Power Measurements for Compute Nodes: Improving Sampling Rates, Granularity and Accuracy

Thomas Ilsche*, Daniel Hackenberg*, Stefan Graul†, Robert Schöne*, Joseph Schuchart*
*Center for Information Services and High Performance Computing (ZIH)
Technische Universität Dresden – 01062 Dresden, Germany
Email: {thomas.ilsche, daniel.hackenberg robert.schoene joseph.schuchart}@tu-dresden.de
†Ingenieurbüro Graul – Hoyerswerdaer Str. 29, 01099 Dresden, Germany
Email: info@ingenieurbuero-graul.de

***Abstract*—Energy efficiency is a key optimization goal for software and hardware in the High Performance Computing (HPC) domain. This necessitates sophisticated power measurement capabilities that are characterized by the key criteria (i) high sampling rates, (ii) measurement of individual components, (iii) well-defined accuracy, and (iv) high scalability. In this paper, we tackle the first three of these goals and describe the instrumentation of two high-end compute nodes with three different current measurement techniques: (i) Hall effect sensors, (ii) measuring shunts in extension cables and riser cards, and (iii) tapping into the voltage regulators. The resulting measurement data for components such as sockets, PCIe cards, and DRAM DIMMs is digitized at sampling rates from 7 kSa/s up to 500 kSa/s, enabling a fine-grained correlation between power usage and application events. The accuracy of all elements in the measurement infrastructure is studied carefully. Moreover, potential pitfalls in building custom power instrumentation are discussed. We raise the awareness for the properties of power measurements, as disregarding existing inaccuracies can lead to invalid conclusions regarding energy efficiency.**

I. Introduction

It is now widely accepted that energy efficiency is a key challenge for High Performance Computing. Similar to time-focused performance optimization, it is an iterative process that relies on metric-based feedback. To that end, the consumed energy or average power consumption is measured in addition to the runtime or throughput of a certain task. While it is not trivial to measure a time duration in distributed computer systems correctly, it is even more challenging to measure energy, which comprises current, voltage, and time measurements. Even low-cost real-time clocks have drifts of 10 parts per million [1] (0.001 %), a level of precision that is much more difficult to achieve for power measurements. In contrast, high-quality power meters may achieve 0.1 % uncertainty under restricted conditions [22]. Furthermore, power measurements are often not readily available and incur additional costs.

In no way does this mean that good energy measurements are impossible but they do require trade-offs between temporal and spatial resolution, accuracy, scalability, cost, and convenience. A good temporal resolution is required to detect and understand effects of rapidly changing power consumption, e.g., alternating short code paths with different power characteristics. Typical temporal resolutions for power measurements range from 1 s to 1 ms. A high spatial resolution helps to understand the energy consumption of different components individually rather than the whole system as a black box. The spatial resolution may range from the power input to a server room to individual voltage lanes of a chip. Scalability is another important property for power measurements in HPC systems. It can be challenging to apply a solution to an entire cluster consisting of hundreds or thousands of nodes leading to a high diversity among power measurement approaches.

This work focuses on custom solutions with high sampling rates, individual component measurements, and good accuracy for individual high performance compute nodes. Section II gives a broader overview of existing power measurement approaches. In Section III, the instrumentation of DC power consumers in computer systems using shunts, Hall effect sensors, and voltage regulators is discussed. This is followed by a description of analog and digital processing in Section IV and Section V, respectively. We describe the calibration and evaluate the accuracy of our measurement implementations in Section VI followed by exemplary use cases in Section VII.

II. Related Work

With the growing importance of power limitations and energy efficiency, a variety of interfaces and tools for power instrumentation have been developed within the last years. Figure 1 gives an overview of the connectors and voltage transformations in a typical system and also highlights possible instrumentation points for power measurements.

On a coarse-grained level, power supply units (PSUs) and power distribution units (PDUs) for different server systems provide power readings that can be accesses from the board management controller (BMC) via IPMI. Such measurements are not only coarse-grained in terms of temporal accuracy (mostly with an update rate of 1 Sa/s or less) and spatial granularity (only the whole node can be measured), they also lack accuracy due to IPMI limitations and the provision of instantaneous measurements. Deep analyses of such an instrumentation are provided in [5] and [3].

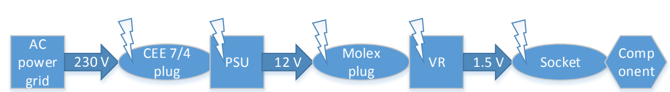


Fig. 1. Overview of possible power transmission and conversion for a single component. Thunderbolts represent possible instrumentation points.

Measurement Type	Temporal Resolution	Spatial Resolution	Accuracy	Scalability	Cost
Shunt at DC	(++) $\approx 70\,\mu s$	(+) per DC plug	(++) $< 1.7\,\%$ (P)	(--)	(-)
Hall effect sensor at DC	(++) $\approx 110\,\mu s$	(+) per DC plug	(+) plausible	(--)	(-)
Voltage regulator	(+++) $\approx 10\,\mu s$	(++) per voltage lane	(-) nonlinearities	(--)	(-)
LMG450 at AC	(-) 50 ms	(-) entire system	(++) $< 0.07\,\% + 0.04\,\%$ (P)	(-)	(o)
PowerPack (shunts) [4]	(+) $\ll 1$ s	(+) DC components	(+) "verified"	(+)	(o)
PowerMon2 [2]	(+) ≈ 1 ms	(+) DC components	(o) $< 6.8\,\%$ (I)	(+)	(+)
PowerInsight [16]	(+) ≈ 1 ms	(+) DC components	(o) avg. 1.8 % (I)	(+)	(o)
HDEEM [6]	(+) 1 ms / 10 ms	(++) node / voltage lanes	(++) $< 2\,\%$ / 3 % (P)	(++)	(+)
PDU (typical) [5]	(--) 1 s	(-) entire system	(-) instantaneous	(++)	(+)
AMD's APM [5]	(o) 10 ms (perturbation)	(+) per socket	(-) systematic errors	(+)	(++)
Intel's RAPL Sandy Bridge [5]	(o) 1 ms (perturbation)	(++) cores, memory per socket	(-) systematic errors	(+)	(++)
Intel's RAPL Haswell [7]	(o) 1 ms (perturbation)	(++) cores, memory per socket	(+) no systematic errors	(+)	(++)

TABLE I. CLASSIFICATION OF DIFFERENT ENERGY MEASUREMENT APPROACHES.

