View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Archivio istituzionale della ricerca - Politecnico di Milano

2012 IEEE Computer Society Annual Symposium on VLSI

A Temperature and Reliability Oriented Simulation Framework for Multi-Core Architectures

Simone Corbetta, Davide Zoni and William Fornaciari
Politecnico di Milano – Dipartimento di Elettronica e Informazione
Via Ponzio 34/5, 20133 Milano, Italy
Email: {scorbetta,zoni,fornacia}@elet.polimi.it

***Abstract*—The increasing complexity of multi-core architectures demands for a comprehensive evaluation of different solutions and alternatives at every stage of the design process, considering different aspects at the same time. Simulation frameworks are attractive tools to fulfil this requirement, due to their flexibility. Nevertheless, state-of-the-art simulation frameworks lack a joint analysis of power, performance, temperature profile and reliability projection at system-level, focusing only on a specific aspect. This paper presents a comprehensive estimation framework that jointly exploits these design metrics at system-level, considering processing cores, interconnect design and storage elements. We describe the framework in details, and provide a set of experiments that highlight its capability and flexibility, focusing on temperature and reliability analysis of multi-core architectures supported by Network-on-Chip interconnect.**

***Keywords*-Simulation; Multi-core; Reliability; Thermal**

I. Introduction

Continuous technology scaling of recent decade leads to an exponential increase in processor performance and power consumption, going as faster as clock rate growth. The transition to multi-core architectures introduced an opportunity for performance to grow faster than power consumption. Shrinking silicon technology and increasing device density are indeed driving the integration capabilities of modern Multi-Processors System-on-Chip (MPSoC) devices. The need for even more performance and integration of cores in a single chip poses new communication challenges, with Network-on-Chip emerging as the appropriate design paradigm to manage increasing performance and reliability requirements [2]. However, on-chip networks are expected to consume significant part of the total chip power, contributing to increasing operating temperature. Also, integration density is limited by the reliability of the circuit, and experimental results have shown that high temperature is responsible for more than 50% of failures in CMOS integrated circuits [3]. The need to consider such a huge amount of architectural design aspects demands for appropriate methodologies for fast and accurate analysis. Simulation represents the most accurate method to extract valuable information from the architecture, while it is very time consuming. On the other side, analytical models usually allow to reduce evaluation time even if the output data is affected by significant errors. Analytical models are usually employed for the analysis of specific parts of the architecture, since they are difficult to extract and their characterization requires low level information from real architecture. In this scenario, this work proposes a new simulation framework for thermal, reliability, performance and power analysis to be used at early design stages, while the extracted information can be used for further localized platform optimizations.

A. Related works

Several proposals can be found in literature for power, performance and thermal estimation of single-core and multi-core processors. Nevertheless, only a few focused on a comprehensive approach to jointly estimate multiple design dimensions. Our work is meant to provide an accurate and holistic design tool that focuses on all the different design aspects for high-performance multi-core architectures based on Network-on-Chip interconnect.

One of the first and accurate proposals to deal with temperature estimation is based on the coupling of power and thermal models. `Wattch` [4] allows for microarchitecture power exploration through a cycle-accurate model of a single-thread single-core processor based on the Alpha ISA, based on `SimpleScalar` to quantify performance metrics. `Wattch` power model is based on access statistics to each microarchitecture block composing the processor, and the extracted experimental data considering an Alpha EV6 architecture are used to annotate power measurements. Power statistics are then used to feed the `HotSpot` thermal model [5] and compute a detailed thermal map of the chip. `HotSpot` represents the *de facto* standard for microarchitecture and architecture research in thermal-aware designs, due to its simplicity, flexibility and accuracy.

The advent of multi-core architectures demands for multi-core simulators that can estimate power, performance and temperature at system-level. For this reason, different frameworks have been proposed in literature, but they generally lack in a suitable methodology to deal with all the design aspects of interest in the multi-core and many-core era. `SESC` [6] is based on the MIPS architecture, providing cycle-accurate simulation of multi-core processors; however it does not support Network-on-Chip architectures, while their relevance is increasing for giga-scale computing platforms [7]. The `Polaris` framework [8] allows to estimate power and area of Network-on-Chip architectures, focusing mainly on the interconnect without providing detailed power consumption for processors and memory hierarchy. Although precise, it is not suitable to be employed in full SoC computer architecture research. Power, area and thermal modeling are accounted for also in the `SST` framework [9]. Their work focuses on large-scale systems, but application traces are emulated, rather than collected from cycle-accurate simulation, with higher simulation rate at the cost of much lower accuracy. Power and thermal models, as well as measurements from real hardware are used in the proposal of [10] based on the `Simics` functional simulator. The interesting achievement in this approach lies on the possibility to develop, analyze and tune different control algorithms for thermal and power management, based on high-level Matlab descriptions. The work

978-0-7695-4767-1/12 $26.00 © 2012 IEEE
DOI 10.1109/ISVLSI.2012.22

IEEE computer society

Table I
STATE-OF-THE-ART MULTI-CORE SIMULATION FRAMEWORKS: FEATURES, ADVANTAGES AND DRAWBACKS COMPARED TO THE PROPOSED FLOW.

