Lighting the Dark Silicon by Exploiting Heterogeneity on Future Processors

Ying Zhang Lu Peng Xin Fu† Yue Hu
Division of Electrical & Computer Engineering
School of Electrical Engineering and Computer Science
Louisiana State University
{yzhan29, lpeng, yhu14}@lsu.edu
†Electrical Engineering and Computer Science
School of Engineering
University of Kansas
xinfu@ittc.ku.edu

ABSTRACT
As we embrace the deep submicron era, dark silicon caused by the failure of Dennard scaling impedes us from attaining commensurate performance benefit from the increased number of transistors. To alleviate the dark silicon and effectively leverage the advantage of decreased feature size, we consider a set of design paradigms by exploiting heterogeneity 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 in the presence of dark silicon.

Categories and Subject Descriptors
C.1 **[PROCESSOR ARCHITECTURE]**: Heterogeneous systems; C.4 **[PERFORMANCE OF SYSTEMS]**: Design studies

General Terms
Design, Experimentation.

Keywords
Dark silicon, emerging device, heterogeneous

1. Introduction
Processor manufacturers have complied with Moore's Law to double the transistor count and performance on each new generation product in past decades. However, as we embrace the deep submicron era, Dennard scaling which describes the continuous decrease on the supply and threshold voltage of a transistor at each new technology node has stalled [8][17], leading to an ever increasing power density on modern processors. On the other hand, the maximum processor power consumption should be always enclosed within a reasonable envelope despite the manufacturing technology, due to physical constraints including heat dissipation and power delivery. Under this limitation, a large portion of integrated transistors on a future processor must be significantly underclocked or even completely turned off in order to satisfy the power constraint and maintain a safe working temperature. This phenomenon, which is termed the "dark silicon", is recognized to be one of the most critical constraints that prevent us from obtaining commensurate performance benefit from the increased number of transistors.

Dark silicon might be exacerbated as Moore's Law continues to dominate the processor development. Figure 1 illustrates the scaling trend of the amount of "dark" transistors according to the ITRS roadmap [3]. As can be seen, the percentage of the dark area on a chip is exponentially expanding at each generation. This results in a chip with up to 93% of all transistors inactive in a few years from now [23]. Therefore, seeking new design dimensions to efficiently utilize the chip-level resource including power and area is important for us to obtain sustainable performance improvement in the future. Prior works have proposed a few solutions to address the dark silicon problem from certain aspects [8][9][17][24][25]. However, most of these works mainly concentrate on a specific solution, lacking general justifications of multiple design options. Considering that an initial guidance to the design of future processors in the presence of dark silicon is highly desired, we conduct a comprehensive assessment of new design dimensions with special concentration on heterogeneity in the early stage of processor manufacturing.

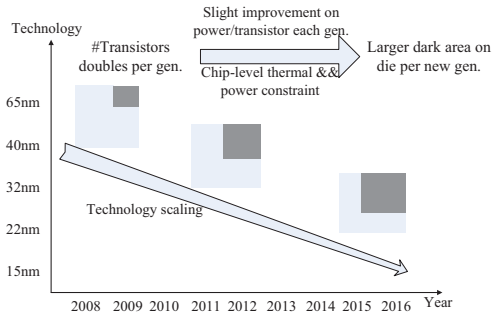


Figure 1. Increasing dark area with technology scaling

Our target processor is a chip multiprocessor (CMP) with fixed power and area budget. The first dimension that will be evaluated is *device heterogeneity*. Since dark silicon is essentially caused by the slow improvement in CMOS device's switch power, emerging low-power materials might be used to build 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 [11]. Due to this limitation, it is inappropriate to use such devices to completely replace the traditional CMOS in processor manufacturing. To effectively alleviate the power constraint without suffering from significant performance degradation, integrating cores made of different materials on the same die emerges as an attractive design option. A few works have justified the feasibility of hybrid-device CMP at circuit level [13][19][20][21] while some of them further demonstrate the advantage of the resultant processors in performance improvement [13]. Nevertheless, these works are mainly conducted on a fixed platform and thus the optimal design configuration which provides desirable balance among disparate evaluation metrics remains an open question. On the other hand, architectural heterogeneity (e.g., including both big and small cores on a processor) has been proved an effective solution to energy efficiency improvement [14][9]. Therefore, jointly applying the device heterogeneity and architec-

DAC'13, May 29 - June 07 2013, Austin, TX, USA.

tural heterogeneity becomes a promising option to further exploit their advantages over conventional designs, hence the second design dimension "*two-fold heterogeneity*". In general, by evaluating the described new design dimensions in detail, our study makes the following key observations:

- We demonstrate that using diverse materials in the chip fabrication is effective in relieving the dark silicon problem. By integrating more cores made of slower and power-saving device and relatively few cores built with faster yet power-consuming device, more processor cores can be booted up. Therefore, the advantages of both materials are leveraged, assisting us to produce processors that deliver impressive energy- and cost-efficiency.
- We observe that architectural heterogeneity is capable of offering higher cost-efficiency in addition to the well-known energy-efficiency over conventional designs, because including small low-power cores is able to reduce the peak chip temperature and thus decreasing the cooling expense. This further confirms the importance of building CMPs with different types of cores in the presence of dark silicon.
- We explore processor designs with two-fold heterogeneity with regards to both manufacturing devices and core architectures. We show that building complex out-of-order cores with power-saving device while manufacturing small in-order cores with relatively power-consuming material is able to deliver extra benefit on energy- and cost-efficiency, thus appearing as the optimal design option.

