Temperature Constrained Power Management Scheme for 3D MPSoC

Arnica Aggarwal, Sumeet S. Kumar, Amir Zjajo, Rene van Leuken
Circuits and Systems Group, Delft University of Technology,
Mekelweg 4, 2628 CD Delft, The Netherlands
arnica.aggarwal@gmail.com, s.s.kumar@tudelft.nl, amir.zjajo@ieee.org, t.g.r.m.vanleuken@tudelft.nl

Abstract

This paper proposes a new temperature constrained power management scheme for 3D MPSoCs that utilizes instantaneous temperature monitoring along with information on the physical structure of the stack to determine operating V-F levels for processing elements (PE). The scheme implements a weighted policy that prevents PEs deep inside the stack from being turned off, maintains operating temperatures stable and within safe margins, and reduces overall execution time by up to 19.55%.

1 Introduction

Progression towards smaller technology nodes has enabled the integration of tremendous amounts of computing power in modern silicon dice. However, the scaling down of feature sizes has exposed issues such as process variation, leakage power consumption, and the limitations of interconnect performance [1]. Total power dissipation and system power density are now at the limits of what current packaging and cooling solutions can support [2]. 3D integration is now emerging as an attractive solution towards sustaining the observed trend of increasing integration densities without significant increase in area footprint. *Through Silicon Vias* (TSV) form the backbone of 3D die-stacking enabling vertical interconnections between tiers in stacks of silicon dice. While their electrical performance has translated to improved system performance [1][2], their use has also aggravated thermal issues [3][4][5], and consequently the reliability of stacked-die chips [5].

Dynamic Voltage and Frequency Scaling (DVFS) is a commonly used architectural-level power management technique that operates *processing elements* (PE) at different *voltage and frequency* (V-F) levels according to their workload [2][6]. Improvements in application performance and effective utilization of power budget are reported in [7] using a temperature constrained DVFS based power management scheme for planar *chip multiprocessors* (CMP). The proposal controls V-F levels of individual processing elements based on their local operating temperature and available chip power budget monitored. However, the proposal cannot be applied to 3D architectures since it does not consider thermal coupling between adjacent PEs - a significant factor in die stacks [8]. A thermal management policy for 3D MPSoCs using inter-tier liquid cooling is proposed in [9]. The work recognizes the variation in thermal conditions between the extremities of deep stacks, highlighting the inefficacy of conventional DVFS approaches that result in deeper PEs turning off more often. The proposal uses a thermal management policy that takes into account the distance of PEs from the cooling liquid inlet port during V-F scaling, and varies the rate of coolant flow based on their temperature. Liquid cooling thus forms the core part of this proposal. A comprehensive thermal management policy for 3D CMPs incorporating temperature aware workload migration and run-time global power-thermal budgeting is presented in [8]. Within the policy, PEs with available temperature budgets executing high *instruction per cycle* (IPC) workloads are scaled to higher V-F levels in order to improve performance after weighing the potential performance benefits that may be obtained against the consequent thermal implications for neighbouring PEs.

This paper presents a new temperature constrained power management scheme for 3D MPSoCs that uses instantaneous temperature monitoring coupled with information on the physical structure of the stack to determine operating V-F levels for PEs. Furthermore, DVFS decisions are aided by a thermal resistance matrix that provides information on thermal relationships between PEs in the stack, and implemented through a weighted policy that prevents PEs on deeper layers from reaching critical temperatures and thus being turned *off*. The scheme is evaluated for per-core and island granularities, and is observed to effectively maintain temperatures of all PEs stable and within safe margins when compared to the conventional 2D DVFS approach.

2 Thermal Modeling

A 3D integrated circuit contains multiple vertically stacked silicon layers, each containing PEs and memory modules. Most compact thermal models use resistor and capacitor to model the steady-state as well as transient temperature response in such circuit, analogous to electrical RC networks [8]. Thermal conductance between two PEs can be calculated using conductance equations. However, due to the flow of heat in different directions, additional information like impedances in different direction and various paths are required to have a direct relation between temperature and V-F level. Figure 1 illustrates a thermal model of a section of 3D die stack where resistor, capacitor and current source denote thermal resistance, thermal capacitance of a PE, and heat transfer rate or power of a PE, respectively. The heat sink is shown at the bottom of the stack which connects to die 1 through a thermal resistance R_{hs}. For a thermal model to be accurate, each thermal cell must be small enough so as for the temperature within it is to be assumed uniform. The heat flows in all directions and through different paths. The ratio of heat flowing in the different directions depends on the ratio of impedances seen in those directions. The difference between ΔT_1 and ΔT_2 is a strong function of material properties, and it increases as thermal resistance between them increases. This becomes more cumbersome in an actual model where a PE node is not only connected to the nodes above or below it, but also on the same plane via $R_{lateral}$.