Framework	Cycle-accurate simulation	NoC support	Power support	Thermal support	Reliability projection	Floorplan exploration	Objectives
Renau et al. (SESC) [6]	✓	✗	✗	✗	✗	✗	multi-core simulation, parallel applications
Soteriou et al. (Polaris) [8]	✗	✓	✓	✗	✗	✗	Network-on-Chip design-space exploration
Hsieh et al. (SST) [9]	✗	✓	✓	✓	✗	✗	microarchitecture, power and thermal
Lis et al. (HORNET) [11]	✗	✓	✓	✓	✗	✗	many-core processors, mainly NoC interconnect
Bartolini et al. [10]	✗	✓	✓	✓	✓	✗	run-time control policies evaluation
Our flow	✓	✓	✓	✓	✓	✓	microarchitecture, NoC, reliability, design-space exploration

proposes itself as a suitable framework for thermal management solutions employing control-theory methodology, although bound to a particular architecture, ISA and floorplan (an Intel©Xeon X7350 system). The work presented in [11] is meant to simulate large-scale architectures, and has the real advantage of exploiting parallel simulation on physical hardware. It can simulate several cores based on the MIPS in-order architecture model. However, the output thermal map refers to the only communication infrastructure, without providing a system-wide perspective (i.e., communication, computation and storage) from a thermal viewpoint. In addition, each of the above approaches lacks in reliability projections, considering temperature-induced failure mechanisms (e.g., stress-migration) as well as NBTI-induced degradation. In this perspective, Table I summarizes the advantages and drawbacks of these works, and reports a comparison against the framework presented in this paper.

B. Novel contributions

The main contribution of this paper is focused on the proposal of a comprehensive simulation framework that allows to explore different design space dimensions during early design stages, focusing on a joint analysis of performance, power, temperature profile and reliability. Among the others, there are three main features exploited by the proposed framework.

1) Thermal analysis: traditional approaches solely focus on performance, but stringent reliability requirements nowadays supersede the performance dimension. Thermal issues are of paramount importance for both performance and architecture lifetime maximization; several works previously presented in literature focus on either processing cores or interconnect design; the proposed simulation framework allows for accurate thermal chip evaluation that jointly accounts for both of them at different levels of detail;

2) Floorplan exploration: floorplan design represents a critical aspect in reliable and high-performance products, since it directly influences both performance and thermal profile. A flexible floorplan estimation tool should consider at least two different aspects: at first the possibility to evaluate different floorplan configurations on a per-core granularity, focusing on the coupling between microarchitecture blocks, and secondly the possibility to evaluate the mutual influence between adjacent cores in a multi-core architecture, with Network-on-Chip interconnect;

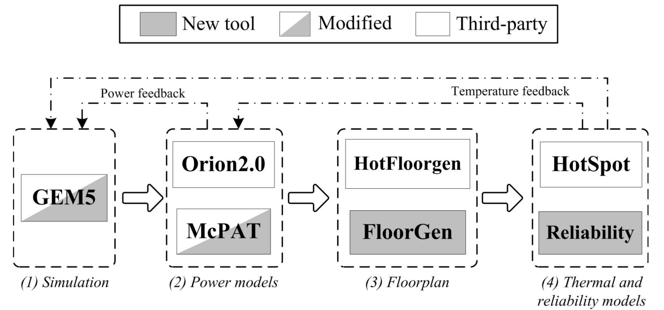


Figure 1. Logical representation of the proposed estimation flow.

3) Reliability: with continuous integration at aggressive technology scales, reliability is becoming a major concern in VLSI designs; early-stage reliability projection at system-level can be of great help to architecture designers to cope with strict requirements. The proposed simulation framework captures two different reliability aspects: Mean-Time To Failure (MTTF) expressions are employed for temperature-induced reliability analysis, and Negative-Bias Temperature Instability (NBTI) degradation is considered for hardware design degradation estimate.

Section II describes and provides an in-depth discussion about the proposed simulation framework, highlighting its capabilities and flexibility. Experimental results are then shown and analyzed in Section III, detailing different use-case scenarios, and a possible use case for each contribution item presented in Section I-B. Conclusions are then given in Section IV.