2. Methodology

2.1 Metric

In this section, we describe the 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 the energy-efficiency and thermal feature, 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 define the cost efficiency as *MIPS/dollar*. The considered cost is composed of the die cost and cooling expense, where the former part can be calculated with the following equations [16]:

$$Die\ cost = \frac{wafer\ cost}{Dies\ per\ wafer \times Die\ yield} \quad (1)$$

$$Dies\ per\ wafer = \frac{\pi \times \left(\frac{wafer\_diam}{2}\right)^2}{Die\ area} - \frac{\pi \times wafer\_diam}{\sqrt{2 \times Die\ area}} - Test\ dies \quad (2)$$

$$Die\ yield = wafer\ yield \times \left\{1 + \frac{Defects\ per\ unit\ area \times Die\ area}{\alpha}\right\}^{-\alpha} \quad (3)$$

Table 1. Parameter values for die cost calculation.

Parameter	Value
Wafer cost	$4900
Wafer diameter	300mm
Wafer yield	0.9
Defects per unit area	0.4/cm^2
Alpha	3

Table 1 lists the values of referred parameters derived from recently released data in industry [5][16]. The cooling cost is computed based on a model that is introduced in a prior work [28]:

$$C_{cooling} = K_c t + c \quad (4)$$

In general, this cost is determined by the peak temperature achieved during the execution. High temperature t corresponds to larger coefficient K_c and results in higher cooling cost as a consequence. Characterizing the cost-efficiency is necessary for computer architects to identify the optimal design configurations, thus deserving careful consideration.

Table 2. Architectural parameters for system components.

Component	Parameter	Value
Big core	Pipeline type	out-of-order
	Processor width	4
	ALU/FPU	4/4
	ROB/RF	160/160
	L1I cache size	32KB
	L1D cache size	32KB
	L1 associativity	4
Small core	Pipeline type	in-order
	Processor width	1
	ALU/FPU	1/1
	L1I cache size	8KB
	L1D cache size	8KB
	L1 associativity	2
Other parameters	L2 cache size	4MB
	L2 associativity	8
	Cache block size	32B
	Technology	22nm
	Frequency (High-K)	3G
	Chip area	100mm^2
	TDP	60W

Table 3. Estimated area and power for system components.

Component	Peak power	Area
Big core	5.6W (High-K)	7.6mm^2
	4.8W (NEMS-CMOS)	
Small core	1.1W(High-K)	1.97mm^2
	0.8W (NEMS-CMOS)	
L2 cache	0.8W/MB	3mm^2/ MB
Interconnect	5W	4mm^2
Other components	11W	23mm^2

Table 4. Selected applications for simulation.

Category	Benchmark Suite	Applications (Kernels)
Homogeneous	SPLASH-2	Barnes, FMM, Radix, Raytrace, Water-spatial, waterNS
	PARSEC	Blackscholes, Swaptions
	ALPBench	MPGDec, MPGEnc
Heterogeneous	Computation-intensive	h264, dealII, namd, sprand, sjeng, omnetpp, gobmk, hmmer, bzip2
	Memory-intensive	mcf, libquantum, milc, leslie3d, perlbench, lbm, soplex, astar

2.2 Simulation Environment and Workloads

We use a modified SESC [18], a widely used cycle-accurate simulator for architectural study, to conduct our investigation. We choose McPat 1.0 [15] for power and area estimation and Hotspot 5.0 [4] for temperature calculation. Note that we assume a 22nm technology in this work, thus we set the system budget based on an Intel Ivy Bridge processor [2]. In specific, the area of the target chip should not exceed 100mm^2 and the maximal power consumption is 60W.

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 listed in Table 2. Table 3 lists the estimated area and peak power for each component on the chip. Given these conditions, the number of cores that can be accommodated is determined by the following expressions:

$$Area\ constraint: N_b \times A_b + N_s \times A_s + A_{all\ other} \leq 100$$

$$Power\ constraint: N_b \times P_b + N_s \times P_s + P_{all\ other} \leq 60$$

where variables N_b and N_s denote the number of big cores and number of small cores respectively. Constants A_b and P_b indicate the area and peak power for a big core as listed in Table 3. Similar interpretations apply to other symbols such as A_s and P_s.

The workloads used for our exploration is based on the specific architecture in study. Multi-threaded programs are generally used for CMPs on which all cores have identical architecture (in the study of device heterogeneity); on the other hand, when both big and small cores are integrated, we consider that "heterogeneous"

Table 5. Features of materials considered in this work.

Material	Features
High-K	Reduce leakage power to 20% of the dynamic power
NEMS-CMOS	OR gate: 20% higher delay, reducing 60% switching power
	SRAM cell: 25% higher delay, saving 85% leakage energy

workloads are more appropriate for the investigation and thus use combinations of programs from SPEC CPU2006 as a substitute. For those parallel applications, the number of threads for execution always equals to the core count of the underlying CMP and all programs are executed till completion in order to guarantee that identical task is performed. We choose a total of 10 programs from SPLASH-2, PARSEC and ALPBench for the simulation. The reason for not including other workloads is that their intrinsic characteristics (e.g., requiring 2^n threads) prohibit the execution on many configurations. As for the SPEC mixes, each of them includes 30 individual programs (the maximum core count in all evaluated configurations). We simulate 100 million instructions after fast-forwarding the initial 1.5 billion for each individual program within a mix. This also ensures that identical tasks are performed across different configurations. Note that when the core count is less than 30, part of programs will be launched after some cores finish their tasks assigned earlier. Also, considering that program feature such as memory intensity determines the computation efficiency on heterogeneous CMPs, we briefly classify the programs from SPEC CPU 2006 into two categories, namely computation-intensive and memory-intensive, based on their L2 miss ratios. Table 4 lists all selected benchmarks used in this study.

