New Trends in Dark Silicon

Jörg Henkel, Heba Khdr, Santiago Pagani, Muhammad Shafique

Chair for Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

E-mail: {henkel, heba.khdr, pagani, muhammad.shafique}@kit.edu

Invited

ABSTRACT

This paper presents new trends in dark silicon reflecting, among others, the deployment of FinFETs in recent technology nodes and the impact of voltage/frquency scaling, which lead to *new less-conservative predictions*. The focus is on dark silicon from a thermal perspective: we show that it is not simply the chip's total power budget, e.g., the Thermal Design Power (TDP), that leads to the dark silicon problem, but instead it is the power density and related thermal effects. We therefore propose to use Thermal Safe Power (TSP) as a more efficient power budget. It is also shown that sophisticated spatio-temporal mapping decisions result in improved thermal profiles with reduced peak temperatures. Moreover, we discuss the implications of Near-Threshold Computing (NTC) and employment of Boosting techniques in dark silicon systems.

1. INTRODUCTION

As Dennard's Scaling is no longer applicable due to the voltage scaling limitations, the on-chip power densities rapidly increase leading to the so-called *dark silicon problem*, i.e., a significant amount of on-chip resources cannot be operated at full performance level at the same time. During recent years, researchers have been exploring the dark silicon problem and its implications on the design of manycore systems. A comprehensive discussion of the state-of-the-art approaches for dark silicon can be found in [1, 2]. Recent studies also leveraged dark silicon to improve the thermal profiles and reliability of manycore systems [3, 4, 5]. Efficient design and management of manycore systems in the dark silicon era require a comprehensive analysis and accurate estimation of the amount of dark silicon.

The work in [6] predicts that dark silicon will be dominant in future technology nodes, and hence it will considerably limit technology scaling and integrating more cores on a chip. This work models dark silicon as a power budget constraint. However, dark silicon is a direct result of high power densities and the related thermal effects. Therefore, there is more than one perspective to be considered when defining and modeling the dark silicon, namely, power budget and temperature. Furthermore, the work in [6] does not study the implication of DVFS and multiple application instances on the estimation of dark silicon. Due to the above-mentioned limitations, the analytical studies of [6] result in over-estimation of dark silicon, for instance, this work predicted that the dark silicon in 22 nm would exceed 50% of the total chip area, which has not been observed in the recent processing systems. *Therefore, there is a need to revise the predictions of dark silicon trends considering advancements in the technology, temperature constraints, and realistic operating scenarios.* Towards this end, this
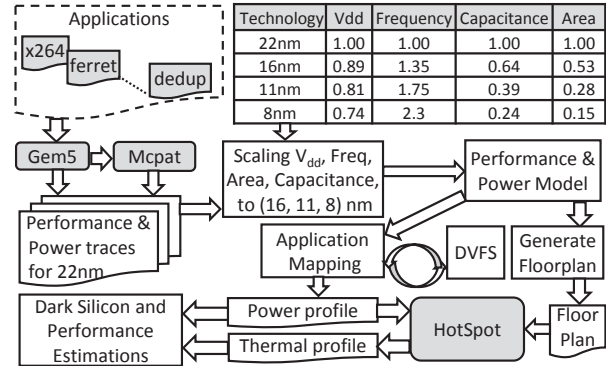
Figure 1: Our Simulation Tool Flow and the adopted scaling factors for future technology nodes according to [9] and [10]. The factors in the table are with respect to 22 nm.

paper proposes new trends in dark silicon, covering different perspectives and constraints.

The novel contributions of this paper are:

- We provide new dark silicon estimations for the future technology nodes under different constraints, specifically, power budget constraints and temperature constraints (Section 3).
- We analyze the influence of using DVFS on dark silicon estimations and the corresponding performance results (Section 3.3).
- We explain some recent techniques on dark silicon management (Section 4).
- For the future technology nodes, we perform a comprehensive comparative analysis of the boosting techniques against constant frequency schemes, both in conventional voltage supply ranges and in the NTC–Near-Threshold Computing (Section 6).

This paper is part of DAC's special session "Dark Silicon: No Way Out?". Other papers in this session are: "Approximate Computing and the Quest for Computing Efficiency" [7], and "Core vs Uncore: The Heart of Darkness" [8].

2. QUANTIFYING DARK SILICON: MODELS AND EXPERIMENTAL SETUP

2.1 Experimental Setup

Figure 1 illustrates the tool flow and scaling factors used in our experiments. For our hardware platform, we consider manycore systems composed of 100, 198, and 361 *out-of-order* Alpha 21264 cores. First, we conduct experimental evaluations for 22 nm using gem5 [11] and McPAT [12]. Then, we use ITRS scaling factors to scale our results from 22 nm to the following technology nodes, i.e., 16 nm, 11 nm and 8 nm. Figure 1 shows a table of the adopted scaling factors. According to our simulations, for 22 nm technology, each core has an area of 9.6mm^2. By applying an area scaling factor from [10], i.e., 53% between nodes (Figure 1), we obtain the following core areas: 5.1mm^2, 2.7mm^2, and 1.4mm^2 for 16 nm, 11 nm and 8 nm, respectively.