II. PROPOSED ESTIMATION FLOW

The proposed estimation flow is composed of tools that are widely used in the computer architecture research community. We provided a set of modifications to the available tools, while a complete set of other tools has been developed from scratch to provide a complete and flexible framework for thermal, performance and reliability analysis.

The logical snapshot of the proposed virtual platform is given in Fig. 1. Steps are executed in a pipeline fashion, i.e. each step provides input to the subsequent one, and requires output from the previous. Four steps are involved: cycle-accurate simulation, power consumption estimation, floorplan generation and

thermal/reliability models. Cycle-accurate simulation is required to provide access and usage statistics to relevant architecture and microarchitecture blocks. During simulation, the most relevant information from the modeled architecture are acquired, e.g. accesses to instruction fetch unit, number of committed integer instructions, number of stall cycles in the pipeline and the like. For processors, this means that we can collect metrics about the status of the hardware pipeline while executing the software application; for NoC routers, on the other hand, we collect raw metrics about network interface accesses and traffic patterns. Memory-related statistics are also collected providing a system-wide set of memory performance and statistics. Accurate power consumption estimates are computed for both processing cores and interconnect primitives, as well as for storage blocks. Last, temperature profile is provided to evaluate the impact of hardware or software design choices on the reliability and power consumption of the entire architecture.

There are also three different feedback paths: temperature feedback exists from the thermal model to the power model for leakage power analysis; an additional temperature feedback is used to back-annotate the simulator, useful for temperature-aware scheduling policies evaluation; last, power back-annotation is used to allow for power management policies to be developed and estimated along the flow. These feedback paths can be employed in a more general fashion to provide an analysis of power-related and temperature/reliability-related design choices, either at design-time or run-time. In Fig. 1, white boxes are those related to third-party software, gray boxes represent tools developed from scratch, while striped boxes are those tools for which relevant modifications have been performed. The rest of this section gives an insight of each step, the tools involved in each stage and the modifications that have been applied to third-party tools.

A. Cycle-accurate simulation

We employ `GEM5` cycle-accurate performance simulator in *syscall emulation* mode to mimic bare-metal execution, and simulate the underlying hardware with precise in-order processor models. We then provide two main contributions: support for clock-toggling and reliability prediction. Current and future multi-core architectures have to face with thermal and reliability issues, as well as performance and power ones. Most of the run-time thermal management and power management techniques rely on dynamic voltage and frequency scaling to control both chip temperature and power consumption, or on clock-gating hardware support to cool down processors. We provide a per-CPU clock-toggling configurable implementation, in which the user can synthesize the desired duty-cycle for thermal/performance tradeoff exploration, either at design-time or run-time. Furthermore, we have added a reliability library to the performance simulator, to dynamically measure the NBTI stress at architectural block level as instructions are executed, such that to provide thermal and reliability directed dynamic instruction scheduling strategies. The NBTI library is based on an off-line characterization of a MIPS-like RTL processor design, as previously done in [20].

B. Power estimation

Power estimation is accounted for using different tools for processor and Network-on-Chip router, as a function of the access statistics provided by the cycle-accurate simulation phase. `McPAT` [12] is used to generate power estimates of core and memory architectures, while `Orion2.0` [13] is employed for computing power contribution of NoC routers and links. Two significant improvements have been made to these tools, mainly related to leakage power estimation and temperature.

The original version of `McPAT` takes as input a single temperature value to compute leakage contribution: the entire chip region is assumed to work at the same operating temperature. This assumption does not fit well with architecture modeling, for two reasons. At first, it is impractical that different regions of the chip experience the same amount of temperature [5], due to the asymmetric load assignment, especially in multi-core architectures. In addition, the chip temperature profile is an aspect of paramount importance for reliable design [14], while the assumption provides an overlay simplistic scenario. The proposed flow, on the other hand, is able to annotate the correct temperature to each microarchitecture block in the processor, thus providing a way to better estimate leakage power contribution. `McPAT` provides a discretized amount of leakage levels, ranging from 300K to 400K temperatures at steps of 10K. The available temperature range is thus reduced to 11 values, providing an impractical scenario for aggressive thermal simulations. In our flow, on the other hand, the temperature range is fully covered, and leakage curve within temperature steps at distance 10K are approximated linearly. This aspect, along with the feedback from thermal model (refer to Section II-C) provides a comprehensive and more accurate estimation.