3. Device Heterogeneity

3.1 New Device and Architectural Implication

The slight improvement in transistor power density is fundamentally caused by the physical characteristics of MOSFET [23]. Due to this limitation, it is intuitive to recognize that breakthroughs in semiconductor technology are the antidote to dark silicon in essence. In this work, we consider two representative emerging devices, namely High-K dielectrical [1] and Nano-electro-mechanical switch (NEMS) [6][11], to exploit the device heterogeneity and combat dark silicon.

High-K dielectrical refers to a device that replaces the silicon dioxide in semiconductor manufacture. The letter K stands for dielectrical constant, indicating how much charge the material can hold. High-K is capable of significantly decreasing the leakage current (i.e., < 1% of SiO_2) and has already been adopted by leading processor manufacturers [1]. In general, as an important substitute of conventional devices in current industry, it deserves a careful evaluation.

The NEMS material, on the other hand, is a candidate for future processor development because it is built on physical switch and is not limited by the drawbacks of MOSFET. NEMS is able to reduce the leakage current by orders of magnitude, however, it demonstrates a significantly longer switch delay compared to conventional devices, implying large performance degradation on the resultant processor. Taking this into consideration, researchers propose a hybrid device that combines NEMS and CMOS together. Dadgour et al. [6] elaborate the features of NEMS-CMOS circuits in detail and demonstrate the potential of this hybrid device in future processor manufacturing. Therefore, we consider NEMS-CMOS as an alternative material in this work. We carefully calibrate the parameters based on recent documents [1][6][11] for High-K and NEMS-CMOS and list the important features in Table 5.

Although the purpose of this section is not to make comparison among emerging devices, a glance at their characteristics can enlighten us on architectural innovation for the next generation CMP.

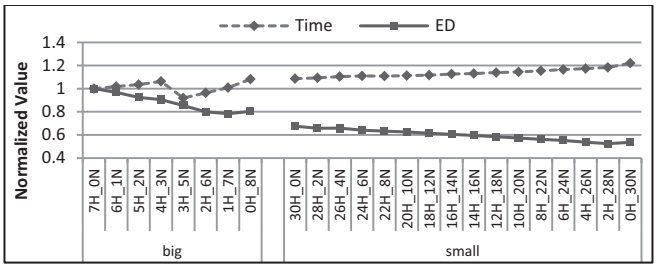


Figure 2. Average execution time and ED of multi-threaded applications running on mix-device CMPs.

Specific to High-K and NEMS-CMOS, the latter material switches at a lower rate than the former one but offering extra saving for both dynamic and leakage energy. Note that using other alternative materials such as Tunnel-FET (TFET) will introduce similar design trade-off. For instance, TFET cannot match the performance of CMOS under normal voltage, but it is beneficial for power saving [19]. Therefore, our conclusion made in this section can be generalized to scenarios where devices other than High-K and NEMS-CMOS are used for processor manufacturing. Nevertheless, this implies that integrating High-K cores and NEMS-CMOS cores on the same chip would deliver a processor that works more efficiently than a CMP manufactured with an exclusive device. Keeping this in mind, we evaluate a set of design configurations, with which a portion of integrated cores are built with High-K while the remaining ones with NEMS-CMOS. We compare such mix-device configurations with CMPs built with a single device alone (i.e., all High-K cores or NEMS-CMOS cores) and aim at identifying the better design choice.

3.2 Result Analysis

3.2.1 Average performance and ED

We consider two categories of CMPs to characterize the impact of device selection. The first group of chip-multiprocessors is composed of big out-of-order cores while the ratio of High-K cores over NEMS-CMOS cores is varying. Based on the power and area constraints depicted in section 2.2, the total number of big cores that can be accommodated on die is either 7 or 8. The reason of the varying core count is as follows. When all cores are manufactured with High-K, the power constraint restricts the maximal number of cores to be 7 although there is enough space for an extra core; as more NEMS-CMOS cores which consume relatively lower power are integrated to replace High-K cores, the area constraint becomes the determinative factor and confines the core count to be 8. On the other aspect, when all cores are small in-order ones, the core count is always limited by the area constraint and should not exceed 30.

We run multi-threaded applications with these configurations for evaluation. Figure 2 plots the average performance and energy-efficiency of these applications. All results are normalized to that corresponding to the *7H_0N* configuration in the "big" category, where the chip contains 7 out-of-order cores made of High-K. Note that in later sections of this paper, we also show results in this normalized fashion. The notation *x*H_*y*N means a total of *x* High-K cores and *y* NEMS-CMOS cores are installed. Also recall that the performance is measured in execution time, thus smaller values indicate better performance. As can be observed, in the "big" category, the execution time gradually increases at first and demonstrates a significant reduction from 4H_3N to 3H_5N, after which the curve rises again. The reason of the performance degradation (e.g., from 7H_0N to 4H_3N, and the segment between 3H_5N and 0H_8N) is that NEMS-CMOS cores execute at a lower rate than the High-K counterparts; therefore, increasing the number of NEMS-CMOS cores tends to prolong the overall execution time. The performance improvement at 3H_5N comes from the extra core in this configuration, with which the applications are executed