A more accurate measurement on a node level can be based on certified and calibrated power meters attached to the PSU, e.g., as required to compare the energy efficiency of computing platforms using benchmarks like SPEC power_ssj2008 [15] or SPEC OMP2012 [17]. We use a calibrated ZES ZIMMER LMG450 for reference measurements. This power meter has a specified uncertainty of 0.07 % + 0.25 W[1] and provides average real power values for well-defined time slices. Still, the low external readout rate of 20 Sa/s and the low spatial granularity (whole system only) limits the application in a fine-grained measurement scenario. Internally, it samples the voltage and current at a much higher rate to achieve its accuracy.

Another non-intrusive option is to rely on energy models and interfaces provided by vendors. Intel's RAPL [19], AMD's APM [11], or performance counter based models [21] provide an estimate of the power consumption of processors, thus increasing the spatial accuracy to the processor level. They also provide a higher update rate (e.g. 1 ms on RAPL and 10 ms on APM) than the previously discussed approaches. However, these models also lack accuracy as described in [5]. Although the RAPL energy counters are updated in 1 ms intervals, a readout in such small intervals is problematic. First, readouts happen on the system under test and thus perturb the measurement. Second, the provided energy values lack any time information. Assuming the readout time to be the update time, inaccuracies occur when attributing the energy to application events or computing average power, especially when the values are read frequently. Additionally, performance counter based models have to be trained for a specific processor instance, as different equally labeled processors may provide different energy profiles due to process variation. Moreover, such models have to be re-evaluated regularly due to aging effects. In contrast to previous implementations, RAPL in the Intel Haswell architecture uses physical measurements that provide a much better accuracy and no more bias towards certain workloads [7].

Fortunately, HPC hardware vendors are starting to recognize the growing need for power measurements and offer convenient and relatively low-cost solutions. Examples are NVIDIA [18], Cray [8], IBM [13], and BULL [6] who now provide software interfaces to gather information from pre-instrumented components. However, these interfaces are usually closed source and information on accuracy and internal specifics like used filters are rarely documented publicly.

[1] With a measuring range of 250 V and 2.5 A [22]

To overcome the obstacles of all these solutions, low level instrumentation frameworks have been developed. PowerPack [4] is a hardware and software framework that can access various types of sensors. In a typical implementation, it uses resistors added to several DC pins and a National Instruments input module. A redundant set of measurements enables accuracy verification. PowerInsight [16] is a solution that is commercially available. It uses sensor modules as Molex adapters and riser cards that are equipped with small Hall effect sensors. Unfortunately, only average errors for current measurement (1.8 %) and voltage measurement (0.3 %) are reported. We assume that cost or size limitations have implied the choice of rather inaccurate Hall effect sensors and only 10-bit analog-digital converters. The sampling rate is reported to be limited to ≈ 1 kSa/s by software overhead. PowerMon2 [2] is a low-cost power monitoring device for commodity computer systems with a measurement rate of up to 1024 Sa/s. It measures up to 8 DC channels with a measurement resistor and a digital power monitor chip that contains both a current sense amplifier and an analog/digital converter. The accuracy is reported as $\pm 0.9\,\%$ for voltage and -6.6 % / +6.8 % (worst-case) for current.

While we employ similar techniques like these measurement frameworks, we focus on pushing the boundaries in terms of sampling rates, while ensuring a verified and high accuracy measurement setup as well as maintaining a good spatial resolution. To that end, we make concessions regarding cost, size, and therefore scalability. Table I puts our efforts into perspective with the other approaches described in this section.

III. DIRECT CURRENT INSTRUMENTATION METHODS

There are various options to instrument the current supply of system components. For those that are connected with a standard Molex connector, an intermediate connector can be used. Components that utilize slots such as DIMMs or PCIe can be measured by using riser cards that provide hooks for measurement probes in the current connectors. This makes the measurement probe modular so that it can be used in different systems under test. Proprietary adapters or hard-wired power supplies require a more intrusive approach. While CPU socket power consumption would be a valuable information, there is no feasible approach to directly instrument the CPU voltage input without major efforts that are unaffordable for typical scientific purposes. Therefore, CPU power consumption has to be instrumented on the mainboard, at the CPU voltage regulators (VRs), or at the input to the CPU VRs.

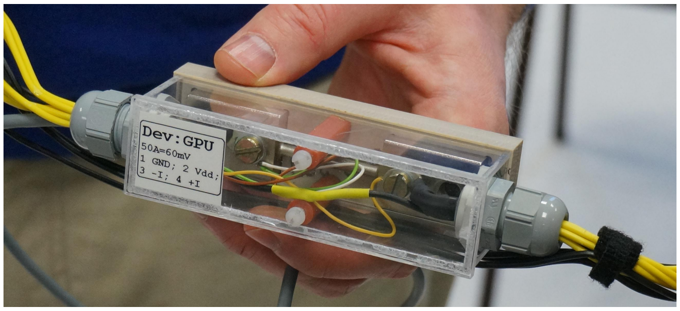


Fig. 2. Measurement shunt inside adapter for 6-pin Molex power connector

We evaluate two common current measurement techniques. Measurement shunts are well-defined resistors that measure current by causing a specific voltage drop. The resistance of a shunt needs to be dimensioned so that the measured components are not affected by the drop of their input voltage. As an alternative, Hall effect sensors use the magnetic field of a current which does not require to tap the current flow and therefore is less intrusive. They also provide a galvanic separation between the measured and the measurement system. Unlike shunts, Hall effect sensors are active sensors requiring a supply voltage independent of the measured line. Their signal can be strong enough to be used without further amplification. Unfortunately, the specific frequency response of Hall effect sensors significantly limits their applicability in the presence of high frequency load swings.

