Power and Thermal Management in Massive Multicore Chips: Theoretical Foundation meets Architectural Innovation and Resource Allocation

Paul Bogdan
Department of Electrical Engineering,
University of Southern California
3740 McClintock Ave.
Los Angeles, CA 90089-2562
pbogdan@usc.edu

Partha Pratim Pande
School of EECS
Washington State University
355 NE Spokane St. EME 102
Pullman, WA 99164-2752
pande@eecs.wsu.edu

Hussam Amrouch, Muhammad Shafique, Jörg Henkel
Chair for Embedded Systems
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany
(amrouch,shafique,henkel@kit.edu

ABSTRACT

Continuing progress and integration levels in silicon technologies make possible complete end-user systems consisting of extremely high number of cores on a single chip targeting either embedded or high-performance computing. However, without new paradigms of energy- and thermally-efficient designs, producing information and communication systems capable of meeting the computing, storage and communication demands of the emerging applications will be unlikely. The broad topic of power and thermal management of massive multicore chips is actively being pursued by a number of researchers worldwide, from a variety of different perspectives, ranging from workload modeling to efficient on-chip network infrastructure design to resource allocation. Successful solutions will likely adopt and encompass elements from all or at least several levels of abstraction. Starting from these ideas, we consider a holistic approach in establishing the Power-Thermal-Performance (PTP) trade-offs of massive multicore processors by considering three inter-related but varying angles, viz., on-chip traffic modeling, novel Networks-on-Chip (NoC) architecture and resource allocation/mapping

On-line workload (mathematical modeling, analysis and prediction) learning is fundamental for endowing the many-core platforms with ***self-optimizing*** capabilities [2][3]. This built-in intelligence capability of many-cores calls for monitoring the interactions between the set of running applications and the architectural (core and uncore) components, the online construction of mathematical models for the observed workloads, and determining the best resource allocation decisions given the limited amount of information about user-to-application-to-system dynamics. However, workload modeling is not a trivial task.

Centralized approaches for analyzing and mining workloads can easily run into scalability issues with increasing number of monitored processing elements and uncore (routers and interface queues) components since it can either lead to significant traffic and energy overhead or require dedicated system infrastructure. In contrast, learning the most ***compact mathematical representation of the workload*** can be done in a *distributed* manner (within the proximity of the observation /sensing) as long as the mathematical techniques are flexible and exploit the mathematical characteristics of the workloads (degree of periodicity, degree of fractal and temporal scaling) [3]. As one can notice, this strategy does not postulate a-priori the mathematical expressions (e.g., a specific order of the autoregressive moving average (ARMA) model). Instead, the periodicity and fractality of the observed computation (e.g., instructions per cycles, last level cache misses, branch prediction successes and failures, TLB access/misses) and communication (request-reply latency, queues utilization, memory queuing delay) metrics dictate the number of coefficients, the linearity or nonlinearity of the dynamical state equations and the noise terms (e.g., Gaussian distributed) [3]. In other words, dedicated minimal logic can be allocated to interact with the local sensor to analyze the incoming workload at run-time, determine the required number of parameters and their values as a function of their characteristics and communicate only the workload model parameters to a hierarchical optimization module (autonomous control architecture). For instance, capturing the fractal characteristics of the core and uncore workloads led to the development of more efficient power management strategy [1] than those based on PID or model predictive control.

In order to develop a compact and accurate mathematical framework for analyzing and modeling the incoming workload, we describe a general *probabilistic approach* that models the statistics of the increments in the magnitude of a stochastic process (associated with a specific workload metric) and the intervals of time (inter-event times) between successive changes in the stochastic process [3]. We show that the statistics of these two components of the stochastic process allows us to derive state equations and capture either short-range or long-range memory properties. To test the benefits of this new workload modeling approach, we describe its integration into a multi-fractal optimal control framework for solving the power management for a 64-core NoC-based manycore platform and contrast it with a mono-fractal and non-fractal schemes [3].

A scalable, low power, and high-bandwidth on-chip communication infrastructure is essential to sustain the predicted growth in the number of embedded cores in a single die. New interconnection fabrics are key for continued performance improvements and energy reduction of manycore chips, and an efficient and robust NoC architecture is one of the key steps towards achieving that goal. An NoC architecture that incorporates emerging interconnect paradigms will be an enabler for low-power, high-bandwidth manycore chips. Innovative interconnect paradigms based on optical technologies, RF/wireless methods, carbon nanotubes, or 3D integration are promising alternatives that may indeed overcome obstacles that impede continued advances of

CASES '16, October 01-07, 2016, Pittsburgh, PA, USA © 2016 ACM. ISBN 978-1-4503-4482-1/16/10...$15.00 DOI: http://dx.doi.org/10.1145/2968455.2974013

the manycore paradigm. These innovations will open new opportunities for research in NoC designs with emerging interconnect infrastructures. In this regard, wireless NoC (WiNoC) is a promising direction to design energy efficient multicore architectures. WiNoC not only helps in improving the energy efficiency and performance, it also opens up opportunities for implementing power management strategies. WiNoCs enable implementation of the two most popular power management mechanisms, viz., dynamic voltage and frequency scaling (DVFS) and voltage frequency island (VFI).

The wireless links in the WiNoC establish one-hop shortcuts between the distant nodes and facilitate energy savings in data exchange [3]. The wireless shortcuts attract a significant amount of the overall traffic within the network. The amount of traffic detoured is substantial and the low power wireless links enable energy savings. However, the overall energy dissipation within the network is still dominated by the data traversing the wireline links. Hence, by incorporating DVFS on these wireline links we can save more energy. Moreover, by incorporating suitable congestion aware routing with DVFS, we can avoid thermal hotspots in the system [4].