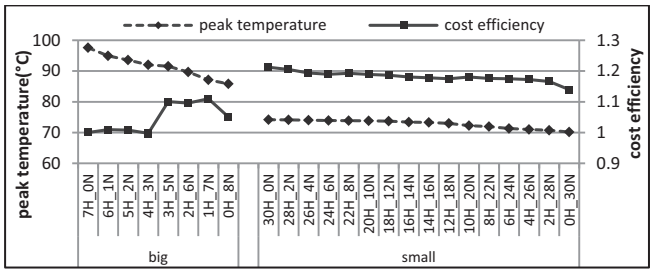


Figure 3. Average peak temperature and cost efficiency of multi-threaded benchmarks running on mix-device CMPs.

with one more thread. Note that in the extreme case where all cores are made of NEMS-CMOS (0H_8N), the processor takes even longer time to finish the execution compared to the 7-core configurations although it is equipped with an extra core. This is because that the slow execution on the master thread becomes the performance bottleneck and elongates the execution duration. As for the "small" category, the execution time gradually increases as more NEMS-CMOS cores are included since the core count is fixed to 30 irrespective of the manufacturing device.

The energy-efficiency demonstrates a different variation from the performance change. In general, the energy-delay product is decreasing as more NEMS-CMOS cores are equipped. This is because that the energy saving from NEMS-CMOS cores outweighs the corresponding performance degradation while running these parallel applications, thus using more such cores is beneficial to improving the energy-efficiency. The only exception is observed at the switch from 1H_7N to 0H_8N in the "big" category (or 2H_28N to 0H_30N in "small"), where the energy-delay demonstrates a slight increase. This is due to the fact that the performance degradation contributes more to the variation of ED for programs with long serial phase. With the 0H_8N configuration, the sequential stages are executed on the NEMS-CMOS cores, thus resulting in significant performance loss and higher ED.

In summary, for a CMP which only consists of big cores, including relatively more NEMS-CMOS cores and a few faster High-K cores is the preferable design paradigm than building a chip with processor cores made of a single device. Specifically, the 3H_5N configuration is able to shorten the execution time by an average of 8.9% while reducing the ED by 14.2% compared to the 7H_0N design. The ED-optimal configuration (i.e., 1H_7N) can save the ED by up to 20.8% with ignorable performance loss in comparison with 7H_0N. For the small-core-oriented architecture, the highest energy-efficiency is delivered by the configuration 2H_28N, meaning the optimal balance between performance and energy consumption is also achieved on a CMP with a large amount of NEMS-CMOS cores and a few High-K cores.

3.2.2 Thermal feature and cost-efficiency

Peak temperature and cost-efficiency are another two important metrics to evaluate a design configuration. We demonstrate the results of these two features for the proposed configurations in Figure 3. As shown in the figure, the temperature drops significantly as we employ more NEMS-CMOS big cores. The reason is that the power density on a NEMS-CMOS core is remarkably smaller than that of a High-K counterpart, thus a NEMS-CMOS core is relatively "cooler" compared to a High-K one. As more cool components are integrated on die, thermal coupling tends to be alleviated and the peak steady temperature is gradually decreased. Therefore, the coolest chip is the one where all cores are manufactured with NEMS-CMOS. On the other aspect, lower temperature results in lower cooling cost. This means that we are essentially trading off "performance" for "low cost" when we replace a NEMS-CMOS core for a High-K core. In this scenario, the cost-efficiency

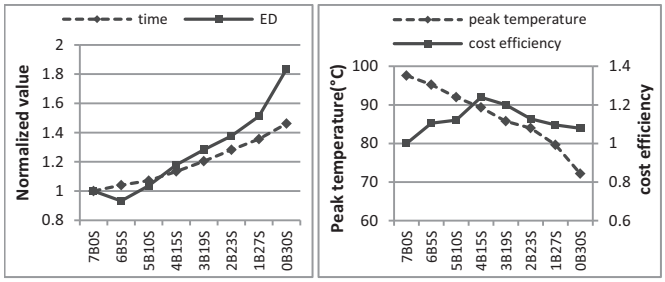


Figure 4. Execution information for computation-intensive workloads on high-K heterogeneous CMPs (a) normalized performance and ED (b) temperature and cost-efficiency.

reaches the peak value at 1H_7N where the performance and cost can be optimally balanced. Note that the increment of cost-efficiency from 4H_3N to 3H_5N is resulted from the performance boost. The curve corresponds to the "small" category is more smooth. The reason is that the in-order cores consume much smaller power than big cores and thus generate less heat. This results in relatively mild temperature variation across configurations. In this situation, the cost-efficiency does not largely vary when we change the manufacturing devices. Nevertheless, generally speaking, it is still reasonable to conclude that hybrid-device CMPs outperform chips built with a single device alone. Furthermore, to achieve the optimal balance among performance, energy consumption and total cost, a CMP should be equipped with more power-saving cores (NEMS-CMOS) and a small amount of faster yet power-consuming (High-K) cores.

4. Two-fold Heterogeneity

4.1 More Observations on Architectural Heterogeneity

Existing works have shown that executing a program on processors with different architecture may result in quite distinctive energy efficiency [14]. For example, a program with fairly low instruction-level parallelism might be more suitable to run on a simple in-order core instead of a big complex one for higher energy efficiency. This observation drives the development of architectural heterogeneous CMPs where integrated cores demonstrate different performance, area, and power features. In this subsection, we use the execution of computation-intensive workloads on a series of High-K heterogeneous CMPs as an example to illustrate that architectural heterogeneity also results in better cost-efficiency. Note that we run SPEC program mixes for the evaluation of architectural heterogeneity.

