A Validation of DRAM RAPL Power Measurements

Spencer Desrochers
University of Maine
spencer.desrochers@maine.edu

Chad Paradis
Garmin Ltd.
cparadis6191@gmail.com

Vincent M. Weaver
University of Maine
vincent.weaver@maine.edu

ABSTRACT

Recent Intel processors support the Running Average Power Level (RAPL) interface, which among other things provides estimated energy measurements for the CPUs, integrated GPU, and DRAM. These measurements are easily accessible by the user, and can be gathered by a wide variety of tools, including the Linux perf_event interface. This allows unprecedented easy access to energy information when designing and optimizing energy-aware code.

While greatly useful, on most systems these RAPL measurements are estimated values, generated on the fly by an on-chip energy model. The values are not documented well, and the results (especially the DRAM results) have undergone only limited validation.

We validate the DRAM RAPL results on both desktop and server Haswell machines, with multiple types of DDR3 and DDR4 memory. We instrument the hardware to gather actual power measurements and compare them to the RAPL values returned via Linux perf_event. We describe the many challenges encountered when instrumenting systems for detailed power measurement.

We find that the RAPL results match overall energy and power trends, usually by a constant power offset. The results match best when the DRAM is being heavily utilized, but do not match as well in cases where the system is idle, or when an integrated GPU is using the memory.

We also verify that Haswell server machines produce more accurate results, as they include actual power measurements gathered through the integrated voltage regulator.

CCS Concepts

•**Hardware** → **Semiconductor memory; Chip-level power issues;** *Hardware validation;* •**Computer systems organization** → *Architectures;*

Keywords

DRAM Power; DRAM Energy; RAPL

MEMSYS 2016 October 3–6, 2016, Washington, DC, USA

ISBN 978-1-4503-4305-3...$15.00

DOI: http://dx.doi.org/10.1145/2989081.2989088

1. INTRODUCTION

In 2011 Intel introduced the Running Average Power Level (RAPL) interface as part of the SandyBridge microarchitecture [6]. This is an advanced powercapping infrastructure which allows the user (or operating system) to specify maximum power limits: the processor can run at the highest possible speed while automatically throttling back to stay within power or thermal bounds. In order to respect these power limits, a processor must be aware of its current power usage. It is often impractical to measure power directly, so instead the processor typically estimates these values using a power model based on performance counters, temperature, and other inputs. As a bonus feature, the results of the power model are made available to the user via a model-specific register (MSR) and can be used when characterizing workloads. RAPL support varies across CPU models, but in general energy measurements are available for the total processor package, for the aggregate total of all cores, for the DRAM, and for the integrated GPU.

RAPL energy results provide much needed feedback when optimizing code for the diverse range of modern computing systems. At the low end, energy use is a key factor in providing long battery life in cell phones and other mobile devices. At the high end (such as in large supercomputing environments) a system may have millions of cores and saving just 1 Watt per core can amount to megawatts (and millions of dollars) in savings.

Usually CPUs are the focus of power optimization as they make up the largest single component of a system's energy budget. Next in line after the processors is the DRAM, which can also consume a large proportion of overall power. We find that on a 16-core Haswell-EP server with 80GB of DDR4 RAM running the Linpack benchmark the CPU cores average 130W and the RAM 13W. In this case the RAM consumes more power than an individual core.

Gathering *actual* power and energy results is difficult in modern systems. Most devices are not instrumented for power measurement, and adding suitable interfaces usually involves intrusive modifications to the system's power distribution network. Some parts of a system, such as the CPU, can be particularly difficult to instrument due to being soldered to the motherboard with the numerous power traces being buried inside a multi-layer circuit board. The difficulties found when attempting actual readings are what make the RAPL interface such an attractive alternative for power measurement.

Before RAPL results can be trusted the interface must be properly validated. There has been initial work in validat-

ing RAPL [8, 11, 28] but this has focused on CPU power measurements. In this work we not only look at CPU values but also those involving the DRAM and GPU.

We extensively instrument various Haswell desktop and server machines for power measurement, detailing the many difficulties encountered along the way. We then run various benchmarks and compare the RAPL results to the actual measurements. We find that the accuracy varies among machines and DRAM manufacturers. RAPL DRAM measurements match more closely when the memory system is under load, and the results are less accurate when memory is idle or not being used by the CPU (such as when the integrated GPU is using the memory bus). We also verify that the server machine with on-board voltage regulation can provide RAPL readings based on actual power measurements rather than the estimates found on desktop systems.

2. RAPL BACKGROUND

The RAPL interface is not well documented; the MSR interface is described in Chapter 14.3 of the Intel Volume3b Documentation [19] but most low-level details are not covered at all.

RAPL provides estimated per-package energy estimates, meaning the totals are at the socket (not per-core) level. Energy measurements available include: total package, Power Plane 0 (PP0) which is the aggregate total of all cores, Power Plane 1 (PP1) which is an implementation-defined part of the uncore (usually the GPU), and the DRAM. Measurement availability varies by chip model; DRAM measurements were originally only available for server systems but starting with Haswell are available on all processors. Similarly, GPU measurements are not available on server versions of the processors.

The estimated energy is updated roughly at 1 millisecond (1kHz) intervals, but there is no timestamp provided. This makes it difficult to get useful results at small timescales [11] as you have no indication of when the current measurement began and thus mapping the reading to executing code is nearly impossible. Work has been done to mitigate this by carefully monitoring when updates happen and starting measurements at the transition [13].

The Haswell-EP hardware has integrated voltage regulators and more advanced RAPL hardware that includes RAPL "DRAM Mode 1" which include actual measurement (rather than the "DRAM Mode 0" pure estimation found on earlier processors) [12].

When reading the MSRs directly it is important to use the proper scaling factor for the results. The minimal energy increment can vary; on regular Haswell this can be read from a register (it is roughly 61μJ) but on Haswell-EP the DRAM increment is documented elsewhere [18] as being fixed at 15μJ. Older versions of the Linux kernel use the wrong scaling value but this has been fixed as of Linux 4.1.

Reading the RAPL MSR registers involves ring-0 access, which is usually handled by an operating system device driver. Linux provides at least three ways to access the values: raw MSR access (via the `/dev/msr` interface), the perf_event subsystem, and the powercapping interface visible under `/sys/class/powercap/intel-rapl/`. For various security reasons (including the system-wide nature of the measurements) reading the values usually requires root permissions (in theory an attacker could use the power metrics to spy on what other users of the CPU are doing).

3. RELATED WORK

Various groups have previously investigated the accuracy of the RAPL counters against actual hardware.

3.1 CPU RAPL Validation

Hähnel et al. [13] investigate CPU RAPL results on a Sandybridge processor and find overall patterns match actual hardware, but there is an offset in the power. They provide only a single graph of a synthetic benchmark in their validation.

Rotem et al. [28] introduced the RAPL interface and provide some low level details on the interface. The only validation provided is one graph of an unspecified benchmark showing a close match of RAPL CPU and package measurements to actual measurements.

Dongarra et al. [8] compare RAPL measurements on a Sandybridge machine using PAPI to actual measurements found using PowerPack [10] on a completely different (non-Sandybridge) microarchitecture. They use LU factorization as a workload.

Demmel and Gearhart [7] validate two Sandybridge machines against RAPL Package with the STREAM [23] benchmark and a full-system wall power meter. They do not look at the DRAM measurements.

Hackenberg et al. [11] validate RAPL (and the similar AMD APM interface) on a variety of Sandybridge machines. They measure at the wall outlet, as well as at the CPU and motherboard level by intercepting the ATX power connectors. They find that RAPL accuracy can vary by workload, and that it can be confused by HyperThreading.

Mazous, Pradelle and Jalby [22] apply statistical validation to RAPL results compared to full system wall outlet measurements on IvyBridge and Sandybridge. They found some anomalies with the RAPL results when only exercising a single core or when operating at maximum frequency.

3.2 DRAM RAPL Validation

The RAPL DRAM interface was first described by David et al. [6]. While concentrating on the power-capping interface, they also describe in detail the underlying power model which presumably is similar to that found in modern Intel chips. A parametric model is built using genetic algorithms based on various inputs and the weights are calibrated by the BIOS as boot. They validate against real hardware using a DIMM riser card and a data acquisition board sampling at 100Hz. They found accuracy of 1% when using a Nehalem server system and a DDR3 1333 4GB memory module.

Khanna et al. [20] describe the weights used in RAPL DRAM measurements. They measure actual DRAM results using a riser with a 5mΩ sense resistor sampled at 100Hz. They find RAPL results within 2.3% of actual measurements.

Hackenberg et al. [12] investigate RAPL on Haswell-EP processors. They find that the DRAM + Package RAPL results correlate well with total system power readings, but do not measure the individual actual power results for CPU or DRAM.

Our measurements that compare DRAM RAPL to actual results show much more variation than some of the previous work. These previous works do not always describe their methodology in sufficient detail to know why their results match more closely than ours.

4. EXPERIMENTAL SETUP

We run experiments on three different Haswell-class machines as described in Table 1. The first desktop, "haswell-i5", is a Lenovo ThinkCentre with a 4-core 2.9GHz i5-4570S Haswell CPU. The "S" series of processors denotes a low-power 65W thermal design envelope. The second desktop, "haswell-i7", is a Lenovo ThinkCentre with a 4-core 3.4GHz i7-4770 processor. Both desktop machines have integrated Intel 4600 HD Graphics. The server machine, "haswell-ep", is a HP ProLiant DL360 Gen9 with two CPU packages totalling 16 cores.

Three different types of DDR3 DRAM and two types of DDR4 DRAM are tested, as shown in Table 2. The relative speeds of the various machine/DRAM combinations while running the STREAM benchmark are shown in Table 3.

The machines are running the Jessie Debian Linux distribution. The desktop machines run Linux 4.1.5 for all DRAM measurements, and a specially patched 4.0.5 kernel for the GPGPU measurements. The server machine is running Linux 4.6-rc2.

4.1 Hardware Measurement Setup

System-wide power is measured using a WattsUpPro? [9] device as seen in Figure 3. Total wall outlet powered is measured at the maximum supported frequency of 1Hz.

CPU power is measured by intercepting the power at the 12V "P4" 4-pin auxiliary ATX connector. This pin primarily powers the CPU [16] but may also power an unknown amount of other parts of the motherboard. This is typically how previous work [5, 21, 27] has measured CPU power, although on another Haswell system we own the connector is specifically marked as "CPU/NIC/USB" so it is possible that these other hardware components are interfering with pure CPU measurements. Due to potentially high currents involved (in the tens of Amps) an ACS715 Hall effect sensor [2] is used for measurement rather than a sense resistor. The Hall effect sensor provides a voltage output that is proportional to the current flowing through the device.

