Microprocessors and Microsystems 40 (2016) 88–101

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Design space exploration for device and architectural heterogeneity in chip-multiprocessors

Ying Zhang[a,*], Samuel Irving[a], Lu Peng[a], Xin Fu[b], David Koppelman[a], Weihua Zhang[c,d], Jesse Ardonne[a]

[a] Division of Electrical & Computer Engineering, School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA 70803, United States
[b] Department of Electrical and Computer Science, University of Houston, Houston, TX 77004, United States
[c] Software School, Fudan University, 201203, China
[d] Parallel Processing Institute, Fudan University, Shanghai, 201203, China

ARTICLE INFO

Keywords:
Heterogeneity
Cost efficiency
Energy efficiency

ABSTRACT

As we enter the deep submicron era, the number of transistors integrated on die is exponentially increased. While the additional transistors largely boost the processor performance, a repugnant side effect caused by the evolution is the ever-rising power consumption and chip temperature. It is widely acknowledged that the shortage of power supplied to a processor will be a major hazard to sustain the generational performance scaling, if the processor design is to follow the conventional approach. To utilize the on-chip resources in an efficient manner, computer architects need to consider new design paradigms that effectively leverage the advantages of modern semiconductor technology. In this paper, we address this issue by exploiting the device-heterogeneity and two-fold asymmetry in the processor manufacturing. We conduct a thorough investigation on these design patterns from different evaluation perspectives including performance, energy-efficiency, and cost-efficiency. Our observations can provide insightful guidance to the design of future processors.

1. Introduction

Processor manufacturers have been able to double transistor count and performance for each new product generation in past decades, as predicted by Moore's Law. However, as we enter the deep submicron era, the continuous decrease of the transistor supply and threshold voltage at each new technology node, known as Dennard Scaling has stalled [18,28], leading to an ever-increasing power density on modern processors. On the other hand, the maximum processor power consumption should always be enclosed within a reasonable envelope, regardless of manufacturing technology due to physical constraints such as heat dissipation and power delivery. Given these limitations, a large portion of the integrated transistors on a future processor must be significantly underclocked or even turned off in order to satisfy power constraints and maintain a safe working temperature. This phenomenon, which has been termed "dark silicon" [18], is recognized as one of the most critical constraints preventing us from obtaining commensurate performance benefits from increasing the number of transistors.

The problem might become exacerbated as Moore's Law continues to dominate processor development. According to the ITRS roadmap [5], the percentage of the chip that cannot be turned on is exponentially expanding with each generation, and up to 93% of all transistors on a chip would be forced inactive in a few years from now. Therefore, seeking new design dimensions to efficiently utilize chip-level resources including power and area is important for us to obtain sustainable performance improvements in the future. In this paper, we conduct a comprehensive assessment of new design dimensions with special concentration on heterogeneity in the early stage of processor manufacturing.

Our target processor is a chip multiprocessor (CMP) with a fixed power and area budget. The first dimension that will be evaluated is *device heterogeneity*. Since the gap between power requirement and supply capability is essentially caused by the slow improvement in a Complementary Metal–Oxide–Semiconductor (CMOS) device's switch power, emerging low-power materials might be used to fabricate processors in order to illuminate the dark area. However, many power-saving devices manufactured with nano-technology manifest a series of drawbacks such as long switch delay [21]. Due to this limitation, it is inappropriate to use such devices to completely replace the traditional CMOS in processor

* Corresponding author.
E-mail addresses: ying.esz.zhang@gmail.com (Y. Zhang), sirvin1@lsu.edu (S. Irving), lpeng@lsu.edu (L. Peng), xfu6@uh.edu (X. Fu), koppel@lsu.edu (D. Koppelman), zhangweihua@fudan.edu.cn (W. Zhang), jardon2@lsu.edu (J. Ardonne).

http://dx.doi.org/10.1016/j.micpro.2015.07.012
0141-9331/

manufacturing. Instead, integrating cores made of different materials on the same die emerges as an attractive design option. A few works have justified the feasibility of a hybrid-device CMP at the circuit level [24,31,33]. On the other hand, *architectural heterogeneity* (e.g., including both big and small cores on a processor) has proven to be an effective way to improve energy efficiency [25]. Therefore, jointly applying device and architectural heterogeneity becomes a promising option compared to conventional designs, hence the second design dimension "two-fold heterogeneity". The third aspect considered in this study is the *operating voltage/frequency* (*v/f*) of processors since it significantly impacts the processor power and thermal characteristics. Finally, the last factor that will be taken into consideration is a recently proposed technique "*computational sprinting*" [28] which allows the system to temporarily exceed the thermal-design power constraint in a burst fashion. In general, by evaluating the described dimensions in detail, we attempt to summarize a set of "principles" that can guide the design of processors in the next generation and beyond. The following is a list of the main observations made in this study.