We first briefly analyze the performance and ED variations which are shown in Figure 4(a) to corroborate conclusions made in prior works. The notation *x*B*y*S indicates that *x* big cores and *y* small cores are integrated on the chip. Recall that the core counts are determined by both area and power constraint as described in section 2.2. From the figure we observe that the total execution time of the computation-intensive workloads keeps increasing as the number of big cores is reduced. This is due to the fact that the execution speed of such programs on big cores is remarkably faster than that on small in-order cores. For example, the relative performance (i.e., time on small core/time on big core) of *dealII* is around 6.02. This means that running a set of programs on a big core sequentially takes even shorter time than running them on a few small cores in parallel. However, the energy-delay product reaches the minimal value when 6 big and 5 small cores are installed on the chip. This is because the energy saving on small cores contributes more to the improvement in energy-efficiency at this point. Nevertheless, this scaling trend proves that architectural heterogeneity is effective in increasing the energy-efficiency.

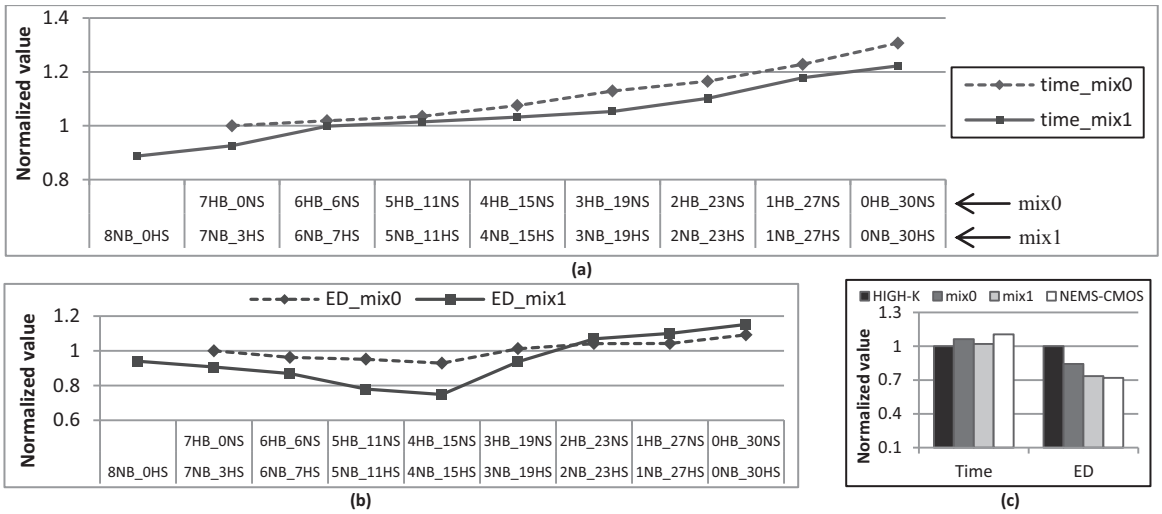


Figure 5. Execution information for computation-intensive workloads running on mix-device heterogeneous CMPs: (a) performance (b) energy-delay product (c) comparison among material-dependent optimal configurations.

Figure 4(b) plots the variations of temperature and cost-efficiency for computation-intensive workloads running on High-K heterogeneous CMPs. As can be observed, the temperature drastically drops as we gradually remove big cores to accommodate more small cores. This is straightforward to understand since small cores are much simpler and consume less power than big cores. The common hotspots in an out-of-order processor such as the instruction issue queue have been eliminated from small cores, thus replacing big cores with small cores is effective to decrease the chip temperature and save the cooling cost. However, computation-intensive workloads favor big cores for better performance, implying that the performance will be degraded as we reduce the number of big cores. In this situation, the interplay between performance and temperature results in a non-monotonic variation of the cost efficiency that it first increases to the peak value at 4B15S and then drops as the big core count is further decreased. In specific, the 4B15S configuration is able to cool the chip by 7.5°C while improving the cost-efficiency by 23.9% compared to the 7B0S organization. In one word, architectural heterogeneity delivers better cost-efficiency compared to homogeneous designs.

4.2 Performance and ED

After justifying the advantage of architectural heterogeneous CMPs with respect to energy-efficiency and cost-efficiency, it is natural for us to introduce the second design dimension, two-fold heterogeneity, with which both device-heterogeneity and architectural asymmetry are jointly adopted. More specifically, we consider a set of configurations where both the material and complexities are different among integrated cores. We assess two kinds of organizations: big High-K cores along with small NEMS-CMOS cores and the opposite.

Figure 5(a) plots the performance scaling of computation-intensive programs with these two design patterns. Note that all results are normalized to that in the 7HB_0NS case. The upper labels on the horizontal axis correspond to the first architecture where big cores are made of High-K and small cores are manufactured with NEMS-CMOS (mix0 or *x*HB_*y*NS); accordingly, the lower labels correspond to the opposite architecture which includes big NEMS-CMOS and small High-K processors (mix1 or *x*NB_*y*HS). As can be observed, configurations with the second pattern, namely *x*NB_*y*HS, always outperform the counterparts from the first category. This can be explained in two aspects. First, since NEMS-CMOS cores are relatively power-saving, the second design pattern accommodates more processors when the core count is power-limited. Due to this reason, the total number of cores is larger in the *x*NB_*y*HS designs, thus these configurations take shorter time to finish executing the program combination. This corresponds to the scenarios where the number of big cores is no smaller than 6. Second, as the constraint factor shifts to chip area, the core counts in both design patterns become identical (from 5B_11S). In this situation, the global execution time basically depends on the performance of small cores because of their larger amounts. For instance, in the 2B_23S configuration, how fast the programs run on small cores determines the overall performance in essence, because the number of small cores is remarkably larger than that of big cores. Since those in-order processors are made of High-K, the chips designed with the second pattern still offer better performance.

