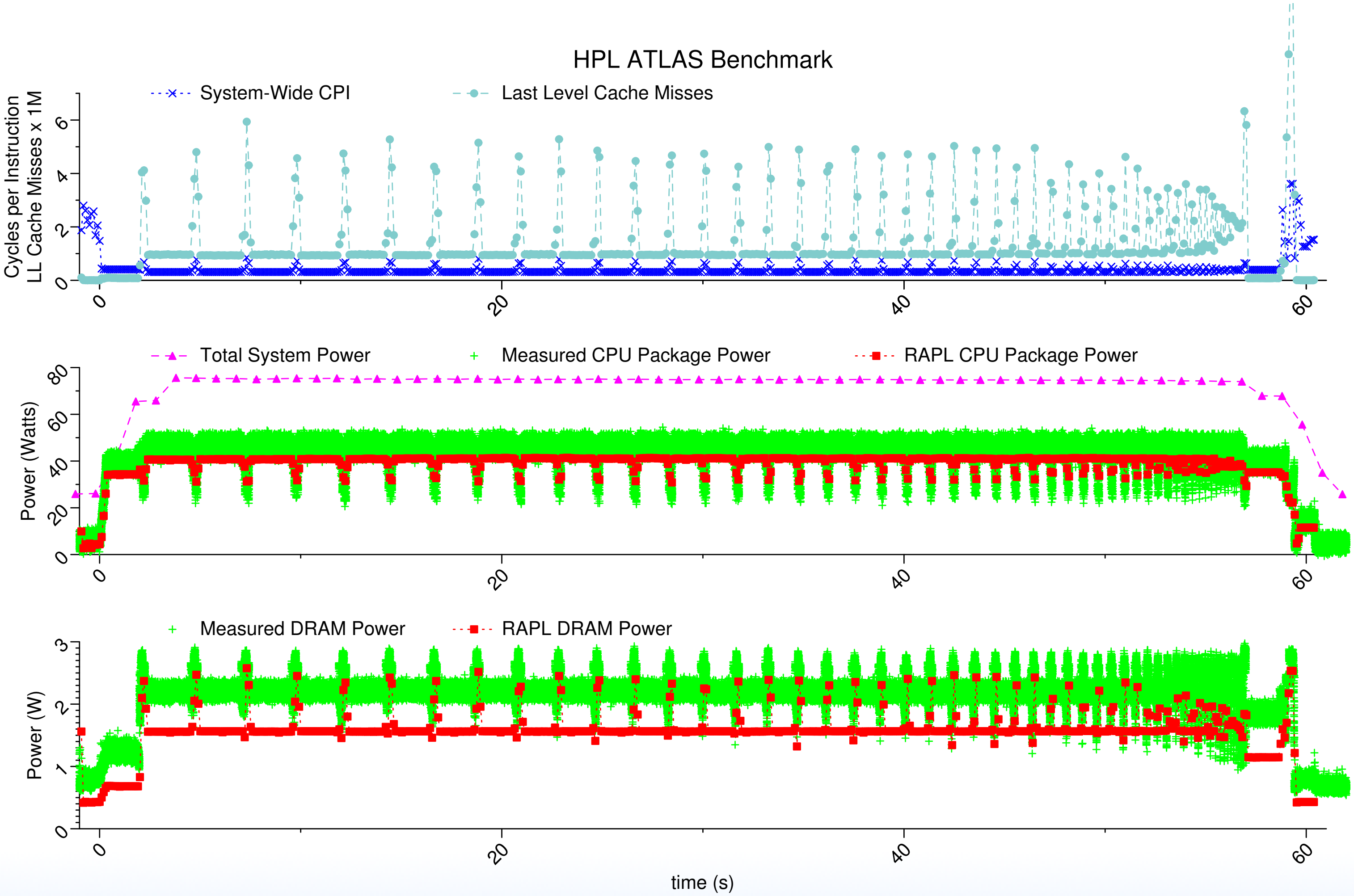
- Lenovo Thinkcentre 4-core 2.9GHz i5-4570S (S means low power 65W envelope)
- Integrated Intel HD Graphics GPU 4600
- 4GB DDR3 RAM
- Debian Jessie, Linux 4.1.5 / 4.0.5 (patched for OpenCL)