Although $R_{lateral}$ is often ignored, this resistance should be considered in deep stacks as the conductivity to the ambient decreases with the depth in a stack. The temperature change at

978-1-4673-1504-3/12/$31.00 ©2012 IEEE

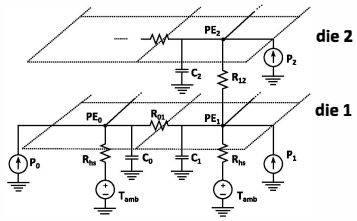

Figure 1: Simplified thermal model of a 3D multi-core system.

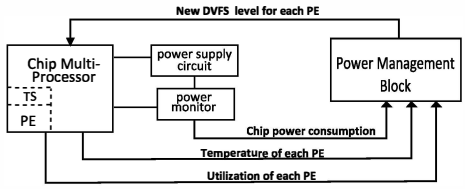

Figure 2: Control loop for power management scheme

a node i and node j due to change in power dissipation at node j can be given by

$$\left.\begin{aligned} \Delta T_j &= x * \Delta P_j * R_{jhs} \Rightarrow x.R_{jhs} = \frac{\Delta T_j}{\Delta P_j}\ ; \quad and \\ \Delta T_i &= \Delta P_j (x * R_{jhs} - y * R_{ij}) \\ &\Rightarrow \underbrace{(x * R_{jhs} - y * R_{ij})}_{\text{Effective resistance}} = \frac{\Delta T_i}{\Delta P_j} \end{aligned}\right\} \quad (1)$$

where x and y denote the fraction of P_1 flowing towards R_{hs} and R_{12} respectively, R_{jhs} is the thermal resistance between node j and heat sink, and R_{ij} is the thermal resistance between node i and j. An *NxN* matrix is created for the values of $(\Delta T_i/\Delta P_j)$ in equation 1 for all PEs, representing the effective thermal resistance between them to form a direct relation between ΔP and ΔT. The change in temperature of PE_i due to change in power dissipation of PE_j, can now be directly given by

$$\Delta T_i = R_{eff_{ij}} * \Delta P_j \Rightarrow \Delta T_i = R_{eff_{ii}} * A * \Delta(V_i^2 * F_i) \quad (2)$$

where A is a constant whose value varies for different PEs according to the characteristics of their workload and can be represented by a generalized value for an intended workload. P is power, V and F are voltage and frequency corresponding to a DVFS level.

3 Power Management Scheme

The temperature of a PE is primarily determined by its power dissipation, as well as its location within the 3D stack and for heterogeneous systems, its area. *Activity factor* (utilization) from PE performance counters, temperature from PE thermal sensors and total chip power obtained from the system are considered as inputs by the proposed *Power Management Block* (PMB) in deciding V-F levels in order to maintain total chip power below a set budget and temperature of PEs under the critical level. This is illustrated in Figure 2. V-F scaling decisions are taken by the algorithm shown in Figure 3. Note that *Control Period* defines the intervals at which the PMB takes inputs and computes new V-F levels. Temperature inputs to the PMB are made available at intervals defined by the *Temperature check period*. The algorithm in Figure 3 is divided into several stages, namely, initial updates, thermal run-out, convergence check, pull up or pull down, and write-back and reset.

Initial Updates: At the beginning of a new control period, the difference between total chip power and local power budget value are computed. In the event that a new temperature check cycle has started, the difference between actual and critical temperature of each PE is updated.

Thermal Runout: This step ensures that temperature of each PE is maintained within the safety margin. Each PE is assigned a weight

$$a * (1 - Util) + b * (normalized R_{eff}[victim PE][i]) \quad (3)$$

where a and b are experimentally determined constants. A less active PE with a strong thermal relation with the victim is considered to have the heaviest weight, and is thus considered for V-F scale down first. If required, the next candidate PE is selected and scaled down. In the event that temperature cannot be brought below the critical, the victim is turned off. In order to prevent repeated fluctuations, when the V-F level of a PE is scaled down due to a victim PE, it is not reinstated until the victim is within the safe temperature margin. Such updates are performed in the initial update stage.