Figure 5(b) demonstrates the variation of the energy-efficiency for the same program set running with considered configurations. Note that the interplay between the performance/energy of different cores makes the variation of ED non-monotonically. For both blending patterns, we note that the energy-delay product gradually decreases at first until the minimal value is reached at 4B_15S, after which the efficiency is getting worse. More specifically, the *x*NB_*y*HS delivers better energy-efficiency than the *x*HB_*y*NS when the configuration is varied from 8 big cores to 3 big cores. This is due to the shorter execution time and less energy consumption on big NEMS-CMOS cores. As small cores begin dominating the chip in 2B_23S and beyond, their relatively large energy consumptions mitigate the performance benefit and make the ED rise again.

To more clearly illustrate the benefit of such two-fold heterogeneity, we identify the most energy-efficient configurations from four different design patterns, namely High-K for all cores, *x*HB_*y*NS (mix0), *x*NB_*y*HS (mix1) and NEMS-CMOS for all cores, and make comparison among these material-dependent optima. For computation-intensive workloads, we choose 6B_5S according to Figure 4(a) and 6B_7S for High-K and NEMS-CMOS, respectively. Note that the evaluation results of architectural heterogeneity with NEMS-CMOS are not included in the paper due to space limitation. Nevertheless, 6B_5S and 6B_7S deliver the optimal energy-efficiency for High-K processors and NEMS-CMOS ones. We then select 4B_15S for HB_NS and NB_HS based on Figure 5(b). We normalize the execution time and ED to those corresponding to the optimal High-K processor and demonstrate the result in Figure 5(c). As can be observed, the CMP with 4 NEMS-CMOS big cores and 15 High-K small cores (4NB_15HS) is the global optimal configuration. It improves the energy-efficiency by 27% with only 4.3% performance degradation compared to the optimal High-K CMP. We conduct similar comparison for memory-intensive workloads and graph the result in the appendix.

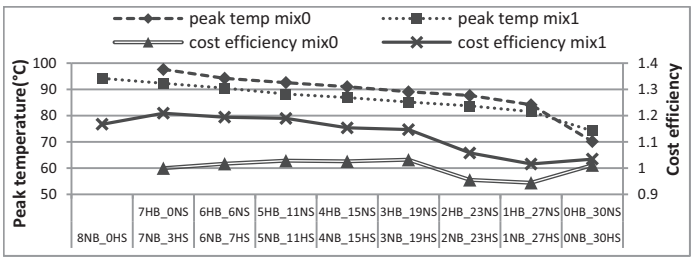


Figure 6. Peak temperature and cost-efficiency of computation-intensive workloads running on mix-device heterogeneous CMPs.

4.3 Thermal Effects and Cost-efficiency

Figure 6 plots the peak temperature and cost-efficiency of these two-fold heterogeneous CMPs while running computation-intensive workloads. As we have observed previously, NEMS-CMOS cores result in lower temperature than High-K cores and small cores are much cooler than big ones. Consequently, the second design pattern (i.e., *x*NB_*y*HS) tends to be cooler than its alternative (*x*HB_*y*NS), because the hotspot on die which is usually located in the out-of-order processor has lower temperature. Recall that the *x*NB_*y*HS also delivers better performance. Therefore, its cost-efficiency is significantly higher than that offered by *x*HB_*y*NS configurations. As can be seen, for computation-intensive workloads, the cost-efficiency reaches the peak value at 7NB_3HS configuration, which improves the efficiency by 20.9% compared to the 7HB_0NS case. For memory-intensive workloads, (graphs are in the appendix), the optimal configuration outperforms the baseline case by up to 66.7%. In conclusion, our observations made in this section demonstrate that the mix1 design paradigm (*x*NB_*y*HS, or big NEMS-CMOS cores along with small High-K cores) stands as the optimal among all evaluated configurations, since it can more efficiently balance the execution performance, energy consumption and total cost.

5. Related Work

Dark silicon emerges as an increasingly important issue that menaces the scaling of Moore's Law in the deep submicron era and beyond. Due to this reason, researchers recently start to investigate this problem and propose several solutions to alleviate the conundrum. A group from UCSD has made significant progress on using dark silicon for processor improvement. They develop conservation cores [24] and Quasi-specific cores [25] for increasing the computation energy-efficiency in different scenarios. In [9], Gupta et al. demonstrate the potential of heterogeneous CMP for energy-efficiency improvement. Systems built with near-threshold voltage processors (NTV) [7][26] are also effective approaches.

While most of these studies focus on a single solution individually, few works make attempt to address the dark silicon problem from a broader perspective. Esmaeilzadeh et al. [8] use an analytical model to predict the processor scaling for next few generations. They demonstrate that dark silicon will be heavily exacerbated as manufacture technology keeps shrinking. Taylor [23] reviews the current status of dark silicon and briefly describes four solutions from the high level. Hardavellas et al. [10] pay specific attention to the server processors and perform an exploration of throughput-oriented processors.