In order to calculate the power values for these technologies, we use the power model explained in Section 2.2. To obtain the temperature values on the cores, we use HotSpot [13] with the following configuration: chip thickness
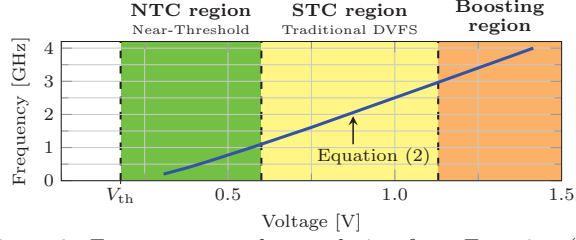
Figure 2: Frequency vs voltage relation from Equation (2) for 22 nm, with $k = 3.7$ and $V_{th} = 178$ mV.

of 0.15 mm, silicon thermal conductivity of $100 \frac{W}{m \cdot K}$, silicon specific heat of $1.75 \cdot 10^6 \frac{J}{m^3 \cdot K}$, a heat sink of 6×6 cm and 6.9 mm thick, heat sink convection capacitance of $140.4 \frac{J}{K}$, heat sink convection resistance of $0.1 \frac{K}{W}$, heat sink and heat spreader thermal conductivity of $400 \frac{W}{m \cdot K}$, heat sink and heat spreader specific heat of $3.55 \cdot 10^6 \frac{J}{m^3 \cdot K}$, a heat spreader of 3×3 cm and 1 mm thick, interface material thickness of 20 um, interface material thermal conductivity of $4 \frac{W}{m \cdot K}$, and interface material specific heat of $4 \cdot 10^6 \frac{J}{m^3 \cdot K}$.

2.2 Power Model

We consider a power consumption model for a core as formulated in Equation (1),

$$P = \alpha \cdot C_{eff}^{app} \cdot V_{dd}^2 \cdot f + V_{dd} \cdot I_{leak}(V_{dd}, T) + P_{ind} \qquad (1)$$

where α represents the activity factor or utilization of the core, C_{eff}^{app} represents the effective switching capacitance of a given application, V_{dd} is the supply voltage, f is the execution frequency, I_{leak} is the leakage current (which depends on the supply voltage and the core's temperature T), and P_{ind} represents the independent power consumption (attributed to keeping the core in execution mode). Furthermore, in Equation (1), $\alpha \cdot C_{eff}^{app} \cdot V_{dd}^2 \cdot f$ represents the dynamic power consumption, while $V_{dd} \cdot I_{leak}(V_{dd}, T)$ represents the leakage power consumption. The simulations results for 22 nm are modeled according to Equation (1) for every application. In this way, the values of C_{eff}^{app}, V_{dd}, I_{leak} and P_{ind} are scaled according to ITRS factors, and the power consumption values for smaller technologies can be accurately estimated.

The original voltage and frequency relation for 22 nm is obtained according to Equation (2),

$$f = k \cdot \frac{(V_{dd} - V_{th})^2}{V_{dd}} \qquad (2)$$

as used in [14], where V_{th} is the threshold voltage and k is a fitting factor modeled from the work in [15]. The physical meaning of Equation (2) is that for a given supply voltage, there is a maximum stable frequency at which a core can be executed. Consequently, when a core needs to be executed at a specific frequency, the minimum required voltage will be given by Equation (2), and running at higher voltages would be power/energy inefficient. Hence, we consider frequency and voltage pairs according to Equation (2). In this way, we arrive at a *cubic* relation between the frequency and the dynamic power consumption. Figure 2 shows the frequency and voltage relation used throughout the paper for 22 nm, and Figure 3 shows how the power model from Equation (1) fits the simulations results from McPAT for an H.264 video encoder from the Parsec benchmark suite [16].

2.3 Application Model

In this paper we use applications from the Parsec benchmark suite [16], running different number of parallel threads. However, along with technology scaling, more cores are integrated into the chip, e.g., as mentioned in Section 2.1, for our experiments we consider up to 361 cores for 8 nm technology. Therefore, mapping only a single Parsec application to these
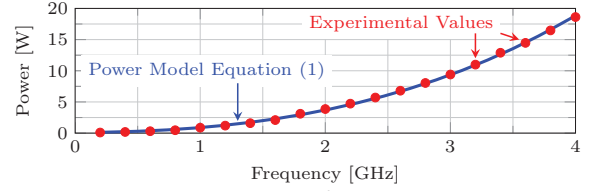

Figure 3: Experimental results for an H.264 video encoder from the Parsec benchmark suite [16] running a single thread and the derived power model from Equation (1).
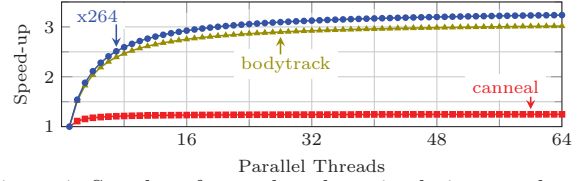