The measurement setup needs to consider that DC components in computer systems draw dynamically regulated currents that can resemble arbitrary patterns based on the (computational) workload of the component or the behavior of intermediate VRs. In addition to the current, the voltage level needs to be measured as well. Assuming a fixed voltage (e.g., 12 V) introduces inaccuracies due to variations in the voltage supply. For instance, the ATX specification allows for up to 5 % variation of voltage [9]. Voltage measurements are usually straight-forward by using data acquisition hardware directly.

In *system A*, we instrumented the output of the power supply unit (PSU) by cutting the cables and rerouting them through Hall effect transducers[2]. This provides information about the 12 V, 5 V, and 3.3 V input into the mainboard. The majority of power is consumed through the 12 V lane, the other lanes have low or almost constant power demand.

An alternative approach to measure closer to the components is to utilize the on-board voltage regulators. These voltage regulators convert the voltage from the PSU (e.g., 12 V) to the respective voltage required by the component (e.g., 1.5 V for DDR3). The involved chips perform a measurement of current using the voltage drop over an inductor. Multiple phase ICs amplify this signal and provide a shared output of the total current to the control IC for voltage positioning. In *system A* we use this summary signal to read the current for each voltage lane of the sockets. The calculation is done by using the formulas given in the datasheet [10]. It has to be noted that the measurements by the voltage regulators are usually designed for a specific operating point and therefore are not necessarily accurate or linear for a wider range of currents. Also, these measurements cover the actual energy used by the consumer, but not the losses of the voltage regulator itself. These losses may be variable depending on the load characteristic, e.g., frequency of load swings.

In contrast, for *system B* we use modular Molex adapters at all PSU outputs, i.e.:

- Two 8-pin connectors supplying power to each of the sockets (including CPU and memory)
- An ATX connector with the 12 V, 5 V, and 3.3 V lines
- Two 6-pin connectors, coupled to a single measurement probe, as external power supply for a GPU card
- A SATA connector with the 12 V and 5 V line
- One 4-pin connector as power supply for all fans

In addition we use two instrumented riser cards for PCIe (12 V and 3 V) as well as a DDR3 DIMM for one memory module. All measurement probes in *system B* use shunt resistors. Figure 2 shows the shunt casing of the GPU Molex adapter. The size of the resistor is necessary to avoid heat transformation due to the large current draws of modern GPUs. Table II summarizes the specifications and instrumentation of our two measurement systems.

IV. Analog Processing and Data Acquisition

The signal from the probes and sensors is processed in three analog steps:

- Amplification into a common voltage range
- Analog low-pass filtering
- Data acquisition (analog/digital conversion)

The signal from current measurement shunts (voltage drop) is usually in the range of millivolts, while the signal from the voltage measurement is > 1 V. Consequently, the signals need to be amplified into a common range to allow a high resolution A/D conversion with a single data acquisition card. We use instrumentation amplifiers[3] with low distortion and high precision. Their programmable gain allows us to calibrate each channel using predefined factors between 0.5 and 500. All current signals are amplified differentially.

[2]LEM LA 100-TP for 12 V and LEM HXS 20-NP for 5 V and 3.3 V lanes

[3]Linear Technology LT1167

	system A	*system B*
Processors	3 × Opteron 6274	2 × Intel Xeon E5-2690
Cores	48	16 (32 threads)
Memory	48 GB DDR3	64 GB DDR3
Instrumentation	all processor voltage regulators (cores, northbridge, RAM)	2 × 12 V input per socket (CPU & RAM, shunt)
	12 V, 5 V, 3.3 V board input (Hall effect sensor)	All other DC Molex plugs (shunt), PCIe, 1 × DDR3 (riser, shunt)
	AC input via LMG450	AC input via LMG450

TABLE II. Specifications of measurement systems

To further condition the signal for A/D conversion, we apply low-pass filters (in addition to the low-pass behavior of the amplifiers) to remove high frequencies that cannot be sampled correctly by data acquisition. The dimensioning of filters depends on the available sampling rate and the frequencies in the signal (variation that is of interest versus noise and effects that are not in the focus of the measurement).

We use two National Instruments data acquisition cards. One NI PCI-6255 can capture up to 80 input signals at an aggregate sampling rate of 1.25 MSa/s. This allows us to use up to 40 measurement points (each with voltage and current) sampled with up to 15 kSa/s. Due to the multiplexing in this card, the amplifiers have to build up the charge within the time between two samples. When utilizing the maximum sampling rate, this is only 800 ns and can lead to cross-talk between different signals. We therefore use a lower sampling rate of 7 kSa/s per channel. It is also possible to reduce the number of actively measured channels and increase the sampling rate. An additional NI PCI-6123 provides 8 inputs sampled simultaneously at 500 kSa/s. This provides an even more detailed view on up to four selected measurement points. The two DC socket measurements and the DDR3 riser measurement are a suitable target for this high resolution. The amplified signals use a common ground plane that is connected to the measurement system ground. A differential data acquisition would require twice as many analog inputs on the NI PCI-6255, whereas the NI PCI-6123 always measures all inputs differentially.

V. Digital Processing, Storage and Analysis

The large amount of generated data makes digital processing challenging, as does the variety of use-cases, such as:

- Correlating application events with full-resolution power measurements.
- Recording total energy consumption of different components for multiple experiments.
- Analyzing long-term data.

Initial testing of the National Instruments data acquisition can be done using LabVIEW, which is a graphical environment that features a range of virtual instruments. However, it is not suitable for long term recording or to correlate the recordings with application events. Therefore, we implemented a data acquisition daemon that can run continuously. Initially, it converts the input signals to the actual measurement voltages, currents and power values. Clients can connect to the daemon via network and request the recording of a selection of channels. While the sampling rate is defined by the daemon configuration, the client can specify a temporal aggregation to reduce the overall data rate. During the experiment, the data is then stored in the daemon's memory. Afterwards, the collected data is transferred to the client for further processing.