As for the hybrid device study, Saripalli et al. [19][20] discuss the feasibility of technology-heterogeneous cores and demonstrate the design of mix-device memory. Wu et al. [27] presents the advantage of hybrid-device cache. Kultursay [13] and Swaminathan [21] 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 to combat dark silicon in the early stage of processor manufacturing. We propose to utilize device heterogeneity and architectural heterogeneity simultaneously to optimally utilize the chip resource and well balance the performance, energy consumption and total cost.

6. Conclusion

As dark silicon has begun to hazard the scaling of Moore's Law and prohibits us benefiting from the increasing number of transistors, new design technologies are in high demand to address this problem. This is especially important in the early stage of processor manufacturing where issues such as architectural organization and device selections need to be carefully considered. For this purpose, our work evaluates a series of design configurations by exploiting the device heterogeneity and architectural asymmetry in the processor manufacturing. Our evaluation results demonstrate that building heterogeneous chip multiprocessors with different materials is more preferable than conventional designs since it can efficiently utilize the chip level resource and deliver the optimal balance among performance, energy consumption and cost.

References

[1] Intel Corporation. High-K and Metal Gate Transistor Research. http://www.intel.com/pressroom/kits/advancedtech/doodle/ref_HiK-MG/high-k.htm

[2] Intel Corporation. Ivy Bridge Products. http://ark.intel.com/products/codename/29902/Ivy-Bridge

[3] International Technology Roadmap for Semiconductors. http://www.itrs.net/

[4] Hotspot 5.0 Temperature Modeling Tool. http://lava.cs.virginia.edu/HotSpot/.

[5] Global Semiconductor Alliance. http://www.gsaglobal.org

[6] H. F. Dadgour and K. Banerjee. Design and analysis of hybrid NEMS-CMOS circuits for ultra low-power applications. In DAC'07.

[7] R. G. Dreslinski, M. Wieckowski, D. Blaauw,D. Sylvester, and T.Mudge. Near-threshold computing: reclaiming Moore's law through energy efficient circuit. Proceedings of the IEEE, special issue on ultra-low power circuit technology, Feb. 2010.

[8] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, D. Burger. Dark silicon and the end of multicore scaling. In ISCA'11.

[9] V. Gupta et al. Using heterogeneous cores to provide a high dynamic power range on over-provisioned processors. In Dark Silicon Workshop in conjunction with ISCA, Jun. 2012.

[10] N. Hardavellas, M. Ferdman, B. Falsafi, A. Ailamaki. Toward dark silicon in servers. In IEEE Computer Society, 2011.

[11] R. Jammy. Materials, process and integration options for emerging technologies. SEMATECH/ISMI symposium, 2009.

[12] P. L-Kamran et al. Scale-out processors. In ISCA'12.

[13] E. Kultursay et al. Performance enhancement under power constraints using heterogeneous CMOS-TFET multicores. In CODES+ISSS'12.

[14] R. Kumar, K. I. Farkas, N. P. Jouppi, P. Ranganathan, D.M. Tullsen. Single-ISA Heterogeneous Multi-Core Architectures: The Potential for Processor Power Reduction. In MICRO'03.

[15] S. Li et al. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In MICRO'09.

[16] J. M. Rabaey, A. Chandrakasan and B. Nikolic. Digital Integrated Circuits, 2nd edition.

[17] A. Raghavan et al. Computational Sprinting. In HPCA'12.

[18] J. Renau et al. SESC Simulator.

[19] V. Saripalli et al. Exploiting heterogeneity for energy efficiency in chip multiprocessors. In IEEE Transactions on Emerging and Selected topics in Circuits and Systems, Jun. 2011.

[20] V. Saripalli, A.K.Mishra, S. Datta and V.Narayanan. An energy-efficient heterogeneous CMP based on hybrid TFET-CMOS cores, in DAC'11.

[21] K. Swaminathan et al. Improving energy efficiency of multi-threaded applications using heterogeneous CMOS-TFET multicores. In ISLPED'11.

[22] S. Swanson et al. Area-performance trade-offs in tiled dataflow architectures. In ISCA'06.

[23] M.B.Taylor. Is dark silicon useful? In DAC'12.

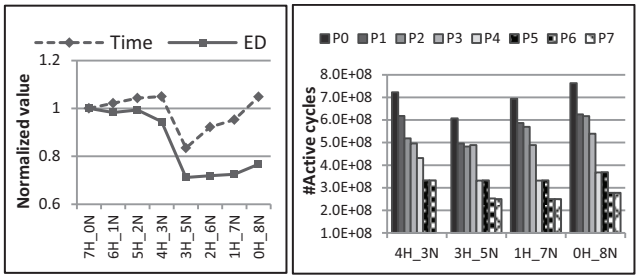
[24] G. Venkatesh, J Sampson, N. Goulding, S. Garcia. Conservation cores: reducing the energy of mature computations. In ASPLOS'10.

[25] G. Venkatesh et al. QSCores: Trading dark silicon for scalable energy efficiency with quasi-specific cores. In MICRO'11.

[26] L. Wang, K. Skadron, and B. H. Calhoun. Dark vs. Dim silicon and near-threshold computing. In Dark Silicon Workshop in conjunction with ISCA, Jun. 2012.

[27] X. Wu et al. Hybrid cache architecture with disparate memory technologies. In ISCA'09.

[28] J. Zhao, X. Dong and Y. Xie. Cost-aware three-dimensional (3D) many-core multiprocessor design. In DAC'10.



(a) (b)

Figure 7. Execution information of MPGEnc: (a) time and ED (b) per-core active cycles while running with selected configurations.