Figure 4: Speed-up factors based on simulations conducted on gem5 [11] and Amdahl's law for three applications from the Parsec benchmark suite [16] executing on a 2 GHz core.

many cores will significantly reduce the overall performance of the system, as an application's performance in general does not scale to such a large amount of cores due to what is known as the parallelism wall. Figure 4 shows speed-up factors with respect to the number of parallel threads for some applications of Parsec running at 2 GHz. Hence, if an application is mapped to all available cores, the activity factor of each core would be very small and the cores become under-utilized. Contrarily, if we map an application only to the amount of cores that satisfies its requirement, the rest of the cores on the chip would be inactive. However, considering such inactive cores as dark cores would mean overestimating the amount of dark silicon. As a result, in this paper we consider multiple application mappings that efficiently utilize the chip's resources, avoiding both under-utilization of the cores and overestimations of dark silicon. In our experiments we consider that every instance of an application can run $1, 2, \ldots, 8$ parallel dependent threads.

3. NEWEST DARK SILICON PREDICTIONS

In this section, we conduct several experiments that provide a comparison of the dark silicon amounts between two cases; modeling dark silicon as a power budget constraint and modeling it as a temperature constraint. Moreover, additional dark silicon estimations are inferred when DVFS is used, along with consideration of application properties; Thread Level Parallelism (TLP) and Instruction Level Parallelism (ILP).

3.1 Dark Silicon as a Power Budget Constraint

Several researchers, like [6], consider dark silicon as a power budget constraint, such that the total power consumption of the cores should not exceed the predefined power budget. TDP is the most commonly used power budget in the dark silicon era, and the state-of-the-art work in dark silicon estimation [6] adopts it.

In this section, we show how the TDP value affects dark silicon estimations. Therefore, we adopt two values of TDP. The first one is quantified when all the cores are executing without exceeding the critical temperature [17]. To guarantee safe operation and avoid damages, the temperature on the chip should remain below the critical temperature. Exceeding this critical temperature triggers Dynamic Thermal Management (DTM) on the chip. Thus, we denote this temperature value as T_{DTM} and we set it to 80 °C in our experiments according to [18]. The second TDP value is quantified to allow at least half of the cores to run when the most power consuming applications are applied. Under these two assumptions, we calculate TDP values for our application and hardware model, resulting in 220 W and 185 W.
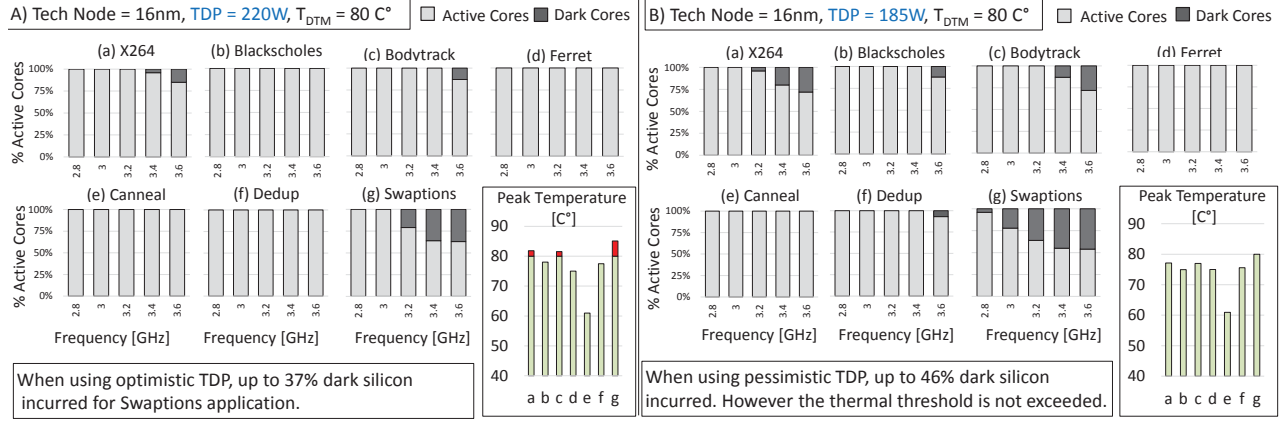
Figure 5: Estimations of dark silicon amounts under two different TDP values.

To evaluate dark silicon under these TDP values, we run several experiments for different Parsec applications on our 100-core chip at 16 nm technology. As described in Section 2.3, we execute multiple instances of each application in order to avoid the parallelism wall. Particularly, in these experiments we execute each application with 8 threads. Moreover, different v/f (voltage/frequency) levels are adopted to show their influence on dark silicon, assuming that the maximum nominal frequency at 16 nm is equal to 3.6 GHz. The resulting amounts of dark cores are depicted in Figure 5. Since different applications consume different amounts of power, different dark silicon amounts are seen in the figure. When executing power hungry applications, we observe that around 37% of the chip stays dark at the maximum v/f levels, when the optimistic TDP (220 W) is adopted. However, when the pessimistic TDP (185 W) is used, the amount of dark silicon reaches up to 46% of the chip. Additionally, we measure the resulting peak temperatures at both TDP values. We notice that the optimistic TDP leads to thermal violations on the chip, and that, in turn, will trigger DTM, which might power down additional cores, resulting in more dark silicon. In contrast, no thermal violations occur when the pessimistic TDP is used under the adopted application scenarios in our experiments.