The DDR3 DRAM is instrumented using a DIMM extender as shown in Figure 2. A DDR3 DIMM has five separate power supplies: V_{DD} (the main supply), V_{DDQ} (I/O driver, directly tied to V_{DD}), V_{REFDQ} and V_{REFCA} (reference voltages), and V_{DDSPD} (supply for the on-board EEPROM). We assume that only the V_{DD} line provides significant power that needs to be measured. We use a JET-5464 DDR3 DIMM Extender with a 3.3mΩ sense resistor [1]. The voltage drop across the sense resistor can be used to calculate the current draw via Ohm's Law $I = \frac{V}{R}$ where V is the voltage drop and R is 3.3mΩ. This current can be passed into the equation $P = IV$ to calculate the power, with this V being the DDR3 RAM supply voltage of 1.5V (which we also measure). The original voltage drop being measured is very small due to the small resistor value, so an INA122 instrumentation amplifier [4] is used to amplify the signal by 300x before measurement.

DDR4 DIMMs have a somewhat different set of voltages. This includes V_{DD} (main supply), V_{TT} (termination supply), $12V$ (not provided on the registered DIMMs we use), V_{PP} (activation power supply), V_{DDSPD} (for the i2c EEPROM), and V_{REFCA} the reference voltage. We use an Adex DDR4-L-CSR extender which allows probing all of those various voltages, but we only look at V_{DD} and V_{PP} which are instrumented with 5mΩ sense resistors. The methodology is similar to that for DDR3, although more complicated as we have to correlate the results from the 1.2V V_{DD} and 2.5V V_{PP} supplies. We again amplify the results; V_{DD} by 100 and V_{PP} by 300.

The DRAM and CPU voltages are logged using a Measurement Computing USB-1208FS-Plus [24] data acquisition board, as shown in Figure 3. A custom Linux utility was developed to log the data, and we sample at 1kHz. A separate machine (in our case a Raspberry Pi) is used to log the various sampled voltages.

Picture of an instrumented desktop machine and an instrumented server can be seen in Figures 1 and 4.

4.2 RAPL Measurement

The RAPL values are gathered using the perf tool that comes with the Linux kernel and uses the perf_event [34] interface. We also gather other hardware performance counter values at the same time, including cycles and cache misses. An example command line used:

```
perf stat -a -e cycles,instructions,
        cache-misses,cache-references,
        uncore_imc/data_reads/,
        uncore_imc/data_writes/,
        power/energy-cores/,
        power/energy-pkg/,
        power/energy-cpu/,
        power/energy-ram/
        ./run_test.sh
```

To allow gathering system-wide measurements as a normal user the `/proc/sys/kernel/perf_event_paranoid` setting is set to "0". When generating phase plots, the additional `-I 100 -x ,` options are passed in to gather regularly sampled results in a CSV format.

4.3 Result Synchronization

The measurements we gather end up on two different machines. The RAPL measurements are collected locally on the machine under test (unfortunately possibly skewing the results due to measurement overhead). The actual power measurements are collected at the same time on a separate machine using the data acquisition board. When collating the results it is necessary to line up the start and stop times of the measurements as closely as possible.

There are various ways to do this and all have their limitations. One common way is to synchronize the clocks of the two machines using NTP (Network Time Protocol). We chose a different approach, where we modify the `perf` tool to toggle the DTR line of a pl2303 USB/serial adapter. This serial port line is connected to one of the inputs on our data acquisition device, allowing our recorded traces to have a clear signal of when `perf` measurements were started on the device under test. This allows our tools to automatically synchronize the two data sets, although there is still some delay as the signal traverses the USB and serial stacks. Since the `perf` tool toggles the DTR line, the tail end of its execution also ends up being included in the power traces. In our experiments the DTR line occasionally glitches when the serial port is opened, so we have additional code that debounces the line in software.

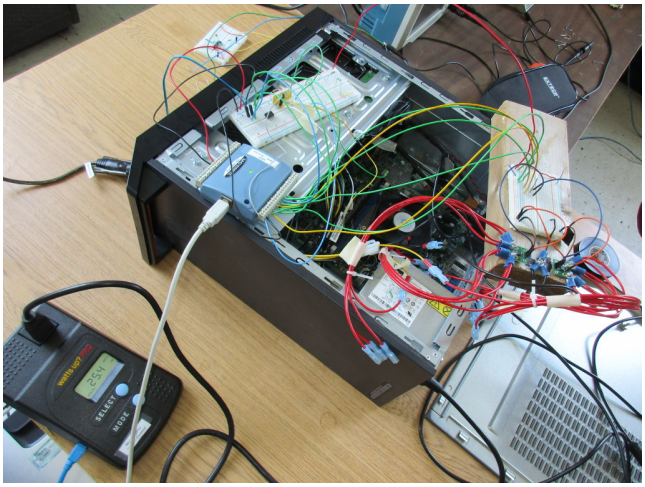
[1]We tried using a Hall effect sensor instead of a sense resistor, but the voltage drop was too much (the system booted but would quickly kernel-panic due to memory errors).

Table 1: Systems used in this paper.

System	CPU	System	BIOS	DRAM
Haswell-i5	i5-4570S, 2.90GHz	Lenovo ThinkCentre E7E	FCKT46AUS 12/16/2013	DDR3
Haswell-i7	i7-4770, 3.40GHz	Lenovo ThinkCentre M83	FBKT72AUS 1/26/2014	DDR3
Haswell-EP	Xeon E5-2640v3, 2.60GHz	HP ProLiant DL360 Gen9	5/6/2015	DDR4

Table 2: DRAM used in this paper.

Type	Manufacturer	Model	Manual	Stats
DDR3	SK Hynix	HMT451U6AFR8C-PB	[30]	4GB 1Rx8 PC3 12800U-11-12-A1
DDR3	Samsung	M378B5173DBO-LKO	[29]	4GB 1Rx8 PC3 12800U-11-12-A1
DDR3	Micron	MT16JTF1G64AZ-1G6E1	[25]	8GB 2Rx8 PC3 12800U-11-13-B1
DDR4	SK Hynix	HMA41GR7MFR4N-TF	[31]	8GB 1Rx4 PC4-2133P-RC0-10
DDR4	Kingston	KTH-PL421/16G	–	16GB 2Rx4 PC4-2133P-RA0-11



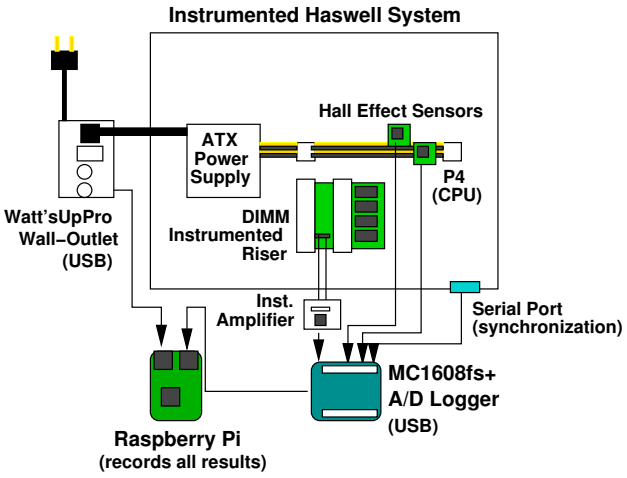


Figure 1: An instrumented desktop machine. The block diagram is representative, the setup varied slightly depending on if the system had a P4 connector and how many DRAM voltages we were measuring.

Table 3: DRAM performance measured by STREAM benchmark (MB/s).

Type	Hardware	copy	scale	add	triad
DDR3	i5 Hynix	7196	7135	8267	8286
	i5 Samsung	7087	7028	8087	8102
	i5 Micron	7591	7551	8563	8588
DDR3	i7 Hynix	6816	6731	7759	7821
	i7 Samsung	7043	6939	7991	8018
	i7 Micron	7622	7549	8457	8502
DDR4	EP Hynix	8312	8014	9028	9268
	EP Kingston	9064	8761	9783	10021

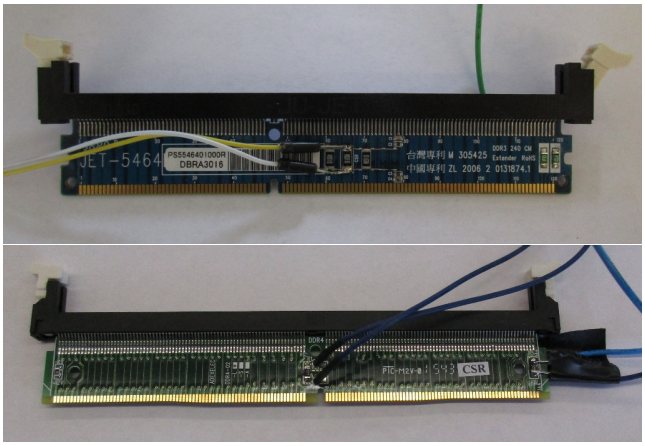


Figure 2: DDR3 and DDR4 DIMM extenders with sense resistors for measuring power.

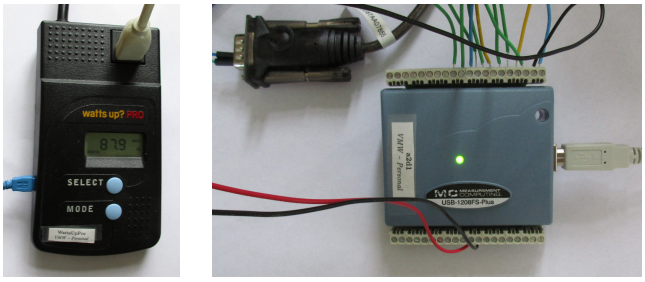


Figure 3: WattsUpPro wall outlet power meter and Measurement Computing USB-1208FS-Plus data acquisition device.

