Variation-Aware Voltage Island Formation for Power Efficient Near-Threshold Manycore Architectures

Ioannis Stamelakos, Sotirios Xydis, Gianluca Palermo, Cristina Silvano
Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria
E-mail: {stamelakos, xydis, gpalermo, silvano}@elet.polimi.it

Abstract—The power-wall problem and its dual utilization-wall problem are considered among the main barriers to feasible/efficient scaling in the manycore era. Several researchers have proposed the usage of aggressive voltage scaling techniques at the near-threshold voltage region, promising significant improvements in power efficiency at the expense of reduced performance values and higher sensitivity to process parametric variations. In this paper, we introduce a variability-aware framework for exploring the potential power-efficiency of the Near Threshold Computing (NTC) under performance constraints. We propose and analyze the usage of fine-grained voltage islands to cope with the increased effect of variability problem in the NTC region. For the considered workloads, we found that the power impact of fine-grained voltage islands formation can be up to 35% for a 128-core chip operating at NTC region, while the adoption of a variability aware technique can bring to a power reduction of up to 43% with respect to a variability unaware technique. Finally, we show that voltage regulator’s complexity, in terms of voltage quantization levels, has a very low effect on the power efficiency at NTC, making in that way the usage of voltage islands a feasible solution for coping with variability. In comparison with the conventional Super-Threshold Voltage Computing (STC), computations at NTC regime are performed in a very energy efficient manner, unfortunately at the expense of reduced performance and high susceptibility to parametric process variations.

In this paper, we investigate the power efficiency potential of manycore architectures at the NTC regime, considering the process variations as well as a power delivery architecture supporting multiple $V_{dd}$ domains, under strict performance constraints originated from multicore architectures at the STC regime. Unlike previous works on variation aware voltage allocation that target the STC regime [12], [13], [14], we propose the formation of voltage islands (VIs) for the minimization of the impact of within-die variations, which are more evident at NTC, in both performance and power. In particular, we developed a framework for manycore architectures operating at NTC to investigate the power efficiency under different workloads, while sustaining the performance when moving from the ST to the NT region. The framework has been parametrized in order to exploit different voltage island formation and to deal with process variability. Additionally, to generalize the analysis, we study four clustered manycore architectural organizations – differing on the number of cores per cluster.

Extensive experimental analysis showed that for the considered workloads, when moving to the NT regime for a 128-core architecture, average power gains close to 65% are delivered while sustaining the performance values obtained by a 16-core architecture at STC. The power impact of fine-grained voltage island formation can be up to 35% for a 128-core chip operating at NTC. Additionally, in comparison with variation unaware techniques, the proposed variation-aware NTC voltage island formation delivers power gain up to 8% considering a single VI per chip and up to 43% when considering the fine-grained multiple VI case, that is able to deal better with variability. Finally, analyzing the $V_{dd}$ distributions at NTC, we demonstrated that the utilization of multiple VIs together with efficient integrated regulators can be considered a feasible option at NTC to efficiently deal with the process variability.

The remainder of the paper is organized as follows. Section II motivates this work by presenting the technical background regarding the NTC paradigm. Section III introduces the proposed framework for variation-aware voltage island formation

I. INTRODUCTION

The continuous technology scaling following Moore’s law [1], has recently raised the era of manycore architectures [2], [3], as the principal strategy for continuing performance growth. However, the end of Dennard’s scaling [4] brings designers in front of the so called power/utilization wall. Projections show that the gap between the number of cores integrated on a chip and the number of cores that can be utilized will continue to widen on future technology nodes [5]. As a result, dark silicon - transistor count under-utilization due to power budget - has been recently emerged as major design challenge that jeopardizes the well established core count scaling path in current and future chip generations.

To address the dark silicon problem, researchers have proposed techniques at the micro-architectural level [6], [7], [8] down to physical and device level [9], [10]. Near-Threshold Voltage Computing (NTC) [11] has been proposed as a promising technique to mitigate the effects of dark silicon, allowing a large number of cores to operate under a given manycore power envelope. NTC takes advantage of the quadratic relation between the supply voltage ($V_{dd}$) and the consumed power, by lowering the operating $V_{dd}$ to a region slightly larger than the transistors’ threshold voltage ($V_{th}$).

