ALEA: A Fine-Grained Energy Profiling Tool

LEV MUKHANOV, Queen's University of Belfast
PAVLOS PETOUMENOS, University of Edinburgh
ZHENG WANG, Lancaster University
NIKOS PARASYRIS, DIMITRIOS S. NIKOLOPOULOS, and BRONIS R. DE SUPINSKI, Queen's University of Belfast
HUGH LEATHER, University of Edinburgh

Energy efficiency is becoming increasingly important, yet few developers understand how source code changes affect the energy and power consumption of their programs. To enable them to achieve energy savings, we must associate energy consumption with software structures, especially at the fine-grained level of functions and loops. Most research in the field relies on direct power/energy measurements taken from on-board sensors or performance counters. However, this coarse granularity does not directly provide the needed fine-grained measurements. This article presents ALEA, a novel fine-grained energy profiling tool based on probabilistic analysis for fine-grained energy accounting. ALEA overcomes the limitations of coarse-grained power-sensing instruments to associate energy information effectively with source code at a fine-grained level. We demonstrate and validate that ALEA can perform accurate energy profiling at various granularity levels on two different architectures: Intel Sandy Bridge and ARM big.LITTLE. ALEA achieves a worst-case error of only 2% for coarse-grained code structures and 6% for fine-grained ones, with less than 1% runtime overhead. Our use cases demonstrate that ALEA supports energy optimizations, with energy savings of up to 2.87 times for a latency-critical option pricing workload under a given power budget.

CCS Concepts: • **Hardware → Platform power issues**; • **Software and its engineering → Software notations and tools;**

Additional Key Words and Phrases: Energy profiling, sampling, energy efficiency, power measurement, ALEA

ACM Reference Format:
Lev Mukhanov, Pavlos Petoumenos, Zheng Wang, Nikos Parasyris, Dimitrios S. Nikolopoulos, Bronis R. de Supinski, and Hugh Leather. 2017. ALEA: A fine-grained energy profiling tool. ACM Trans. Archit. Code Optim. 14, 1, Article 1 (March 2017), 25 pages.
DOI: http://dx.doi.org/10.1145/3050436

This article is an extension of a conference paper. The preliminary version, "ALEA: Fine-Grain Energy Profiling with Basic Block Sampling" by Lev Mukhanov, Dimitrios S. Nikolopoulos, and Bronis R. de Supinski, appeared at the 2015 International Conference on Parallel Architectures and Compilation Techniques (PACT'15).
This research was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through grant agreements EP/L000055/1 (ALEA), EP/L004232/1 (ENPOWER), EP/M01567X/1 (SANDeRs), EP/M015823/1, EP/M015793/1 (DIVIDEND), and EP/K017594/1 (GEMSCLAIM), and by the European Commission under the Seventh Framework Programme through grant agreements FP7-610509 (NanoStreams) and FP7-323872 (SCORPIO).
Authors' addresses: L. Mukhanov; email: l.mukhanov@qub.ac.uk; P. Petoumenos; email: ppetoume@inf.ed.ac.uk; Z. Wang; email: z.wang@lancaster.ac.uk; N. Parasyris; email: nik.parasyr@gmail.com; D. S. Nikolopoulos; email: d.nikolopoulos@qub.ac.uk; B. R. de Supinski; email: b.de-supinski@qub.ac.uk; H. Leather; email: hleather@inf.ed.ac.uk.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2017 ACM 1544-3566/2017/03-ART1 $15.00
DOI: http://dx.doi.org/10.1145/3050436

1. INTRODUCTION

In an era dominated by battery-powered mobile devices and servers hitting the power wall, increasing the energy efficiency of our computing systems is of paramount importance. Although performance optimization is a familiar topic for developers, few are even aware of the effects that source code changes have on the energy profiles of their programs. Therefore, developers require tools that analyze the power and energy consumption at the source-code level of their programs.

Prior energy accounting tools can be broadly classified into two categories: tools that directly measure energy using on-board sensors or external instruments [Chang et al. 2003; Flinn and Satyanarayanan 1999; Ge et al. 2010; Kansal and Zhao 2008; Keranidis et al. 2014; McIntire et al. 2007], and tools that model energy based on activity vectors derived from hardware performance counters, kernel event counters, finite state machines, or instruction counters in microbenchmarks [Bertran et al. 2013; Manousakis et al. 2014; Contreras and Martonosi 2005; Isci and Martonosi 2003; Curtis-Maury et al. 2008; Shao and Brooks 2013; Tsoi and Luk 2011; Tu et al. 2014; Wilke et al. 2013; Pathak et al. 2012; Schubert et al. 2012]. All of these tools can associate energy measurements with software contexts via manual instrumentation, context tracing, or profiling. However, these approaches have their limitations.

Direct power measurement. Using direct power measurement instruments, tools can accurately measure hardware component–level and systemwide energy consumption. State-of-the-art external instruments such as the Monsoon power meter [Brouwers et al. 2014], direct energy measurement and profiling tools [Ge et al. 2010; Flinn and Satyanarayanan 1999], and internal energy and power sensors such as Intel's Running Average Power Limit (RAPL) [Rotem et al. 2012] or on-board sensors [Cao et al. 2012] sample at a rate between 1kHz and 3.57kHz. This rate is unsuitable for energy accounting of individual basic blocks, functions, or even longer code segments that typically run for short time intervals. Thus, we cannot characterize the power usage of these fine-grained code units and will miss many optimization opportunities that can only be discovered through fine-grained energy profiling.

Activity vector–based measurement. Tools that model energy consumption from activity vectors extracted from hardware performance counters can produce estimates for fine-grained code units but suffer from several shortcomings. Their accuracy is architecture- and workload dependent [Curtis-Maury et al. 2008; Contreras and Martonosi 2005; Isci and Martonosi 2003; Schubert et al. 2012]. Thus, they are only suitable for a limited set of platforms and applications. Further, targeting a new platform or workload requires expensive benchmarking and training to calibrate the model. Finally, as we show later, these tools are not accurate enough to capture the power variation caused by the execution of different code regions.

To overcome the limitations of prior work, this article presents ALEA, a new probabilistic approach for estimating power and energy consumption at a fine-grained level. Specifically, we propose a systematic constant power model (CPM) to estimate the power consumption of coarse-grained code regions with latency greater than the minimum feasible power-sensing interval and an improved variation-aware power model (VPM) that estimates power more accurately by identifying high-frequency variations at the fine-grained level through additional profiling runs. The CPM approximates the power consumption of a region under the assumption that power is constant between two successive readings of the system's power sensors. Its accuracy depends on the minimum interval that is allowed between two such readings. The VPM relaxes this assumption and identifies power variation over intervals significantly shorter than the minimum feasible power-sensing interval. We validate the CPM by comparing the results of profiling with real energy measurements on two different architectures using

a set of diverse benchmarks. The experimental results show that the CPM is highly accurate, with an average error below 2% for coarse-grained regions. However, for fine-grained regions, this model can produce much higher errors. The VPM successfully handles these cases, allowing energy accounting at a frequency that is up to two orders of magnitude higher than on-board power-sensing instruments, with the error of its estimates at 6%, which is 92% lower than those of the CPM.

Finally, we present two use cases of ALEA: fine-grained memoization in financial option pricing kernels to optimize performance under power caps and fine-grained DVFS to optimize energy consumption in a sequence alignment algorithm. These use cases show that our approach can profile large real-world software systems to understand the effect of optimizations on energy and power consumption.

The contributions of this article are the following:

—A new method for accurate energy accounting in code blocks with finer granularity than that of power-sensing instruments on real hardware
—Two first-of-their-kind, portable probabilistic models for accurate energy accounting of code fragments of granularity down to 10μs, or two orders of magnitude finer than direct power measurement instruments
—Low-latency and nonintrusive implementation methods for fine-grained, online energy accounting in sequential and parallel applications
—Tangible use cases that demonstrate energy-efficiency improvements in realistic applications, including energy reduction by 2.87$\times$ for an industrial strength option pricing code executed under a power cap and up to 9% reduction in energy consumption in a sequence alignment algorithm.

The rest of this article is structured as follows. Section 2 provides a short overview of our profiling approach. Section 3 details our energy sampling and profiling models, whereas Section 4 examines the key aspects of their implementation. Section 5 demonstrates the accuracy of the proposed probabilistic models. Section 6 details opportunities and limitations of ALEA, including the effect of hysteresis on physical power measurements and an analysis of ALEA's accuracy compared against power models based on hardware performance counters. Section 7 presents how ALEA can be used to improve energy efficiency in realistic applications. Section 8 discusses previous work in this area, and Section 9 summarizes our findings.

2. OVERVIEW OF ALEA

ALEA provides highly accurate energy accounting information at a fine granularity to reason about the energy consumption of the program. It collects power and performance data at runtime without modifying the binary. ALEA uses probabilistic models based on its profiling data to estimate the energy consumption of code blocks.

2.1. Definitions

The following terms are used throughout the article:

—*Code block*: Region of code that the user defines at the final profiling stage and that can include one or more basic blocks or procedures
—*Sampling interval*: Time between two consecutive samples
—*Power-sensing interval*: Interval during which a single power measurement is taken in a computing system
—*Fine-grained block*: Code block that has an average latency less than the power-sensing interval
—*Coarse-grained block*: Code block that has an average latency of at least the power-sensing interval.

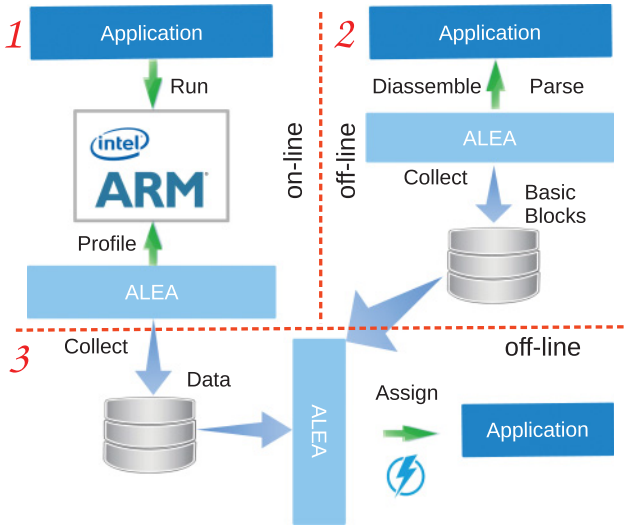

Fig. 1. Workflow of ALEA.