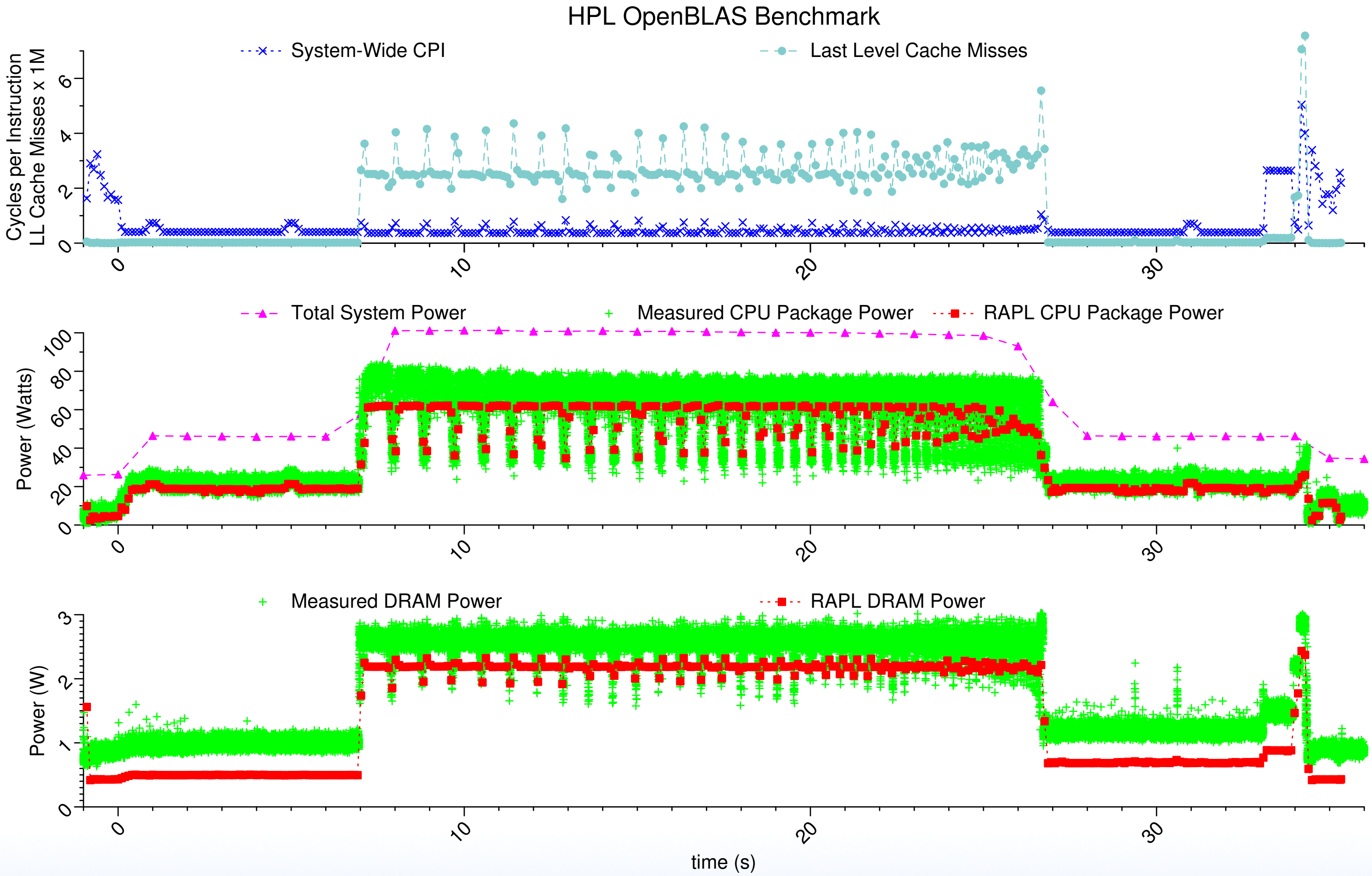
Idle


Idle/Sleep Benchmark
System-Wide CPI
Last Level Cache Misses
Cycles per Instruction
LL Cache Misses x 10k
0
2
4
6
8
Total System Power
Measured CPU Package Power
RAPL CPU Package Power
Power (Watts)
10
20
30
40
Measured DRAM Power
RAPL DRAM Power
Power (W)
0.0
0.5
1.0
1.5
2.0
time (s)