This workflow allows for an unperturbed experiment without data processing at the client. The C++ client library supports additional digital filters for noise reduction at high sampling rates as well as multiplexing readings from several data acquisition cards. To correlate the power measurement data with application events, the tracing infrastructure Score-P [14] is supported via metric plugins [20] that connect to the daemon and integrate the power measurements into the application trace after the execution.

One of the most challenging aspect is the synchronization of timestamps from the data acquisition system and the system under test. Considering that measurement values are only valid for intervals of $100\,\mu s$ and shorter, the accuracy of NTP synchronized clocks is not sufficient. Precise GPS clocks would be an option but require additional hardware. In our implementation, the metric plugin runs a synthetic load pattern on the system under test at the beginning and end of the experiment. This pattern is detected in the power measurement series which results in two pairs of timestamps from the data acquisition system and system under test that is used as baseline for a linear interpolation to translate all power measurement timestamps to fit in the application measurement. In practice, this synchronization is usually accurate to $50\,\mu$s. Measurements at 500 kSa/s may require additional manual realignment.

In addition to serving clients, the daemon also sends a continuous stream of aggregated measurement values to a persistent storage infrastructure [12]. It is configured to reduce the data to 20 average values per second. While the storage infrastructure is not suited for handling raw data with more than 1 kSa/s, it provides a rich set of tools and APIs to analyze the measurements when a high time resolution is not required. Furthermore, a web-based GUI allows to visualize past measurements as well as live monitoring.

VI. Measurements Calibration and Verification

A. Calibration

We calibrate our measurement setup to achieve good accuracy. A signal generator is used as input to the amplifiers, and two calibrated voltmeters measure the input and output of each amplifier. The calibration factor is an 8-bit number, resulting in inaccuracies around $\pm 0.21\,\%$ for each amplifier. The measurement shunts are calibrated using a large sliding resistor as constant load, again measuring the voltage drop and current with calibrated voltmeters. We have observed up to 10 % deviation from the specification of our measuring resistances. Considering small resistances of down to $1.2\,\mathrm{m}\Omega$, these can likely be additional contact resistances. The accurate values of the measurement resistances are important for computing the current from the measured voltage drop correctly.

B. Verification technique

We run a set of micro-kernels at different thread configurations and CPU frequencies to generate a variety of workload points to compare our measurement with the reference measurement. The set includes kernels with both static and dynamic power consumptions (see [5]). All load configurations are run for 10 seconds of which the average power consumption of the inner 9 seconds is used. This hides parts of the measurement noise but is necessary to avoid errors due to difference in readout rates and time synchronization between the fine-grained measurement and the reference measurement. We focus on the CPUs and main memory due to their variability and large fraction of the total power consumption. During our verification, there was no dedicated GPU installed in *system B*.

For *system A*, we compare the VR measurements with the Hall effect sensor measurements at the 12 V board input, which should correlate closely. In *system B*, we measure the

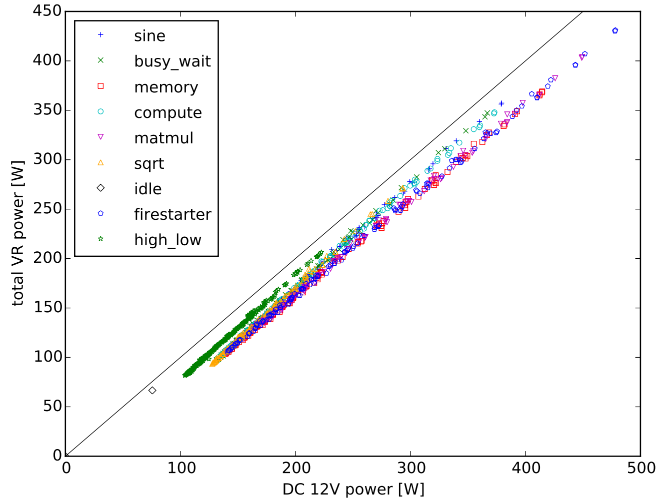


Fig. 3. Measurement points of the 12 V board input measurement compared to the sum of all voltage regulator measurements on *system A* under different load characteristics.

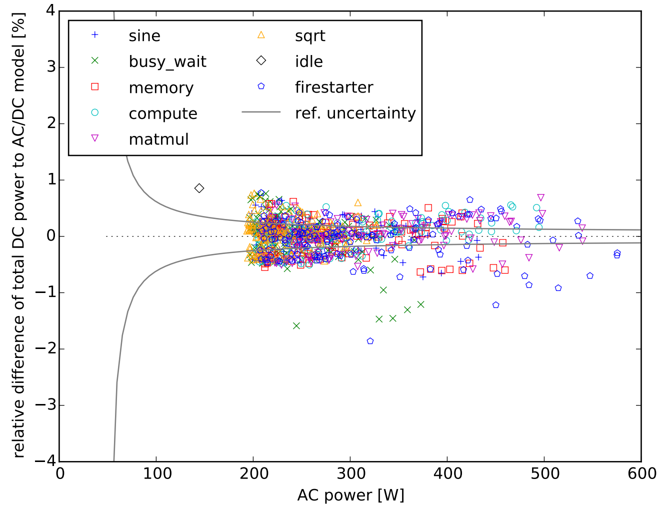


Fig. 4. Relative difference of the sum of board input Hall effect measurements compared to the quadratic PSU model of the AC reference measurement on *system A*. Includes the uncertainty bound of the reference measurement.

per-socket DC power consumption with a calibrated reference power meter (LMG450) and additional connectors. In addition, we correlate the sum of DC measurements of both systems with AC measurements using the reference power meter.