2.2. Workflow

Figure 1 illustrates the three steps of the ALEA workflow: profiling, disassembling and parsing, and energy accounting.

Profiling. The instrumented program is profiled using several representative input datasets. During each run, ALEA periodically samples instruction addresses and records readings from a power measurement instrument such as an on-board energy sensor or a software-defined power model (e.g., based on hardware counters).

Disassembling and parsing. After each run, ALEA disassembles the binary and identifies basic blocks in the code.

Energy accounting. As we show in Section 5, the default model of ALEA (the CPM, see Section 3.1) provides accurate energy and power estimates for coarse-grained blocks.

To obtain accurate energy estimates for specific fine-grained code blocks, the user should choose to use our second model (the VPM). The VPM overcomes the coarse granularity of the power sensors and provides accurate energy estimates over intervals with latency down to 10μs. Finally, ALEA probabilistically estimates the execution time and power consumption of each code block.

3. ENERGY ACCOUNTING BASED ON A PROBABILISTIC MODEL

This section details our probabilistic energy accounting models. We first present our execution time profiling method in Section 3.1, followed by our sampling method in Section 3.2. We then describe our constant and variation-aware power models in Section 3.3. We extend our methodology to multithreaded code in Section 3.4. We close this section with an analytic study of the accuracy of our energy accounting models.

3.1. Execution Time Profiling Model

Our execution time profiler randomly samples the program counter multiple times per run. Based on the total execution time of the program and the ratio of samples of a code block to the total number of samples, we estimate the time that a block executes. More formally, we define a processor clock cycle as a unit of the finite population (U). It has been demonstrated that the execution time of code block blm can be estimated

based on the probability to sample this block at a random cycle [Mukhanov et al. 2015; Graham et al. 1982]:

$$\hat{t}_{blm} = \hat{p}_{blm} \cdot t_{exec} = \frac{n_{blm} \cdot t_{exec}}{n}, \tag{1}$$

where t_{blm} is the the total execution time estimate of instances of blm, t_{exec} is the program's total execution time, $\hat{p}_{blm}$ is the probability to sample block blm, n_{blm} is the number of samples of any instruction from blm, and n is the total number of samples. Equation (1) captures that the probability of sampling a code block at a random clock cycle is the ratio of its execution time to the program's total execution time.

3.2. Systematic Sampling

We use systematic sampling, which approximates random sampling. It selects the first sample (unit) randomly from the bounded interval [1, $length\,of\,sampling\,interval$] and other samples are selected from an ordered population with an identical sampling interval. We subsequently sample in the same run every $sampling\,interval$ clock cycle. The random selection of the first sample ensures that sampling is sufficiently random. Even if the sampling interval aligns with a pattern in the execution of the program, we can still obtain accurate estimates by aggregating multiple profiling runs, each one with a different initial sampling time.

3.3. Probabilistic Power Models

We apply the same probabilistic approach that we use for time profiling to profile power and energy. We consider power consumption as a random variable that follows a normal distribution and is a characteristic associated with the clock cycle population. We simultaneously sample the program counter and power consumption, which we assign to the sampled code block.

Assuming n_{blm} samples of block blm, we estimate its mean power consumption as

$$\widehat{pow}_{blm} = \frac{1}{n_{blm}} \cdot \sum_{i=1}^{n_{blm}} pow^{i}_{blm}, \tag{2}$$

where pow^{i}_{blm} is the power consumption associated with the ith sample of block blm.

We estimate the energy consumption of blm as

$$\hat{e}_{blm} = \widehat{pow}_{blm} \cdot \hat{t}_{blm}. \tag{3}$$

Our model can derive energy estimates for any system component (e.g., processors, DRAM, on-chip GPU) for which a hardware or software-defined power sensor exists. The rest of this article considers only processor energy accounting. The model implies that we measure the instantaneous power consumption of a sampled code block. However, power meters can only average measurements over a certain interval, and power measurement can itself have a latency of hundreds of microseconds. We next discuss our two approaches to approximate the actual power consumption of the code block.

3.3.1. The Constant Power Model. Figure 2 shows an example of the power measurement process of our CPM. When we sample the instruction pointer (the blue line), we interrupt the profiled threads, which causes the power consumption to drop quickly to the idle state, where it stays while all threads are paused. If we then measure power, this behavior would impact our measurement significantly. Thus, we measure power before sampling the current instruction pointer.

This model provides accurate power estimates for coarse-grained blocks. However, the power-sensing interval necessarily adjusts power consumption for other code blocks with potentially different power behavior. The CPM, as well as all other existing energy

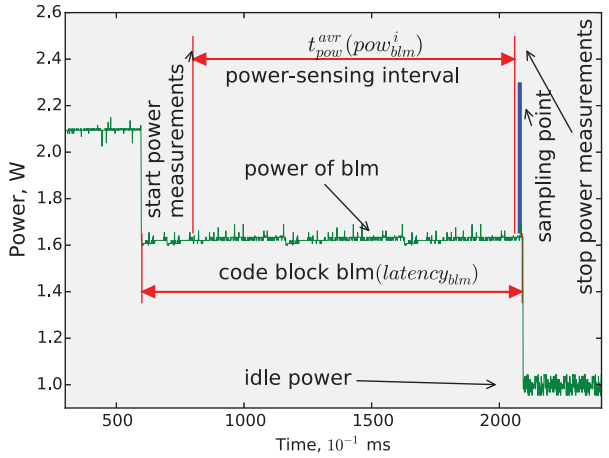

Fig. 2. Power measurements for the CPM.

profiling approaches, assumes that power remains constant throughout the power-sensing interval. If the power level varies during this interval, these energy profiling techniques will not capture that power variation.

3.3.2. The Variation-Aware Power Model. Accurately measuring power for fine-grained code blocks requires a more sophisticated approach, which our VPM provides. We isolate each block's power consumption from that of the neighboring code blocks through multiple power measurements for the exact same sequences of code blocks. By pausing (or not) the execution of the last block in this sequence, we control whether it contributes (or not) to the power measurement. We subtract the power associated with a sequence that omits the block from that of one that includes it to obtain the power and energy consumption of that block.

The top plot in Figure 3 shows this process in more detail. The VPM partially decouples measuring power from sampling. We initiate the power measurement at point `A` in the plot. The actual sample is taken after a certain amount of time, at point `B`. Sampling causes the program to pause, bringing the power consumption down to pow_{idle}. At the end of the power-sensing interval, point `C`, we obtain $pow^{no\_delay}_{A\to C}$.

Every time point `B` occurs at the beginning of an iteration of code block *blm*, the power reading is roughly the power consumption of the preceding blocks, excluding the contribution of *blm*. To obtain their power consumption including the contribution of *blm*, we delay sampling for an interval equal to *blm*'s latency for half of the samples.

The bottom half of Figure 3 shows that this delay causes half of the samples that would be taken at `B` to be taken at `B'` (*blm*'s end). The power-sensing interval covers the same preceding blocks as in the case of sampling at `B` but also includes *blm*. The difference of the two power readings gives us the energy and power consumption of *blm*:

$$e^{i}_{blm} = pow^{delay}_{A\to C} \cdot t^{avr}_{pow} - pow^{no\_delay}_{A\to C} \cdot t^{avr}_{pow} + pow_{idle} \cdot latency^{i}_{blm}, \tag{4}$$

$$pow^{i}_{blm} = \frac{pow^{delay}_{A\to C} - pow^{no\_delay}_{A\to C}}{latency^{i}_{blm}} \cdot t^{avr}_{pow} + pow_{idle}, \tag{5}$$

where $pow^{delay}_{A\to C}$ denotes the power measurements taken when we sample the last instruction of *blm* at `B'`, $pow^{no\_delay}_{A\to C}$ denotes power measurements taken when the first

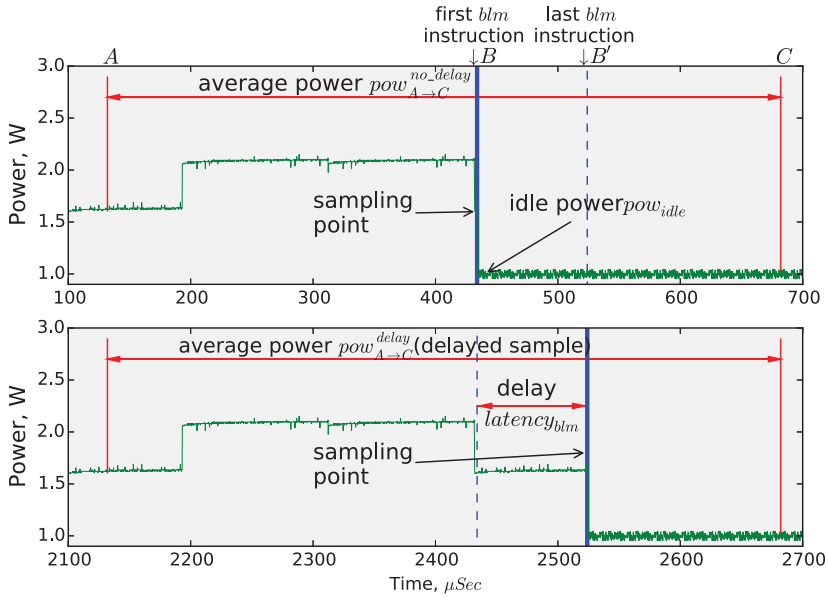

Fig. 3. Power measurements for the VPM.

instruction of the code block was sampled at B, pow_{idle} is the idle power consumption obtained after interrupting the profiled threads, and t_{pow}^{avr} is the power-sensing interval.

Our approach assumes that the power readings correspond to observations of a random variable with a normal distribution. Thus, we can average $pow_{A\to C}^{delay}$ and $pow_{A\to C}^{no\_delay}$ over all samples to estimate the power consumption for blm as

$$\widehat{pow}_{blm} = \frac{\widehat{pow}_{A\to C}^{delay} - \widehat{pow}_{A\to C}^{no\_delay}}{\widehat{latency}_{blm}} \cdot t_{pow}^{avr} + pow_{idle}, \tag{6}$$

where $\widehat{latency}_{blm}$ is the average latency of the code block.

For the VPM to produce accurate results, all power readings associated with the same sample, delayed or not, must cover the power consumption of the same code blocks. The time between the start of the power measurement and the sampling point must be constant, as must the time between the sampling point and the end of the measurement. We carefully implement the VPM to meet these requirements as much as possible. The only source of error that we cannot control is variations of the execution path and the power consumption before the targeted code block. Given the regular nature of most applications, these variations are not a significant source of error. In Section 5.2, we validate VPM and show that it accurately estimates power consumption.