C. Temperature estimation

Temperature estimation is an essential phase in the computer architecture research, due to the increasing relevance of temperature-aware designs. `HotSpot` [5] is the most widely used temperature model in computer architecture research, thanks to its simplicity and the accuracy of its estimation. This model requires a chip floorplan, and a set of power measurements to compute steady-state or transient temperature analysis solving RC equivalent circuits [5]. The main improvement we propose in this work is related to a flexible and customizable floorplan generation tool coupled to `HotFloorplan`, part of the `HotSpot` model release. Since `HotFloorplan` does provide single-core floorplan only, we developed `FloorGen` to generate the floorplan for the desired multi-core architecture. We focused on two main aspects: to provide flexibility to generate any desired floorplan, and to provide the flexibility to generate the floorplan at any desired level of detail. Up to now, we target only 2D mesh topologies, based on the Alpha-21364 network architecture [15], but we are able to generate the floorplan of each core according to user-defined requirements: the user can thus specify core floorplan, and let the tool generate a multi-core architecture with core replication. An interesting support in this direction has been made to integrate the output from `HotFloorplan` with `FloorGen`. Notice that this step is entirely decoupled from the cycle-accurate simulation, since core access statistics are not affected by the floorplan; indeed the wires are assumed to be dense in their respective microarchitecture block. However, the floorplan has direct impact on the power consumed by router links. For this reason, there is a cooperation between this phase and the one presented in Section II-B.

Table II
ARCHITECTURE AND TECHNOLOGY PARAMETERS FOR PROCESSOR AND ROUTER.

Processor core	3GHz, in-order Alpha-21264 core
Int-ALU	4 integer ALU functional units
Int-Mult/Div	4 integer multiply/divide functional units
FP-Mult/Div	4 floating-point multiply/divide functional units
L1 cache	64kB 2-way set assoc. split I/D, 2 cycles latency
L2 cache	1.75MB per bank, 8-way associative
Router	3-stage wormhole switched (Garnet network [17])
Topology	2D-mesh, based on Alpha21364 network processor
Technology	45nm at 1.1V

D. Reliability analysis

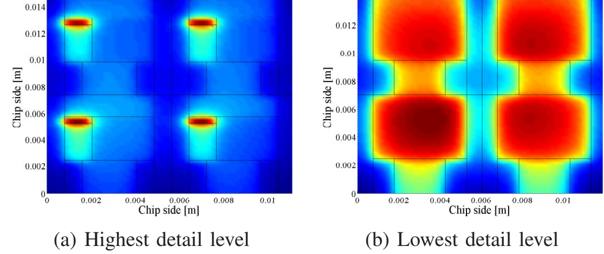
The last step from Fig. 1 is used to compute reliability projection, under two different flavors. Temperature-dependent reliability estimate is done through MTTF estimation of different mechanisms: electromigration, stress-migration and thermal cycling. MTTF for these processes is known to be exponentially dependent on temperature, and our library provides an easy way to perform reliability-directed design optimizations with direct input from the simulated architecture. On a second instance, NBTI-induced degradation is computed along the line during cycle-accurate simulation, and statistical quantities are summarized at the end of the estimation flow. The sum of the two contributions makes the proposed framework suitable for aggressive reliability projections and hardware/software estimation.

III. EXPERIMENTAL RESULTS

Purpose of this section is to highlight the flexibility of the proposed framework to estimate different aspects that are of utmost interest to the authors. Three main results are discussed: Section III-A shows how the proposed flow can be used to easily analyze the impact of multi-core floorplan on the temperature profile of the chip, as well as the impact of floorplan details on the estimation process; Section III-B presents a simple exploration of the thermal/performance trade-off in 16-cores architectures, providing analysis of the impact of clock toggling on temperature profile; last, Section III-C presents NBTI-directed degradation as a function of the running application. In the experiments, we consider as reference architecture an Alpha-21364 network processor [15]. This is composed of tiles, organized along a 2D-mesh topology: each tile is composed of an in-order version of the Alpha-21264 processor core, with private L1 cache and shared distributed L2 cache; the router is used to interface the processing core to the local and shared memory. The router is a 3-stage state-of-the-art architecture inspired from [16], and local L2 cache bank surrounds both processing core and router as in [15]. The main features and technology parameters are summarized in Table II.

A. Thermal-aware floorplan design exploration

The floorplan organization represents a design dimension of paramount importance when dealing with thermal and reliability issues. For instance, the topological organization of the unit blocks in a core processor greatly impacts thermal hotspot generation in the chip. Moreover the floorplan design needs to consider different detail levels ranging from the functional unit organization in a single core to a complex multi-core floorplan accounting for routers, communication links and computational logics.