This work was partially supported by the EC under the grant HARPA FP7-612069
at NTC under given performance constraints, while Section IV experimentally analyzes its effectiveness. Finally, Section V concludes the paper.

II. NTC: MOTIVATION AND CHALLENGES

Near-threshold voltage operation relies on the aggressive tuning of the $V_{dd}$ of the integrated circuit very close to the transistors’ threshold voltage $V_{th}$, to a region where still $V_{dd} > V_{th}$. This decrement of the supply voltage increases the potential for energy efficient computation, e.g. by reducing $V_{dd}$ from the nominal 1.1 V to 500 mV, energy gains of $10 \times$ are reported [11]. While near-threshold region is not more power efficient than the theoretical limits provided in the sub-threshold/ultra-low voltage region, where $V_{dd} < V_{th}$, NTC has gained a lot of attention due to low energy operation at higher performance and easier adoption across multiple application platforms in comparison with the sub-threshold circuits. NTC is the region that delivers interesting trade-offs regarding to energy efficiency and transistor delay, since super-threshold $V_{dd}$ quickly reduces energy efficiency while sub-threshold $V_{dd}$ leads to slower transistors. However, NTC comes together with two major drawbacks, i.e. reduced performance and increased sensitivity to process variations.

Performance reduction at NTC is exposed through the limited maximum achievable clock frequency. This is an implicit effect due to the reduction of the $V_{dd} - V_{th}$ difference, applied when moving to the NTC region. Performance degradation can be compensated by exploiting trade-off points of higher task parallelism at lower clock frequencies. Thus, an important open question for NTC to be examined is the following: Is the inherent parallelism of the existing applications enough to retain the performance levels of super-threshold design with lower power consumption, thus making it worth going to near-threshold operation? Pinckney et al. [15] studied the limits of voltage scaling together with task parallelization knobs to address the performance degradation at NTC by considering a clustered micro-architectural template with cores sharing the local cache memory [16]. They showed that under realistic application/architecture/technology features (i.e. parallelization efficiency, inter-core communication, $V_{th}$ selection etc.) the theoretical energy optimum point ($\frac{d\text{Energy}}{dV_{th}} = 0$), moves from the sub-threshold to the near-threshold region. Considering a single supply voltage per die, the energy optimum point can be found within an interval of 200 mV higher $V_{th}$, thus implicitly defining the upper limits of the NTC region.

The second important challenge for manycore architectures operating at the NTC regime is their increased sensitivity to process variations. The transistor delay is heavily affected by the variation of $V_{th}$ at NT voltages compared to the one in super-threshold voltages [17], [18]. In addition, failure rate of conventional SRAM cells is increased in low voltage operation [19]. As a consequence, the operating frequency of the cores varies considerably, reducing the yield. In addition, variation’s effects on the total power of the chip have to be carefully considered, due to the exponential dependency of leakage current upon $V_{th}$. Karpuzcu et al. presented Various-NTV [20], a micro-architectural model that adopts proper gate-delay and SRAM cell type models to capture the increased sensitivity of manycore chips to process variations at NTC. For mitigating variation effects on performance at NTC, the EnergySmart architecture and thread assignment methodology have been recently proposed in [21]. By assuming on-chip voltage regulators of low-efficiency, EnergySmart adopts single voltage - multiple frequency islands to cope with variability.

In this paper, we focus our study on the NTC design space defined though [11] and [21]. Specifically, we target power efficient NTC manycore architectures that sustain STC performance levels considering their increased sensitivity to process variation. While previous work [11] addresses the problem by considering single per chip voltage domain, in this paper we differentiate our approach by exploring multiple voltage domain NTC architectures through variation aware-voltage island formation techniques. Although in [21] some concerns are raised regarding the usage of on-chip regulators for multiple voltage islands in NTC, we motivated our research by the recent advancements in cost-effective fully-integrated voltage regulators (FIVR) [22]. FIVR regulators are expected to achieve very high voltage efficiency, enabling the integration of multiple power rails. We show that the workload characteristics at NTC enable fine-grained $V_{dd}$ tuning, by delivering narrow $V_{dd}$ distributions with very low power consumption (Section IV).

