PROFET: Modeling System Performance and Energy Without Simulating the CPU

Milan Radulovic
Barcelona Supercomputing Center
(BSC) & Universitat Politècnica de
Catalunya (UPC)
Barcelona, Spain
milan.radulovic@bsc.es

Rommel Sánchez Verdejo
BSC & UPC
Barcelona, Spain
rommel.sanchez@bsc.es

Paul Carpenter
BSC
Barcelona, Spain
paul.carpenter@bsc.es

Petar Radojković
BSC
Barcelona, Spain
petar.radojkovic@bsc.es

Bruce Jacob
University of Maryland
College Park, Maryland, USA
blj@umd.edu

Eduard Ayguadé
BSC & UPC
Barcelona, Spain
eduard@ac.upc.edu

KEYWORDS

Memory bandwidth; Memory access latency; DRAM; MCDRAM; Performance; Power; Energy; Modeling

ACM Reference Format:

Milan Radulovic, Rommel Sánchez Verdejo, Paul Carpenter, Petar Radojković, Bruce Jacob, and Eduard Ayguadé. 2019. PROFET: Modeling System Performance and Energy Without Simulating the CPU. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts)*, June 24–28, 2019, Phoenix, AZ, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3309697.3331502>

1 INTRODUCTION

Application performance on novel memory systems is typically estimated using a hardware simulator. The simulation is, however, time consuming, which limits the number of design options that can be explored within a practical length of time. Also, although memory simulators are typically well validated, current CPU simulators have various shortcomings, such as simplified out-of-order execution, an obsolete data prefetcher and a lack of virtual-to-physical memory translation, all of which can make a huge difference between the simulated and actual memory system.

This study proposes PROFET (PROFiling-based EsTimation of performance and energy), an analytical model that predicts how an application's performance, power and energy consumption would change when it is executed on a new memory system. The method is based on instrumentation of an application running on actual hardware, so it already takes account of CPU microarchitectural details such as the real (and not publicly disclosed) data prefetcher and out-of-order engine. Therefore, it can be used to model various platforms as long as they support the required application profiling. PROFET was initially developed for the Sandy Bridge platform, and later we evaluated it for the Knights Landing (KNL) server. Adjustment of the PROFET model to the KNL system was trivial,

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 bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6678-6/19/06.

<https://doi.org/10.1145/3309697.3331502>

as it required changes to only a few hardware parameters, such as the reorder buffer size.

We evaluated PROFET on two actual platforms: Sandy Bridge-EP E5-2670 with four DRAM configurations DDR3-800/1066/1333/1600, and Knights Landing Xeon Phi with DDR4 and 3D-stacked MCDRAM. The evaluation results show that PROFET's predictions are **highly-accurate**, typically with only 2% difference from the performance, power and energy consumption measurements on the actual hardware. We also compare PROFET's performance predictions with simulation results for the Sandy Bridge-EP E5-2670 system with ZSim and DRAMSim2, and PROFET shows significantly better accuracy over the simulator. PROFET is also faster than the hardware simulators by *three orders of magnitude*, so it can be used to analyze production HPC applications, on arbitrarily sized systems.

We release the PROFET source code as open source [1]. The released PROFET model is ready to be used on high-end Intel platforms, and we would encourage the community to use it, adapt it to other platforms, and share their own evaluations.

2 PROFET OVERVIEW

The main idea behind the PROFET is that we can understand the effect of changing the memory system by understanding how the application moves from one bandwidth-latency curve to another. We illustrate this idea using the DDR4 and MCDRAM memories on Intel's Knights Landing platform. This platform has two memory systems, so there are two bandwidth-latency curves, shown together on the same plot in Figure 1.

When an application (or application phase) executes on the DDR4 main memory, it will be positioned at some point on the

Figure 1: High-level view of the transition from DDR4 to high-bandwidth MCDRAM memory on the KNL platform.

DDR4 curve; e.g. (BW_{used}^{DDR4} , lat_{mem}^{DDR4}) illustrated in Figure 1. Analogously, when the same application is executed on the MCDRAM memory, it will be positioned at some point on the MCDRAM curve, e.g. (BW_{used}^{MCDRAM} , lat_{mem}^{MCDRAM}). We see that the application in Figure 1 benefits from running on the MCDRAM through a lower memory latency (MCDRAM point is lower).

2.1 PROFET inputs

The inputs to PROFET are the memory bandwidth–latency curves and the power parameters, some basic parameters of the CPU under study, and the application profiles. These inputs can all be easily obtained on mainstream platforms and many emerging platforms. The outputs from PROFET will be the predicted performance, power and energy consumption on the target memory system.