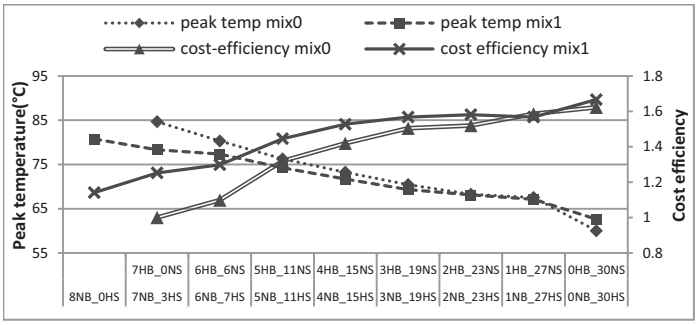
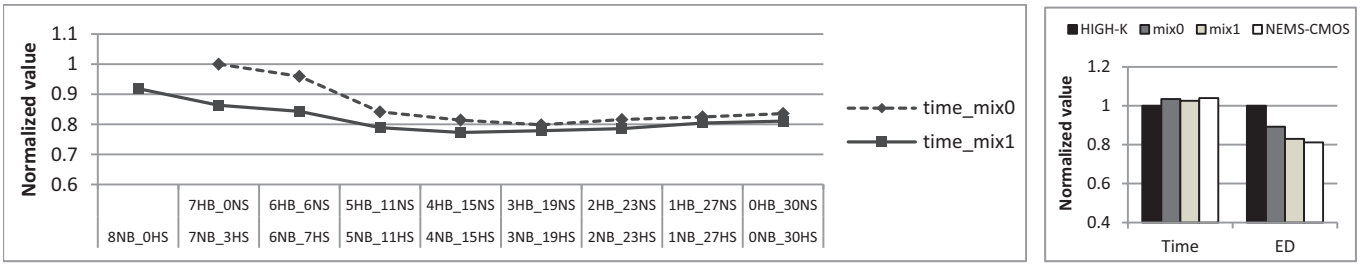


Figure 9. Peak temperature and cost-efficiency of memory-intensive workloads running on mix-device heterogeneous CMPs.



(a) (b)

Figure 8. Execution information for memory-intensive workloads running on mix-device heterogeneous CMPs: (a) performance (b) comparison among material-dependent optimal configurations.

APPENDIX

Case Study for Device Heterogeneity

To further understand the performance scaling trend shown in Figure 2, we choose a representative application (*MPGEnc*) from the program set for analysis and demonstrate the results in Figure 7. Note that we only show the results on CMPs with big cores. The MPGEnc benchmark implements a parallel version of MPEG-2 encoder. In this application, the threads are respectively forked and joined at the beginning and end of the encoding for each frame. Each thread is responsible for encoding a set of macroblocks of a frame while thread 0 always operates on its dedicated buffer. The task assigned to each thread is not identical, thus the time spent by each thread also varies. Plot (a) demonstrates the performance and ED scaling while Plot (b) shows the active cycles of each core during the execution of this program with four configurations. The total execution time is determined by the main thread running on the first processor (P0), and the performance of the parallel stage can be generally estimated from the active cycles of P1. As can be observed, since the number of threads is increased from 7 to 8, the 3H_5N configuration takes much shorter time than 4H_3N to finish the encoding due to the acceleration in parallel stage, hence the remarkable performance improvement at 3H_5N. For the latter three configurations where the core counts are identical, the performance degradation is caused by the decreasing of faster cores (High-K). In specific, the 1H_7N organization includes only one High-K core (P0) while three such cores are equipped in 3H_5N; as a consequence, the parallel stage needs longer time to complete on the CMP configured as 1H_7N, thus lowering the overall performance. On the other hand, the performance degradation from 1H_7N to 0H_8N essentially stems from the slow execution of the sequential stage. This is especially critical for programs with long initialization and finalization.

More Results of Mix-device Heterogeneous CMP

We have shown that mix-device heterogeneous CMP is benefitial to improving the energy- and cost-efficiency for computation-intensive workloads. In this subsection, we will present the result of memory-intensive workloads in order to further justify the conclusion that the design paradigm mix1 is the globally optimal. Figure 8(a) demonstrates the performance comparison between mix0 and mix1 while Figure 8(b) illustrates the performance and energy-efficiency comparison among four material-dependent optimal configurations. Generally, we observe a similar trend that the mix1 design paradigm is more preferable than mix0 by delivering better performance. However, compared with the scaling behavior shown in Figure 5(a), Figure 8(a) demonstrates that memory-intensive workloads favor more small cores, hence more total number of cores, for shorter execution time. The reason is that running memory-bound programs on big cores will not significantly accelerate the execution as opposed to computation-intensive ones. Therefore, executing more programs concurrently can effectively reduce the time for completing all tasks compared to running them sequentially on few big cores. On the other hand, from Figure 8(b), we observe a trend similar to that shown in Figure 5(c). Specifically, the most energy-efficient configuration in the mix1 category outperforms the optimal High-K CMP by 17% in energy-efficiency with less than 4% performance loss. Figure 9 plots the thermal and cost-efficiency results for memory-intensive workloads running on mix-device heterogeneous CMPs. Not surprisingly, the mix1 design paradigm results in a cooler chip than mix0 in most cases, thus delivering up to 66.7% higher cost-efficiency compared to the baseline configuration. In one word, our conclusion that building big out-of-order cores with NEMS-CMOS and manufacturing small in-order cores with High-K is able to achieve the optimal balance among performance, energy consumption and total cost also holds for the memory-intensive applications.