C. Verification of measurements in system A

1) VR measurements versus 12 V board input: We directly compare the power measured by the 12 V board instrumentation and the sum of all voltage regulator measurements. There is at least one additional chip on the board supplied by the 12 V input that we cannot measure and have to assume to have a negligible or constant power consumption. Moreover, the VR measurements only cover the consumption of the supplied components, not the VR losses. However, we would expect the two measurements to correlate closely. Unfortunately, the results shown in Figure 3 reveal that the measurement points do not map well from VR power to 12 V power. It is especially noteworthy that different workloads, that stress different VRs unevenly, have distinct characteristics. This cannot be fully attributed to different VRs having varying efficiencies. We were unable to build a plausible model that would map VR power correctly to 12 V power. This may be due to the errors in the VR current measurement, as those are not originally designed to provide precise linear measurements but rather accurately measure specific points. As a consequence, we do not advise to consider these measurements as correct absolute numbers or when comparing voltage lanes. They can still be useful for understanding the relative effect of an algorithmic or configuration change on a single voltage lane.

2) Total board input versus AC reference measurement: We have more confidence in the Hall effect based DC board input measurements. The current sensor itself has a specified uncertainty of 0.45 % for the 12 V measurement that is dominating the power consumption. The total uncertainty is also affected by the amplifier calibration and data acquisition uncertainty, both of which applies to current and voltage measurement. We thus compare the resulting sum of 12 V, 5 V and 3.3 V DC power measurements with the AC reference measurement. These measurements correlate strongly, resulting in a plausible quadratic model[4] that accounts for the losses in the PSU. While the model may hide some calibration issues, the difference can still reveal bias errors of our measurement. Note that we have excluded those workloads with dynamic power consumption because they show additional variance of the PSU efficiency and power factor that would require separate investigation and discussion outside the scope of this paper. Figure 4 shows the relative difference between the measured DC power and this model for our set of verification measurements. The figure also shows the uncertainty bound of the reference measurement, meaning that any deviation inside this bound can stem from either of the measurements. This does not mean that all values should be within that bound—it merely shows that the reference measurement is not perfect and gives an impression of its influence on our results. The plot reveals no systematic errors towards a certain workload and the remaining variation is below 2 %. It is impossible to determine whether the variation stems from the measurement or variable consumption of the fans in the system[5].

D. Verification of measurements in system B

1) 12 V DC measurements: For the verification of the 12 V socket shunt measurements in *system B*, additional adapters are inserted between the PSU and our custom shunt measurement adapters for each socket. With this arrangement, we measure the same power domains with a reference measurement and our custom shunt measurement simultaneously (see Figure 5).

Figure 6 shows that our measurement is always within 1.7 % of the reference measurement. The absolute error is also <2.3 W, which is almost completely within the uncertainty of the reference measurement. This is a stronger verification than the one shown in Figure 4, as no model is applied that may hide systematic errors. We therefore have high confidence in the DC shunt measurements.

[4] $P_{AC} = 0.00011 * P_{DC}^2/W + 1.0015 * P_{DC} + 48.3W$

[5] Fanspeed in *system A* is configured at constant maximum setting. Fans are supplied by the PSU directly and not included in DC measurements.

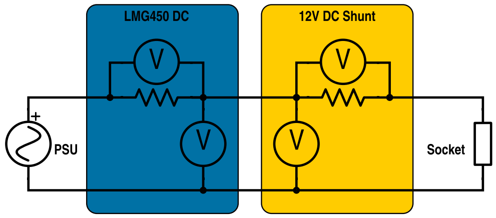


Fig. 5. Schematic diagram of the verification setup for *system B*.

2) Total DC measurements versus AC reference measurement: In *system B*, we measure all DC consumers, except for the fan in the PSU itself. This allows us to compare the total DC power consumption with the AC power consumption. Similarly to *system A*, we apply a quadratic model[6] of the PSU efficiency based on regression on our comparison measurement. The relative difference between total DC measurements and the modeled DC power consumption based on the AC measurement is shown in Figure 7. The remaining noise is less than 1 % in this case and as expected there are no systematic errors based on the selected workload.

In such a complex setup, verification and calibration is an ongoing process. Just like any other electrical measurement device, calibration should be done regularly, e.g. in 12 month intervals. This aspect is often overlooked, especially in scalable solutions. It is unrealistic to hope that a deployed calibrated measurement system within a large HPC system remains calibrated over the lifetime of the system, e.g. several years. In our setup, we keep a close look on any irregularities occurring during normal operation. This does happen in practice; over the course of 3 years, two amplifiers failed, tight screws in solid copper blocks loosened over time, and 50 Hz signals suddenly appeared on the reference ground plane. This is also the consequence of our highly customized and complex equipment that enables energy efficiency studies that would be impossible to perform in a simpler setup.

[6] $P_{AC} = 0.00026 * P_{DC}^2/W + 0.99988 * P_{DC} + 14.7W$

VII. Application Power Traces

The fine-grained measurements detailed above can be used for a wide variety of different energy efficiency analyses. Due to its high spatial and temporal resolution, these measurements can be employed to demonstrate effects that are not visible with less detailed measurement solutions. We demonstrate one use case to highlight the importance of this fine-grained measurement. Figure 8 depicts a Vampir screenshot of application traces of two runs of the SPEC OMP 371.applu benchmark on *system B* that only differ in the `KMP_BLOCKTIME` setting, which was either left at its default (200 ms, blue background) or set to zero (white background). While the traces show a slight performance advantage of the parallel region between the two barriers (dark blue parallel `do`) in the default case, it also demonstrates the difference of the power consumption over time between the two runs.

Disabling the thread blocktime leads to immediate sleeping of threads in a barrier, thus allowing the processor to enter a sleep state. This is demonstrated for socket 0 where it appears that all threads on that socket finish the first parallel `do` loop at least 50 ms earlier than the threads running on socket 1 before entering a barrier to perform a global thread synchronization. In the default case, power consumption drops slightly by ≈30 W since busy wait consumes less power than heavy computation. However, with blocktime disabled, all threads enter a sleep state, allowing the processor to enter a deep C-state and dropping power consumption by about 100 W. A similar behavior can be observed for the second parallel `do` loop (dark blue), where the early arrival of some of the threads running on socket 1 in a barrier has a notable impact on power consumption of this socket if the blocktime is disabled.