III. FRAMEWORK FOR VI FORMATION AT NTC

The overall exploration framework for variation-aware VI formation at NTC is shown in Figure 1. It accepts as main inputs the performance, power and area characterization curves of the targeted application at the STC regime. The STC characterization is performed by scaling the number of cores of the underlying manycore architecture template and then scaling accordingly the application’s degree of parallelization. The Sniper multicore simulator [23] and the McPAT power modeling framework [24] have been used for the performance and power characterization, respectively. Designer/architect specific constraints are provided regarding the minimum allowed performance, $L_{min}$ and the maximum core count constraint, $C_{max}$ for the NTC manycore. Given the aforementioned inputs, the proposed exploration framework generates the VI configurations and the corresponding $V_{dd}$ allocation decisions per VI, for a manycore architecture with $C_{max}$ number of cores operating at the NTC regime and satisfying the performance constraint $L_{min}$ under parametric process variation. In the remainder of this section, we describe the basic components of the proposed methodology.

A. NTC Frequency Calculation for Sustaining STC Performance

Core to core frequency and leakage variations can be mitigated by varying voltage/frequency levels of independent voltage islands. There are four different power management schemes supporting voltage/frequency islands, i.e. Single-Voltage/Single-Frequency (SVSF) for all cores,
Single-Voltage/Multiple-Frequencies (SVMF), Multiple-Voltages/Multiple-Frequencies (MVMF) and Multiple-Voltages/Single-Frequency (MVSF). While the SVSF scheme requires overdesigned power management decisions, the SVMF [21] and MVMF schemes mitigate variations by exposing the system heterogeneity to the programmer, i.e. application execution on different clock frequencies across the cores. In order to alleviate the aforementioned inefficiencies, we adopt the MVSF power management scheme that enables to mitigate process variations by tuning the $V_{dd}$ allocation of the VIs while exposing to the programmer an iso-frequency view of the overall manycore platform.

Under this configuration, we first calculate the clock frequency of the platform at NTC regime, $f_{N,C}$, that satisfies the performance constraint, $L_{\text{min}}$. $f_{N,C}$ will be then used by the VI formation and variability aware $V_{dd}$ allocation module to configure the MVSF decisions. Let $L_{C_{\text{max}}}$ be the performance, in terms of latency, at the STC regime of a manycore architecture with $C_{\text{max}}$ number of cores, running at $f_{\text{STC}}$. At STC region, $L_{\text{min}} - L_{C_{\text{max}}} > 0$ is the available slack in latency due to the higher parallelism degree of the architecture, that can be exploited to run the application in lower frequency. Utilizing this positive slack, the $f_{NTC}$ is calculated as follows:

$$f_{NTC} = \frac{L_{C_{\text{max}}}}{L_{\text{min}}} \times f_{\text{STC}} \quad (1)$$

B. Tiled Manycore Architecture Description

While the framework considers the underlying manycore architecture as a free exploration parameter, we focus our study on tile-based architectures. Figure 2 shows an abstract view of the tile-based manycore architecture, as well as the different intra-tile organizations. We consider four intra-tile architectures by varying the number of cores per tile and the memory configuration of the last level cache (LLC) per tile. Each core owns a private instruction and data cache (PS). The LLC (LLS) is shared among the different cores composing a tile. The Intel Nehalem processor [25] configuration for the core and the PS has been adopted. While the PS size remains constant across the different intra-tile configurations, the size of the (LLS) is scaled according to the number of cores in the tile, keeping constant the total chip area. We use the following abbreviations for differentiating manycore architectures based on four tile types: (i) $S1$: each core owns a Last Level LLS, (ii) $S2$: LLS is shared between 2 adjacent cores, (iii) $S4$: LLS is shared among 4 adjacent cores, (iv) $S8$: LLS is shared among 8 adjacent cores. While $S4$ and $S8$ resemble the cluster organizations proposed in [16], [21], we also explored more fine-grained clusters, i.e. $S1$ and $S2$. Tile's type defines the minimum VI granularity supported by each manycore configuration. Thus, for a $S1$ manycore platform the finest granularity of a voltage domain is $i$ core per VI.
C. Micro-architectural Process Variation Model at NTC