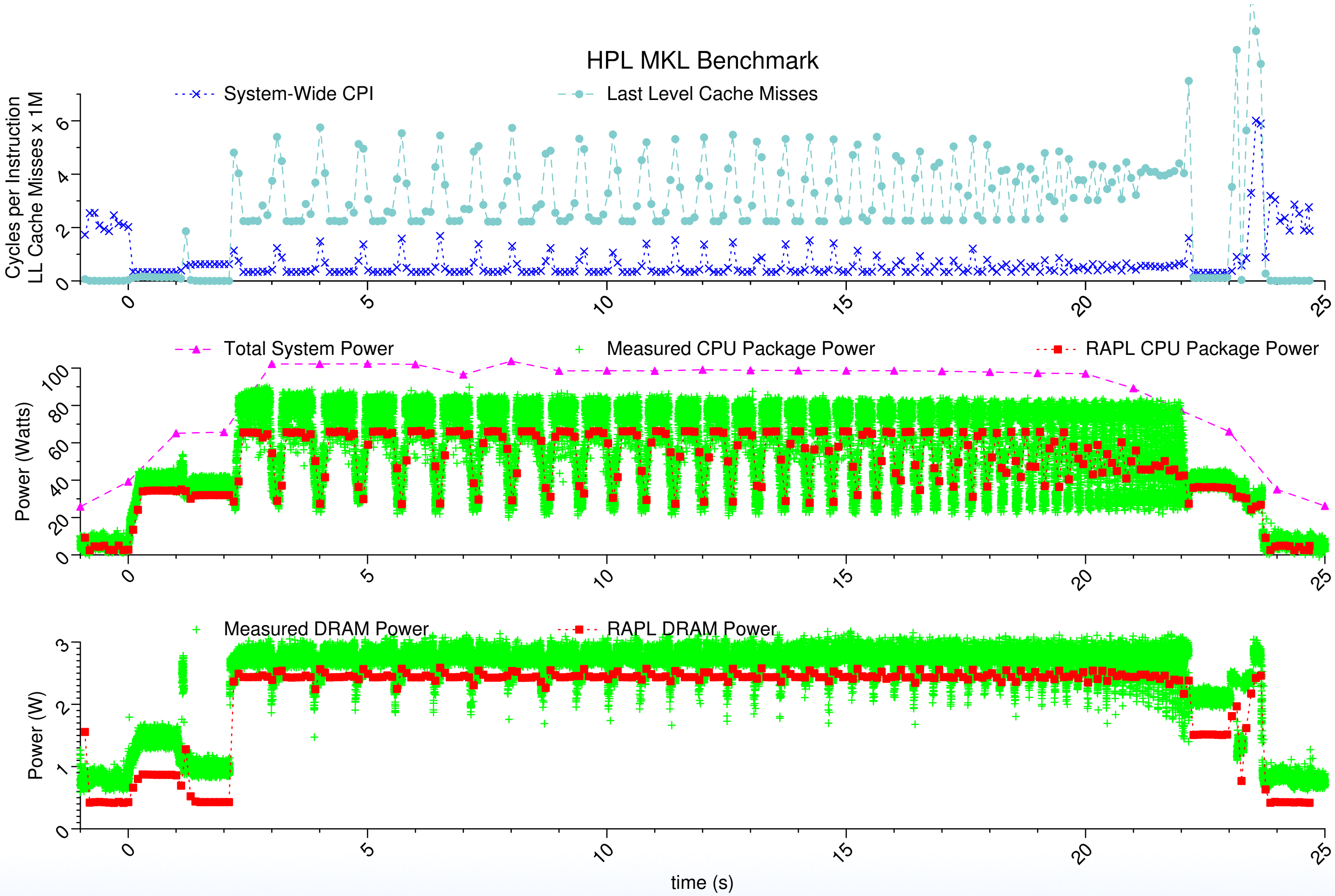
Linpack ATLAS



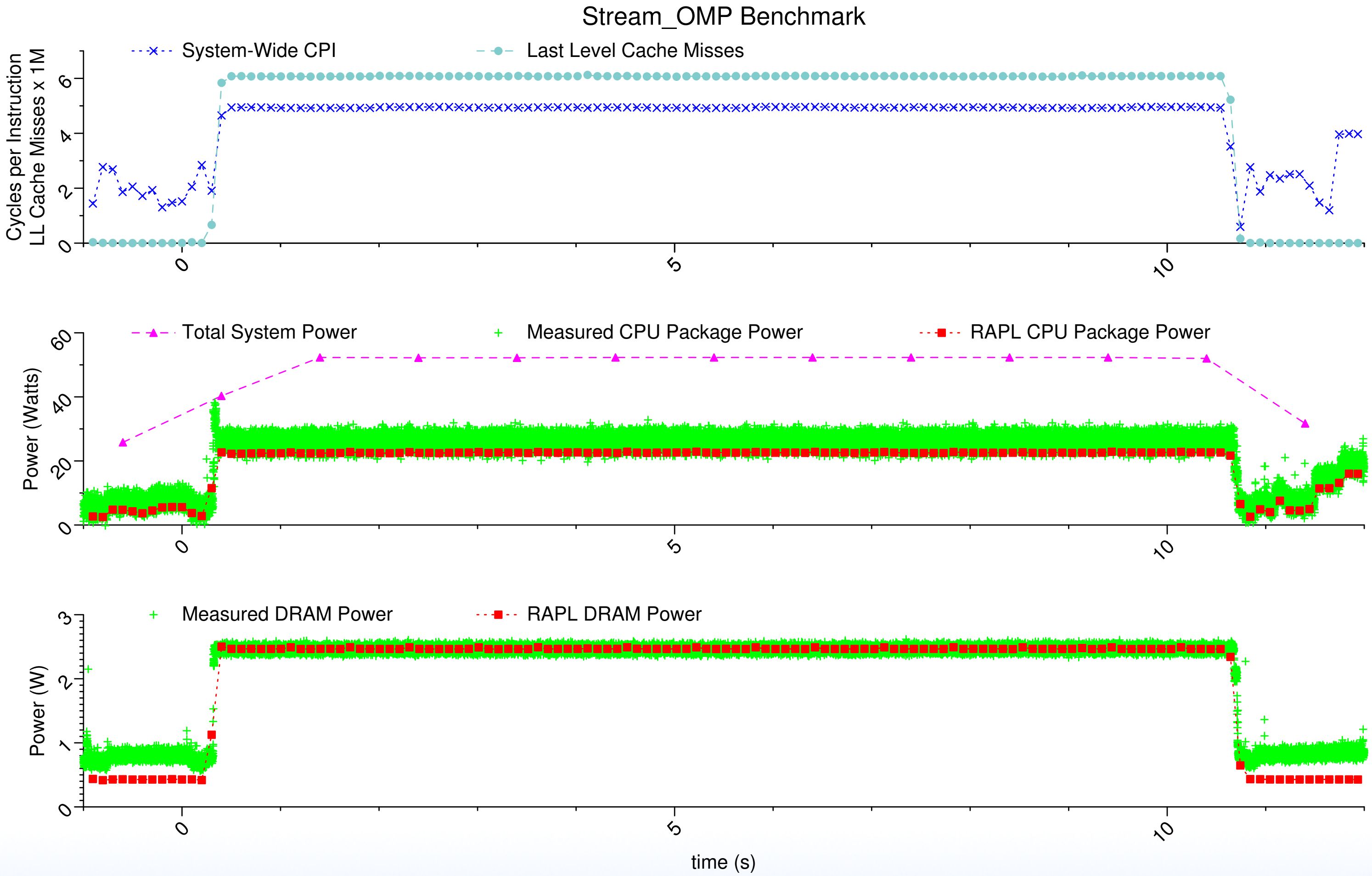
Linpack OpenBLAS



Linpack MKL

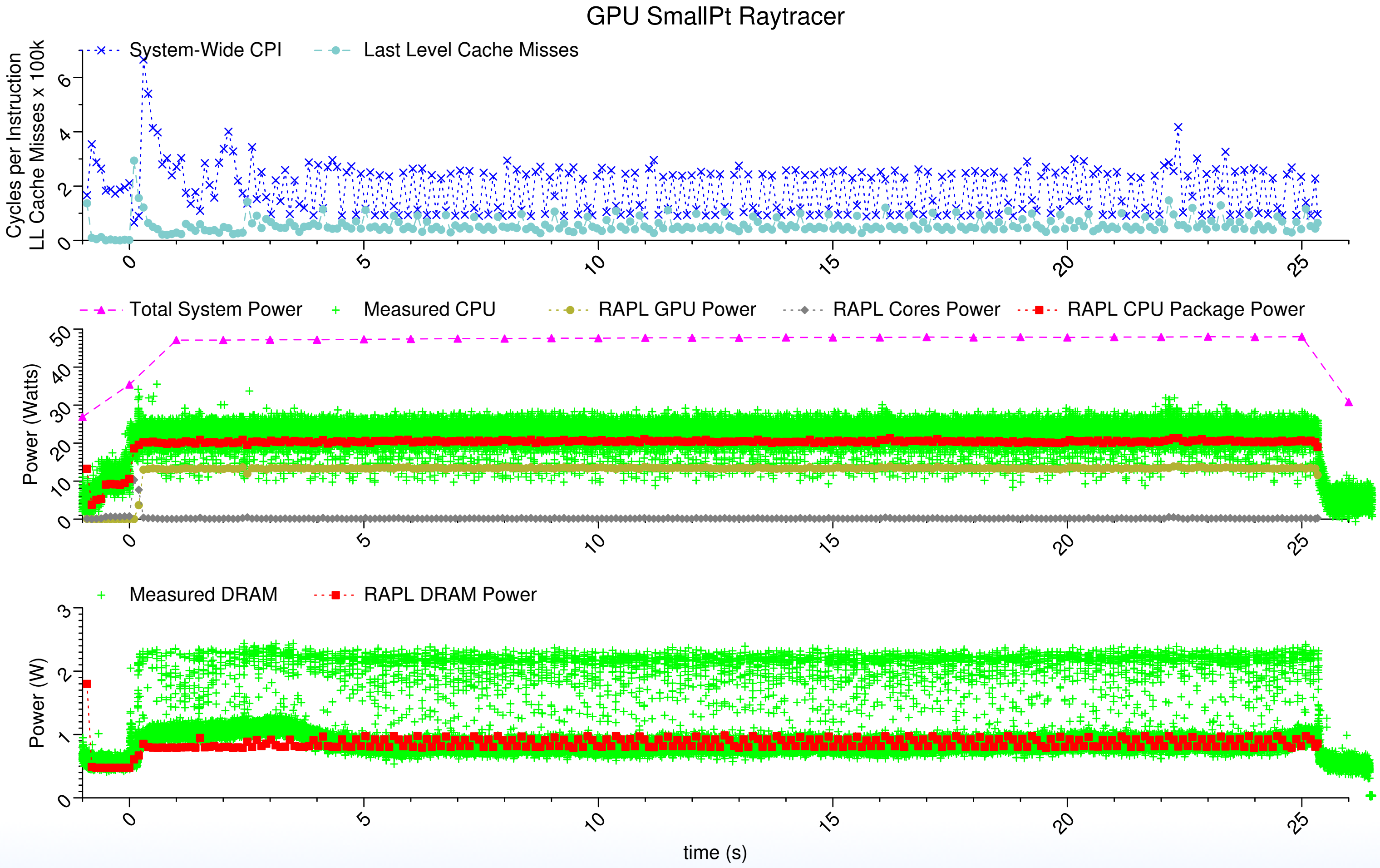

HPL MKL Benchmark
System-Wide CPI
Last Level Cache Misses
Cycles per Instruction
LL Cache Misses x 1M
0
2
4
6
0
5
10
15
20
25
Total System Power
Measured CPU Package Power
RAPL CPU Package Power
Power (Watts)
0
20
40
60
80
100
0
5
10
15
20
25
Measured DRAM Power
RAPL DRAM Power
Power (W)
0
1
2
3
0
5
10
15
20
25
time (s)