At about 270 ms into the depicted time frame, a spike in power consumption is visible for the narrow green band in the timeline. This area is detailed in Figure 9 and shows a parallel region (green) in which all threads are active before a new parallel region is started (brown), in which the thread activities start with different offsets, hence leading to another drop and

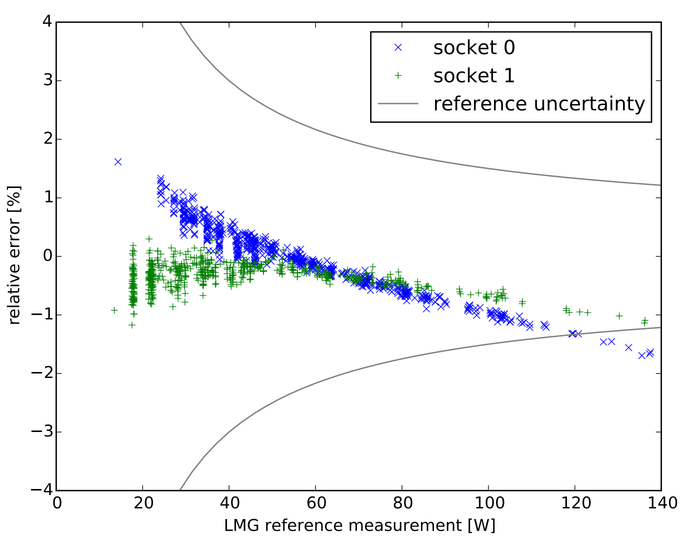


Fig. 6. Relative difference of the 12 V per-socket shunt measurements compared to 12 V per-socket reference measurements on *system B*. Includes the uncertainty bound of the reference measurements.

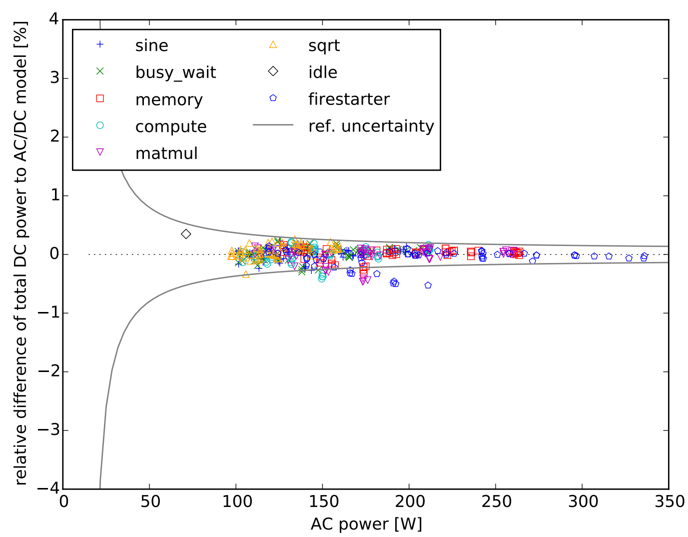


Fig. 7. Relative difference of the total DC shunt measurements compared to a PSU model of the AC reference measurement on *system B*. Includes the uncertainty bound of the reference measurement.

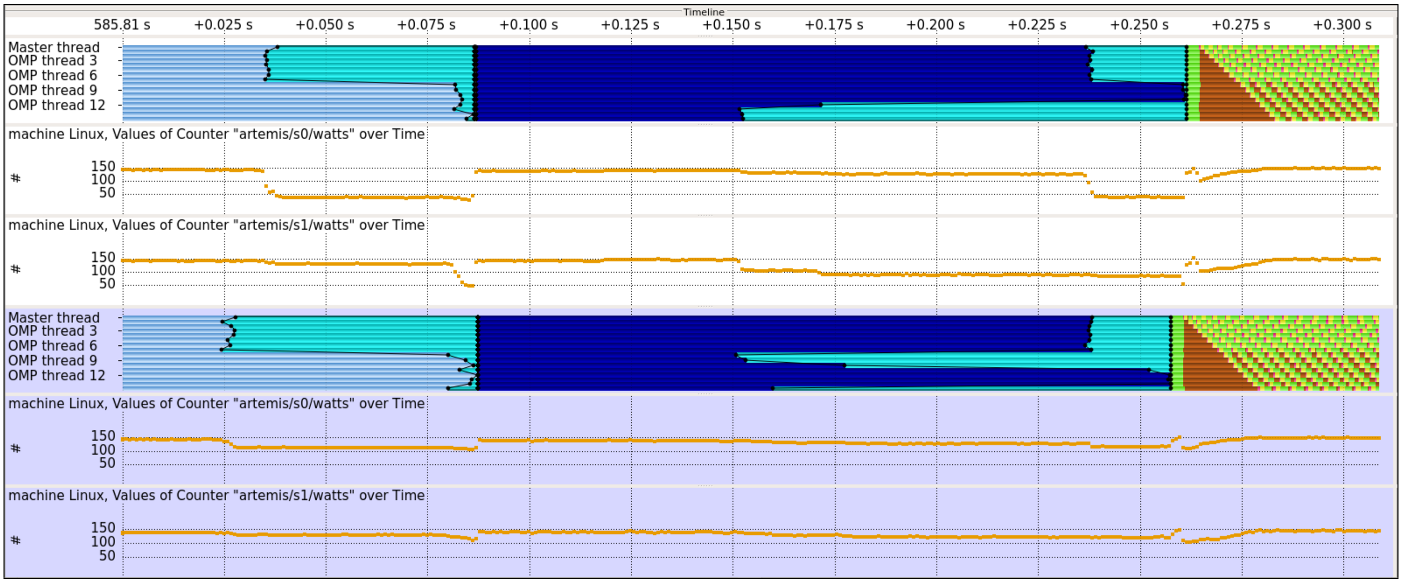


Fig. 8. Screenshot of Vampir comparing a section of two identical runs of SPEC OMP 371.applu on *system B* with different settings for `KMP_BLOCKTIME` (white background: set to zero; blue background: default) showing the timeline of the 16 threads each and the power measurements for sockets 0 and 1. OpenMP regions from left to right: parallel `for` (light blue), implicit barrier (cyan), parallel `for` (dark blue), implicit barrier, parallel region (green), barrier (brown).

slow increase of power consumption over about 15 ms. The measurements are able to provide enough temporal accuracy to allow for a detailed analysis of even such short regions.