In order to capture the process variation at the NTN regime, we integrate the Various-NTV [20] microarchitectural model within the proposed framework. While Various-NTV reuses the spherical distance function in [26] for modeling the intradie spatial correlations, it heavily extends [26] by updating the STC micro-architectural delay and SRAM cell models to reflect in a more accurate manner the higher sensitivity of NTN on process variation. We used ArchFP [27] tool to automatically generate the floorplan of the targeted manycore architectures. Based on the provided manycore floorplan, Various-NTV generates the corresponding variation maps accounting for the within-die (WID) and die-to-die (D2D) process variations. Figure 3 shows the floorplan of the S8 manycore architecture with 128 cores (Figure 3(a)), together with a sample instance of its $V_{th}$ variation map (Figure 3(b)). Assuming $B$ as the set of component blocks found in the floorplan and $D$ the set of dies, we now define $V_{th}^{(i,j)}$, $i \in B, j \in D$ that corresponds to the $V_{th}$ of the architecture’s component $i$ in sample die $j$. Once extracted, $V_{th}^{(i,j)}$ is used for allocating to each component the lowest possible $V_{dd}^{(i,j)}$ for sustaining $f_{NTC}$ frequency constraint given that:

$$f_{NTC} \propto \frac{(V_{dd}^{(i,j)} - V_{th}^{(i,j)})^\beta}{V_{dd}^{(i,j)}}$$

where $\beta$ is a technology-dependent constant ($\approx 1.5$).

D. Voltage Island Formation & Variability Aware $V_{dd}$ Allocation

The final phase of the proposed framework first performs the generation of the VIIs and then for each VI solution computes the per island $V_{dd}$ assignment that satisfies the $f_{NTC}$ derived by the latency performance constraint. The VI formation procedure explores all valid granularities of rectangular voltage islands in both vertical and horizontal directions, by grouping together adjacent cores. We use the notation $r \times c$, i.e. $r$ rows - $c$ cores per row, to indicate the type of the evaluated VI granularity. Depending on the type of the underlying tiled architecture (S1, S2, S4, S8), different constraints are included in the VI formation, since not all the core groupings are valid. For example, in the case of a S8 tiled manycore, LLS is shared among 8 cores, thus the finest granularity that we can be evaluated is $2 \times 4$: 2 rows, 4 cores per row.

For the $j^{th}$ die, $j \in D$, each VI, $k \in VI$, operates in its own $V_{dd}^{(k,j)}$, tuned for the VI$k,j$ group of processors and memories. In VI$k,j$, the core with the highest $V_{th}^{(k,j)}$, $i \in B, j \in D$ determines the $V_{dd}$ for the specific voltage island, to satisfy the VI’s critical path timing. The trade-off by moving towards more severe grainy VI granularities is that, we reduce area cost since less voltage regulation logic is allocated at the expense of degrading the power efficiency of the manycore in respect to the finest possible granularity. For $B_k, k \in VI$, the set of resources found in VI$k$ and from eq. 2, we calculate $V_{dd}^{(k,j)}$ according to the following relation:

$$V_{dd}^{(k,j)} = \max_{i \in B_k,j \in D} [V_{dd}^{(i,j)}]$$

Given the $V_{dd}$ allocation per VI, $V_{dd}^{(k,j)}$, $k \in VI, j \in D$, and the power characterization for the manycore with $C_{max}$ number of cores at STC, we can calculate the power of each component in NTN. For $i \in B_k, j \in D, k \in VI$, the dynamic, $DP$ and leakage, $LP$, power scaling factors are:

$$S_{DP}^{(i,j,k)} = \left( \frac{V_{dd}^{(k,j)}}{V_{ddSTC}} \right)^2 \times \left( \frac{f_{NTC}}{f_{STC}} \right)$$

$$S_{LP}^{(i,j,k)} = \exp \left( \frac{V_{thSTC} - V_{th}^{(i,j)} + DIBL}{n \times V_{thermal}} \right)$$

where $DIBL$ is the coefficient modeling the Drain-Induced Barrier Lowering effect, $V_{thermal}$ is the thermal voltage, and $n$ is the sub-threshold slope coefficient. The DIBL effect is a deep-submicron effect and is related to the reduction of the threshold voltage as a function of the drain voltage. DIBL is enhanced at higher drain voltage and tends to be more severe with process scaling to shorter gate lengths. Lowering supply voltage provides an exponential reduction in sub-threshold current resulting from the DIBL effect. Figure 4 shows the impact of DIBL effect on the reduction of leakage power in manycore architectures at NTN regime. As shown, by moving from STC multicore (16 cores) to NTN manycore (128 cores) architecture configurations, the DIBL effect accounts for a significant portion of the total power of the system.