Memory system profiling is done via bandwidth–latency curves, for the baseline and target memory systems, as illustrated in Figure 1. **Memory power parameters** characterize the memory systems in terms of the power consumption in various operational modes, idle state and power-down states, as well as the energy consumption for various operations such as read and write transfers, row buffer hits and misses. These figures are typically provided by the memory device manufacturers.

CPU parameters are needed, alongside the application profiling, to characterize the relationship between memory system latency and execution time. This relationship is dependent on the processor's ability to hide memory latency by overlapping memory accesses with independent instructions. The PROFET performance model therefore requires some basic parameters of the processor under study, such as the capacities of the re-order buffer and miss information status holding register (MSHR).

Application profiling is done on the baseline memory system, and consists of executing the application and profiling it using hardware performance counters. *Application performance profiling* obtains the number of CPU cycles, number of instructions, number of last-level cache misses and the read and write memory bandwidths. *Application power profiling* measures the total power consumption using integrated or external power measurement infrastructure, and memory-related power parameters using performance counters.

2.2 Performance, power and energy estimation

The **PROFET performance model** determines the application's position on the bandwidth–latency curve for the baseline memory system, e.g., DDR4 curve in Figure 1. The PROFET then estimates the application's position on the memory bandwidth–latency curve for the target memory system, e.g., MCDRAM curve in Figure 1, and uses it to predict the application's performance on the target memory system. The **PROFET power model** estimates the power consumption of the target memory system using the application performance profiling and the memory power parameters. Finally, the **PROFET energy model** is done based on the output of the performance and power models.

3 EVALUATION

We evaluated PROFET on a set of SPEC CPU2006 benchmarks and scientific HPC applications selected from the Unified European Application Benchmark Suite (UEABS). The performance evaluation

is done on two actual platforms: Sandy Bridge-EP E5-2670 with four DRAM configurations DDR3-800/1066/1333/1600, and Knights Landing Xeon Phi with DDR4 and 3D-stacked MCDRAM. PROFET's power and energy estimates are evaluated versus measurements of an external power meter taken on the actual Sandy Bridge server.

The evaluation results show that PROFET's predictions are very accurate: the average difference from the performance, power and energy measured on the actual hardware is only 2%, 1.1% and 1.7%, respectively. PROFET provides accurate estimations even when the baseline and target memory systems have fundamentally different bandwidth–latency curves, with different lead-off latency and with an n -fold difference in the available bandwidth.

We also compare the performance estimates of PROFET with ZSim+DRAMSim2 hardware simulators. The comparison is done based on the actual measurements on the Sandy Bridge platform. PROFET shows significantly better accuracy while being three orders of magnitude faster than the hardware simulators.

4 CONCLUSIONS

This study presents PROFET, an analytical model that quantifies the impact of the main memory on application performance and system power and energy consumption. PROFET is based on memory system profiling and instrumentation of an application execution on a real platform. The outputs from PROFET are the predicted performance, power and energy consumption on the target memory.

PROFET is evaluated on two actual platforms: Sandy Bridge-EP E5-2670 and Knights Landing Xeon Phi platforms with various memory configurations. The evaluation results show that PROFET's predictions are very accurate — the average difference from the performance, power and energy measured on the actual hardware is typically only about 2%.

We also compare PROFET's performance predictions with simulation results for the Sandy Bridge-EP E5-2670 system with ZSim and DRAMSim2 and PROFET shows significantly better accuracy over the simulators. PROFET is faster than the hardware simulators by *three orders of magnitude*, so it can be used to analyze production HPC applications, arbitrarily sized systems, and numerous design options, within a practical length of time. Also, PROFET does not require detailed CPU modeling since the CPU functionality, including actual data prefetcher and out-of-order engine, is already accounted for by the application profiling. Finally, PROFET can be used on various platforms as long as they support the required application profiling. PROFET was initially developed for the Sandy Bridge platform, and later we evaluated it for the KNL server. Adjustment of PROFET to the KNL system was trivial, requiring changes to only a few hardware parameters.

We release PROFET's source code and all input data required for memory system and application profiling [1]. The released model is ready to be used on high-end Intel platforms, and we would encourage the community to use it, adapt it to other platforms, and share their own evaluations.

REFERENCES

- [1] Milan Radulovic, Rommel Sanchez Verdejo, Paul Carpenter, Petar Radojković, Bruce Jacob, and Eduard Ayguadé. 2019. PROFET — Analytical model that quantifies the impact of the main memory on application performance and system power and energy consumption. <https://github.com/bsc-mem/PROFET>.