The value of measurements with a sampling rate of 500 kSa/s is demonstrated with a synthetic program that changes its load in very short intervals. Figure 10a depicts the execution of such a synthetic workload on *system A* alongside the VR socket 0 core measurements. This shows that it is still possible to clearly identify different regions of code having distinct power consumptions at scales of $\approx 10\,\mu s$. For even shorter load changes, the amplitude decreases as a result of a low-pass characteristic in the system. A similar measurement executed on *system B* using the socket 12 V DC shunts is displayed in Figure 10b. In this case, regions of $\approx 70\,\mu s$ can be observed without significant amplitude drop. The low-pass effect is stronger before the voltage regulator than in the previous measurement. Similar to the *system A* 12 V Hall effect sensor, regions of $\approx 120\,\mu s$ length can be observed without dampening. Considering the much higher bandwidth of the Hall effect sensor (200 kHz, -1 dB), this also likely reflects the actual change in power consumption and not an effect of our measurement system.

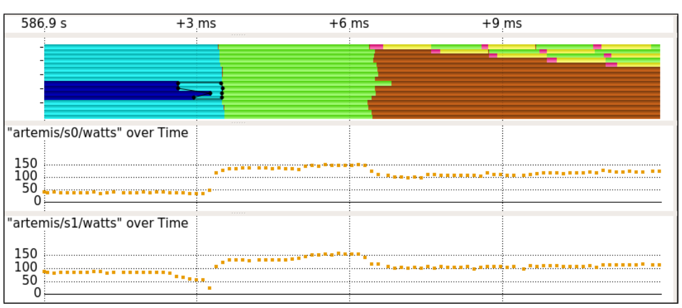
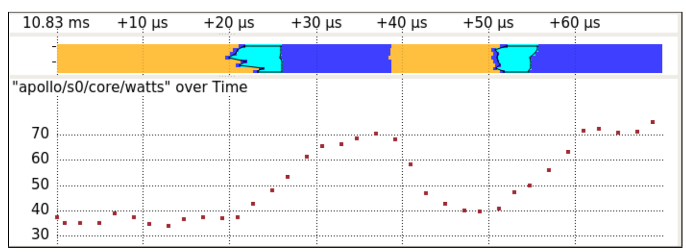
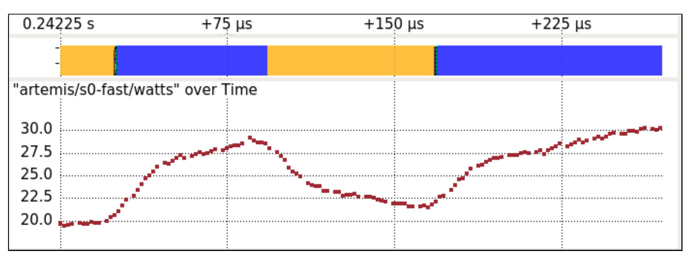


Fig. 9. Screenshot of Vampir of a section of the SPEC OMP 371.applu benchmark detailing the spike in power consumption at +0.260 s in Figure 8.

These experiments show the limits of the actually achievable temporal granularity at the specific measurement points. Further increasing the sampling rate would not reveal more details of power variance from processor operations but instead reflect high-frequency effects of the voltage regulator operation. Nevertheless, it is possible to trace the execution of a parallel application at those time scales. When combined, this presents a unequaled tool to understand the energy consumption of an application, not only for manual optimizations, but also for building accurate energy models of very short application functions.



(a) 12 μs load changes and the core power consumption (VR) on *system A*



(b) 70 μs load changes and the socket power consumption on *system B*

Fig. 10. Screenshots of Vampir displaying a synthetic workload and the power consumption measured with 500 kSa/s. Low load (sqrt): orange, thread synchronization: cyan, High load (compute): dark blue.

VIII. Conclusion and Future Work

This paper describes possible approaches to measure the power consumption of high performance compute nodes. We discuss limitations of related work and present a custom approach for per-component measurements that pushes the limits regarding temporal resolution while providing high accuracy and thorough verification. This solution is rather costly and limited in scalability, but provides power consumption details for application regions with runtimes in the order of only tens of microseconds. Our experiments and verification show that measurements at the voltage regulators provide the best temporal and spatial resolution, but suffer from limited accuracy. Measurement probes inserted at the DC input of the mainboard are slightly more coarse-grained, but reveal power consumption details in the order of $100\,\mu$s. For this, Molex adapters can be used to build a modular measurement infrastructure. Both shunts and Hall effect sensors can be accurate. We tend to prefer shunts for their non-distorted frequency response, but they do require good amplifiers and calibration.

Our application traces with power consumption metrics show that this novel infrastructure can enable to a deeper understanding of how systems and applications use energy. While our work is highly customized and complex, the experiences presented can help building similarly powerful measurement setups. With vendors recognizing the growing importance of this topic, such detailed measurements should be more easily accessible in the future.

While we put the focus on CPU and socket power consumption in this paper, the full range of DC instrumentation allows for a broader view on power consumption in compute nodes. PCIe instrumentation enables high resolution measurements for GPUs and network cards. Other upcoming topics are the power consumption of disk I/O as well as PSU efficiencies. Moreover, DDR3 riser cards help to separate the power consumptions of memory and CPUs. In future work we plan to combine power consumption recordings with system events such as interrupts or processor state changes to gain a more profound understanding of these features. The measurement system will be used to bolster a wide area of research in energy-efficient computing, both from the application as well as the system point of view.