If the latency of profiled code blocks is small enough, then the accuracy of power sensors could be insufficient to evaluate the difference between $\widehat{pow}_{A\to C}^{delay}$ and $\widehat{pow}_{A\to C}^{no\_delay}$ (see Section 5.3). The VPM may obtain erratic power measurements due to insufficient accuracy of the power sensors. To overcome this problem, we force upper and lower bounds on VPM power readings to correspond to the peak power consumption of the processor and pow_{idle}, respectively.

3.4. Multithreaded Code

Multithreaded applications execute many code blocks in parallel on different cores. To profile applications running on the same processor package at the code block level, we

Code blocks	Samples	Time	Energy	Power
core1:cb100,core2:cb100	163	21.99 +/-0.44	67.95 +2.71/-2.65	3.09 +/-0.06
core1:cb100,core2:sys	94	0.95 +/-0.19	1.66 +0.42/-0.39	1.75 +/-0.08
core1:cb161,core2:cb161	91	0.92 +/-0.18	2.29 +0.54/-0.51	2.48 +/-0.07

Fig. 4. Profiling results for the heartwall benchmark on the Exynos platform.

define a vector of code blocks that execute in parallel on different cores as

$$\overrightarrow{blm} = blm_{thr_1}, blm_{thr_2}, \ldots, blm_{thr_l}. \tag{7}$$

We use the same time and energy profiling models as in sequential applications. The only difference is that we associate multithreading samples with vectors of code blocks. We distribute the execution time and energy across code block vectors as

$$\hat{t}_{\overrightarrow{blm}} = \hat{p}_{\overrightarrow{blm}} \cdot t_{exec} = \frac{n_{\overrightarrow{blm}} \cdot t_{exec}}{n}, \tag{8}$$

$$\widehat{pow}_{\overrightarrow{blm}} = \frac{1}{n_{\overrightarrow{blm}}} \cdot \sum_{i=1}^{n_{\overrightarrow{blm}}} pow^{i}_{\overrightarrow{blm}}. \tag{9}$$

We examine all running threads collectively during sampling because they share resources. Such shared resources include caches, buses, and network links, all of which can significantly increase power consumption under contention. We could apportion power between threads based on dynamic activity vectors that measure the occupancy of shared hardware resources per thread [Manousakis et al. 2014]. However, these vectors are difficult to collect and to verify as current hardware monitoring infrastructures do not distinguish between the activity of different threads on shared resources.

Figure 4 shows the results of profiling the heartwall benchmark (Rodinia suite) on the Exynos platform (see Section 4.1), including 95% confidence intervals for each estimate. core1:cb100, core2:cb100 denotes the code block with identification 100 executed simultaneously on the first and second cores. Similarly, cb161 is the code block with identification 161, and sys is a code block from the pthread_cond_wait() function of the glibc library; it executes a mutex lock that keeps a core in sleep mode until the corresponding (pthread_cond_signal) occurs. We see that cb100 executed in parallel consumes almost twice as much power than when it is co-executed with the sys block. In the second case, the second core is sleeping and consumes almost no power. The difference in power consumption between parallel execution of cb161 and cb100 is related to the cache access intensity of these blocks. Our earlier work shows that dynamic power consumption grows with this intensity [Mukhanov et al. 2015].

The number of sampled code block vectors depends on the number of code blocks in a parallel region, and it grows exponentially with the number of cores. Thus, for highly parallel regions containing multiple code blocks, the number of code block vectors could render the analysis of profiling results slow, as we should account for the energy consumption of all vectors. In practice and in all applications with which we have experimented, strong temporal locality of executing code blocks keeps the number of vectors of code blocks that execute concurrently (and thus are sampled) low. For example, in heartwall, three code block vectors dominate energy consumption (Figure 4), whereas the total number of sampled code block vectors is about 100 in this benchmark.

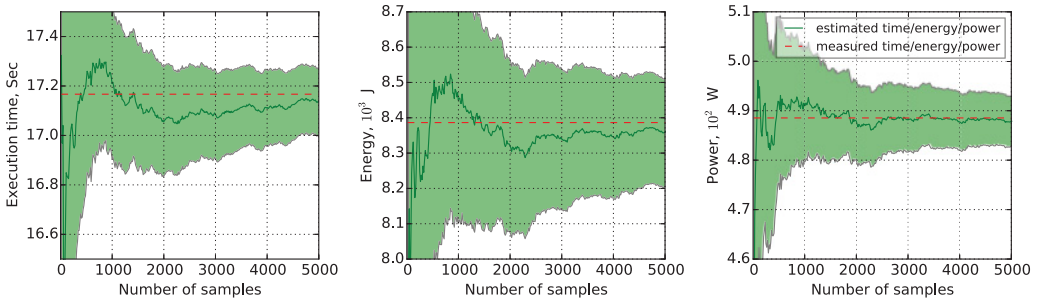

Fig. 5. Estimated versus measured execution time, energy, and CPM power for a code block of `heartwall` (Sandy Bridge).

3.5. Analysis of Model Accuracy and Convergence

ALEA provides confidence intervals for its time, power, and energy estimates. We have shown in earlier work [Mukhanov et al. 2015] how to construct confidence intervals for the time and CPM power estimates.

Using the confidence intervals for execution time and power, we can also evaluate the worst-case upper and lower bounds for energy consumption:

$$\hat{t}^{l}_{\overrightarrow{blm}} \cdot \widehat{pow}^{l}_{\overrightarrow{blm}} \leq e_{\overrightarrow{blm}} \leq \hat{t}^{u}_{\overrightarrow{blm}} \cdot \widehat{pow}^{u}_{\overrightarrow{blm}}, \tag{10}$$

where $\hat{t}^{l}_{\overrightarrow{blm}}$ ($\hat{t}^{u}_{\overrightarrow{blm}}$) is the lower (upper) bound of the confidence interval for the time estimates and $\widehat{pow}^{l}_{\overrightarrow{blm}}$ ($\widehat{pow}^{u}_{\overrightarrow{blm}}$) is the lower (upper) bound of the power confidence interval. We can increase the total number of samples to tighten the confidence interval for execution time. We can increase the number of samples of each code block to narrow the confidence interval for power estimates. Thus, to narrow both confidence intervals, we can increase the total number of samples and the number of samples per block.

To demonstrate the accuracy improvements of our time, energy, and CPM power estimates as the number of samples increases, Figure 5 shows their evolution for a coarse-grained code block from the `heartwall` benchmark. We run this benchmark on our Sandy Bridge server (see Section 4.1). The dark green line represents the time (left), energy (middle), and power (right) estimates, whereas the lighter green area shows the 95% confidence interval for each estimate. The red dashed line in each plot corresponds to direct measurements for the block averaged over several runs.

The minimum number of samples for the block, collected during the first run of the experiment, is 16. The error of the energy estimate from the 16 samples, compared to direct measurement, is 12%. This error stabilizes below 1.7% with 110 samples and following that drops slowly and becomes lower than 1.5% after 421 samples and lower than 1.2% after 2,300 samples. Other benchmarks exhibit similar behavior.

In the case of the VPM, the width of the confidence interval for power estimates depends on the number of samples of the first instruction and the number of samples of the last instruction. The VPM needs twice as many samples to achieve the accuracy of the CPM.

We could reduce the length of the sampling interval to obtain more samples during each profiling run. However, sampling uses interrupts that distort program execution, thus introducing errors in the time and energy estimates. We adopt a less intrusive but costlier solution that performs more profiling runs and combines their samples into a single set. This approach improves the accuracy of our energy estimates and also increases the probability of sampling fine-grained code blocks. In practice, each run of an application is unique and the total execution time of the application changes slightly

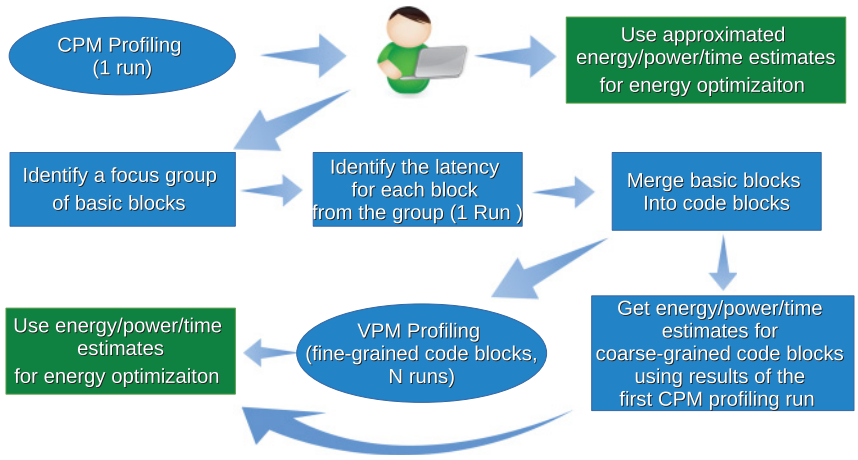

Fig. 6. The ALEA profiling process.

from run to run, even though the application is executed under the same conditions. To address this issue, we average the total execution time over the runs.

4. IMPLEMENTATION

ALEA uses an online module that runs as a separate control process. This module samples the application's instruction pointer and power levels. It collects profiling information at random points transparently to the application. The sampling process has four stages. It first measures power and then attaches to the program threads through `ptrace`. Next, it retrieves the current instruction pointer and finally detaches from the threads. The control process sleeps between samples to minimize application impact.

The online module transfers the results of profiling, including sampled instruction addresses and power measurements, to an offline module that derives energy estimates. The offline module uses `objdump` to disassemble the profiled program and its own parsers to identify basic blocks and instructions inside these blocks. It uses this information to assign samples to specific blocks. Thus, ALEA can estimate the execution time for each basic block, using the probabilistic model described earlier. To enable accurate energy and power profiling, the user should identify basic blocks to be merged into coarse-grained blocks if the default power model (the CPM) is used. The VPM can be applied to fine-grained blocks (with latency as low as 10μs, see Section 5.3); however, the user should point ALEA to these blocks. The user obtains the latency of each basic block by using an additional run and dynamic instrumentation tools, or direct instrumentation. Finally, we use DWARF debugging information to associate ALEA's estimates with specific lines in the source code.