Two observations can be inferred from this analysis:

OBSERVATION 1. **Considering temperature in dark silicon estimations:** *Modeling dark silicon as a TDP constraint may lead either to underestimation of dark silicon, like in Figure 5-A, or to overestimation of dark silicon, like in Figure 5-B. Therefore, to provide more accurate analysis of dark silicon, temperature needs to be considered in estimating dark silicon.*

OBSERVATION 2. **Importance of DVFS for dark silicon estimations:** *Dark silicon is reduced significantly by scaling down the v/f levels. However, in the state-of-the-art technique [6], the assumption was to run at the maximum feasible v/f that respects a predetermined core power budget. In this case, dark silicon might be overestimated. To avoid this problem, we should account for different v/f levels.*

The effects of these two factors, *temperature and DVFS*, on dark silicon are deeply analyzed in Sections 3.2 and 3.3.

3.2 Dark Silicon as a Temperature Constraint

As motivated in the previous section, temperature needs to be considered in estimating dark silicon. Therefore, we repeat the same experiments shown in Figure 5, but now we compare the effects of using TDP or temperature as the dark silicon constraint. Namely, we assume that the maximum temperature among all cores needs to stay below T_{DTM}. Thus, we conduct two scenarios. In the first one, we map multiple instances of each application until the total power
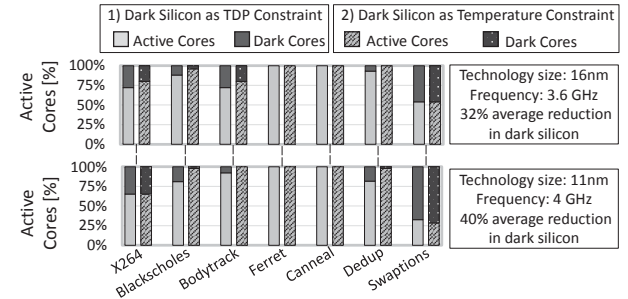

Figure 6: Comparison between dark silicon amounts under TDP and a temperature constraint for different technologies.

consumption reaches TDP, while in the second one, we stop mapping applications once the peak temperature reaches T_{DTM}. The results of these two scenarios for technology nodes 16 nm and 11 nm are presented in Figure 6.

These results show how the amount of dark silicon is reduced by modeling it as a temperature constraint, compared to the case of modeling it as a power budget constraint. However, this reduction varies from one application to another, depending on the thermal headroom that exists. Moreover, the average reduction in dark silicon is also different from one technology node to another.

Additionally, we conduct the same experiment for 8 nm, but the dark silicon reduction is smaller compared to the other technologies. The reason is that at 8 nm technology the power densities are very high, due to the exponential increase of dynamic power along with increasing v/f levels (we use 4.4 GHz for 8 nm). On the other hand, at 8 nm more v/f levels are available and we can use DVFS to reduce the amount of dark silicon, as described in Section 3.3.

3.3 DVFS in Dark Silicon

The amount of dark silicon amount decreases by scaling down the v/f levels, as shown in Section 3.1. However, this also reduces the system's performance. In other words, applying DVFS represents a trade-off between the amount of dark cores and the system's performance. This trade-off differs from one application to another according to their characteristics, i.e., their Thread Level Parallelism (TLP) and Instruction Level Parallelism (ILP). For high TLP applications, the performance is improved more by increasing the number of threads (active cores) rather than increasing the v/f levels. In contrast, the performance of high ILP applications improves significantly by increasing the v/f levels. To evaluate the impact of DVFS on both the performance and dark silicon amounts for different application's characteristics, we run experiments for two scenarios. (1) We use different v/f levels and different number of parallel threads for different applications considering their
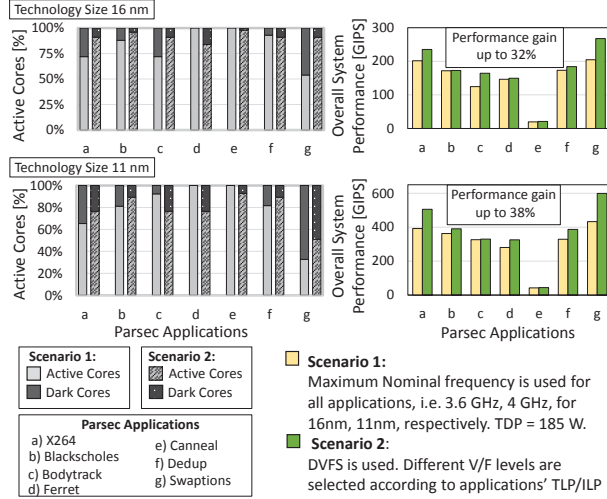
Figure 7: Resulting overall system performance and dark silicon amounts, with and without DVFS, for 16 nm and 11 nm. Using DVFS according to the applications' characteristics achieves a performance gain up to 38% at 11 nm.

characteristics. (2) We use the nominal frequency (3.6 GHz and 4 GHz for 16 nm and 11 nm, respectively) and 8 threads for each application. Both scenarios consider TDP (185 W) as the dark silicon constraint.