Convergence Check: Power value is considered as converged if total chip power is between *98%* and *100%* of power budget value. If this is not the case, V-F level pull up or pull down is required.

Pull Up/Pull Down: To scale the system up or down depending on the allocated power budget, a weighted equation is considered.

$$\begin{aligned} &(c * Util) + (d * normalized\_temp\_margin) + \\ &(e * normalized\_height) + (f * normalized\_area) \end{aligned} \quad (4)$$

where c, d, e and f are experimentally determined constants. A highly active PE that is cooler, situated close to heatsink and with a larger area is the preferred choice for V-F upscaling. However, scaling is performed only if the new temperature after scaling is below the safety margin. This upscaling is performed iteratively until no more PEs can be pulled up, or if the total power reaches the *98%* window of convergence with the budget value. In the event that the budget has been exceeded, the pull down stage is invoked in order to converge. For V-F downscaling, the PE with the smallest weight is selected and the pull down is iteratively performed until no more PEs can be pulled down, or if the total power falls below the budget value. At each instance of pull up and pull down at a PE, the difference between its actual and critical temperature is updated.

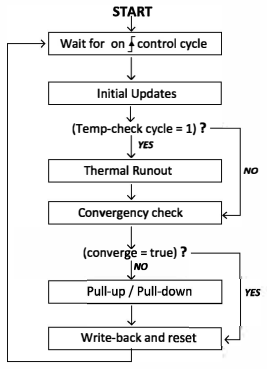

Figure 3: Flowchart showing stages in PMB

Write-Back and Reset: The chosen V-F values and the on/off state signals for each PE are implemented on the PEs and internal parameters are reset.

4 Experimental Results

In order to evaluate the proposed scheme, the *auto.basicmath* application from the *MiBench* benchmark suite was used [11]. A trace of the application's activity profile was generated using *SimpleScalar* [12], while *3D-ICE* [3] was used to obtain temperature information of each PE in the stack based on the floorplan illustrated in Figure 4. These were provided as inputs to a *SystemC*-based simulation of the proposed PMB. For simulations, a *2%* window of convergence was used to maximize power budget utilization and reduce fluctuations in V-F levels. The temperature margin is dependent on temperature sensor accuracy, and was determined experimentally as *2K*. A power budget of *160W* was imposed the check the ability of the PMB in converging to the set budget value, while a temperature constraint of *320K* was imposed on all PEs to examine the effectiveness of the proposed algorithm under harsh temperature conditions. Each PE in the 12-core MPSoC was considered to execute the same task, each however with a different start time. Task migration to compensate for performance losses was not considered in the setup in order to expose the actual effect of the scheme on performance. Per-core DVFS was implemented based on the floorplan shown in Figure 4. Each PE was operated at one of six V-F levels: *0.8V/700MHz, 0.855V/800MHz, 0.907V/900MHz, 0.956V/1000MHz, 1.003V/1100MHz and 1.048V/1200MHz.* A control period of *60,000 cycles*, corresponding to $50\mu s$ at maximum frequency was selected based on the time required for voltage transitions to complete. Since such transitions are of the order of tens of nanoseconds, the selected control period results in a negligible overhead during switching of V-F levels [13]. The value of A in equation 2 was determined experimentally as *0.0083055* for these six DVFS levels.

Two parallel simulation setups were used, each using the same convergence algorithm, similar conditions for V-F scaling and similar constraints on power and temperature. However, one of the setups implemented the proposed power management approach while the other used a conventional DVFS scheme for 2D chips where the temperature of each PE is considered independently. Figure 5 presents the sum of frequencies of all PEs in the stack obtained with both approaches. The new approach is observed to increase the aggregate frequency attainable under the set temperature and power constraints. As a consequence, tasks complete in a shorter time when compared to the 2D approach. Figure 6 indicates that even with improved performance, the new approach reduces total power