Figure 6 illustrates the process of profiling from the user perspective. During the first run of ALEA, the user identifies execution time, energy, and power estimates for each basic block using the CPM. From this point, the user can do additional runs and select basic blocks where more accurate estimates are needed. This step can also be automated with hotspot analysis, using, for example, clustering and classification techniques similar to the ones used in architectural simulators [Sherwood et al. 2001]. The user should point to ALEA which basic blocks should be merged into code blocks using addresses of these blocks. If latency of a merged code block is greater than the power-sensing interval, then the results of the first profiling run could be used to get accurate power and energy estimates for this code block. If the latency of the code block

is less than the power-sensing interval and no lower than 10μs, then the user needs to run the VPM for the specific code block and pass to ALEA the addresses of the first and last instructions and the latency of the block. The user can initiate any desirable number of profiling runs using the CPM or VPM to achieve the desirable accuracy.

We note that ALEA can be implemented in the kernel space, which would reduce its overhead and allow sampling at a higher frequency than a user-level implementation. Nonetheless, we opted for a user-level implementation to enable easier deployment of the tool in environments where the user cannot have access to the kernel.

4.1. Platforms and Energy Measurement

ALEA builds on platform-specific capabilities to measure power. It can use direct power instrumentation, integrated power sensors, or software-defined power models. ALEA operates under the sampling rate constraints of the underlying sensors but is not limited by them since its probabilistic models can estimate power at finer code block granularity. To demonstrate the independence of our approach from the specifics of the hardware, we use two platforms that differ in architectural characteristics and support for power instrumentation.

Our first platform, a Xeon Sandy Bridge server, uses RAPL for on-chip software-defined modeling of energy consumption, which we obtain by accessing the RAPL MSRs [Rotem et al. 2012]. The server includes four Intel Xeon E7-4860 v2 processors, each with 12 physical cores (24 logical cores) per processor, 32KB I and D caches per core, a 3MB shared L2 cache, and a 30MB shared L3 cache per package. The system runs CentOS 6.5 and has a maximum frequency of 2.6GHz.

Our second platform, an ODROID-XU+E board [Hardkernel 2016], has a single Exynos 5 Octa processor based on the ARM Big.LITTLE architecture, with four Cortex-A15 cores and four Cortex-A7 cores. In our experiments, we only use the Cortex-A15 cores at their maximum frequency of 1.6GHz. Each core has 32KB I and D caches, whereas the big cores also share a 2MB L2 cache. We measure power through the board's integrated sensors. The system runs Ubuntu 14.04 LTS.

4.2. Power-Sensing Interval

Our Intel Sandy Bridge server provides energy estimates through RAPL, but not power measurements. We measure processor power consumption on it by dividing energy measurements taken at the start and end of the power-sensing interval by the size of this interval. The energy counter is updated only once per millisecond, which is the server's minimum feasible power-sensing interval. Our Exynos platform includes the TI INA231 power meters [TI 2013], which directly sample power consumption for the whole system-on-chip and several of its components. These samples are averaged over a user-defined interval. The minimum available interval on the Exynos is 280μs.

Shorter power-sensing intervals improve CPM estimates for the fine-grained blocks. Thus, our CPM experiments use the minimum power-sensing interval on both platforms.

The VPM assumes that we can precisely control when power is measured. Unfortunately, on Intel processors, RAPL energy counters are updated by the hardware at unknown points and transparently to software, with an update frequency of 1ms. Thus, we cannot use the VPM for Intel processors with RAPL. On the Exynos platform, even though we initiate the power measurement, a small variable delay occurs between the system call to read the power counters and the actual reading. Still, we can assume that this delay remains fixed on average and converges to its mean value as the number of samples increases.

In theory, the granularity of the power meters does not limit the VPM, so the power-sensing interval is not critical. In practice, our experiments show that the maximum

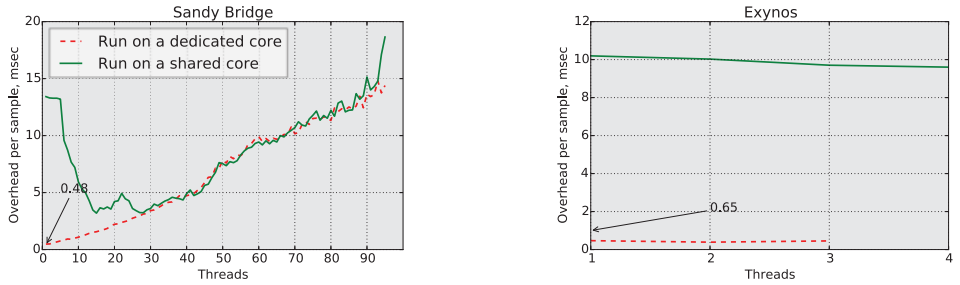

Fig. 7. Overhead per sample versus number of threads with ALEA running on a dedicated or a shared core.

overall latency of the first stage of sampling in the VPM is about 650μs. At the first stage, ALEA calls the interface that starts power measurements and the attach interface that stops all threads. At the second stage, the tool calls the interface that takes the power reading, delivers the instruction pointer for each thread, and then detaches from all threads. While in the CPM the attach interface is called after the power measurement, in the VPM it is part of the measurement. For our power reading to contain any information about the profiled code block, the power-sensing interval must be longer than the first-stage sampling latency. For this reason, we set the power-sensing interval to 1,100μs.

4.3. Sampling Interval and Overhead

The attach, deliver, and detach phases of the tool introduce overhead. The attach phase interrupts the execution of all threads. Threads continue to execute after completion of the detach phase.

We use a set of benchmarks from the NAS suite (`BT`, `CG`, `IS`, and `MG`) to illustrate the profiling overhead of the ALEA online module, with a core dedicated to each application thread. We choose benchmarks that scale well so that we can evaluate the overhead of ALEA under a difficult scenario in which the application utilizes the processor resources fully and is sensitive to sharing them with ALEA.

Figure 7 shows the overhead of ALEA, which clearly depends on the number of application threads, as the tool has to deliver the instruction pointer for each thread. With more threads, the delivery overhead necessarily increases. The profiler can either share a core with the program or use a dedicated core. We observe higher overhead when the tool and the program share a core, which is caused by a delay in the delivery phase. This delay grows significantly, up to 10×, when the tool attaches to the application thread on the shared core due to the overhead of switching between an application thread and the ALEA profiler. The profiler leverages the OS load-balancing mechanisms to exploit idle cores, if available. The benchmarks occasionally leave cores idle, and this slack grows as the benchmark uses more threads. Linux dynamically migrates profiled threads from a core used by ALEA to an idle one to eliminate the context-switching overhead, which leads to an overhead reduction as the number of threads grows on the Sandy Bridge platform when ALEA uses a shared core. Still, if the benchmarks use more than 30 threads, the context-switching overhead becomes negligible in comparison to the overhead caused by the attach, deliver, and detach phases.

These findings suggest that we should choose the sampling interval based on the number of threads and the availability of a dedicated core to reduce the bias in the estimates. However, for most programs, the thread count is not constant, so the overhead and the optimal sampling interval vary during execution. We cannot change the sampling interval dynamically, as our probabilistic model assumes a uniform sampling

process. Instead, the sampling interval should reflect the average overhead measured over a whole run of the application. ALEA reports the average overhead for each profiling run by measuring the runtime of the sampling phases. Thus, the user can easily find a sampling interval with an acceptable overhead. By default, ALEA uses a sampling interval that keeps overhead under 1% when all available cores are used by an application. It also estimates the optimal sampling interval that gives the minimum overhead and the maximum number of samples after a profiling run.

5. VALIDATION

In this section, we validate the accuracy of our approach under extreme scenarios, where the power consumption changes rapidly and widely. We use multiple benchmarks on both hardware platforms to establish the generality of our results. We validate ALEA using loops with coarse-grained blocks for each platform. We then show that the CPM (Section 5.1) can accurately estimate the energy consumption of coarse-grained blocks, whereas the VPM (Section 5.2) does so even for fine-grained blocks.

5.1. CPM Validation

To validate the accuracy of ALEA's energy consumption estimates, we use 10 parallel benchmarks from the NAS and `Rodinia` suites. We use a wide range of benchmarks to achieve good coverage of loop features in terms of execution time and energy consumption, including loops with distinct power profiles and power variation between their samples. For each benchmark, we compare CPM time and energy estimates for each loop with direct measurements. These direct measurements use instrumentation at the entry and exit points of the loops. On the Sandy Bridge platform, we instrument the code using the `rdtsc` instruction and the RAPL interface for time and energy, respectively. On Exynos, we use `clock_gettime()` to measure time and the `ioctl()` interfaces to access the INA231 power sensors.

We compile the `Rodinia` benchmarks using the flags specified by the benchmark suite and run them with the default inputs. The benchmarks from NAS are built using option `CLASS=B` and run with the default options. We use the abbreviation `suite.benchmark` to indicate the benchmarks on our graphs. For example, `R.sc` corresponds to the `StreamCluster` benchmark from the `Rodinia` suite.

5.1.1. Sandy Bridge. We run each benchmark using 96 threads on the Sandy Bridge platform. ALEA shares a core with an application thread. We use a 3-second sampling interval to achieve a runtime overhead of under 1%. The average latency of most instrumented loops is less than this sampling interval.

We combine samples from multiple runs for each benchmark to achieve accurate estimates. We collected 500 runs for each benchmark but found the accuracy with 180 to be sufficient. Using more runs has little effect, so our remaining experiments use 180 runs.

Figure 8 shows how the error of energy estimates changes with the number of combined runs. Ignoring `StreamCluster`, the mean error is as high as 32% when we profile each benchmark only once. Predicting energy for `StreamCluster` requires more than one run. The total execution time of one instrumented loop is shorter than the sampling interval, so the loop may not be sampled during a run.