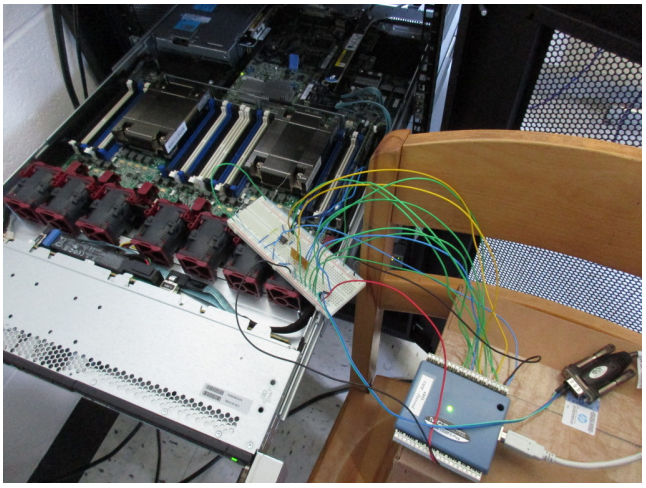


Figure 4: Instrumented server machine.

4.4 DAQ Measurement Accuracy

With such a complex measurement setup there are various sources of error that are hard to control for. There is the slowness of the serial port which can make trace synchronization inexact. The power converters have some inefficiency when reducing 12V to the voltages used by the DRAM (it is unclear if RAPL accounts for this or not; our measurements are beyond the converter). Our test setup introduces various capacitances and resistances by using a breadboard as well as long probe wires. Variations in various resistors (the ones controlling the gain in the instrumentation amplifiers, as well as the one being used to sense current) could make the gain calculations inaccurate. The INA122 amplifier may show non-linear gain at low voltages (we were specifically worried about this, but as shown in Figure 5 the variation is small). The data acquisition device analog/digital converters may not be perfectly calibrated. The BIOS and firmware on the various machines might be configuring the RAPL interface with different parameters. Many of the steps in the measurements are temperature dependent yet we did not regulate or measure temperature across runs. The state of the computers themselves could be different; it is possible different power states across runs. Also the physical memory addresses where code and data live are likely different across runs and across reboots. Any of the above could be the source of error and it will be extremely time consuming to eliminate all possible sources of error.

4.5 Sampling Frequency

For our overall results we sample the data acquisition board at 1kHz. This conveniently matches the internal 1kHz frequency of RAPL, but there are other reasons we chose it. The board can sample 4 channels of results up to about a frequency of 13kHz. We did some experiments to see what the optimal frequency was, with the primary tradeoff being the size of the resulting data files. Other external factors limit the sample size too, for example the instrumentation amplifier gain response becomes nonlinear above around 10kHz. The results of our experiments are shown in Table 4. We find that when measuring total aggregate results, 100Hz and above seems to do a reasonable job. This varies with the behavioral complexity of the underlying benchmark.

Table 4: Measured average power on Haswell-i5 Hynix DRAM of same run with different sample frequencies.

Benchmark	10kHz	5kHz	1kHz	500Hz	100Hz	50Hz	10Hz	5Hz	1Hz
sleep	0.540W	0.540W	0.540W	0.540W	0.537W	0.536W	0.539W	0.545W	0.571W
stream	2.34W	2.34W	2.34W	2.34W	2.34W	2.34W	2.32W	2.30W	2.20W
gcc-papi	—	—	1.26W	1.26W	1.25W	1.25W	1.24W	1.24W	1.26W
hpl-atlas	—	—	2.00W	2.00W	2.00W	2.00W	1.99W	1.97W	2.00W
hpl-mkl	—	—	2.24W	2.23W	2.23W	2.23W	2.23W	2.23W	2.18W
hpl-openblas	—	—	1.69W	1.69W	1.69W	1.68W	1.69W	1.70W	1.63W
opencl	—	—	1.02W	1.03W	1.03W	1.03W	1.14W	1.15W	1.21W
ksp	—	—	1.50W	1.50W	1.49W	1.48W	1.48W	1.51W	1.29W

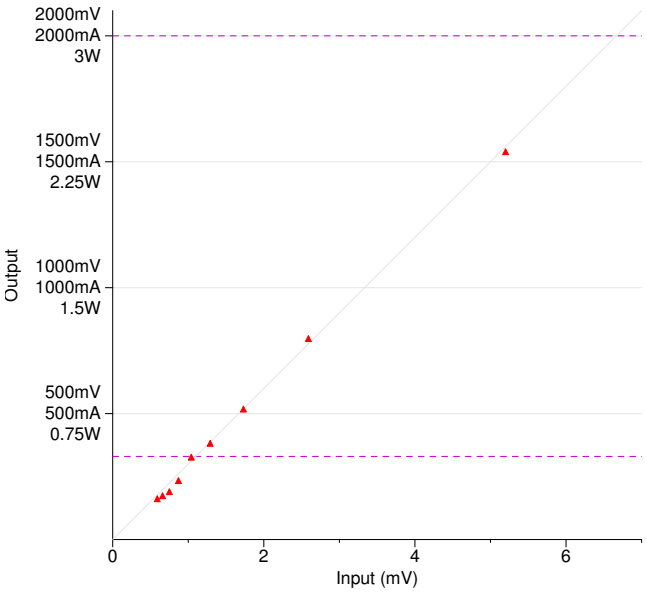


Figure 5: Measured INA122 Gain with 680 Ohm resistor (grey line is 300x expected) at 1.5V common mode. The dashed lines outline the typical operating range of the DDR3 DIMMs measured.

4.6 DIMM Extender Overhead

The DIMM extenders themselves can affect the performance of the machines, due to increased signal path lengths and voltage drops due to the sense resistors. We investigate whether it was possible to notice these effects in our RAPL measurements. Table 5 shows the percent difference in RAPL DRAM energy values with the extender in to with the extender removed. There does not seem to be a clear correlation in the results (which are probably lost in the noise), but oddly on the EP machine (which has actual power measurement circuitry) it seems that removing the extender actually makes performance and energy consumption *worse*.

4.7 RAPL Measurement Overhead

When generating the phase plots we only gather perf results at 10Hz (100ms) resolution. This is a relatively low frequency, as the RAPL counters update at 1kHz. The perf tool has a convenient "print every interval" (`-I`) mode but it is hard-coded to not allow measurement faster than 100ms. We found that by removing the limit and trying to gather data at 100Hz caused a noticeable 0.5W jump in power consumption due to measurement/interrupt overhead. We investigated writing a custom tool that would use the perf_event interface's sampling/`mmap()` ring-buffer recording mode to provide lower-overhead access, but when we tried to record at 1kHz the kernel's interrupt throttling kicked in due to the performance interrupts taking up over 25% of CPU time. For now we are using the lower (10Hz) sampling frequency. Possible ways to avoid this would be to use a different performance interface such as LIKWID [33] or to read the MSRs directly.

4.8 Benchmarks

We investigate a variety of benchmarks of interest to us, including some commonly used in high-performance computing.

sleep: For a baseline we look at an idle system, which is just recording system behavior when a "sleep" command is issued. Note that we do have a full Debian Jessie environment running so the system is not truly idle (i.e. we are not running in single-user mode with all unnecessary processes killed). This is because we are interested in the power behavior of a real-life system that is sitting unused.

stream: In order to exercise the DRAM we look at the STREAM [23] benchmark which tests a machine's memory performance. STREAM performs operations such as copying bytes in memory, adding values together, and scaling values by another number. We use the OpenMP version of the benchmark to try to use all of the cores in the system.

HPL Linpack: To exercise the CPU we use the high-

Table 5: % Difference in RAPL DRAM Energy after removing extender.

Hardware	sleep	stream	gcc-papi	hpl-atlas	hpl-mkl	hpl-openblas	opencl	ksp
i5 Hynix	-0.2%	0.0%	0.4%	0.2%	-3.0%	-2.1%	-1.3%	-0.8%
i5 Micron	-6.7%	-2.1%	-6.0%	0.1%	-1.4%	-3.1%	-1.7%	0.0%
i5 Samsung	3.4%	0.7%	-2.0%	0.7%	0.4%	0.0%	0.1%	-0.5%
i7 Hynix	0.0%	0.2%	3.1%	2.4%	3.0%	3.0%	—	—
i7 Micron	6.8%	—	2.9%	1.6%	1.2%	1.7%	—	0.7%
i7 Samsung	13.3%	2.4%	-4.8%	1.6%	2.0%	3.5%	—	-0.7%
EP Hynix	-0.7%	-2.2%	-0.7%	0.5%	1.8%	1.7%	—	—
EP Kingston	1.9%	3.0%	3.2%	2.6%	1.0%	1.0%	—	—

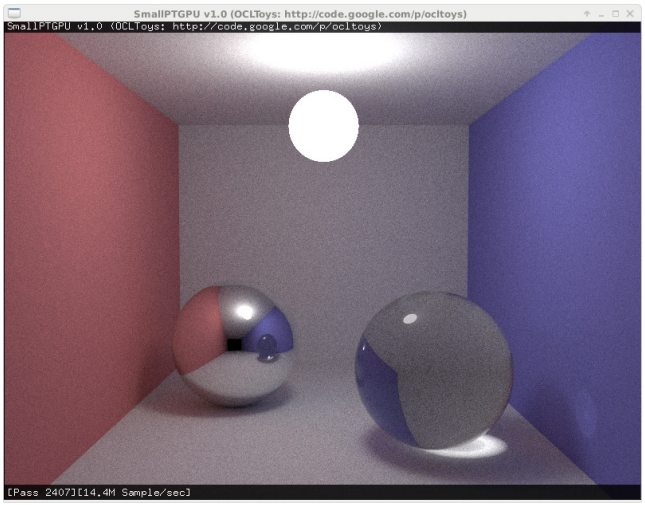


Figure 6: OpenCL raytracing benchmark output.

Figure 7: Kerbal Space Program OpenGL gaming benchmark.

performance Linpack HPL benchmark. We use it with three different BLAS (linear algebra) libraries:

- The version of Automatically Tuned Linear Algebra Software (ATLAS) [35] that ships with Debian Linux,
- OpenBLAS [1] optimized for Haswell processors (including using the new FMA fused-multiply-add) instruction, and
- a statically linked version that comes with Intel's MKL libraries [15].

HPL is configured with a problem size of N=15000 and to use a 2x2 grid of processors, which gives high performance for all of the BLAS implementations and uses nearly 4GB of memory.

gcc PAPI: For a less intense integer benchmark we use the gcc compiler. The version of gcc used is 4.9 that comes with Debian Linux. The test compiles version 5.4.1 of the PAPI library [26] with a four-way parallel make.

