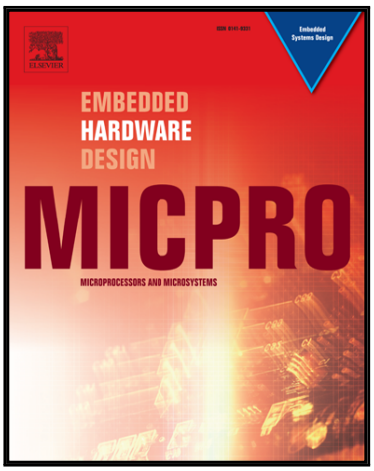

PowerTap: All-digital Power Meter Modeling for Run-time Power Monitoring

Davide Zoni, Luca Cremona, Alessandro Cilardo, Mirko Gagliardi, William Fornaciari

PII: S0141-9331(18)30230-8
DOI: 10.1016/j.micpro.2018.07.007
Reference: MICPRO 2723

To appear in: *Microprocessors and Microsystems*

Received date: 24 April 2018
Accepted date: 19 July 2018

Please cite this article as: Davide Zoni, Luca Cremona, Alessandro Cilardo, Mirko Gagliardi, William Fornaciari, PowerTap: All-digital Power Meter Modeling for Run-time Power Monitoring, *Microprocessors and Microsystems* (2018), doi: 10.1016/j.micpro.2018.07.007

PowerTap: All-digital Power Meter Modeling for Run-time Power Monitoring

Davide Zoni[a,*], Luca Cremona[a], Alessandro Cilardo[b], Mirko Gagliardi[b], William Fornaciari[a]

[a] *Politecnico di Milano - DEIB, 20133 Milan, Italy*
[b] *Università degli Studi di Napoli Federico II - DIETI and CeRICT, 80125 Napoli, Italy*

Abstract

The power consumption is a key metric to design computing platforms. In particular, the variety and complexity of current applications fueled an increasing number of run-time power-aware optimization solutions to dynamically trade the computational power for the power consumption. In this scenario, the online power monitoring methodologies are the core of any power-aware optimization, since the incorrect assessment of the run-time power consumption prevents any effective actuation. This work proposes *PowerTap*, an all-digital power modeling methodology for designing online power monitoring solutions. In contrast with state-of-the-art solutions, *PowerTap* adds domain-specific constraints to the data-driven power modeling problem. *PowerTap* identifies the power model iteratively to balance the accuracy error of the power estimates and the complexity of the final monitoring infrastructure. As a representative use-case, we employed a complex hardware multi-threaded SIMD processor, also considering different operating clock frequencies. The RTL implementation of the identified power model targeting an *Xilinx Artix 7 XC7A200T FPGA* highlights an accuracy error within 1.79% with an area overhead of 9.95% (LUT) and 3.87% (flip flops) and an average power overhead of 12.17 mW regardless of the operating conditions, i.e., number of software threads and operating frequency.

Keywords: Dynamic Power, Power Modeling, Power Monitoring, run-time power optimization, RTL methods, Low power

1. Introduction

The power consumption represents a major obstacle to any advancement in computing technologies, limiting the performance of both embedded and high performance computing (HPC) platforms. On one hand, embedded and portable devices operate within tight power budget constraints to prolong their battery lifetime. On the other hand, HPC platforms, that aim to maximize the performance, are becoming hot-spot limited since the performance increase is restricted by both the maximum junction temperature and the cost of the required cooling systems. While the literature contains several ad-hoc solutions to optimize the power and energy metrics of both the on-chip interconnect [1, 2] and the cache hierarchy [3, 4], the power consumed by the compute unit cores, i.e. microprocessors and accelerators like GPUs, represents a major component of the power budget in such systems, particularly in embedded and mobile platforms. As a consequence, the research community has explored an increasing number of online power monitoring techniques aimed at optimizing the trade-off between power and performance [5, 6, 7, 8]. Unlike special-purpose hardware, general-purpose units like microprocessors and GPUs pose significant challenges both because they are inherently less power efficient and more difficult to characterize by means of closed power models, due to the strong dependence on the software workloads. Furthermore, because general-purpose units are meant to provide the largest degree of flexibility to software applications, they are usually overprovisioned in terms of hardware resources and a significant portion of their subcomponents stay idle, depending on the requirements of the specific application (or application phase) being run [9]. On the other hand, the well-known *dark silicon* problem makes it impossible to concurrently power all the parts of the computing device due to the impossibility of dissipating the full amount of generated heat [10]. In this scenario, online power-aware optimization techniques may play a key role in that they allow the dynamic tuning of the available computing capacity aimed at maximizing the energy efficiency under given thermal constraints.

However, the effectiveness of such optimization techniques is critically subject to the employed power monitoring method as the incorrect assessment of the power state of the system strongly affects the quality of the actuation with a negative impact on the power efficiency on the platform. At run-time, the power consumption can be read out as either a *direct measurement* or an *indirect estimate*. The *direct measurement* is achieved by means of analog sensors providing highly accurate power values at high temporal resolution. However, such solution suffers from a severe scalability issue that limits the deployment of more than few sensors even in complex designs and the use of complex mixed analog-digital design methodologies to implement them. This fact also prevents the identification of the thermal hot-spots at run-time, thus negatively impacting the reliability of the computing platform [11]. In contrast, the *indirect estimate* is achieved by means of a power model of the target architecture that is fed with the platform statistics at run-

*Corresponding Author
Email addresses: davide.zoni@polimi.it (Davide Zoni), luca2.cremona@mail.polimi.it (Luca Cremona), acilardo@unina.it (Alessandro Cilardo), mirko.gagliardi@unina.it (Mirko Gagliardi), william.fornaciari@polimi.it (William Fornaciari)