Increasing the number of combined runs reduces the error. On average, the error decreases to 2.1% after 180 profiling runs. Even for `StreamCluster`, the error is only slightly above average at 2.3%. Regarding the estimation of the total energy consumption of the program, the average error across all benchmarks is 1.6% for 180 runs. The energy estimates errors of ALEA are thus comparable to ALEA's runtime overhead. For completeness, we also profile a 94-thread version of each benchmark with ALEA

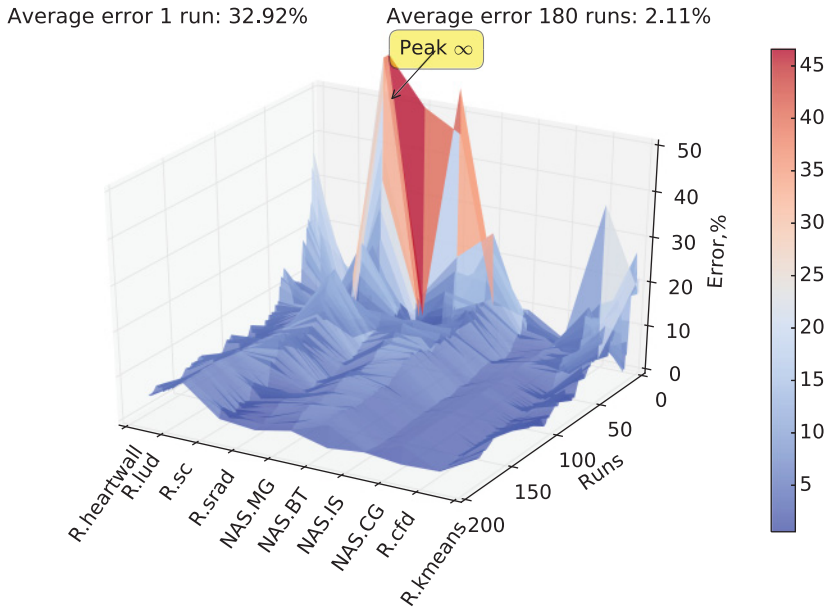

Fig. 8. Average error of energy estimates across all loops on Sandy Bridge.

using a dedicated core. We used a sampling interval of 3 seconds to keep overhead low (under 1%). The minimum error of 1.9% with 155 runs is again comparable to the runtime overhead.

5.1.2. Exynos. We similarly validate ALEA on the Exynos platform. We used the same experimental setup, but we use two threads for the benchmarks to give ALEA a dedicated core and set the sampling interval to 500ms with the aim of keeping the overhead slightly below 1%. Figure 9 presents the trade-off between the number of runs and the average error in our energy estimates for the loops. The sampling interval is longer than the average latency of almost all instrumented loops. The average error of the energy estimates is about 21% for a single profiling run but decreases to 1% with 50 or more profiling runs. The whole program absolute error for energy, averaged across all benchmarks, is 1.1% for 50 profiling runs.

ALEA exhibits its highest error in the energy estimates of `heartwall` from the `Rodinia` suite. A loop at line 57 in avimod.c has a total execution time of 0.31 seconds, which is less than the sampling interval. Thus, it was sampled only once during the first profiling run, and the error in terms of both execution time and energy exceeded 120%. However, this error falls below 5% within 50 runs. Although ALEA can theoretically achieve any desirable accuracy, in practice the lower limit of the error is about 1% due to the instrumentation overhead and measurement interference.

We repeat our experiments using four threads for each benchmark, with ALEA sharing a core. The results show the same patterns, with the average error of the energy estimates at about 1.3% and stabilizing after 110 profiling runs.

5.2. VPM Validation

Next, we validate the estimates produced by our VPM. These experiments only use the Exynos platform due to RAPL's limitations (see Section 4.2). We design the VPM to capture fine-grained program power behavior. Thus, we validate it on programs with significant power variation between fine-grained blocks: loops with latency less than

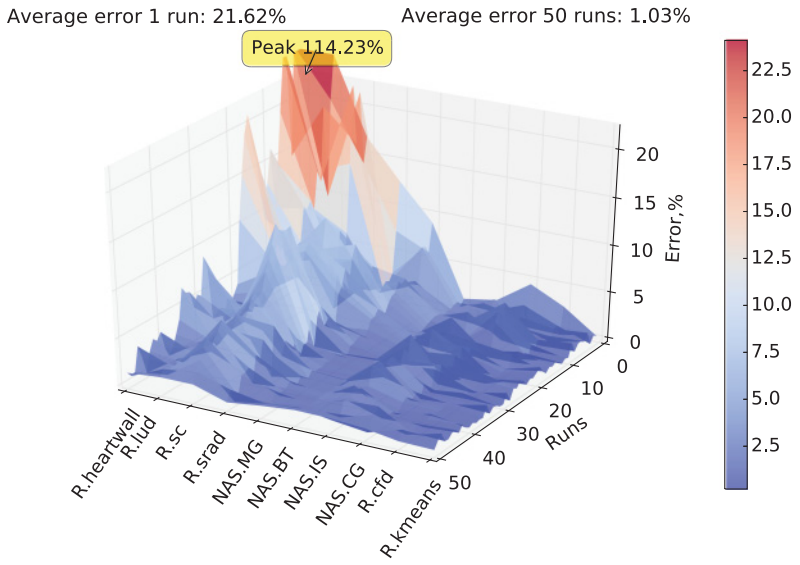

Fig. 9. Average error of energy estimates across all loops on Exynos.

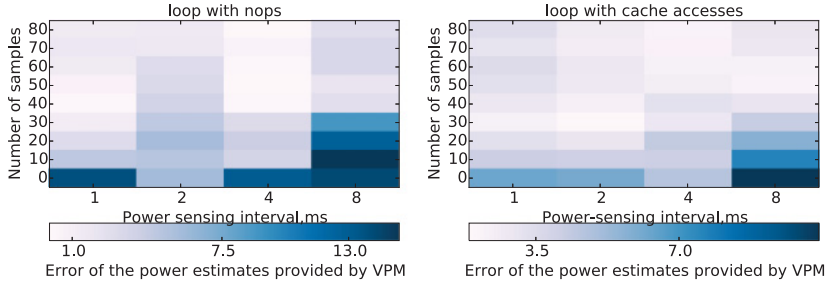

Fig. 10. Power estimates provided by VPM for the microbenchmark.

the power-sensing interval. We use a 1,100μs power-sensing interval for the reasons discussed in Section 4.2. On the other hand, to validate the VPM, we must use loops with known power consumption and latency. To measure them directly, latency must be above the minimum power-sensing interval (280μs).

5.2.1. Context-Driven Power Variation. Our experiments show that processor dynamic power consumption is primarily affected by cache and memory access intensity [Mukhanov et al. 2015]. Thus, we use a microbenchmark that consists of two loops. The first loop contains a block with `nop` instructions, consuming 1.68 Watts. The second loop contains a block with memory access instructions engineered to hit in the L1 cache and consumes 2.54 Watts. We control the latency of each loop through the number of iterations. We set the latency of each loop close to 400μs, which is within the range of latencies discussed previously.

We profile a two-thread version of the microbenchmark 1,000 times (ALEA runs on a dedicated core) to collect at least 80 samples, which ensures at least a 95% confidence interval of the power measurements within 1% of the mean. Figure 10 shows the trade-off between the number of samples and the average error of the power estimates provided by the VPM, which we compare to direct power measurements. In this figure,

the darkness of the color indicates the average error of the power estimates. We change the power-sensing interval from 1ms to 8ms and use a 1-second sampling interval to keep overhead under 1%.

Figure 10 demonstrates that the average error of VPM power estimates decreases with the number of samples. Forty samples are enough to estimate the power consumption of the loops with an error less than 4%, even if we use an 8ms power-sensing interval, which is 20× more than the latency of these loops. The average error of VPM power estimates is about 3.5% for each power-sensing interval when the number of samples is close to 80 (approximately 3.5% for the energy estimates).

Even though the VPM provides precise power estimates, interference from power measurements impacts its accuracy. Although we expect the average error to be higher for a 4ms interval than a 2ms interval, our results show the opposite. We speculate that this happens because of OS interference, but performing kernel modifications to validate this hypothesis was beyond the scope of this article.

We use the same experiment and 1,000 runs to test the CPM's accuracy. Similarly to the VPM, we vary the power-sensing interval from 1ms to 8ms. Regardless of the interval, the average error of CPM power estimates is 45.5% after 1,000 runs (approximately 45.5% for the energy estimates), which is 12× higher than the average error of the VPM. Clearly, the VPM can successfully handle power variations across fine-grained blocks, which the CPM and other known power measurement and modeling techniques cannot.

5.2.2. Concurrency-Driven Power Variation. We next investigate how the CPM and VPM handle power variations due to switching from executing parallel coarse-grained loops to executing sequential fine-grained loops. The sequential loops consume less power since fewer cores are active. The following benchmarks have sequential fine-grained loops that are surrounded by parallel coarse-grained loops: `Srad (Rodinia)` and `StreamCluster (Rodinia)`. For those fine-grained loops, we artificially inflate their execution time to allow direct measurements.

Similarly to the microbenchmark experiments, we run each benchmark 1,000 times. We use two threads for each benchmark and run ALEA on a dedicated core. Similarly to the context-driven power variation experiment, we use a 1-second sampling interval to keep the overhead negligible. Our experiments show that the average error of VPM power estimates is 5.4% (approximately 5.5% for energy estimates), whereas the average error of CPM estimates is 15.7% (approximately 16% for energy estimates), which is almost 3× higher.

We repeat the experiments with four application threads (so ALEA must share a core) and a 1ms power-sensing interval. Our results demonstrate that the average error of VPM power estimates is 6% (approximately 6% for energy estimates), whereas the average error of CPM estimates is 25% (approximately 25% for energy estimates). Again, the VPM can handle programs with fast changing power behavior better than the CPM and produces accurate estimates.

5.3. Limitations of ALEA

ALEA is a tool for fine-grained energy profiling. The probabilistic model of the tool makes it possible to estimate the execution time of a code block with high accuracy even if the average execution time of the block is a few microseconds, whereas the sampling interval can be several seconds. However, our base power model (the CPM) assumes that power remains constant during the power-sensing interval. The VPM relaxes this assumption and captures power variation over intervals that are shorter than the power-sensing interval. Nonetheless, the resolution of this model is also limited.

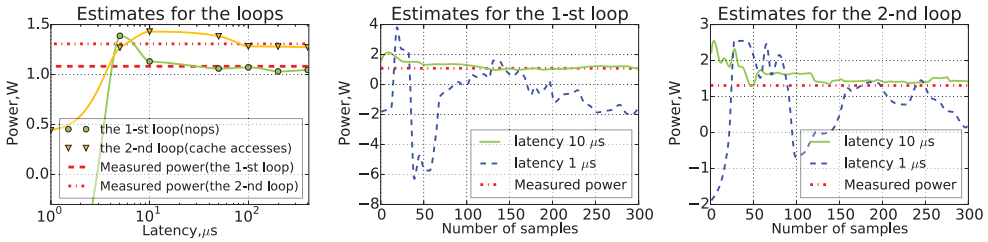

Fig. 11. Power estimates provided by the VPM for the microbenchmark with reduced latencies of the loops.