OpenCL SmallptGPU2: It is difficult to obtain power measurements for the integrated GPU, as it is on-die and there is no way to intercept the input voltages. There are additional (non-power related) hardware performance counters available for the integrated GPU [17] but as of yet the Linux support for reading these is not complete.

We use SmallptGPU2 [3], an OpenCL ray-tracer whose result is shown in Figure 6. For an OpenCL implementation we use Beignet [14] which is developed for the Intel HD series of integrated GPUs. We use the default ray-trace setup, ending after 250 iterations of tracing.

OpenGL KSP: For an OpenGL intensive video game benchmark we use the game Kerbal Space Program [32] as shown in Figure 7. We record a 20s long snapshot of behavior while launching a rocket in-game.

5. RESULTS

We measure actual power measurements on the various Haswell systems and compare these results to those returned by RAPL.

5.1 Haswell/DDR3 Aggregate Results

The results from the DDR3 measurements are presented in Tables 6, 7 and 8, as well as Figure 8.

On the i5 machine the results tend to match within 20% on the 4GB DIMMs but a bit worse on the 8GB DIMM. The hpl_mkl benchmark does particularly well, perhaps Intel used it as one of the workloads when calibrating the interface. In all cases the worst behavior is from benchmarks which have significant idle time (such as sleep), where RAPL

Table 6: Hynix 4GB DDR3 results

Benchmark	Haswell i5					Haswell i7				
	Avg. Power		Energy			Avg. Power		Energy		
	Measured	RAPL	Measured	RAPL	% Diff	Measured	RAPL	Measured	RAPL	% Diff
sleep	0.53W	0.42W	5.27J	4.20J	-20.3%	0.21W	0.48W	2.10J	4.76J	127%
stream	2.3W	2.5W	24.3J	25.6J	5.35%	2.1W	2.6W	23.5J	28.2J	20.0%
hpl-openblas	1.7W	1.6W	56.2J	51.7J	-8.00%	1.4W	1.7W	54.4J	67.6J	24.3%
hpl-atlas	2.0W	1.6W	119J	95.3J	-20.0%	1.4W	1.7W	73.6J	89.5J	21.6%
hpl-mkl	2.2W	2.3W	52.6J	54.5J	3.61%	1.9W	2.3W	42.7J	51.6J	20.8%
gcc	1.3W	1.1W	13.7J	11.8J	-13.9%	0.81W	1.1W	7.27J	10.2J	40.3%
ksp	1.5W	1.3W	29.5J	25.0J	-15.3%	0.91W	1.3W	18.2J	25.9J	42.3%
openCL	1.1W	0.81W	13.6J	10.4J	-23.5%	n/a	n/a	n/a	n/a	n/a

Table 7: Samsung 4GB DDR3 results

Benchmark	Haswell i5					Haswell i7				
	Avg. Power		Energy			Avg. Power		Energy		
	Measured	RAPL	Measured	RAPL	% Diff	Measured	RAPL	Measured	RAPL	% Diff
sleep	0.65W	0.46W	6.60J	4.64J	-29.7%	0.13W	0.42W	1.27J	4.20J	230%
stream	2.2W	2.5W	23.7J	26.0J	9.70%	2.0W	2.6W	21.0J	27.6J	31.4%
hpl-openblas	1.7W	1.6W	55.2J	51.9J	-5.98%	1.3W	1.8W	50.2J	68.1J	35.7%
hpl-atlas	1.9W	1.6W	113J	97.0J	-14.2%	1.3W	1.7W	68.2J	90.5J	32.7%
hpl-mkl	2.2W	2.3W	50.9J	53.7J	5.50%	1.8W	2.3W	39.4J	51.7J	31.2%
gcc	1.4W	1.1W	15.5J	12.3J	-20.6%	0.81W	1.2W	7.31J	11.1J	51.8%
ksp	1.5W	1.2W	29.6J	24.9J	-15.9%	0.85W	1.3W	16.9J	25.7J	52.1%
openCL	1.2W	0.82W	14.6J	10.3J	-29.5%	n/a	n/a	n/a	n/a	n/a

Table 8: Micron 8GB DDR3 results

Benchmark	Haswell i5					Haswell i7				
	Avg. Power		Energy			Avg. Power		Energy		
	Measured	RAPL	Measured	RAPL	% Diff	Measured	RAPL	Measured	RAPL	% Diff
sleep	2.0W	0.90W	20.2J	8.98J	-55.5%	0.33W	0.84W	3.33J	8.38J	152%
stream	2.5W	3.1W	25.2J	30.6J	21.4%	2.7W	3.3W	26.6J	32.4J	21.8%
hpl-openblas	2.4W	2.1W	77.0J	66.8J	-13.2%	1.9W	2.3W	69.7J	85.0J	22.0%
hpl-atlas	2.4W	2.1W	142J	124J	-12.7%	1.9W	2.2W	96.9J	114J	17.6%
hpl-mkl	2.8W	3.0W	59.9J	63.7J	6.34%	2.5W	3.0W	51.0J	61.2J	20%
gcc	2.2W	1.6W	22.2J	16.1J	-27.5%	1.1W	1.6W	9.45J	13.9J	47.1%
ksp	2.3W	1.7W	45.2J	34.3J	-24.1%	1.2W	1.7W	23.4J	34.4J	47.0%
openCL	2.1W	1.2W	27.3J	16.0J	-41.4%	n/a	n/a	n/a	n/a	n/a

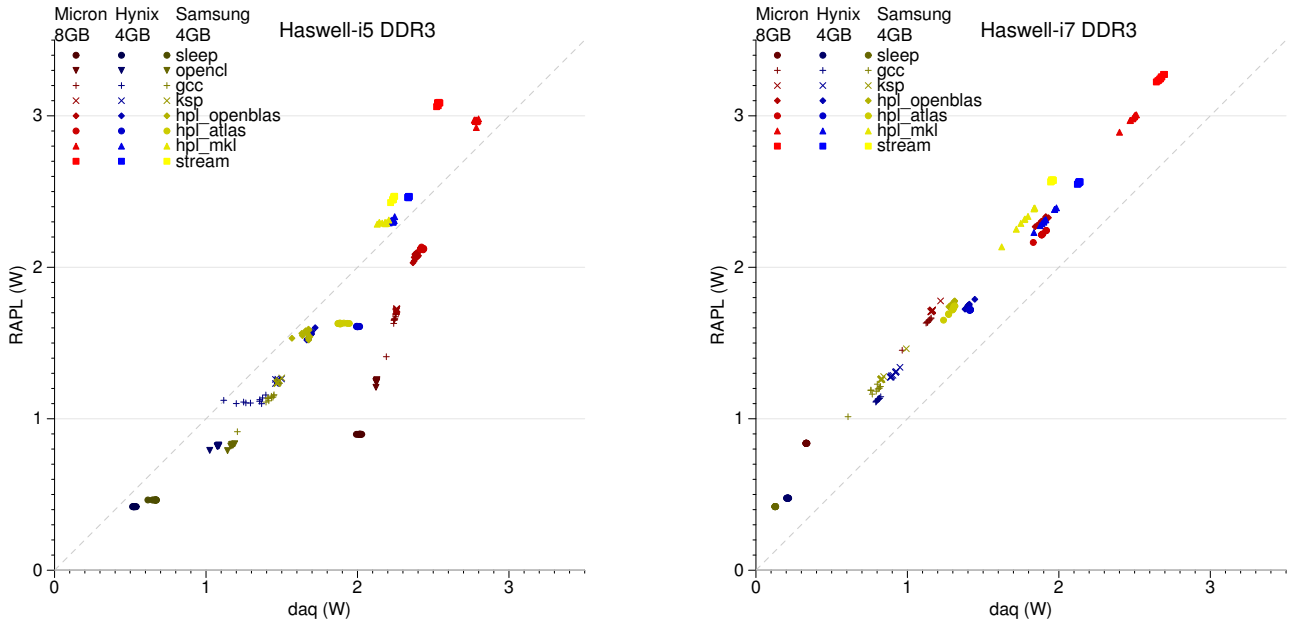


Figure 8: Haswell i5 and i7 scatter plots. In general the RAPL results are similar to those of the actual measurement, with a constant voltage offset. The i5 system seems to undercount at low voltages, and the 8GB DIMM is not modeled well.

consistently underestimates the results. The same behavior is also seen on the GPU intensive benchmarks, where RAPL possibly is not accounting for memory or DMA accesses that bypass the CPU. RAPL has difficulty estimating the behavior of the 8GB DIMM.

On the i7 machine the results are very different, even though the same DIMMs are being used. In this case RAPL consistently *over*-estimated the amount of energy being used, with the power offset by a roughly constant amount. Again the worst case seems to be the cases where the system is most idle.

The RAPL estimates themselves are fairly consistent with the same DIMM/benchmark combination having similar estimates across the different machines.

5.2 Haswell/DDR3 Phase Plots

The aggregate results in the previous section show that the overall results can be over 20% off between RAPL and actual measurements. This might not be a problem if our performance analysis only depends on relative power measurements rather than absolute ones. To look at the relative measurements we collect phase graphs by periodically sampling the machine's behavior and plotting this over time.

In addition to the RAPL and actual DRAM energy results we use the CPU hardware performance counter to gather a number of other metrics, including the cycles per instruction (CPI) and last level cache (LLC) misses. In general the RAPL DRAM power follows the LLC rate, and the RAPL package power follows the CPI metric.

For the GPU benchmarks we additionally measure the RAPL core and RAPL GPU values. Finally we take actual hardware measurements of total system power, the P4 ATX connector (which should be closely related to package power), as well as the actual DIMM power.

The results shown are gathered with the i5 system and the 4GB Hynix DIMM. The actual measurements are taken at 2kHz while the RAPL results are gathered at 10Hz.

5.2.1 CPU Benchmark Results

In Figure 9 we show the results of an idle system. No attempts were made to limit the number of background jobs running, or in any way artificially limiting the background noise. We wanted to measure the power behavior of a typical system sitting unused. It turns out that this setup has surprisingly high CPI and cache variability.

The CPU RAPL and actual power measurements match each other fairly well, although RAPL seems to underestimate the power slightly (but this could be due to the P4 connector powering devices other than the CPU, as well as losses in the power converters that convert the 12V input to the much lower voltages used by the CPU).

As seen with the aggregate measurements, the DRAM RAPL values are much lower than actual values, possibly RAPL has trouble estimating power if the DIMM has entered a low-power mode.

Figure 10 shows the results when DRAM is being stressed by a multi-core aware OpenMP version of the STREAM benchmark.