STREAM

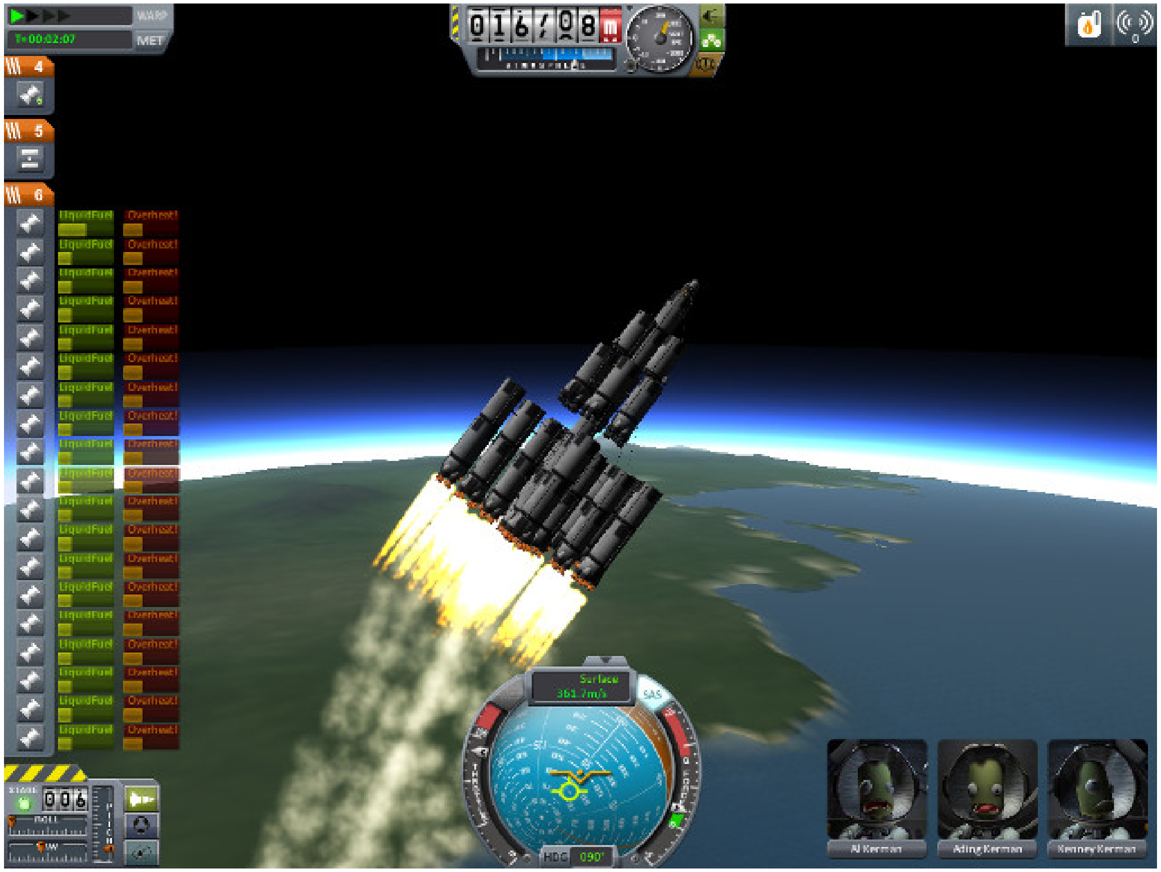


Stream_OMP Benchmark
System-Wide CPI
Last Level Cache Misses
Cycles per Instruction
LL Cache Misses x 1M
0
2
4
6
Total System Power
Measured CPU Package Power
RAPL CPU Package Power
Power (Watts)
0
20
40
60
Measured DRAM Power
RAPL DRAM Power
Power (W)
0
1
2
3
0
5
10
time (s)