IV. Experimental Results

In this section, we experimentally evaluate the efficiency of the proposed framework. Without loss of generality, we consider that the performance $L_{min}$ corresponds to a 16 core multicore in the STC regime, while the constraint $C_{max}$ targets a 128 core many-core chip at NTN, at 22nm technology node. Maximum $V_{dd}$ was set to 1.05V and the frequency to 3.2 GHz for the STC regime, according to parameter values derived from [28] for conservative technology scaling. From Various-NTV, we extract 100 different variation maps by using a 24x16 grid based on the core/cache granularity. The most significant parameters and their values are summarized in
Table I. For our experiments we used five applications from the SPLASH-2 benchmark suite [29] by using the "large dataset" provided within Sniper [23]. The applications exhibit different speedup scaling degrees, i.e. close to ideal (radiosity), medium (barnes, water-nsq) and limited (raytrace, water-sp).

Additionally, we examined an average case workload, that aggregates in the execution sequence the five aforementioned applications. Specifically, this average case workload is like executing all the aforementioned applications, one after the other and then treating it as a single benchmark. In that way, we manage to see what happens in an average case since it includes benchmarks that scale well and others that don’t scale well. We present results regarding the power efficiency delivered by adopting the proposed approach and we provide a sensitivity oriented analysis regarding parameters of the voltage regulation structure.

A. Analysis of Power Gains at NTC Regime

1) Moving from 16 STC Cores to 128 NTC Cores: Figure 5 shows the power consumption when moving from 16 cores at STC to 128 cores at NTC for each benchmark. Radiosity delivers the highest power gains since it scales almost ideally in terms of performance as the number of cores increases. For the radiosity benchmark we observed a 95% decrement in power while for the barnes and water-nsq that exhibit a medium scaling behavior, we observe a reduction of 77%. As shown, the scaling behavior of the applications with respect to increasing the number of cores, heavily affects the power efficiency at NTC, since the application takes advantage of the available performance slack when moving to a large number of cores. Thus, besides frequency we can aggressively scale down $V_{dd}$ as well, and reduce power drastically, taking advantage of $V_{dd}$'s quadratic relation with dynamic power and its linear plus DIBL relation with leakage current. Raytrace and water_sp exhibit lower scaling degrees, limiting performance boost when their load is split and distributed to 128 cores. In this case, $V_{dd}$ assignment acquires higher values restricting power gains. The last column of Figure 6 depicts the power gains delivered in the average workload. Although benchmarks that don’t scale well are included, a near threshold $V_{dd}$ (~0.4V) is acquired, delivering a 65% power reduction with respect to the 16 core STC manycore.

2) Variation Aware versus Overdesign NTC operation: We compared the power gains delivered by the proposed variation aware VI formation versus an overdesign approach to mitigate variation effects. From the $V_{th}$ distribution, we calculate the $V_{dd}$ of architectural components according to Eq. 2, with $V_{th}$’s overdesign value being equal to $\mu V_{th} + 3\sigma V_{th}$. Figure 6 reports the gains of the variability aware approach over the overdesign one. The histograms with the singleVI annotation represent power gains when having only one VI, and as a consequence one $V_{dd}$ for the whole chip. Under a singleVI configuration, the variation aware approach achieves power gains around 5%, for all the available cluster architectures ($S_i$, $i \in \{1, 2, 4, 8\}$). On the contrary, the histograms with the finestVI annotation show the power gains achieved by considering the finest VI granularity possible for each architecture. Since S1 enables the finest $1 \times 1$ VI granularity to be exploited, it delivers the highest gains over the overdesign approach, that range between 34-42%. In the rest of the architectures, namely (S2, S4, S8), the gains vary between 29-34%, 25-28% and 18-23%, respectively.