The total CPU package measurements match closely the CPI results from the performance counters, and the DRAM results match closely the last level cache misses. Again, the CPU RAPL estimates read a bit lower than the actual measurements. While under high utilization the DRAM RAPL results closely match measured results, but when memory utilization drops toward idle the RAPL values read low.

Figure 11, 12 and 13 show Linpack running with various BLAS libraries. Despite being the same benchmark, the underlying BLAS libraries lead to markedly different phase behaviors. The phase behavior is also much more complex than the other benchmarks we investigate. In some of the figures it appears as though the total CPU package power is higher than the wall-outlet power measurement; this is just an artifact due to the much lower sampling frequency of the WattsUpPro? device.

In the ATLAS results (Figure 11) there are periodic spikes in cache misses which correspond to increased memory power usage as well as dips in cpu power usage. The DRAM RAPL measurements seem to be consistently lower than measured, even when the memory system is busy. This could be a measurement artifact due to the higher sampling rate of the hardware measurement compared to the lower rate that we sample the RAPL counters.

The OpenBLAS results (Figure 12) have different underlying behavior to the ATLAS results, but the power values show similar trends, with the DRAM results being consistently lower.

The Intel MKL results (Figure 13) again have similar trends, with the DRAM results being lower.

5.2.2 GPU Benchmark Results

Figure 14 shows the results when the GPU is being used for OpenCL raytracing calculations. According to the RAPL results the actual cores are almost completely idle and contributing very little power. The GPU is using the bulk of the power, and there is an interesting 5W reported by the package not accounted for by the GPU, perhaps some other aspect of the uncore.

The DRAM behavior is complex and the RAPL readings do not seem to capture this, possibly due to the low sampling frequency. Another possibility is that the GPU is doing extensive DMA transfers which might not be accounted for by the RAPL model.

Figure 15 shows the results when the GPU is being used to play a 3D video game. The power profile is very similar to that of the OpenCL demo with slightly more CPU being used (though the game is only using 1 core). Again, the DRAM RAPL count seems to not be accounting for GPU interactions.

5.2.3 DRAM GFLOPS/W

One use of power measurement is to compare the power efficiency of equivalent algorithms. The three HPL benchmarks are all calculating the same result, so the GFLOPS/Watt metric can be used to compare which one has the most efficient RAM power usage. These results are shown in Table 9.

Even though the aggregate results returned by RAPL and actual measurements do not match exactly, they return similar rankings for which HPL implementation to use. On both i5 and Haswell-EP either power measurement method will show you that OpenBLAS is best, followed by mkl, then by ATLAS. Things are more complex on i7 as OpenBLAS and mkl have similar power efficiencies, but both power measurement methodologies reflect this.

At least in the case of HPL we find that even though the results do not exactly match, RAPL and actual measure-

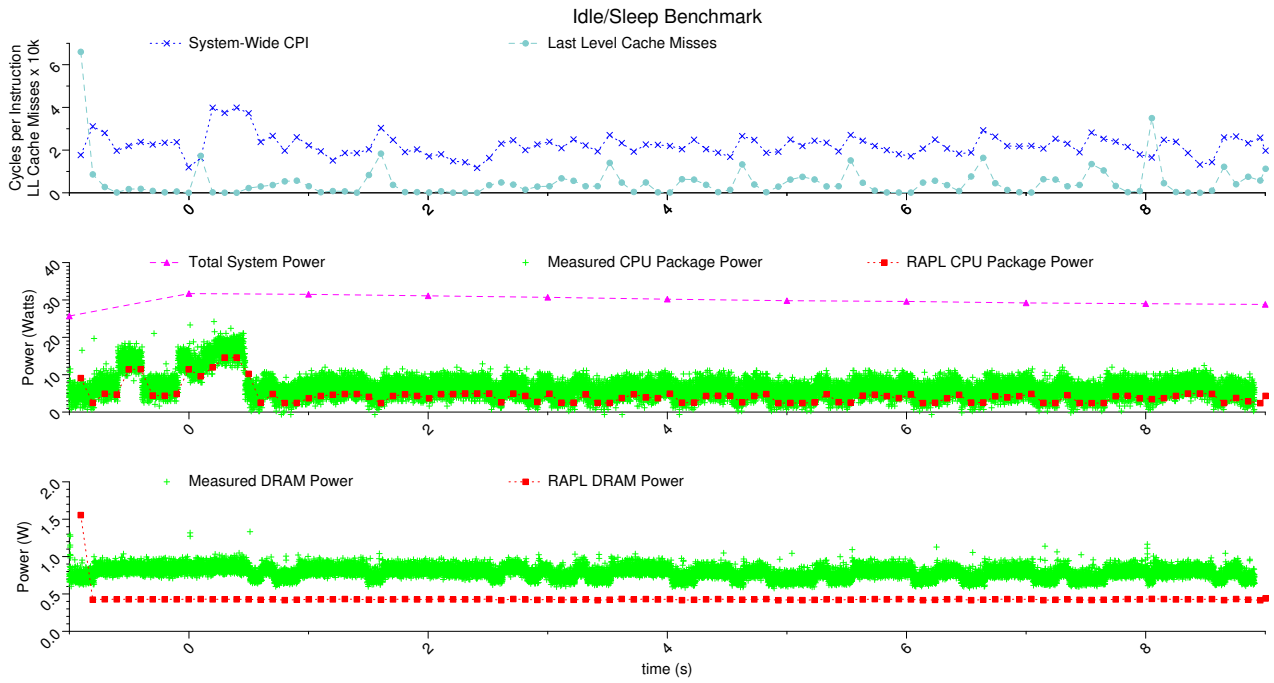


Figure 9: Power measurements for an idle system (perf is run on a call to the `sleep` command). While CPU actual vs estimated is close, RAPL DRAM measurements underestimate the power used.

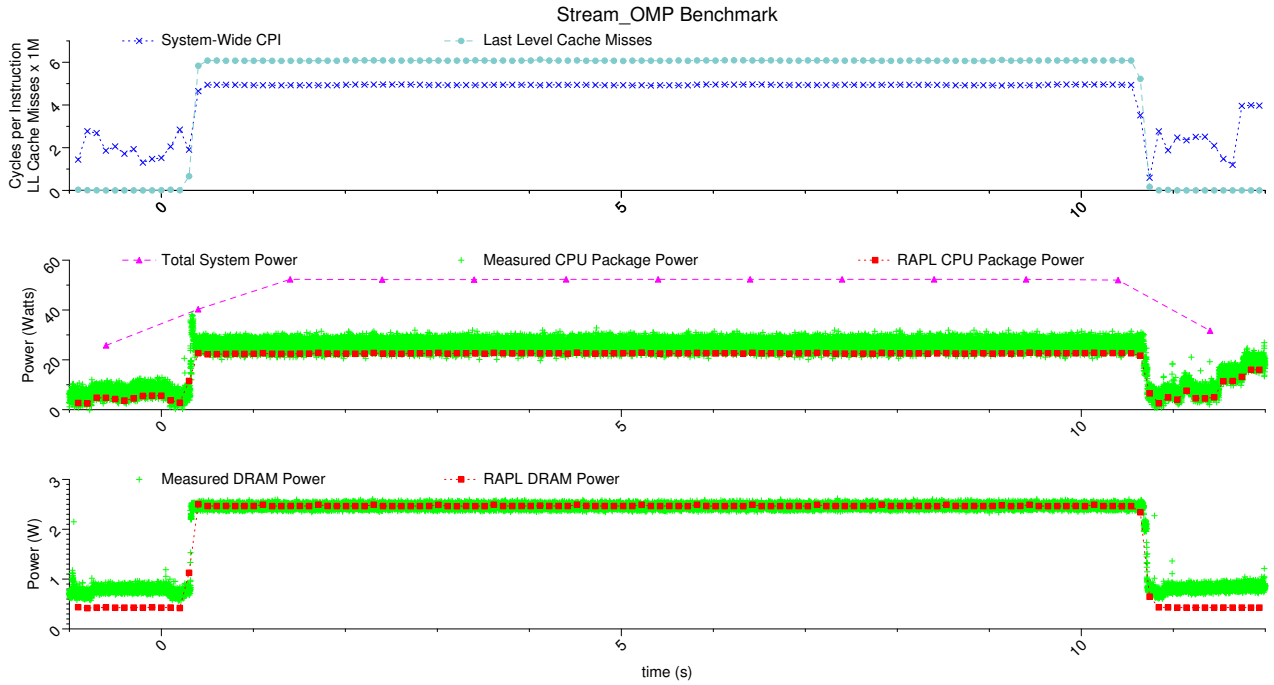


Figure 10: Power measurements while running an OpenMP version of the memory-intensive STREAM benchmark. The DRAM measurements match estimated RAPL results when under heavy memory stress, but when memory usage drops RAPL again underestimates the power.