Figure 7 illustrates how using DVFS considering the application's characteristics decreases the amount of dark cores in some applications and increases it for others, but always improves the overall system performance. The average gain of the overall performance is up to 40% for 11 nm. An additional experiment is conducted for 8 nm technology, where the resulting average performance from using Scenario 2 was 1.5x the performance from Scenario 1.

4. DARK SILICON MANAGEMENT

As illustrated in the previous sections, temperature plays a major role in the dark silicon problem and considering it reduces the amount of dark silicon. However, application mappings can affect the peak temperature on the chip. For instance, two mappings of the same application under the same settings, i.e., v/f levels and number of threads, may result in different peak temperatures for each, as shown in the thermal analysis performed in [5, 19]. To benefit from mapping effects on temperature, the *DaSim* technique [5], that considers dark silicon as a temperature constraint, proposed what is called *dark silicon patterning*, which determines the positions of the threads on the chip such that the peak temperature is reduced. Reducing the peak temperature gives the ability to turn on additional cores. An example of this work is shown in Figure 8, which illustrates how dark silicon is reduced using a good dark silicon pattern, and more applications could be mapped to the chip compared to a scenario that contiguously maps the applications to the chip without considering the concept of *dark silicon patterning*.

Besides choosing a good pattern, a more recent work [19], namely *DsRem*, jointly determines the number of active cores for each application and their v/f levels, such that the overall performance is maximized. We compare *DsRem* [19] with a TDP-based mapping policy, namely *TDPmap*, which maps the applications using the same number of threads (8 threads for each application) and assigns the maximum v/f level to their cores. Once TDP is reached, no more applications can be mapped. On the other hand, *DsRem* first computes the optimal settings of applications under TDP, then it heuristically modifies them, either to avoid potential thermal violations or to exploit any available thermal headroom. Figure 9 depicts the results of this comparison in terms of dark silicon amounts and overall system performance.
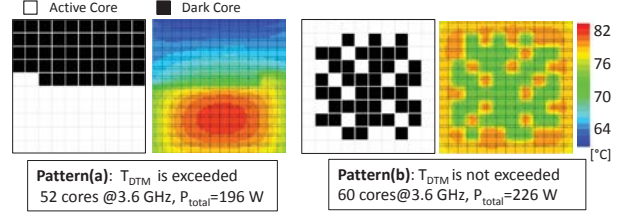

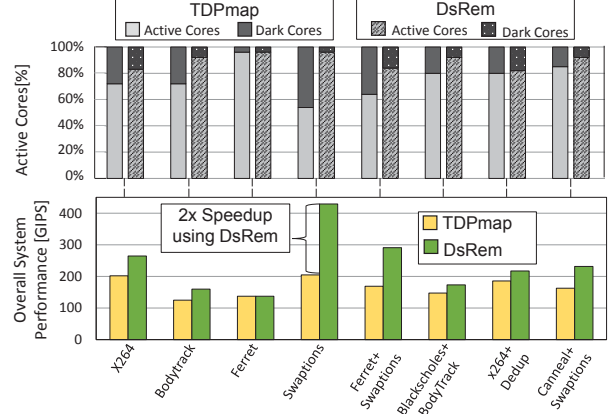Figure 8: Different application mappings and their resulting thermal profile.

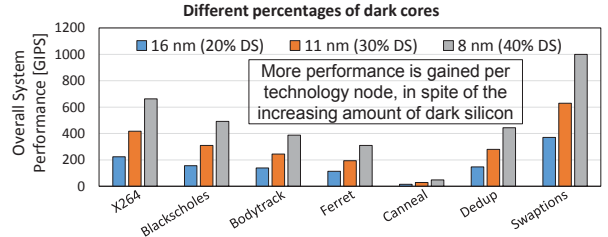Figure 9: Evaluation of *DsRem* scheme on 16 nm technology.

Figure 10: Resulting overall system performance using the TSP technique for 16 nm, 11 nm, and 8 nm, under different dark silicon percentages.

5. TSP–THERMAL SAFE POWER

As shown in the previous sections, both temperature and DVFS need to be considered in the dark silicon era. A major step in this direction is through efficient power budgeting techniques, such as the new power budget concept called Thermal Safe Power (TSP) [20], which results in a higher total system performance compared to using TDP as a constraint. TSP is an abstraction that provides safe power constraints as a function of the number of active cores, i.e., it guarantees that the maximum temperature among all cores remains below the temperature threshold when the power consumption of all active cores is below the TSP values for the given number of cores. As the number of active cores grows, the TSP values decrease, which in turn means executing cores at lower v/f levels. That is, we again observe the trade-off between v/f level and the number of dark cores.