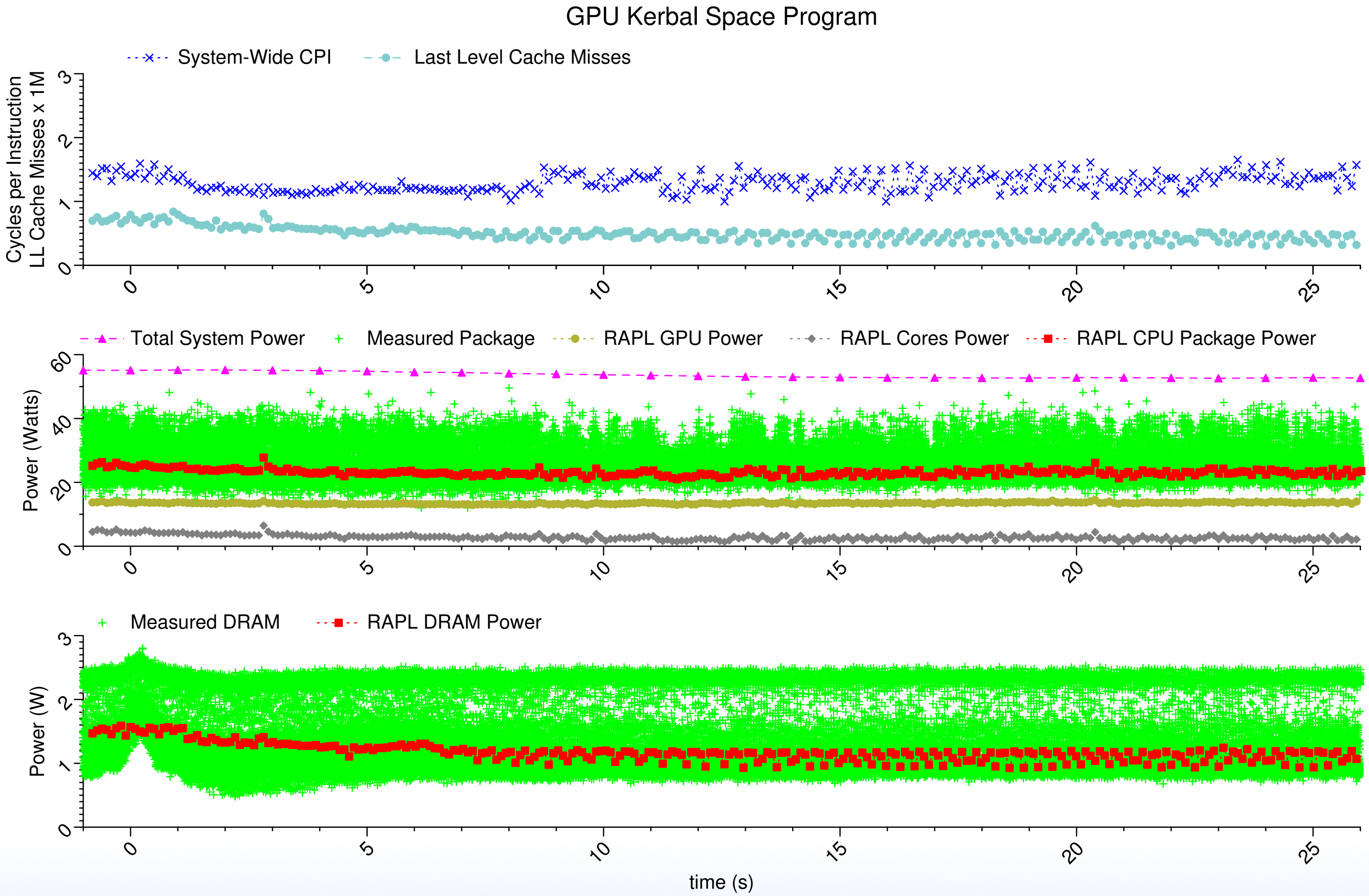
GPU: OpenCL SmallPt Raytracer



GPU: Kerbal Space Program



GPU: Kerbal Space Program Results



Validation? DRAM GFLOPS/W

(Top actual, bottom RAPL)

Benchmark	Time (s)	$GFlops$	Energy (J)	Average Power (W)	$\frac{GFlops}{W}$
Sleep	9.7	—	7.7	0.79	—
			4.2	0.43	—
STREAM	12.7	—	27.5	2.16	—
			26.6	2.09	—
HPL-ATLAS	61.2	40.9	131.3	2.15	19.0
			96.2	1.57	26.1
HPL-OpenBLAS	36.1	113.9	69.0	1.91	59.6
			53.2	1.47	77.5
HPL-mkl	25.5	106.8	62.0	2.43	44.0
			53.9	2.11	50.6
OpenCL-raytrace	26.1	—	24.8	0.95	—
			22.3	0.85	—
OpenGL-kerbal	26.7	—	36.9	1.38	—
			31.2	1.17	—

Challenges with high-speed RAPL

- Reading the MSR values directly (`/dev/msr`):
  Reading in a tight loop, trying to get 100Hz or better.
  Problem in userspace, the process can get scheduled out.

- Using the `perf` tool:
  It has to be hacked to let you measure more than 10Hz.
  Gathering at 100Hz causes 0.5W increase in power.

Using advanced methods

- Using perf_event sampling interface:
  A high-speed ring buffer calling to userspace when full.
  Sampling interface fails due to interrupt load.
  Kernel will throttle if it detects sampling using more than 25% of CPU on performance interrupts (1KHz reads triggered this).
  Use of NMI interrupts by interface has a lot of overhead.

Future Work: Why Not a Server Machine?

- Working on it.
- Cannot find DDR4 DIMM extender with power sensor
- Servers have much harder to instrument power connectors

Questions?

`vincent.weaver@maine.edu`

More details can be found in our tech report:

`http://web.eece.maine.edu/~vweaver/papers/tech_reports/2015_dram_rapl_tr.pdf`