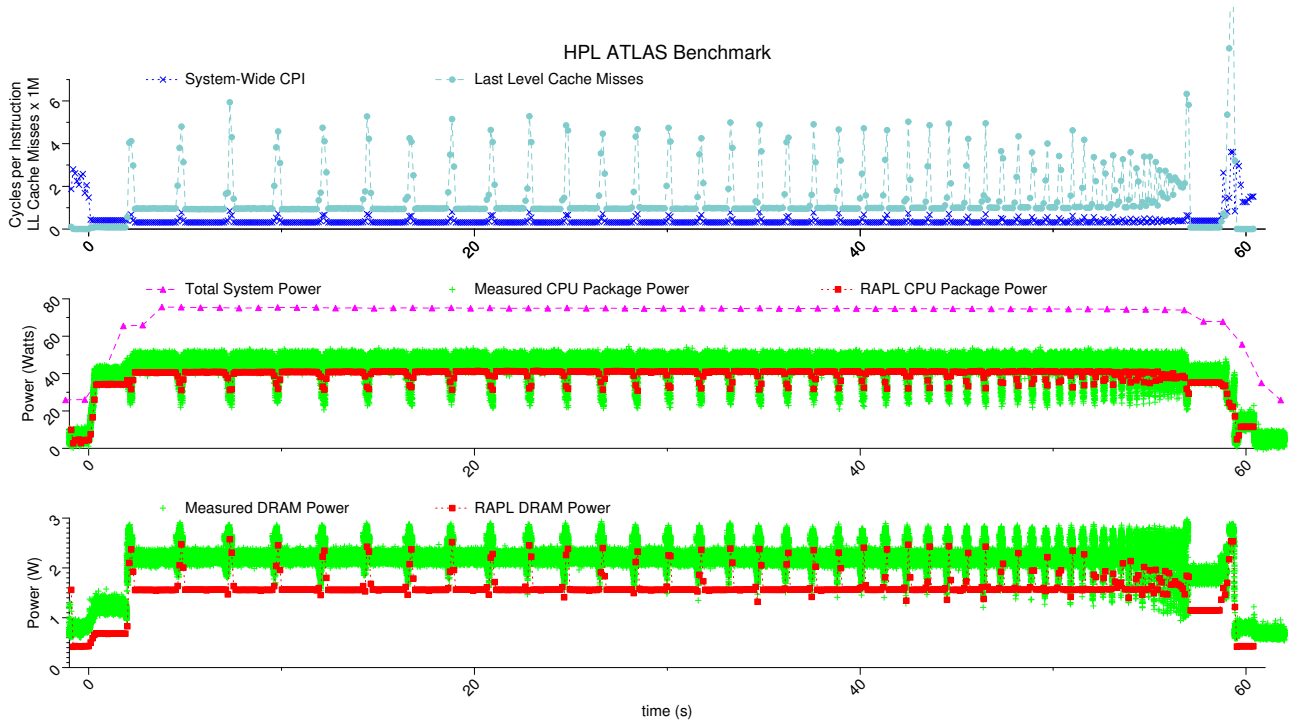


Figure 11: Linpack (HPL) using Atlas BLAS. The periodic spikes in cache misses correspond with rises in DRAM power but dips in CPU power. It appears that package power is higher than total system power, but this is an artifact of the low sampling period of the WattsUpPro meter.

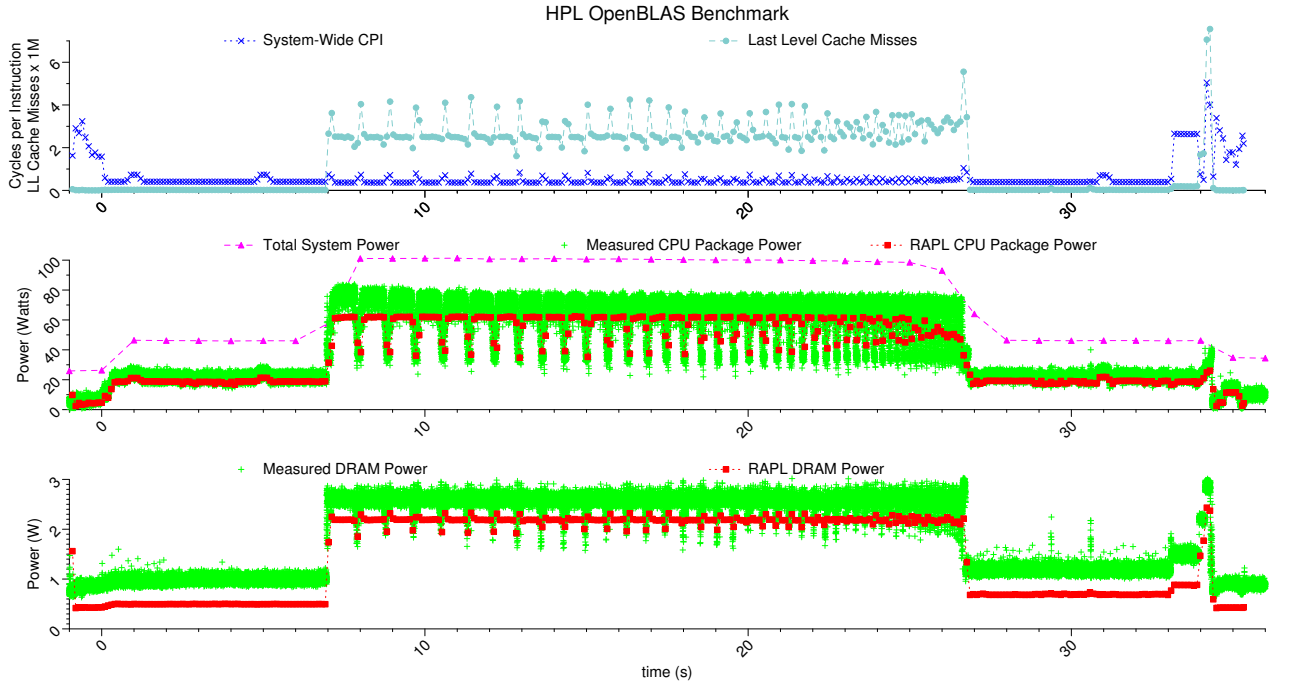


Figure 12: Linpack (HPL) using OpenBLAS. The DRAM estimated RAPL power is consistently less than total power. It appears that package power is higher than total system power, but this is an artifact of the low sampling period of the WattsUpPro meter.

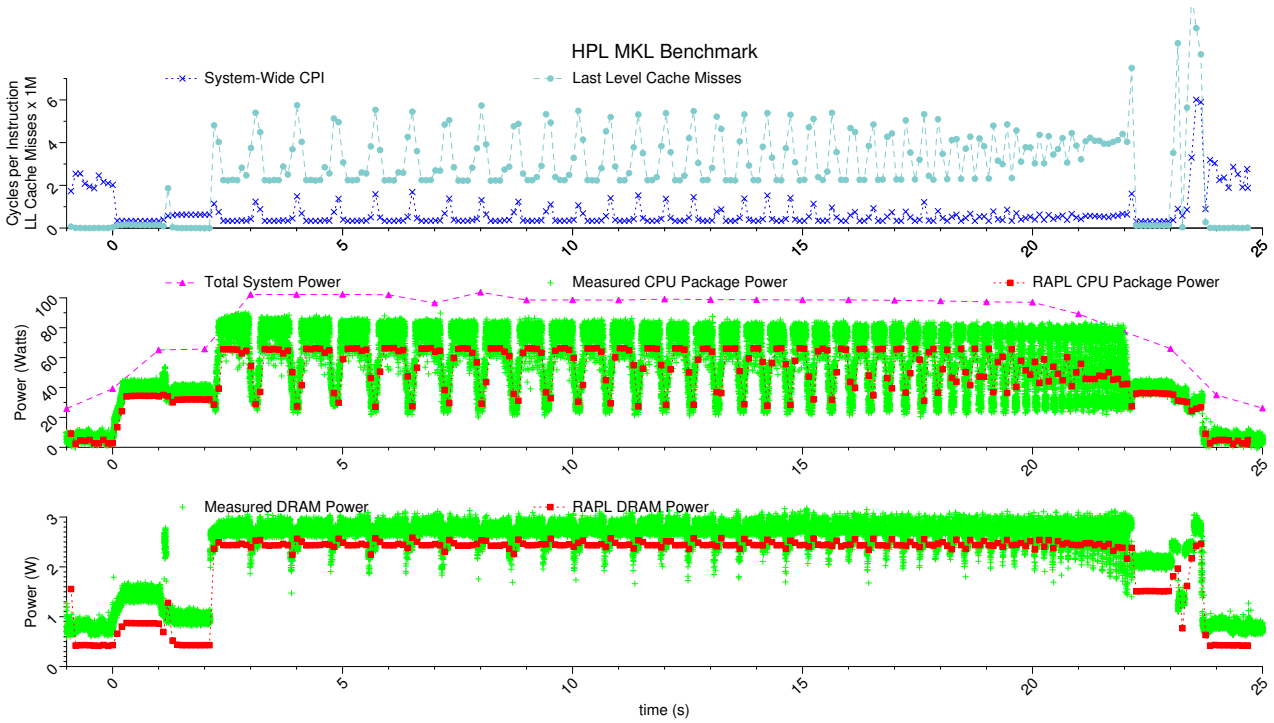


Figure 13: Linpack (HPL) using Intel MKL BLAS. The DRAM estimated RAPL power is lower than measured when the DRAM is less active. It appears that package power is higher than total system power, but this is an artifact of the low sampling period of the WattsUpPro meter.

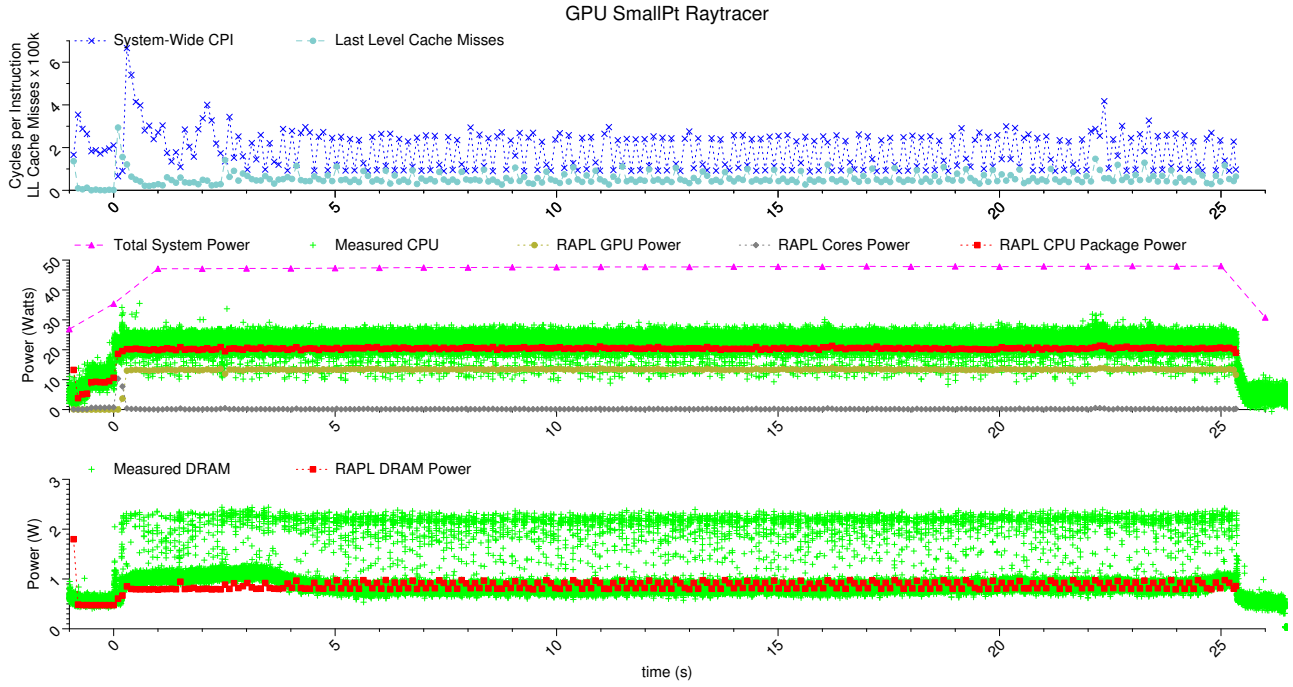


Figure 14: Smallpt OpenCL Raytracer. The majority of package power is consumed by the GPU, with the CPU cores mostly idle. The complex DRAM power behavior is not captured well by RAPL.