Acknowledgements

This work is supported in parts by the German Research Foundation (DFG) in the Collaborative Research Center 912 "Highly Adaptive Energy-Efficient Computing" and the Bundesministerium für Bildung und Forschung via the research project Score-E (BMBF 01IH13001). The authors would like to thank Mario Bielert and Maik Schmidt for their support.

References

[1] Atmel: AT03155: Real-Time-Clock Calibration and Compensation, http://www.atmel.com/Images/Atmel-42251-RTC-Calibration-and-Compensation_AP-Note_AT03155.pdf

[2] Bedard, D., Lim, M.Y., Fowler, R., Porterfield, A.: Powermon: Fine-grained and integrated power monitoring for commodity computer systems. In: IEEE SoutheastCon (2010), DOI: 10.1109/SECON.2010.5453824

[3] Diouri, Mohammed E. M. et.al.: Solving Some Mysteries in Power Monitoring of Servers: Take Care of Your Wattmeters! In: Energy Efficiency in Large Scale Distributed Systems (2013), DOI: 10.1007/978-3-642-40517-4_1

[4] Ge, R., Feng, X., Song, S., Chang, H.C., Li, D., Cameron, K.W.: PowerPack: Energy Profiling and Analysis of High-Performance Systems and Applications. IEEE Transactions on Parallel and Distributed Systems (TPDS) (2010), DOI: 10.1109/TPDS.2009.76

[5] Hackenberg, D., Ilsche, T., Schöne, R., Molka, D., Schmidt, M., Nagel, W.E.: Power measurement techniques on standard compute nodes: A quantitative comparison. In: International Symposium on Performance Analysis of Systems and Software (ISPASS) (2013), DOI: 10.1109/ISPASS.2013.6557170

[6] Hackenberg, D., Ilsche, T., Schuchart, J., Schöne, R., Nagel, W.E., Simon, M., Georgiou, Y.: HDEEM: High Definition Energy Efficiency Monitoring. In: International Workshop on Energy Efficient Supercomputing (E2SC) (2014), DOI: 10.1109/E2SC.2014.13

[7] Hackenberg, D., Schöne, R., Ilsche, T., Molka, D., Schuchart, J., Geyer, R.: An Energy Efficiency Feature Survey of the Intel Haswell Processor. In: International Parallel and Distributed Processing Symposium Workshops (IPDPS) (accepted) (2015)

[8] Hart, A., Richardson, H., Doleschal, J., Ilsche, T., Bielert, M., Kappel, M.: User-level Power Monitoring and Application Performance on Cray XC30 supercomputers. Cray User Group CUG (2014), https://cug.org/proceedings/cug2014_proceedings/includes/files/pap136.pdf

[9] Intel Corporation: ATX Specification - Version 2.2 (2004), http://www.formfactors.org/developer\specs\atx2_2.PDF

[10] International Rectifier: IR3529 DATA SHEET, TMXPHASE3 PHASE IC, http://www.irf.com/product-info/datasheets/data/ir3521mpbf.pdf

[11] Jotwani, R., Sundaram, S., Kosonocky, S., Schaefer, A., Andrade, V., Constant, G., Novak, A., Naffziger, S.: An x86-64 core implemented in 32nm SOI CMOS. In: IEEE International Solid-State Circuits Conference (ISSCC) (2010), DOI: 10.1109/ISSCC.2010.5434076

[12] Kluge, M., Hackenberg, D., Nagel, W.E.: Collecting Distributed Performance Data with Dataheap: Generating and Exploiting a Holistic System View. Procedia Computer Science (2012), DOI: 10.1016/j.procs.2012.04.215

[13] Knobloch, M., Foszczynski, M., Homberg, W., Pleiter, D., Böttiger, H.: Mapping fine-grained power measurements to HPC application runtime characteristics on IBM POWER7. Computer Science - Research and Development (2013), DOI: 10.1007/s00450-013-0245-5

[14] Knüpfer, Andreas et.al.: Score-P: A Joint Performance Measurement Run-Time Infrastructure for Periscope, Scalasca, TAU, and Vampir. In: Tools for High Performance Computing (2012), DOI: 10.1007/978-3-642-31476-6_7

[15] Lange, K.D.: Identifying Shades of Green: The SPECpower Benchmarks. Computer (2009), DOI: 10.1109/MC.2009.84

[16] Laros III, J.H., Pokorny, P., Debonis, D.: Powerinsight - A commodity power measurement capability. In: International Green Computing Conference (IGCC) (2013), DOI: 10.1109/IGCC.2013.6604485

[17] Müller, Matthias S. et.al.: SPEC OMP2012 - An Application Benchmark Suite for Parallel Systems Using OpenMP. In: OpenMP in a Heterogeneous World (2012), DOI: 10.1007/978-3-642-30961-8_17

[18] NVIDIA: NVML Reference Manual v340, http://docs.nvidia.com/deploy/pdf/NVML_API_Reference_Guide.pdf

[19] Rotem, E., Naveh, A., Rajwan, D., Ananthakrishnan, A., Weissmann, E.: Power-Management Architecture of the Intel Microarchitecture Code-Named Sandy Bridge. IEEE Micro (2012), DOI: 10.1109/MM.2012.12

[20] Schöne, R., Tschüter, R., Hackenberg, D., Ilsche, T.: The VampirTrace Plugin Counter Interface: Introduction and Examples. In: EuroPar Parallel Processing Workshops (2010), DOI: 10.1007/978-3-642-21878-1_62

[21] Singh, K., Bhadauria, M., McKee, S.A.: Real time power estimation and thread scheduling via performance counters. Computer Architecture News (2009), DOI: 10.1145/1577129.1577137

[22] ZES ZIMMER Electronic Systems GmbH: 4 Channel Power Meter LMG450 User manual (2011)