To explore how the VPM captures power variation over an interval with a latency of a few microseconds or nanoseconds, we employ a microbenchmark similar to the one with which we validated the VPM. The microbenchmark contains two loops: the first loop executes only `nop` instructions, and in each iteration of the second loop we access the same 368 bytes fetched into the L1 cache after their initialization. We test the resolution limit of the VPM by progressively reducing the execution time of the loops. We control loop execution time by changing the number of iterations. We ensure that each iteration executes the same instructions and the same number of L1 cache accesses in particular. Thus, loop power consumption should remain constant when we reduce the iteration count, but the difficulty of estimating it accurately will progressively increase. We do not use the parallel version of the microbenchmark, as the system calls starting and synchronizing threads inside the parallel regions already take hundreds of microseconds.

Figure 11 shows the results for the microbenchmark with loop latencies between 1 and 400 μs. Dashed red lines correspond to the power consumption of each loop found in the experiments in Section 5.2.1. The leftmost plot depicts how the power estimates for the loops change with their latency. The middle plot shows how the power estimates for the loop with `nops` change with the number of samples, whereas the rightmost plot demonstrates the same for the loop with cache accesses. Although we cannot directly validate the power consumption since the loop latencies are less than the minimum power-sensing interval (280μs), we expect that the difference in power consumption between these loops should remain constant for all loop latencies.

We observe the expected difference for latencies of at least 10μs (leftmost subplot of Figure 11). Although ALEA can theoretically achieve any desirable accuracy, in practice the lower limit of the error is about 1% due to the instrumentation overhead and measurement interference. The error of the power estimates for 10μs is 6.9%. This latency is the maximum resolution of the power estimates provided by the VPM. On the other two subplots, we also observe that the power estimates for the loops with this latency clearly converge to the measured power as the number of samples grows. For latencies below 10μs, the estimates never converge due to the limited accuracy of the power sensors. To capture power variations over intervals with 1μs latency, the sensors should provide accuracy of measurements within 10^{-4} Watts if a 1,100μs power-sensing interval is used. Moreover, the time required to measure power varies over samples, and this variation exceeds 1μs, which negatively impacts the power estimates.

ALEA could not reliably capture power variation between code blocks with execution times less than 10μs. Nonetheless, this interval is still 110 times shorter than the used power-sensing interval and 28 times shorter than the minimum feasible power-sensing interval in our platforms. As we show later, this resolution is sufficient to support energy optimization. To the best of our knowledge, ALEA is the first tool that provides accurate power estimates at this resolution.

6. DISCUSSION

6.1. Hysteresis in Power Measurements

Physical power measurements are affected by bypass capacitance, inductance, and impedance components [Gupta et al. 2007; Smith et al. 1999]. A voltage regulator module delivers power to a processor through a power distribution network (PDN). This network contains impedance and inductance components located at different levels: motherboard, package, and die. By executing different code blocks, a processor produces variations in current demand. Due to current spikes and impedance components, a voltage drop can occur across the network [Bertran et al. 2014]. In turn, the inductance components introduce voltage fluctuations because of the `di/dt` effect [Joseph et al. 2003; Zhang et al. 2013; Das et al. 2015]. Thus, voltage variations occur.

The voltage regulator module adjusts the voltage level to mitigate these variations. However, it cannot always change voltage quickly enough. To maintain voltage at safe levels when the current increases significantly, the PDN uses decoupling capacitors close to the processor (or package or motherboard). On the Exynos board, the power sensors (INA231) take power measurements on a power rail that connects the power management integrated circuits (PMIC), `MAX77802`, which provide a voltage regulator, with the chip. Decoupling capacitors, inductance and impedance components located between the power sensors and the die, affect instantaneous power dissipation. Thus, the effect of a code block on the measured power is not instantaneous and hysteresis occurs between execution of the block and its effect on the measured power consumption. Moreover, capacitance components also smooth out instantaneous power dissipation.

ALEA's probabilistic sampling can theoretically capture power shifts during instruction execution at any granularity. In practice, several factors, including software overheads and the overheads and measurement error of the power sensing instruments, prevent such very fine granularity. The length of the hysteresis (i.e. the typical response time of the PDN) is orders of magnitude less than the minimum code block duration for which ALEA can probabilistically estimate power. The literature reports PDN response times in the range of 75ns [Grochowski et al. 2002], versus ALEA's 10μs minimum code block length and minimum power-sensing intervals on the order of hundreds of microseconds on the Exynos and Intel platforms.

6.2. Power Modeling with Performance Counters

One of the possible solutions to break the coarse granularity of power sensors and mitigate hysteresis effects is to predict power consumption from hardware activity vectors. Previous research has extensively investigated the use of regression models on activity vectors of hardware performance counters to estimate power consumption. These models assume that processor power consumption correlates with the frequency of certain performance impacting events. Performance counters can be accessed with low overhead on most modern processors, which makes them appropriate for fine-grained power modeling. We modify ALEA to use a regression model for power modeling following Shen et al. [2013]. We use PAPI to measure nonhalt core cycles, retired instructions per CPU cycle, floating point operations per cycle, last-level cache requests per cycle, and memory transactions per cycle over the power-sensing interval and then correlate these measurements to the sampled power consumption. We build the regression model for each platform individually on the basis of 10 benchmarks used for validation of the CPM and VPM with varying thread counts. We use a sampling interval that keeps runtime overhead under 1%. We add L1 cache accesses to the model because its power consumption is highly sensitive to this event and omitting it loses significant accuracy.

In this experiment, we use linear regression to predict power consumption for samples that we use for training and compare the results of prediction to the power measurements taken for these samples. The average error of the power estimates provided by the model is about 10% on both platforms, which is acceptable and consistent with the results demonstrated by Shen et al. [2013]. However, the regression models produce anomalies by predicting the power of cache access intensive code blocks to be lower than the power of code blocks with `nop` instructions. Thus, accuracy of the modeling could be insufficient to capture the effect of code blocks on power consumption, even though the average error of the linear regression is relatively small.

Intel's RAPL provides access to a digital power meter that predicts power consumption on the basis of performance event counters, similar to linear regression models [Rotem et al. 2012]. However, RAPL uses about 100 internal microarchitecture counters [Rotem et al. 2011]. Unfortunately, no more than 12 performance counters can be accessed by system software at once through the performance monitoring unit (PMU). On the Exynos platform, the PMU module allows the user to sample only 7 performance counters at once. Thus, it is likely that RAPL achieves better accuracy by having fast hardware-supported access to many more performance counters than our model.

7. USE CASES

We now present practical examples of how ALEA enables energy-aware optimizations. We first explore the effect of energy optimizations on power consumption in a latency-critical financial risk application where we apply the VPM. We then demonstrate results for energy optimization based on runtime DVFS control using the CPM.

7.1. Energy Optimization of a Latency-Critical Workload

We applied ALEA to profile and optimize a real-time and latency-critical option pricing workload. An option pricing model allows financial institutions to predict a future payoff using the current price and volatility of an underlying instrument. However, the current price is valid only for a limited interval, which implies that a decision about buying or selling the instrument is time critical. Binomial option pricing is a popular pricing model that enables accurate prediction with low computational requirements [Georgakoudis et al. 2016]. We use a commercial binomial option pricing kernel and a realistic setup that delivers requests to a worker thread running the kernel. The setup consists of two threads: the first thread listens to a TCP/IP socket, receives messages, and fills a queue with the messages; the second thread parses the messages and transfers requests to the worker thread. A third thread, the worker thread, processes option pricing requests.

The option pricing kernel uses the current price, volatility, and number of nodes in the binomial tree as input parameters. The latency of the kernel depends only on the number of nodes, as variation of the current price or volatility does not affect control and data flows of the kernel. Accuracy of the prediction is also primarily affected by the number of nodes in the binomial tree. Although the current price and volatility could change in option pricing requests, in practice financial institutions use the minimal number of nodes that guarantees an acceptable quality of the prediction to reduce a pricing kernel response time. Consistent with this, the latency of the kernel should be a constant if the number of nodes does not vary. Following private communication with a leading financial institution, we used 512 nodes for realistic runs. The latency of the binomial option pricing kernel, which is about 177μs, is less than the power-sensing interval (280μs) for this number of nodes on the Exynos platform.

According to the results of energy profiling, two code blocks of the kernel with corunning system threads consume more than 98% of the energy spent per request: the first code block builds a binomial tree, and the second block computes the value of the

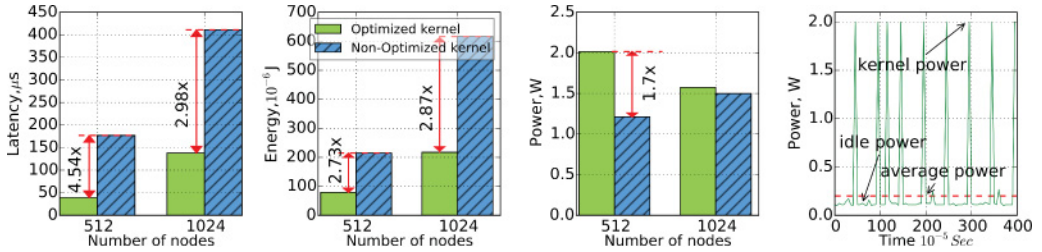

Fig. 12. Latency, energy, and power for the binomial option pricing kernel. The rightmost plot presents power changes during a run of the workload for the optimized kernel (512 nodes).

option used to estimate the future payoff. To build the binomial tree, the first code block calls the exponential function in a loop, which is expensive in terms of energy and performance. The parameters of the function depend on an iteration of the loop and the underlying volatility, which is constant per request. To improve energy efficiency, we modified the kernel to dynamically identify the most frequently occurring volatility in option pricing requests and build at runtime an array containing the results of the exponent call for each iteration and this volatility (an example of memoization). In the optimized version of the kernel, we use the precomputed results instead of calls of the exponential function when the volatility of a request conforms with the most frequently occurring volatility.

Our experiments show that in 99.9% of the samples, the first system thread executes the `do_futex_wait` system call and the second thread executes the `epoll_wait` system call when the worker thread runs the kernel. Both system calls keep a core in sleep mode waiting for an incoming network message (the first thread) or a message from the queue (the second thread). The highest power consumption is observed when the worker thread runs the kernel.