3) Analysis of VI Granularity on NTC Power Efficiency: Figure 7 shows the impact of the different voltage island configurations in terms of power consumption at the NTC regime. The voltage island formation that has been analyzed includes all the possible combination in terms of power-of-two of the cores. We restrict the voltage island granularities to an aspect ratio between 1 and 1/4 (considering a voltage island that has been analyzed includes all the possible combination in terms of power-of-two of the cores. We restrict the voltage island granularities to an aspect ratio between 1 and 1/4 (considering a voltage island configuration $c \times r$ the aspect ratio is $c/r$). In each case, we considered the tile size as the smallest possible voltage island. The constant trend over all the workloads and architectures, is that finer the granularity of the voltage island higher the power savings. In fact, selecting smaller voltage islands, we can cope with the variability in a more aggressive way by using a fine-grained tuning of the $V_{dd}$ over the entire chip. The advantage passing from the single voltage island to the finest voltage island depends on the different architectural...
configuration: 30-35% for 128s1, 24-30% for 128s2, 19-24% for 128s4 and 14-18% for 128s8. In addition, the impact of the different voltage island configurations (when composed of the same number of cores), such as passing from $4 \times 4$ to $2 \times 8$ or passing from $1 \times 4$ to $2 \times 2$, is very limited. Despite the global trend, analyzing in more detail the power behavior over all the four clustered architectures, we noticed that the 128s1 architecture is not constantly the best configuration across the various VI granularities. In some cases the best configuration shifts over the 128s2 architecture (see water-sp in Figure 7). This phenomenon depends on the application scalability over the different clustered architectures.

### B. Voltage Regulation Oriented Analysis

The analysis conducted so far considers the ability to ideally deliver all the requested voltage levels. Since this is not a realistic scenario according to current state-of-art power supply architectures, hereafter we analyze the impact of the on-chip voltage regulator resolution on power efficiency.

We analyzed three different voltage regulator resolutions, delivering voltage with a precision of (i) 12.5mV, (ii) 25mV and (iii) 50mV. Adopting the aforementioned schemes, we demonstrate the effect of allocating integrated regulators in the NTC region (from $[V_{th}] \rightarrow [V_{th} + 200mV]$) that includes respectively 16, 8, and 4 voltage quantization levels. Figure 8 presents the average power overhead for each one of the voltage regulator precisions. Power overhead refers to the normalized difference between the power consumed in the ideal case (voltage regulator delivering arbitrary $V_{dd}$ values) and the power with the specific value of voltage precision. The results are the average values for all the benchmarks and all the four architectures that we investigated. As expected, the higher is the resolution the smaller is the overhead since we are closer to the ideal case, passing from a 12% at 50mV to less than 3% at 12.5mV. For the applications that exhibit ideal or medium scaling with respect to increasing the number of cores, such as radiosity, barnes and water-nsq, the overhead of 12% can be efficiently compensated by the low-power consumption at NTC regime. On the opposite, for the case of applications with limited scaling, such as raytrace and water-sp, an integrated voltage regulation scheme that provides high resolution of the delivered $V_{dd}$ is preferable.

Finally, Figure 9 shows the $V_{dd}$ probability distribution considering the 12.5mV as the regulator’s granularity for the *barnes* application running on the 128core architecture and operating at NTC, across all the examined S7-S8 tile types. We can observe that the $V_{dd}$ distribution is very concentrated across the mean value $\mu = 0.388V$ with $\sigma = 0.071V$. The
Vdd distributions of the rest of the applications resemble the depicted one on Figure 9 with similar σ and slightly shifted μ values, according to the scaling behavior of the application. Narrow Vdd distributions together with low power consumption at NTC, instruct that the integrated voltage regulation circuitry can efficiently supply the requested power without the need of allocating multiple levels of voltage supplies, showing the efficiency of moving towards multiple VIs for supporting NTC operation on manycore chips.

V. CONCLUSION

In this paper, we presented a variability-aware framework for exploring the power-efficiency of Near-Threshold Computing. Motivated by recent advancement on power delivery systems, we proposed the utilization of voltage island formation combined with the operation at the near-threshold region as an effective technique for building power efficient manycore architectures that sustain performance values delivered by conventional super-threshold computing. Through extensive experimentation, we showed the optimization potentials of moving towards near-threshold voltage computation, exposing its high dependency on both workload characteristics and underlying architectural organization.

REFERENCES