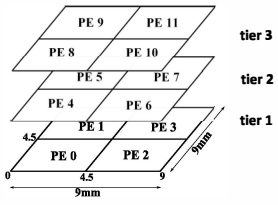

Figure 4: 3D stack with 3 tiers and total of 12 PEs

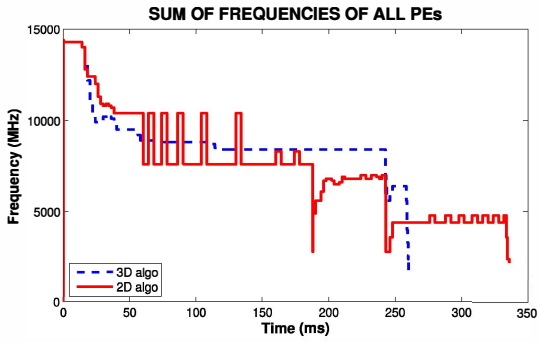

Figure 5: Sum of frequencies of all PEs illustrating an overall increase in aggregate frequency, and reduction in execution time.

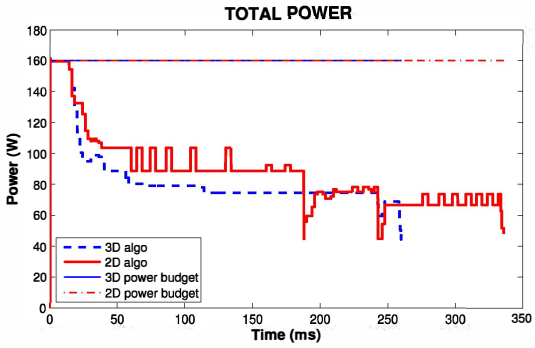

Figure 6: Total Chip Power Dissipation

dissipation of the MPSoC. Further more, the power levels are observed to be more stable than those from the 2D approach. Table 1 summarizes the performance losses over the ideal case for both approaches. Note that the reported losses do not include overheads for turning PEs *off* and *on* and for switching V-F levels. However, these losses are negligible when compared to the performance improvement obtained from V-F scaling. The observed difference in simulation time is explained by the turning off of tier 1 PEs in the 2D approach until their temperature returned to sub-critical levels. Since the V-F levels in the 2D approach are controlled independently for each PE, this was possible only when PEs on the upper tiers had completed their tasks and switched to the off state. The total simulation time does not include the time for which PEs remain idle after completing their tasks.

Figure 7 illustrates the operating V-F levels for *PE0* on the lowest tier of the stack. It may be observed that while the 2D approach allowed PEs on upper tiers to operate at higher V-F levels, PE0 was constantly switched between the on and off states. The proposed approach however achieves a balance between the performance losses across all three tiers while maintaining PEs in lower tiers at stable operating V-F levels with fewer transitions than the 2D approach. While techniques such as task migration may alleviate the performance losses due to critical PEs in the off state, frequent migration of tasks to cooler PEs would result in uneven aging across the stack. Consequently, cooler PEs may fail earlier than those that are turned off more often.

The proposed approach was also applied to vertical voltage

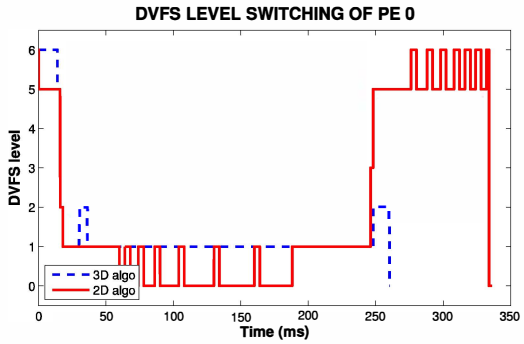

Figure 7: Operating V-F levels of PE 0. Note stable operating V-F profile for the new approach and constant switching with 2D approach.

islands as shown in Figure 8. The weight of an island was taken to be the average of its constituent PEs. Table 2 provides a comparison between per-core and voltage island based approaches. The granularity and depth of islands can essentially be altered in a deep stack to achieve benefits of islands as well as per-core approach. Implementations of such a scheme would also need to consider thermal relationships between islands in order to control temperatures effectively. As a result, islands higher up in the stack could achieve better performance, while considering their thermal relationships would allow for V-F levels to be effectively scaled down should thermal conditions on lower dice require it.