We apply the VPM to profile the kernel while reproducing a real workload of option pricing requests from an actual trading day in the NYSE. Figure 12 shows latency, energy, and power consumption of the optimized and nonoptimized versions of the kernel. By applying memoization, we managed to reduce latency of the kernel by 2.98 times and energy consumption by 2.87 times with binomial trees of 1,024 nodes. We also reduced latency by 4.54 times for trees with 512 nodes. However, the energy reduction in this case was only 2.73 times. We explain this by a power spike that accompanied the latency reduction: the difference in the power consumption between the optimized and nonoptimized versions for 512 nodes is 1.7 times, whereas this difference is only 4% for 1,024 nodes. The explanation for the spike is that the number of cache accesses executed per second is higher in the optimized kernel with 512 nodes. These results demonstrate that energy and performance optimizations could lead to a significant rise in power consumption—this is a critical parameter for datacenter operators. ALEA provides the user with an option to choose an acceptable level of performance and energy optimization that meets a power budget. For example, on our Exynos platform, the user should not use the optimized version for binomial trees with 512 nodes and power capped at 2 Watts. As follows, ALEA enables users to optimize code under controlled power consumed at the fine-grained level without applying more intrusive power capping techniques such as DVFS [Petoumenos et al. 2015].

ALEA reveals a further interesting finding. The average power consumption over the entire workload run with the optimized kernel (512 nodes) is 0.2 Watts (rightmost plot of Figure 12); this is 10 times less than the power consumption of the kernel. The difference is due to power variations during the workload run: the minimum power consumption (0.12 Watts) is observed when all threads stay idle waiting for

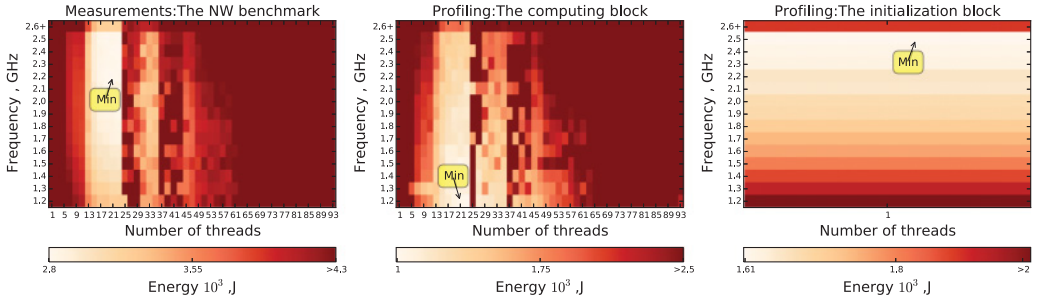

Fig. 13. Energy consumption versus frequency and number of threads for the Needleman-Wunsch benchmark.

incoming requests, and the maximum power consumption (2 Watts) is observed when the kernel is running. Furthermore, the average power consumption over the entire run with the nonoptimized version is 0.28 Watts, which is 40% higher than the average power consumption of the workload for the optimized version (0.2 Watts). In contrast, the power consumption of the optimized kernel is 1.7 times higher than the power consumption of the nonoptimized kernel. This observation suggests that the average power consumption of the entire workload cannot characterize the power consumption of the key kernel in the workload.

7.2. Energy Optimization with Runtime DVFS Control

The frequency that minimizes energy consumption depends on how intensely the cores interact with parts of the system not affected by frequency scaling, especially the main memory. Most existing approaches control frequency at the phase granularity, based on performance profiling, performance counter information, or direct energy measurements. We show how to use ALEA to drive decisions at any DVFS-supported granularity.

We use the Needleman-Wunsch benchmark from the `Rodinia` suite with a 40,000 $\times$40,000 matrix (approximately 6GB) as its input, executed on the Sandy Bridge platform. We first identify the optimal frequency and amount of parallelism for the whole program by taking energy measurements for the entire program. We test all combinations of frequency and number of threads. The leftmost plot of Figure 13 shows the results. The x-axis is the number of threads (1 to 96), the y-axis is frequency (1.2 to 2.6+ GHz), and the darkness of the color indicates energy consumption. The highest point on the y-axis, 2.6+GHz, corresponds to enabled Intel Turbo Boost, in which case the processor can automatically raise the frequency above 2.6GHz. The minimum energy consumption (2,801 Joules) is observed when using 21 threads at 2.2GHz.

ALEA finds two distinct code blocks that consume 99% of the total energy. The first code block uses one thread to initialize the arrays used by the benchmark. The second code block processes these arrays in parallel. We profile these blocks using the CPM at different frequencies and with different thread counts for the second block. The middle and the rightmost plots of Figure 13 show the profiling results for the second and first blocks, respectively. The two blocks consume their minimum energy at different frequencies, 2.5GHz for the initialization block and 1.2GHz for the computing block.

The blocks have different memory access patterns, which explains their energy profiles. Both read data from an array much larger than the caches with similar frequency. However, the initialization block has high temporal and spatial locality, so most memory loads are serviced by the caches, processor stalls are few, and performance depends only on frequency. Thus, limiting the frequency too much increases execution time

more than power reduction compensates. Alternatively, the computing block reuses data once or not at all, reading data from completely different pages in each iteration. Most accesses result in cache and TLB misses, stalling the processor and making it run effectively at the speed of main memory. Lowering the frequency of the processor has little effect on performance, so the optimal choice is to use the lowest frequency.

With this knowledge, we dynamically set the frequency to the optimal one at the beginning of each block. Compared to using a single frequency, optimal for the program as a whole, we reduced energy consumption by almost 9% (2,581 Joules). In terms of peak power, which is reached when executing the parallel computing group, the improvement is significant at 41.6%, whereas performance degradation is only 3.5%.

Several works consider different approaches to reduce energy consumption through runtime DVFS [Wu et al. 2005]. However, most of these works use analytical models to predict the voltage and frequency setting, which reduces energy consumption without a significant performance degradation. Central to such approaches is that the analytical models typically rely on parameters that have to be discovered for each platform individually. Moreover, the identified parameters do not guarantee the optimal voltage/frequency in terms of energy consumption. ALEA mitigates these restrictions by assigning accurate power and energy estimates to source code for each voltage/frequency set.

8. RELATED WORK

PowerScope [Flinn and Satyanarayanan 1999], an early energy profiling mechanism, profiles mobile systems through sampling of the system activity and power consumption, which it attributes to processes and procedures.

The tool does not use a probabilistic model and does not address randomness of the sampling process to enable fine-grained energy profiling. PowerScope assumes that for most samples, the program executes only one procedure during the sampling interval. This assumption holds only for coarse-grained procedures with latency greater than the sampling interval, which is approximately 1.6ms in PowerScope. Therefore, the tool cannot perform energy accounting at a fine-grained level. Furthermore, the accuracy of energy estimates is limited by the minimum feasible power-sensing interval provided by hardware sensors. In this article, we demonstrate that ALEA effectively mitigates all of these restrictions by following a probabilistic approach.

Several tools for energy profiling use manual instrumentation to collect samples of hardware event rates from hardware performance monitors (HPMs) [Curtis-Maury et al. 2008; Bertran et al. 2013; Isci and Martonosi 2003; Li et al. 2013; Contreras and Martonosi 2005; Schubert et al. 2012]. These tools empirically model power consumption as a function of activity rates. These rates attempt to capture the utilization and dynamic power consumption of specific hardware components. HPM-based tools and their models have guided several power-aware optimization methods in computing systems. However, they often estimate power with low accuracy. Further, they rely on architecture-specific training and calibration.

Eprof [Pathak et al. 2012] models hardware components as finite state machines with discrete power states and emulates their transitions to attribute energy use to system call executions. JouleUnit [Wilke et al. 2013] correlates workload profiles with external power measurement to derive energy profiles across method calls. JouleMeter [Kansal and Zhao 2008] uses postexecution event tracing to map measured energy consumption to threads or processes. Although useful, these tools can only perform energy accounting of coarse-grained functions or system calls, which severely limits the scope of power-aware optimizations that can be applied at compile time or runtime.

Other approaches measure power consumption of distinct components. PowerPack [Ge et al. 2010] uses manual code instrumentation and platform-specific hardware

instrumentation for component-level power measurement to associate power samples with functions. NITOS [Keranidis et al. 2014] measures the energy consumption of mobile device components with a custom instrumentation device. Similarly, LEAP [McIntire et al. 2007] measures energy consumption of networked sensors with custom instrumentation hardware. These tools profile power at the hardware component level, thus capturing the contributions of non-CPU components, such as memories, interconnects, and storage and networking devices. ALEA complements these efforts. ALEA's sampling method can account for energy consumed by any hardware component between code blocks, whereas its statistical approach overcomes the limitations of coarse and variable power sampling frequency in system components.

Other energy profiling tools build instruction-level power models bottom-up from gate-level or design-time models to provide power profiles to simulators and hardware prototyping environments [Tu et al. 2014; Tsoi and Luk 2011]. These inherently static models fail to capture the variability in instruction-level power consumption due to the context in which instructions execute. Similarly, using microbenchmarks [Shao and Brooks 2013] to estimate the energy per instruction or per code blocks based on their instruction mix does not capture the impact of the execution context.

9. CONCLUSION

We presented and evaluated ALEA, a tool for fine-grained energy profiling based on a novel probabilistic approach. We introduced two probabilistic energy accounting models that provide different levels of accuracy for power and energy estimates, depending on the targeted code granularity. We demonstrated that ALEA achieves highly accurate energy estimates with worst-case errors under 2% for coarse-grained code blocks and 6% for fine-grained ones.

ALEA overcomes the fundamental limitation of the low sampling frequency of power sensors. Thus, ALEA is more accurate than existing energy profiling tools when profiling programs with fast changing power behavior. ALEA opens up previously unexploited opportunities for power-aware optimization in computing systems. Its portable user space module does not require modifications to the application or the OS for deployment. Using ALEA, we profiled two realistic applications and evaluated the effect of optimizations enabled uniquely by ALEA on their energy and power. Overall, ALEA supports simpler, more targeted application optimization.

ACKNOWLEDGMENTS

We are grateful to Jose R. Herrero (UPC), Pavel Ilyin (Soft Machines), and the anonymous reviewers for their insightful comments and feedback.