- We demonstrate that the on-chip resources can be more efficiently utilized by using diverse materials in the chip fabrication. By integrating more cores made of slower power-saving devices and less cores built with faster yet power-consuming devices, more processor cores can be booted up, thus delivering better energy- and cost-efficiency.
- We explore processor designs with two-fold heterogeneity with regards to both manufacturing devices and core architectures. We show that by building complex out-of-order cores using power-saving devices while in conjunction with small in-order cores using relatively power-consuming material, we are able to deliver extra energy- and cost-efficiency benefits.
- We examine the impact of the voltage/frequency setting on the overall performance, energy- and cost-efficiency of the target processor. Our evaluations demonstrate that the most promising design pattern remains the same (i.e., building big cores with power-saving devices and small cores with faster devices) although appropriately setting the operating voltage/frequency can effectively increase the performance and efficiency of other configurations.
- We enable the computational sprinting technique on the target system and investigate its implication on the design pattern selection. The results show that this technique is capable of delivering better performance and execution efficiencies than regular configurations. Moreover, as for the distribution of the extra power in the sprinting phase, an "even" distribution (i.e., increase the frequency of all cores by an amount) is more preferable than "prioritized" distribution which gives all extra power to a few cores (e.g., the big cores).

2. Related work

The problem of power supply shortage for activating transistors (i.e., dark silicon) emerges as an increasingly important issue that jeopardizes the scaling of Moore's Law in the deep submicron era and beyond. For this reason, researchers recently started to investigate this problem and propose several solutions. Esmaeilzadeh et al. [18] use an analytical model to predict processor scaling for the next few generations and show that the percentage of unused transistors will be expanding as manufacturing technology keeps shrinking. Turakhia et al. [36] propose an iterative optimization based approach to investigate the optimal number of cores of each type with given area and power budget for heterogeneous CMPs, where cores with different architectures are made of identical devices. Hardavellas et al. [19] pay specific attention to the server processors and perform an exploration of throughput-oriented processors. Systems built with near-threshold voltage processors (NTV) [14] are also effective approaches.

As for the hybrid device study, Saripalli et al. [31] discuss the feasibility of technology-heterogeneous cores and demonstrate the design of mix-device memory. Wu et al. [38] presents the advantage of hybrid-device cache. Kultursay [24] and Swaminathan [33] respectively introduce a few runtime schemes to improve performance and energy efficiency on CMOS-TFET hybrid CMPs. Our work deviates from the aforementioned in that we conduct a more comprehensive study in the early stage of processor manufacturing. We propose to utilize architectural and device heterogeneity simultaneously to optimally utilize the on-chip resources and balance the performance, energy consumption and total cost. Additionally, in comparison to our previous work [42], this study extends the investigation to more important design factors and aims at drawing more comprehensive conclusions.

3. Methodology

3.1. Metrics

In this section, we describe metrics for the evaluation of different configurations. Note that we characterize multiple aspects including performance, energy efficiency, thermal features and cost-efficiency for each design configuration in order to make a comprehensive investigation.

We choose the total execution time for performance evaluation. For energy-efficiency and thermal features, we use energy-delay product (ED) and peak temperature for assessment. Besides these three extensively discussed metrics, we also include cost-efficiency as the fourth factor for investigation. In this work, we mainly concentrate on the operating cost which is essentially determined by the temperature during execution. The cost efficiency is defined as MIPS/dollar, a widely used metric in computer engineering studies that quantifies the efficiency in delivering performance at a specific cost [6,37,38]. The cooling cost is computed based on a model introduced in a prior work [41]:

$$C_{\text{cooling}} = K_c t + c \tag{1}$$

Note that both K_c and c are cooling cost parameters. K_c is a coefficient associated with the temperature and c is a fitted parameter dependent on the temperature range as well. In general, this cost is determined by the peak temperature achieved during execution. Note that K_c is a variable which is highly related to the steady temperature. High temperature t corresponds to a larger coefficient K_c and results in higher cooling cost consequently. Characterizing the cost-efficiency is necessary for computer architects to identify the optimal design configurations, thus deserving careful consideration.

3.2. Simulation environment and workloads

We use a modified SESC [29], a widely used cycle-accurate simulator for architectural study, to conduct our investigation. We choose McPat 1.0 [26] for power and area estimation and Hotspot 5.0 [32] for temperature calculation. Note that we assume the technology is 22 nm in this work, thus we set the system budget based on an Intel Ivy Bridge processor [3]. The area of the target chip should not exceed 100 mm^2 and the maximal power consumption is 60 W.

Recall that our design space includes configurations which integrate both big and small cores on the same chip. For this purpose, we assume a complex out-of-order core and a simple in-order core whose parameters are summarized from recent commercial processors [3,4,20] and are listed in Table 1. Given these conditions,