Our goal in this section is to estimate the system performance for future technology nodes by using TSP. Namely, for a given number of active cores (or dark cores), we compute TSP accordingly and find the v/f levels that satisfy TSP for each application scenario. Figure 10 presents results of an experiment that evaluates the total system performance under different dark silicon percentages, specifically, 20%, 30%, and 40%, for 16 nm, 11 nm, and 8 nm, respectively. As shown in the figure, the total performance keeps increasing with future technologies. This increment from 11 nm to 8 nm is on average 60%. Moreover, having more dark cores does not always imply that we will have a lower performance, and this depends on specific application characteristics.
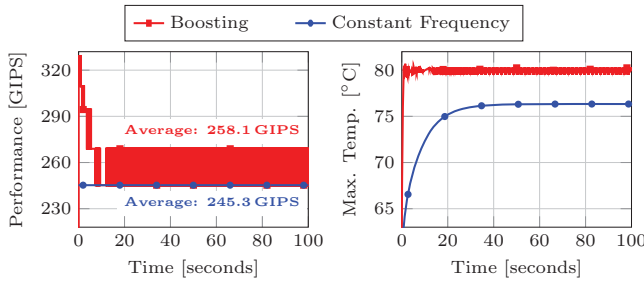
Figure 11: Transient simulation results for 12 instances of an H.264 video encoder, running 8 parallel threads each, in 16 nm. The total performance is measured in Giga-Instruction Per Second (GIPS). We show the maximum temperature among all cores.
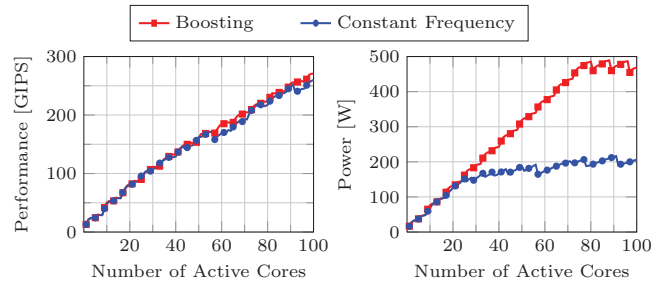

Figure 12: Resulting total performance and total power consumption for an H.264 video encoder in 16 nm with respect to the number of active cores. Each instance of the application runs up to 8 threads, and we consider a new application instance every 8 active cores.

6. STC/NTC VS BOOSTING

In this section we compare and discuss two different approaches and their implications in dark silicon. Namely, we evaluate and compare Boosting techniques vs executing at a constant voltage/frequency, both under Super-Threshold Voltage Computing (STC) and Near-Threshold Computing (NTC). Executing in the STC or NTC region will depend on the selected constant voltage/frequency that results in the highest performance while satisfying the dark silicon constraints. Specifically, we consider an electrical power constraint of 500 W, and a critical temperature of 80°C.

Boosting techniques have been widely adopted in commercial manycore systems, e.g., Intel's Turbo Boost [18, 21, 22, 23] and AMD's Turbo CORE [24]. They allow the system to execute cores at high voltage/frequency levels during short time intervals, normally exceeding standard operating power budgets like TDP. Because doing this increases the power consumption, thus increasing the chip's temperature, once the temperature reaches a predefined threshold, the system must cool-down at nominal operation or use some closed-loop control to oscillate around the threshold (prolonging the boosting time). For our experiments, we use a closed-loop control as used in Intel's Turbo Boost [22, 23], with a control period of 1 ms. That is, every 1 ms the system verifies that the temperature on all cores is below or above the predefined threshold of 80°C, and the frequency on all cores is increased or decreased one step (200 MHz) accordingly.

STC represents the conventional voltage supply region, where V_{dd} usually takes values above 0.6 V. Contrarily, NTC [14, 25] focuses on reducing the supply voltage V_{dd} to values near the threshold voltage V_{th}, in order to reduce the power and energy consumption. This power and energy reduction comes at the cost of decreasing the execution frequency. The trade-off here is that, for example, running an application under a single thread scenario at frequency $2f$ will consume about $4x$ the power of running the same application under two parallel threads at frequency f. Nevertheless, since applications are never perfectly parallelized, the resulting total performance for the latter case will generally be less than the total performance of the former case. Moreover, such an effect is more evident when we consider more threads, as seen in the example in Figure 4.