REFERENCES

Ramon Bertran, Alper Buyuktosunoglu, Pradip Bose, Timothy J. Slegel, Gerard Salem, Sean Carey, Richard F. Rizzolo, and Thomas Strach. 2014. Voltage noise in multi-core processors: Empirical characterization and optimization opportunities. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-47)*. IEEE, Los Alamitos, CA, 368–380. DOI: http://dx.doi.org/10.1109/MICRO.2014.12

Ramon Bertran, Marc Gonzalez Tallada, Xavier Martorell, Nacho Navarro, and Eduard Ayguade. 2013. A systematic methodology to generate decomposable and responsive power models for CMPs. *IEEE Transactions on Computers* 62, 7, 1289–1302.

Niels Brouwers, Marco Zuniga, and Koen Langendoen. 2014. NEAT: A novel energy analysis toolkit for free-roaming smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems (SenSys'14)*. ACM, New York, NY, 16–30.

Ting Cao, Stephen M. Blackburn, Tiejun Gao, and Kathryn S. McKinley. 2012. The yin and yang of power and performance for asymmetric hardware and managed software. In *Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA'12)*. IEEE, Los Alamitos, CA, 225–236.

Fay Chang, Keith I. Farkas, and Parthasarathy Ranganathan. 2003. Energy-driven statistical sampling: Detecting software hotspots. In *Proceedings of the 2nd International Conference on Power-Aware Computer Systems (PACS'02)*. 110–129.

Gilberto Contreras and Margaret Martonosi. 2005. Power prediction for Intel XScale processors using performance monitoring unit events. In *Proceedings of the 2005 International Symposium on Low Power Electronics and Design (ISLPED'05)*. ACM, New York, NY, 221–226.

Matthew Curtis-Maury, Ankur Shah, Filip Blagojevic, Dimitrios S. Nikolopoulos, Bronis R. de Supinski, and Martin Schulz. 2008. Prediction models for multi-dimensional power-performance optimization on many cores. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (PACT'08)*. ACM, New York, NY, 250–259.

S. Das, P. Whatmough, and D. Bull. 2015. Modeling and characterization of the system-level power delivery network for a dual-core ARM Cortex-A57 cluster in 28nm CMOS. In *Proceedings of the 2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED'15)*. 146–151. DOI: http://dx.doi.org/10.1109/ISLPED.2015.7273505

J. Flinn and M. Satyanarayanan. 1999. PowerScope: A tool for profiling the energy usage of mobile applications. In *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*.

Rong Ge, Xizhou Feng, Shuaiwen Song, Hung-Ching Chang, Dong Li, and Kirk W. Cameron. 2010. PowerPack: Energy profiling and analysis of high-performance systems and applications. *IEEE Transactions on Parallel and Distributed Systems* 21, 5, 658–671.

Giorgis Georgakoudis, Charles J. Gillan, Ahmed Sayed, Ivor Spence, Richard Faloon, and Dimitrios S. Nikolopoulos. 2016. Methods and metrics for fair server assessment under real-time financial workloads. *Concurrency and Computation: Practice and Experience* 28, 3, 916–928. DOI: http://dx.doi.org/10.1002/cpe.3704

Susan L. Graham, Peter B. Kessler, and Marshall K. Mckusick. 1982. Gprof: A call graph execution profiler. *ACM SIGPLAN Notices* 17, 6, 120–126. DOI: http://dx.doi.org/10.1145/872726.806987

E. Grochowski, D. Ayers, and V. Tiwari. 2002. Microarchitectural simulation and control of di/dt-induced power supply voltage variation. In *Proceedings of the 8th International Symposium on High-Performance Computer Architecture*. 7–16. DOI: http://dx.doi.org/10.1109/HPCA.2002.995694

M. S. Gupta, J. L. Oatley, R. Joseph, G. Y. Wei, and D. M. Brooks. 2007. Understanding voltage variations in chip multiprocessors using a distributed power-delivery network. In *Proceedings of the 2007 Design, Automation, and Test in Europe Conference and Exhibition*. 1–6. DOI: http://dx.doi.org/10.1109/DATE.2007.364663

Hardkernel. 2016. ODROID-XU+E. Retrieved February 13, 2017, from http://www.webcitation.org/6f2nShdcN.

Canturk Isci and Margaret Martonosi. 2003. Runtime power monitoring in high-end processors: Methodology and empirical data. In *Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-36)*. IEEE, Los Alamitos, CA, 93.

R. Joseph, D. Brooks, and M. Martonosi. 2003. Control techniques to eliminate voltage emergencies in high performance processors. In *Proceedings of the 9th International Symposium on High-Performance Architecture (HPCA-9)*. 79–90. DOI: http://dx.doi.org/10.1109/HPCA.2003.1183526

Aman Kansal and Feng Zhao. 2008. Fine-grained energy profiling for power-aware application design. *SIGMETRICS Performance Evaluation Review* 36, 2, 26–31.

Stratos Keranidis, Giannis Kazdaridis, Virgilios Passas, Giannis Igoumenos, Thanasis Korakis, Iordanis Koutsopoulos, and Leandros Tassiulas. 2014. NITOS mobile monitoring solution: Realistic energy consumption profiling of mobile devices. In *Proceedings of the 5th International Conference on Future Energy Systems (e-Energy'14)*. ACM, New York, NY, 219–220.

Dong Li, Bronis R. de Supinski, Martin Schulz, Dimitrios S. Nikolopoulos, and Kirk W. Cameron. 2013. Strategies for energy-efficient resource management of hybrid programming models. *IEEE Transactions on Parallel and Distributed Systems* 24, 1, 144–157.

Ioannis Manousakis, Foivos S. Zakkak, Polyvios Pratikakis, and Dimitrios S. Nikolopoulos. 2014. TProf: An energy profiler for task-parallel programs. *Sustainable Computing: Informatics and Systems* 5, 1–13. DOI: http://dx.doi.org/10.1016/j.suscom.2014.07.004

Dustin McIntire, Thanos Stathopoulos, and William Kaiser. 2007. Etop: Sensor network application energy profiling on the LEAP2 platform. In *Proceedings of the 6th International Conference on Information Processing in Sensor Networks (IPSN'07)*. ACM, New York, NY, 576–577.

Lev Mukhanov, Dimitrios S. Nikolopoulos, and Bronis R. de Supinski. 2015. ALEA: Fine-grain energy profiling with basic block sampling. In *Proceedings of the 24th International Conference on Parallel Architectures and Compilation Techniques (PACT'15)*.

Abhinav Pathak, Y. Charlie Hu, and Ming Zhang. 2012. Where is the energy spent inside my app?: Fine grained energy accounting on smartphones with Eprof. In *Proceedings of the 7th ACM European Conference on Computer Systems (EuroSys'12)*. ACM, New York, NY, 29–42.

P. Petoumenos, L. Mukhanov, Z. Wang, H. Leather, and D. S. Nikolopoulos. 2015. Power capping: What works, what does not. In *Proceedings of the 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS'15)*. 525–534. DOI:http://dx.doi.org/10.1109/ICPADS.2015.72

E. Rotem, A. Naveh, A. Ananthakrishnan, E. Weissmann, and D. Rajwan. 2012. Power-management architecture of the Intel microarchitecture code-named Sandy Bridge. *IEEE Micro* 32, 2, 20–27. DOI:http://dx.doi.org/10.1109/MM.2012.12

Efi Rotem, Alon Naveh, Doron Rajwan, Avinash Ananthakrishnan, and Eli Weissmann. 2011. Power management architecture of the 2nd generation Intel Core microarchitecture, formerly codenamed Sandy Bridge. In *Proceedings of the 2011 IEEE Hot Chips 23 Symposium (HCS'11)*.

Simon Schubert, Dejan Kostic, Willy Zwaenepoel, and Kang G. Shin. 2012. Profiling software for energy consumption. *Proceedings of the 2012 IEEE International Conference on Green Computing and Communications.* 515–522.

Yakun Sophia Shao and David Brooks. 2013. Energy characterization and instruction-level energy model of Intel's Xeon Phi processor. In *Proceedings of the 2013 International Symposium on Low Power Electronics and Design (ISLPED'13)*. IEEE, Los Alamitos, CA, 389–394.

Kai Shen, Arrvindh Shriraman, Sandhya Dwarkadas, Xiao Zhang, and Zhuan Chen. 2013. Power containers: An OS facility for fine-grained power and energy management on multicore servers. *ACM SIGPLAN Notices* 48, 4, 65–76. DOI:http://dx.doi.org/10.1145/2499368.2451124

Tomothy Sherwood, Erez Perelman, and Brad Calder. 2001. *Basic Block Distribution Analysis to Find Periodic Behavior and Simulation Points in Applications*. Technical Report. University of California at San Diego, La Jolla, CA.

L. D. Smith, R. E. Anderson, D. W. Forehand, T. J. Pelc, and T. Roy. 1999. Power distribution system design methodology and capacitor selection for modern CMOS technology. *IEEE Transactions on Advanced Packaging* 22, 3, 284–291. DOI:http://dx.doi.org/10.1109/6040.784476

TI. 2013. *High- or Low-Side Measurement, Bidirectional CURRENT/POWER MONITOR with 1.8-V I2CTM Interface*.

Kuen Hung Tsoi and Wayne Luk. 2011. Power profiling and optimization for heterogeneous multi-core systems. *ACM SIGARCH Computer Architecture News* 39, 4, 8–13.

Chia-Heng Tu, Hui-Hsin Hsu, Jen-Hao Chen, Chun-Han Chen, and Shih-Hao Hung. 2014. Performance and power profiling for emulated Android systems. *ACM Transactions on Design Automation of Electronic Systems* 19, 2, Article No. 10.

Claas Wilke, Sebastian Götz, and Sebastian Richly. 2013. JouleUnit: A generic framework for software energy profiling and testing. In *Proceedings of the 2013 Workshop on Green in/by Software Engineering (GIBSE'13)*. ACM, New York, NY, 9–14.

Q. Wu, M. Martonosi, D. W. Clark, V. J. Reddi, D. Connors, Y. Wu, J. Lee, and D. Brooks. 2005. A dynamic compilation framework for controlling microprocessor energy and performance. In *Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-38)*. IEEE, Los Alamitos, CA, 271–282. DOI:http://dx.doi.org/10.1109/MICRO.2005.7

Xuan Zhang, Tao Tong, Svilen Kanev, Sae Kyu Lee, Gu-Yeon Wei, and David Brooks. 2013. Characterizing and evaluating voltage noise in multi-core near-threshold processors. In *Proceedings of the 2013 International Symposium on Low Power Electronics and Design (ISLPED'13)*. IEEE, Los Alamitos, CA, 82–87.

Received May 2016; revised September 2016; accepted November 2016