It should be noted that for large system size the hardware overhead in terms of on-chip voltage regulators and synchronizers is much more in DVFS than in VFI. WiNoC-enabled VFI designs mitigate some of the full-system performance degradation inherent in VFI-partitioned multicore designs, and it also help in eliminating it entirely for certain applications [5]. The VFI-partitioned designs used in conjunction with a novel NoC architecture like WiNoC can achieve significant energy savings while minimizing the impact on the achievable performance.

On-chip power density and temperature trends are continuously increasing due to high integration density of nano-scale transistors and failure of Dennard Scaling as a result of diminishing voltage scaling. Hence, all computing is temperature-constrained computing and therefore, employing thermal management techniques that keep chip temperatures within safe limits along with meeting the constraints of spatial/temporal thermal gradients and avoid wear-out effects [8]is key.

We introduced the novel concept of *Dark Silicon Patterning*, i.e. spatio-temporal control of power states of different cores [9] Sophisticated patterning and thread-to-core mapping decisions are made considering the knowledge of process variations and lateral heat dissipation of power-gated cores in order to enhance the performance of multi-threaded workloads through dynamic core count scaling (DCCS). This is enabled by a lightweight online prediction of chip's thermal profile for a given patterning candidate. We also present an enhanced temperature-aware resource management technique that, besides *active* and *dark* states of cores, also exploit various *grey* states (i.e., using different voltage-frequency levels) in order to achieve a high performance for mixed ILP-TLP workloads under peak temperature constraints. High ILP applications benefit from high V-f and boosting levels, while high TLP applications benefit from

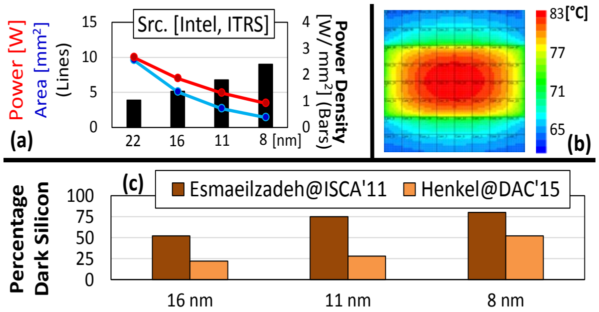

Figure 1: Trends for power density, temperature, dark silicon[10]

As the scaling trends move from multi-core to many-core processors, the centralized solutions become infeasible, and thereby require distributed techniques. In [6], we proposed an agent-based distributed temperature-aware resource management technique called *TAPE*. It assigns a so-called agent to each core, a software or hardware entity that acts on behalf of the core. Following the principles of economic theory, these agents negotiate with each other to trade their power budgets in order to fulfil the performance requirements of their tasks, while keep the $T_{Peak} \leq T_{critical}$. In case of thermal violations, task migration or V-f throttling is triggered, and a penalty is applied to the trading process to improve the decision making.

ACKNOWLEDGEMENTS

PB acknowledges the support by USA National Science Foundation (NSF) under CNS/CPS 1453860 and CyberSEES 1331610 grants. PP acknowledges support from USA NSF grants CCF-1514269, CCF-1162202 and Army Research Office grant W911NF-12-1-0373. JH's work is supported in parts by the German Research Foundation (DFG) as part of the Transregional Collaborative Research Centre Invasive Computing (SFB/TR 89 – http://invasic.de) and in parts by the priority program Dependable Embedded Systems (SPP 1500 –http://spp1500.itec.kit.edu.

REFERENCES

[1] P. Bogdan, R. Marculescu, and S. Jain, "Dynamic Power Management for Multidomain System-on-Chip Platforms: An Optimal Control Approach," *ACM Transactions on Design Automation of Electronic Systems*, 18, 4, Article 46, October 2013.

[2] P. Bogdan, "A cyber-physical systems approach to personalized medicine: challenges and opportunities for noc-based multicore platforms," *Proc. of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, March 09-13, 2015, Grenoble, France.

[3] P. Bogdan, "Mathematical Modeling and Control of Multifractal Workloads for Data-Center-on-a-Chip Optimization," *Proc. of the 9th International Symposium on Networks-on-Chip* (NOCS), 2015.

[4] S. Deb, et al., "Design of an Energy Efficient CMOS Compatible NoC Architecture with Millimeter-Wave Wireless Interconnects", *IEEE Transactions on Computers*, Vol.62, pp.2382-2396, Dec. 2013.

[5] J. Murray, et. al., "Performance Evaluation of Congestion-Aware Routing with DVFS on a Millimeter-Wave Small World Wireless NoC", *ACM Journal of Emerging Technologies in Computing Systems* (JETC), Volume 11 Issue 2, November 2014.

[6] R. G. Kim, et. al., "Wireless NoC for VFI-Enabled Multicore Chip Design: Performance Evaluation and Design Trade-offs", *IEEE Transactions on Computers*, Vol. 65, Issue 4, pp. 1323–1336.

[7] T. Ebi, et al. "TAPE: Thermal-aware agent-based power economy multi/many-core architectures", *ICCAD*, 2009.

[8] H. Amrouch et al., "Towards interdependencies of aging mechanisms", *IEEE/ACM 33rd Intl. Conf. on Computer-Aided Design* (ICCAD), San Jose, USA, pp. 478-485, 2014.

[9] M. Shafique et al., "Variability-aware dark silicon management in on-chip many-core systems", DATE, 2015.

[10] M. Shafique et al., "The EDA challenges in the dark silicon era: temperature, reliability, and variability perspectives", *DAC*, 2014.