Figure 11 and Figure 12 show the details of our experiments for an H.264 video encoder from Parsec for 16 nm. Figure 11 shows the transient simulations when running 12 application instances with 8 parallel threads for each instance. Here we can see that boosting operates oscillating around the critical temperature, while the constant frequency approach remains a few degrees below the critical temperature due to the available voltage/frequency steps, that is, running at a the next available voltage/frequency would violate the critical temperature. Figure 12 shows the resulting total performance and total power consumption for different number of active cores. Furthermore, Figure 13 presents results for several Parsec applications for 11 nm, showing only a selected number of representative application scenarios. We also conducted equivalent experiments for
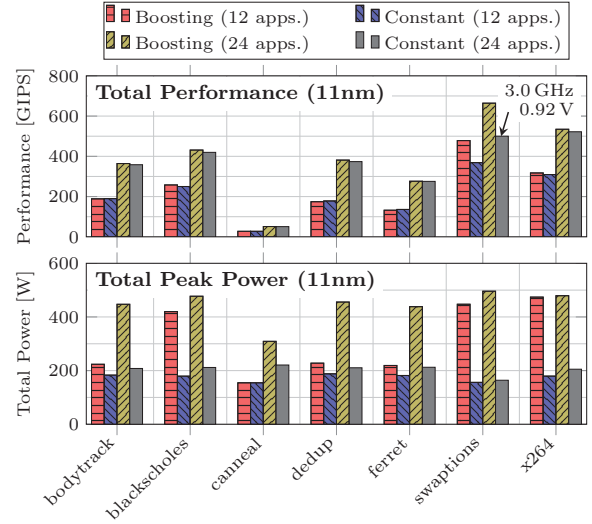

Figure 13: Total performance and total power consumption for several Parsec applications running 8 parallel threads and different numbers of application instances, in 11 nm. Among all tested cases, the minimum utilized voltage and frequency for satisfying the thermal constraints was 0.92 V and 3.0 GHz, respectively, which is still in the STC region.

16 nm and 8 nm (omitted due to space constraints), arriving at similar results to those presented in Figure 13.

For all the experiments in 16 nm, 11 nm, and 8 nm technologies (Figures 11–13), both for boosting and the constant frequency approach, the minimum utilized voltage and frequency for satisfying the dark silicon constraints was 0.84 V and 3.0 GHz, respectively, at 8 nm, which is still in the STC region. Thus, we conduct an additional experiment for the constant frequency approach that shows the total energy consumption when executing in the STC and NTC region for a resulting similar performance. The results are shown in Figure 14. For ISO performance, the figure shows that using NTC by running at low voltages/frequencies and several threads can be very energy efficient when the application's performance scales with the number of threads. From the evaluated cases, *canneal* does not scale well with more threads, thus running at NTC consumes more energy.

Finally, two very important observations can be made from the experiments in this section, summarized as follows.

OBSERVATION 3. **Boosting vs Constant Frequencies:** *Although using boosting techniques results in a higher average performance than using constant frequencies for all tested cases, the performance gain from using boosting is very small and arguably unjustified when considering the big increments to the total peak power consumption. Therefore, in order to alleviate the dark silicon problems by running the system in a more power and energy efficient fashion, executing at constant frequencies is a better approach.*
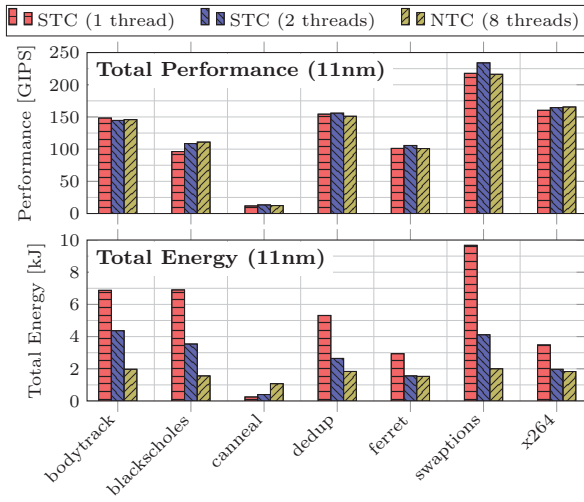
Figure 14: Total performance and total energy consumption in 11 nm for several Parsec applications running on the STC and NTC region. We execute 24 application instances in all cases. For NTC, each application instance runs 8 threads at 1 GHz and 0.46 V. For STC, each application instance runs 1 and 2 threads, and the frequencies are chosen in the STC region to match the performance achieved with NTC.

OBSERVATION 4. **STC vs NTC:** *When the goal is to maximize performance under dark silicon constraints, cores will be generally executed at constant frequencies in the STC region. Thus, NTC is not a necessary technique for dealing with dark silicon under this scenario in current and future scaling technologies. NTC is therefore better suited to handle the dual problem of minimizing power or energy under performance constraints, as shown in Figure 14.*

7. CONCLUSIONS

In this paper, new dark silicon trends were explored through extensive experiments and analysis that covered different perspectives like temperature constraints, FinFETs-based scaling factors, and consideration of advanced processor features. These trends show that modeling dark silicon only as a power budget constraint results in over-estimations of dark silicon. On the other hand, consideration of the peak temperature constraint during the modeling provides reduced amount of dark silicon. Furthermore, using DVFS provides the ability to increase the overall system performance and further decrease the amounts of dark silicon. Additionally, an analysis of boosting techniques and constant frequency schemes, both in STC and NTC regions, was performed. The results show that boosting techniques achieve a slightly higher overall performance at the cost of high peak power and energy consumptions, implying that running at constant frequencies in STC is a better approach in dark silicon to achieve sustainable performance. Furthermore, NTC is only needed when minimizing energy under performance constraints, but not for maximizing performance in the dark silicon era.