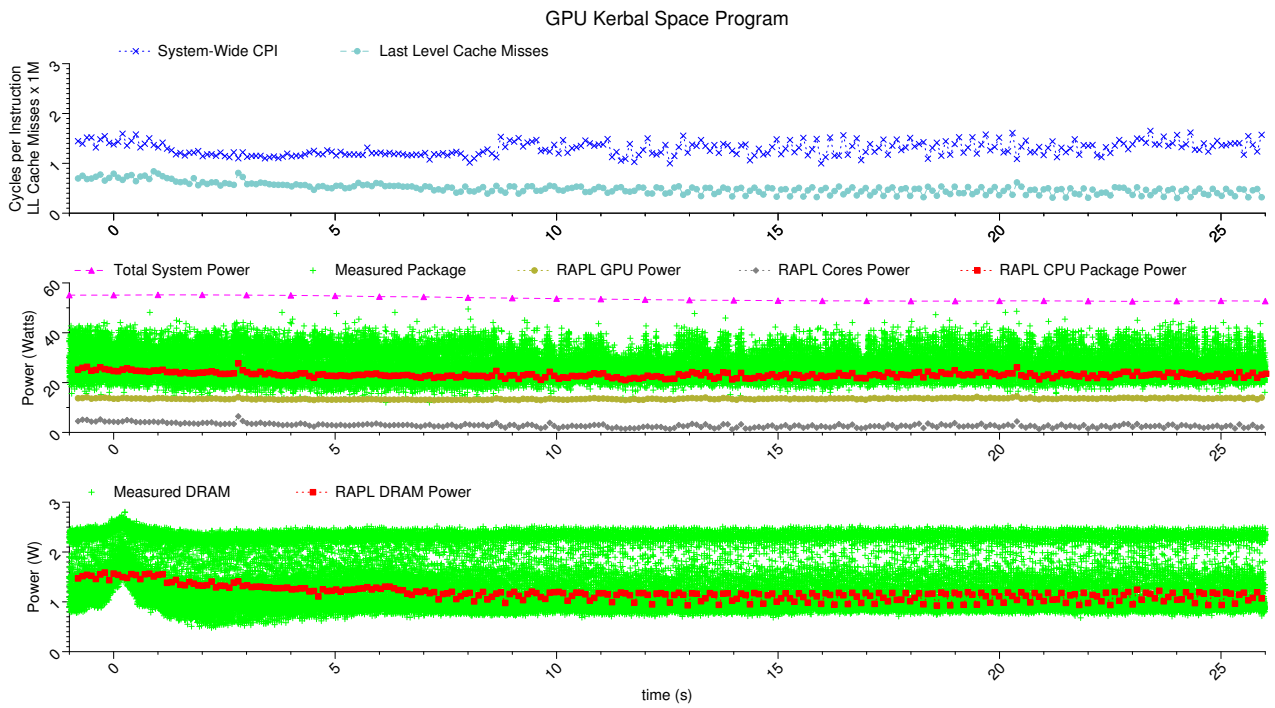


Figure 15: Kerbal Space Program, a 3D/GPU intensive Game. The majority of package power is consumed by the GPU, with the CPU cores mostly idle. The complex DRAM behavior is not captured well by RAPL.

Table 9: DRAM GFLOPS/W

Hardware	ATLAS			OpenBLAS			mkl		
	GFLOPS	$\frac{GFLOPS}{W}$ Measured	$\frac{GFLOPS}{W}$ RAPL	GFLOPS	$\frac{GFLOPS}{W}$ Measured	$\frac{GFLOPS}{W}$ RAPL	GFLOPS	$\frac{GFLOPS}{W}$ Measured	$\frac{GFLOPS}{W}$ RAPL
i5-hynix	40.8	20.4	25.3	112	66.3	72.3	106	47.3	46.1
i5-samsung	40.6	21.6	24.9	117	74.5	76.5	110	49.8	47.6
i5-micron	41.5	17.1	19.7	120	50.0	57.5	116	41.7	39.7
i7-hynix	46.6	33.0	27.1	82.1	58.6	47.2	110	58.5	48.2
i7-samsung	46.2	35.3	26.6	84.7	65.2	48.4	114	63.3	48.7
i7-micron	47.3	24.8	21.1	97.4	51.0	41.8	125	50.0	41.7
ep-hynix	43.5	21.9	18.8	178	97.3	82.4	130	53.1	46.4
ep-kingston	44.0	14.7	14.2	194	63.8	64.7	136	35.2	35.5

Table 10: DDR4 results

Benchmark	Haswell-EP – Hynix 8GB					Haswell-EP – Kingston 16GB				
	Avg. Power		Energy			Avg. Power		Energy		
	Measured	RAPL	Measured	RAPL	% Diff	Measured	RAPL	Measured	RAPL	% Diff
sleep	0.34W	0.60W	3.43J	6.05J	76.4%	0.46W	0.64W	4.64J	6.41J	38.4%
stream	3.0W	3.3W	28.7J	31.3J	9.06%	4.6W	4.7W	40.2J	40.4J	0.50%
hpl-openblas	1.9W	2.2W	52.8J	61.9J	17.2%	3.1W	3.1W	83.8J	82.4J	-1.67%
hpl-atlas	2.0W	2.3W	111J	129J	16.2%	3.0W	3.1W	166J	172J	3.61%
hpl-mkl	2.5W	2.8W	49.6J	56.4J	13.7%	3.9W	3.8W	74.9J	73.8J	-1.47%
gcc	1.4W	1.7W	13.5J	16.2J	20.0%	2.2W	2.3W	21.0J	22.1J	5.24%

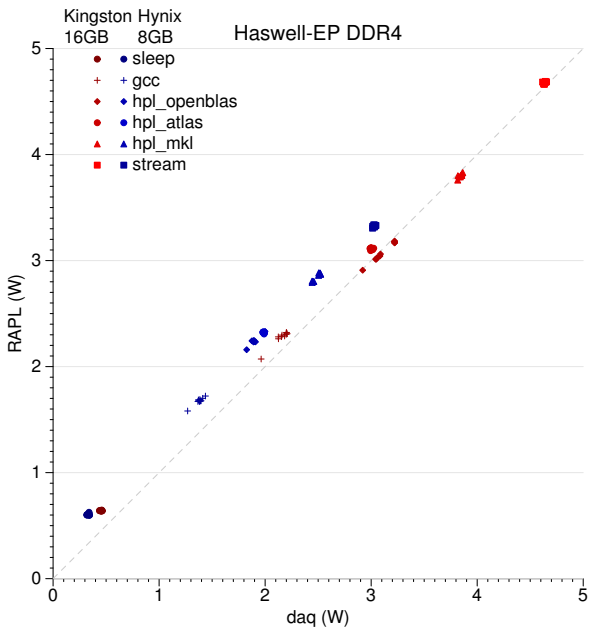


Figure 16: Haswell-EP scatter plot. On server machines the voltage regulator provides actual measurements to the RAPL analysis, so the RAPL results match the measured values much better than on the desktop machines.

ment can both be used to achieve the same results when gauging DRAM power efficiency.

5.3 Haswell-EP/DDR4 Aggregate Results

The DDR4 results are show in Table 10 and Figure 16. These should be more accurate than the DDR3 ones as the Haswell-EP has "Mode 1" actual power measurements happening.

The Hynix results are all within 20% except for the totally idle sleep results, which seem to be high as per with DDR3. There seems to be a consistent 0.3W offset on all the readings that explains the differences in results. No GPU benchmarks were run so there is no comparison of idle with GPU as we did with DDR3.

The Kingston results match much more closely than the Hynix ones. Excepting the sleep results, all values match to within 5%.

5.4 Haswell-EP/DDR4 Phase Results

Results from six benchmarks running on the Haswell-EP with Hynix DDR4 RAM are shown in Figure 17. Each plot shows the power used by V_{DD}, the power used by V_{PP} (which is too low to register), and the DRAM RAPL values. The actual readings are sampled at 1kHz and the RAPL values at 10Hz.

The idle plot has an interesting wave pattern going on which could use some more investigation; this is a case where a higher sample frequency would be useful in the analysis. The rest of the plots are similar to their DDR3 equivalents.

The RAPL results closely follow the actual results with a slightly positive offset, probably the same 0.3W offset seen in the DDR4 aggregate total results.

6. FUTURE WORK

We would like to extend this work to a wider variety of machines. Various other Intel processors support DRAM RAPL results:

- Sandybridge-EP – we have a Sandybridge-EP machine, an HP Proliant system. However the firmware does not allow gathering DRAM RAPL results.
- Broadwell – most broadwell systems take DDR3 memory in the SODIMM format. We have obtained an instrumented SODIMM extender in order to conduct measurements.
- Skylake – Skylake systems use DDR4 memory so we should be able to use our Haswell-EP methodology to gather results there.
- Knights Landing – these accelerator boards have DRAM RAPL support and it would be interesting to validate these.

In addition we would like to expand our results by measuring with multiple DIMMs installed, enabling monitoring of NUMA workloads.

7. CONCLUSION

We find that DRAM RAPL results can provide useful results with much less hassle than manually instrumenting the memory system. Our attempt to validate the results found that total aggregate results seem to vary widely depending on the CPU and the type of DIMM being used, but in general measurements match within 20%. Even better, the actual relative phase behavior of the results seems to follow the estimated RAPL values. The largest divergences seem to happen when the system is idle, as well as when the CPU is idle and the GPU is talking to memory.

On Haswell-EP server machines the RAPL results include actual power measurements, and indeed we find these results to be closer than the pure estimates on desktop machines. The results are close, or else offset by a constant amount.

Despite the differences in aggregate measurements, the RAPL counters do track overall program behavior and can be a useful measurement methodology especially when compared to the alternatives of either complex hand-instrumentation of every machine of interest or else having no memory energy information at all.