	2D (x)	new (y)	(x-y)
Total simulation time	336.05ms	260.35ms	65.7ms (19.55% of x)
Avg. OFF-time on tier 1	106.5ms	0ms	106.5ms
Avg. performance loss on tier 1 (including time in OFF state)	78.38%	38.48%	39.9%
Avg. performance loss on tier 2 (including time in OFF state)	29.28%	37.80%	-8.52%
Avg. performance loss on tier 3 (including time in OFF state)	0%	29.34%	-29.34%

Table 1: Performance losses for conventional 2D DVFS and new approach

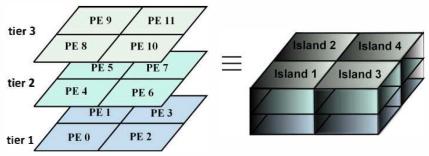

Figure 8: Voltage island partitioning for the 3D stack

Conclusions

A new temperature constrained power management scheme for 3D-MPSoC is proposed which takes into account not only the activity factor of PEs, but also their positional details, available instantaneous temperature margin and area. An effective thermal resistance matrix is generated at intervals determined by the temperature check period in order to maximize utilization of available instantaneous temperature margin. In a 3D stack with hundreds of PEs, voltage islands may become essential and more practical an approach due to the overhead of level shifters and voltage converters required to implement DVFS schemes. In such a scenario, the thermal relationships between islands becomes an important factor in effectively monitoring and managing temperature. The proposed approach takes these relationships into account while scaling closely related islands for maintaining temperature within a safe margin below the critical value. The approach showed an improvement of up to 19.55% in total execution time by considering these interdependencies for scaling V-F levels and preventing PEs deeper in the stack from being turned off. Since the thermal model and effective resistance matrix for the stack are derived based on the target floorplan, the proposed scheme is applicable to other stacked architectures as well.

Per-Core	Voltage Islands
Scaled as and when necessary	PEs in an island are bound to operate on same V level
Higher performance on PEs closer to heat sink	Similar performance throughout the island
Performance losses may differ according to temperature	Similar performance losses on an island
Larger overhead of level shifters and voltage converters	Depends on the granularity of voltage islands

Table 2: Per-Core DVFS versus Voltage Island based DVFS

References

[1] Joohee, K *et al.*, 'High-frequency scalable electrical model and analysis of a through silicon via (TSV),' *IEEE Transactions on Components, Packaging and Manufacturing Technology,* Volume. 1, No. 2 (2011), pp. 181-195.

[2] Keating, M. *et al.*, "Low Power Methodology Manual: For System-on-Chip Design," Springer Publishing (2007).

[3] Sridhar, A. *et al.*, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on , vol., no., pp.463-470, 7-11 Nov. 2010

[4] Jain, A. *et al.*, "Thermal modeling and design of 3D integrated circuits," Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on , vol., no., pp.1139-1145, 28-31 May 2008

[5] Chong Sun *et al.*, "Three-dimensional multiprocessor system-on-chip thermal optimization," Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2007 5th IEEE/ACM/IFIP International Conference on , vol., no., pp.117-122, Sept. 30 2007-Oct. 3 2007

[6] Herbert, S.; Marculescu, D.; , "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," Low Power Electronics and Design (ISLPED), 2007 ACM/IEEE International Symposium on , vol., no., pp.38-43, 27-29 Aug. 2007

[7] Xiaorui Wang; *et al.*, "Adaptive Power Control with Online Model Estimation for Chip Multiprocessors," Parallel and Distributed Systems, IEEE Transactions on , vol.22, no.10, pp.1681-1696, Oct. 2011

[8] Changyun Zhu *et al.*, "Three-Dimensional Chip-Multiprocessor Run-Time Thermal Management," Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on , vol.27, no.8, pp.1479-1492, Aug. 2008

[9] Sabry, M.M. *et al.*, "Thermal analysis and active cooling management for 3D MPSoCs," Circuits and Systems (ISCAS), 2011 IEEE International Symposium on , vol., no., pp.2237-2240, 15-18 May 2011

[10] Ayala, J.L. *et al.*, "'Invited paper: Thermal modeling and analysis of 3D multi-processor chips," Integr. VLSI J. 43, 4 September 2010, 327-341.

[11] Guthaus, M.R. *et al.*, "MiBench: A free, commercially representative embedded benchmark suite," Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on , vol., no., pp. 3- 14, 2 Dec. 2001

[12] Austin, T. *et al.*, "SimpleScalar: an infrastructure for computer system modeling," Computer , vol.35, no.2, pp.59-67, Feb 2002

[13] Wonyoung Kim *et al.*; , "System level analysis of fast, per-core DVFS using on-chip switching regulators," High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on , vol., no., pp.123-134, 16-20 Feb. 2008