In a broader perspective, the emergence of advanced computing paradigms like Invasive Computing [26] provide new opportunities and incentives to improve the computing efficiency in the dark silicon era. Towards this end, accurate estimation of dark silicon, thermal-aware dark silicon management, and exploration of performance vs. power/energy tradeoffs in the STC and NTC regimes are crucial.

References

[1] M. Shafique, S. Garg, J. Henkel, and D. Marculescu, "The EDA challenges in the dark silicon era: Temperature, reliability, and variability perspectives", in *51st Design Automation Conference (DAC)*, 2014, pp. 185:1–185:6.

[2] M. Shafique, S. Garg, T. Mitra, S. Parameswaran, and J. Henkel, "Dark silicon as a challenge for hardware/software co-design: Invited special session paper", in *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2014, pp. 13:1–13:10.

[3] D. Gnad, M. Shafique, F. Kriebel, D. Sun, and J. Henkel, "Hayat: Harnessing dark silicon and variability for aging deceleration and balancing", in *52nd Design Automation Conference (DAC)*, 2015.

[4] F. Kriebel, S. Rehman, D. Sun, M. Shafique, and J. Henkel, "ASER: Adaptive soft error resilience for reliability-heterogeneous processors in the dark silicon era", in *51st Design Automation Conference (DAC)*, 2014.

[5] M. Shafique, D. Gnad, S. Garg, and J. Henkel, "Variability-aware dark silicon management in on-chip many-core systems", in *Design, Automation and Test in Europe (DATE)*, 2015.

[6] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling", in *the 38th International Symposium on Computer Architecture (ISCA)*, 2011, pp. 365–376.

[7] S. Venkataramani, S. Chakradhar, K. Roy, and A. Raghunathan, "Approximate computing and the quest for computing efficiency", in *52nd Design Automation Conference (DAC)*, 2015.

[8] H. Cheng, J. Zhan, J. Zhao, Y. Xie, J. Sampson, and M. J. Irwin, "Core vs. uncore: The heart of darkness", in *52nd Design Automation Conference (DAC)*, 2015.

[9] "International technology roadmap for semiconductors (ITRS)", http://www.itrs.net.

[10] R. Borkar, M. Bohr, and S. Jourdan, "Advancing moore's law in 2014."

[11] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 simulator", *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, Aug. 2011.

[12] S. Li, J.-H. Ahn, R. Strong, J. Brockman, D. Tullsen, and N. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures", in *42nd IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2009, pp. 469–480.

[13] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "Hotspot: A compact thermal modeling methodology for early-stage vlsi design", *IEEE Transactions on VLSI Systems*, vol. 14, no. 5, pp. 501–513, May 2006.

[14] N. Pinckney, K. Sewell, R. G. Dreslinski, D. Fick, T. Mudge, D. Sylvester, and D. Blaauw, "Assessing the performance limits of parallelized near-threshold computing", in *49th Design Automation Conference (DAC)*, 2012, pp. 1147–1152.

[15] A. Grenat, S. Pant, R. Rachala, and S. Naffziger, "5.6 adaptive clocking system for improved power efficiency in a 28nm x86-64 microprocessor", in *International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 106–107.

[16] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: Characterization and architectural implications", in *17th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2008, pp. 72–81.

[17] S. Huck, "Measuring processor power tdp vs. acp."

[18] Intel Corporation, "Dual-core intel xeon processor 5100 series datasheet, revision 003", August 2007.

[19] H. Khdr, S. Pagani, M. Shafique, and J. Henkel, "Thermal constrained resource management for mixed ILP-TLP workloads in dark silicon chips", in *52nd Design Automation Conference (DAC)*, 2015.

[20] S. Pagani, H. Khdr, W. Munawar, J.-J. Chen, M. Shafique, M. Li, and J. Henkel, "TSP: Thermal safe power - efficient power budgeting for many-core systems in dark silicon", in *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, Oct 2014, pp. 10:1–10:10.

[21] E. Rotem, A. Naveh, D. Rajwan, A. Ananthakrishnan, and E. Weissmann, "Power-management architecture of the intel microarchitecture code-named sandy bridge", *Micro, IEEE*, vol. 32, no. 2, pp. 20–27, March 2012.

[22] J. Casazza, "First the tick, now the tock: Intel microarchitecture (nehalem)", Intel Corporation, White Paper, 2009.

[23] Intel Corporation, "Intel turbo boost technology in intel CoreTM microarchitecture (nehalem) based processors", White Paper, November 2008.

[24] S. Nussbaum, "AMD trinity APU", in *Hot Chips: A Symposium on High Performance Chips*, 2012.

[25] D.-C. Juan, S. Garg, J. Park, and D. Marculescu, "Learning the optimal operating point for many-core systems with extended range voltage/frequency scaling", in *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2013, pp. 8:1–8:10.

[26] J. Teich, J. Henkel, A. Herkersdorf, D. Schmitt-Landsiedel, W. Schröder-Preikschat, and G. Snelting, "Invasive computing: An overview", in *Multiprocessor System-on-Chip*. Springer New York, 2011, pp. 241–268.