All of the tools and raw data used in this report can be found in a git repository linked from our website: http://web.eece.maine.edu/~vweaver/projects/rapl/rapl_validation.html

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. SSI-1450122. Spencer's contribution was sponsored by the University of Maine Center for Undergraduate Research (CUGR). The authors would also like to thank Nicholas Nethercote and Thomas Ilsche for their comments on an earlier version of this document.

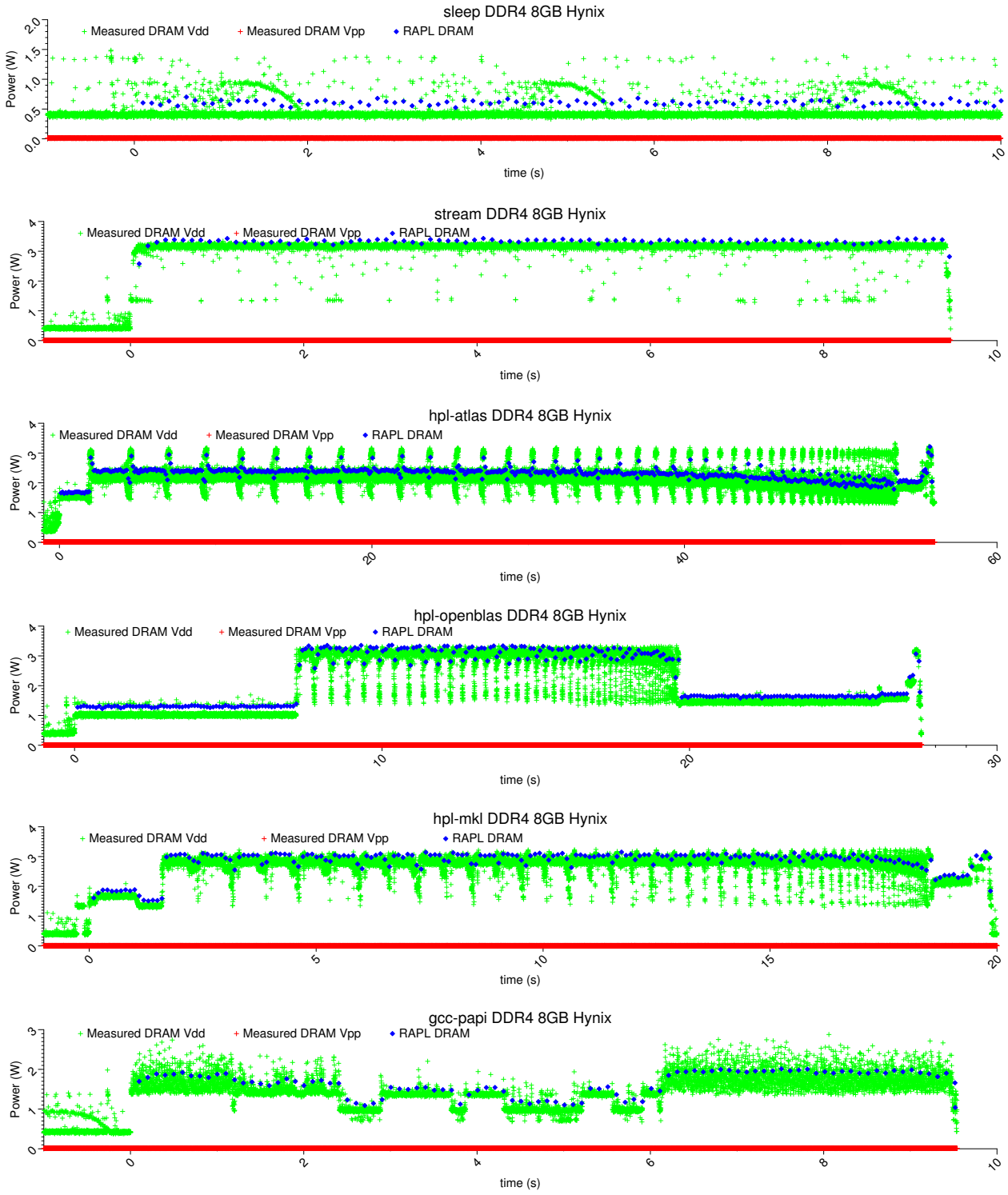


Figure 17: DDR4 Phase Plots

9. REFERENCES

[1] OpenBLAS an optimized BLAS library website. http://www.openblas.net/.

[2] Allegro MicroSystems LLC. *ACS715: Automotive Grade, Fully Integrated, Hall Effect-Based Linear Current Sensor IC with 2.1 kVRMS Voltage Isolation and a Low-Resistance Current Conductor Lightweight Profiling Specification*, 2013.

[3] D. Bucciarelli. Smallptgpu2. http://davibu.interfree.it/opencl/smallptgpu2/smallptGPU2.html.

[4] Burr-Brown. *INA122: Single Supply, MicroPower Instrumentation Amplifier*, Oct. 1997.

[5] H. Chen, S. Wang, and W. Shi. Where does the power go in a computer system: Experimental analysis and implications. In *International Green Computing Conference*, pages 1–6, July 2011.

[6] H. David, E. Gorbatov, U. Hanebutte, R. Khanna, and C. Le. RAPL: Memory power estimation and capping. In *ACM/IEEE International Symposium on Low-Power Electronics and Design*, pages 189–194, Aug. 2010.

[7] J. Demmel and A. Gearhart. Instrumenting linear algebra energy consumption via on-chip energy counters. Technical report, Electrical Engineering and Computer Sciences, University of California at Berkeley, June 2012.

[8] J. Dongarra, H. Ltaief, P. Luszczek, and V. Weaver. Energy footprint of advanced dense numerical linear algebra using tile algorithms on multicore architecture. In *Proc. of the 2nd International Conference on Cloud and Green Computing*, Nov. 2012.

[9] Electronic Educational Devices. Watts Up PRO. http://www.wattsupmeters.com/, May 2009.

[10] R. Ge, X. Feng, S. Song, H.-C. Chang, D. Li, and K. Cameron. PowerPack: Energy profiling and analysis of high-performance systems and applications. *IEEE Transactions on Parallel and Distributed Systems*, 21(6), May 2010.

[11] D. Hackenberg, T. Ilsche, R. Schoene, D. Molka, M. Schmidt, and W. E. Nagel. Power measurement techniques on standard compute nodes: A quantitative comparison. In *Proc. IEEE International Symposium on Performance Analysis of Systems and Software*, Apr. 2013.

[12] D. Hackenberg, R. Schöne, T. Ilsche, D. Molka, J. Schuchart, and R. Geyer. An energy efficiency feature survey of the Intel Haswell processor. In *Proc of the 11th Workshop on High-Performance, Power-Aware Computing*, May 2015.

[13] M. Hähnel, B. Döbel, M. Völp, and H. Härtig. Measuring energy consumption for short code paths using RAPL. In *Proc. Greenmetrics Workshop*, June 2012.

[14] Intel. Beignet. http://www.freedesktop.org/wiki/Software/Beignet/.

[15] Intel. *Intel, Math Kernel Library (MKL)*.

[16] Intel. Voltage regulator-down (vrd) 11.0 processor power delivery design guidelines for desktop lga775 socket. http://www.intel.com/content/dam/doc/design-guide/voltage-regulator-down-11-0-processor-power-delivery-guide.pdf, Nov. 2006.

[17] Intel. *Open Source Intel® HD Graphics Programmer's Reference Manual (PRM) Observability Performance Counters for Intel® Core ™ Processor Family*, 2013.

[18] Intel. *Intel® Xeon® Processor E5-1600 and E5-2600 v3 Product Families, Volume 2 of 2, Register Data Sheet*, June 2015.

[19] Intel Corporation. *Intel® 64 and IA-32 Architectures Software Developer's Manual Volume 3: System Programming Guide*, June 2015.

[20] R. Khanna, F. Zuhayri, M. Nachimuthu, C. Le, and M. Kumar. Unified extensible firmware interface: An innovative approach to DRAM power control. In *Proc. International Conference on Energy Aware Computing*, Nov. 2011.

[21] S. Khoshbakht and N. Dimopoulos. Relating application memory activity to processor power. In *Proc. International Conference on Parallel Processing*, pages 849–857, Oct. 2013.

[22] A. Mazouz, B. Pradelle, and W. Jalby. Statistical validation methodology of CPU power probes. In *Proc. of 1st International Workshop on Reproducibility in Parallel Computing*, Aug. 2014.

[23] J. McCalpin. STREAM: Sustainable memory bandwidth in high performance computers. http://www.cs.virginia.edu/stream/, 1999.

[24] Measurement Computing. *USB-1208FS-Plus Analog and Digital I/O User's Guide*, 2014.

[25] Micron. *2GB, 4GB, 8GB (x64, DR) 240-Pin DDR3 UDIMM Features*, 2008.

[26] P. J. Mucci, S. Browne, C. Deane, and G. Ho. PAPI: A portable interface to hardware performance counters. In *Proc. Department of Defense HPCMP User Group Conference*, June 1999.

[27] M. C. no, S. Catalán, R. Mayo, and E. Quintana-Ortí. Reducing the cost of power monitoring with DC Wattmeters. *Computer Science – Research and Development*, 30(2):107–114, May 2015.

[28] E. Rotem, A. Naveh, D. Rajwan, A. Anathakrishnan, and E. Weissmann. Power-management architecture of the Intel microarchitecture code-named Sandy Bridge. *IEEE Micro*, 32(2):20–27, 2012.

[29] Samsung Electronics. *240pin Unbuffered DIMM based on 4Gb D-die*, Aug. 2013.

[30] SK hynix. *DDR3 SDRAM Unbuffered DIMMs Based on 4Gb A-Die*, 2013.

[31] SK hynix. *DDR4 SDRAM Registered DIMM Based on 4Gb M-Die*, 2013.

[32] Squad. *Kerbal Space Program*.

[33] J. Treibig, G. Hager, and G. Wellein. LIKWID: A lightweight performance-oriented tool suite for x86 multicore environments. In *Proc. of the First International Workshop on Parallel Software Tools and Tool Infrastructures*, Sept. 2010.

[34] V. Weaver. perf_event_open manual page. In M. Kerrisk, editor, *Linux Programmer's Manual*. Dec. 2013.

[35] R. C. Whaley and J. Dongarra. Automatically tuned linear algebra software. In *Proc. of Ninth SIAM Conference on Parallel Processing for Scientific Computing*, 1999.