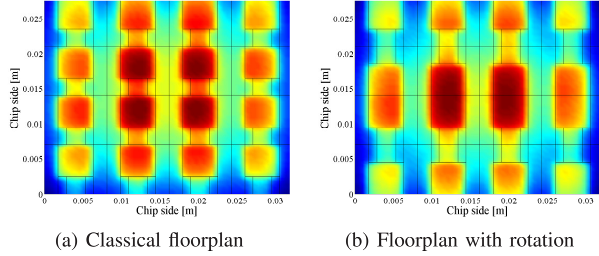
(a) Highest detail level (b) Lowest detail level

Figure 2. Different details level for the same floorplan give different temperature estimation, due to the approximation given by the thermal model.

Table III
ESTIMATION MISMATCH BETWEEN DETAILED AND RAW FLOORPLAN.

Microarchitecture block	**Temperature** [K]		**Absolute error** [K]
	Detailed	*Raw*	
Instruction Fetch Unit	328.44	327.26	1.18
MMU	335.58	327.33	8.25
Execution Unit	326.57	327.87	1.30
Load/Store Unit	327.12	327.51	0.39
L2	325.14	325.37	0.23
router	326.76	326.99	0.23

Design-time thermal evaluation relies upon the flexibility of the estimation procedure itself, while different levels of details can raise up different observations about the quality of the solution. For this reason, our framework is able to capture the power and thermal profile of the processor at two different granularity levels, according to the needs of the designer: the higher the details the higher the number of temperature samples to be controlled, while lower details tend to average adjacent temperature readings. Fig. 2 shows the thermal map of a quad-core architecture in two different flavors: a detailed floorplan and a simple, raw floorplan. The detailed floorplan has been found by integrating the framework with `HotFloorplan` as previously described in Section II. The map shows normalized temperature values through colors: from red (highest temperatures) to blue (lowest temperatures). In the first case, the processing core is composed of several microarchitecture blocks, and temperature measurements are given separately for each block in the processor; in the latter case, on the other hand, the core is seen as a black-box and the temperature is computed by considering the average power consumption of the blocks in the entire box area. As it can be seen, the level of detail in the floorplan gives slightly different thermal profiles. Table III compares the temperature readings from each microarchitecture block in the two scenarios; since the thermal model is based on a grid view of the chip area, we mapped each microarchitecture block to the grid cells they belong to, and compare them. Data in Table III is given for the only leftmost top core in the architecture for lack of space, but similar conclusions can be obtained for each core in the quad-core architecture. As it is clear from the quantitative comparison, the detailed floorplan gives much larger insight on what effectively happens in each block; in addition, although an analysis on the temperature as done for the simple raw floorplan is enough for average-cases analysis, it cannot be employed for aggressive thermal management solutions.



(a) Classical floorplan (b) Floorplan with rotation

Figure 3. Impact of tiles rotation on the processor temperature profile.

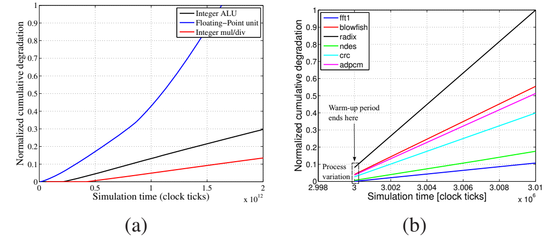
The success of multi-core architectures is driven also by their ability to tune the temperature profile of the chip as a function of the task-to-core allocation strategy in use. Furthermore, multi-core architectures move the focus from a per-core floorplan optimization to a system-wide view. Our framework can help computer architects to evaluate floorplan design choices early at the design stage, and compare different solutions beforehand. Starting from a 16-cores 2D-mesh chip, Fig. 3 shows how a rotation of the cores in the central region can have severe effects on the temperature profile. As done before, temperature map is given according to a color scale, reporting normalized temperature values. In particular, Fig. 3(a) shows the thermal map of the original floorplan, based on the Alpha-21364 network processor, while Fig. 3(b) shows the thermal map with the modified floorplan: the tiles belonging to the second and third row have been mirrored along the horizontal axis. The thermal map shows a difference in the temperature distribution, with a higher coupling at the centre of the chip between cores.

B. Thermal/performance trade-off

The complexity of multi-core architectures represents a great issue to deal with in order to provide accurate trade-off optimizations, i.e. thermal/performance or power/performance. Chip temperature represents a constraint of paramount importance in current multi-core design, while at the same time a constrained chip temperature can degrade performance. Finding a suitable thermal/performance trade-off is relevant topic, since it allows to fine-tune the performance on a per-core basis in order to provide optimal system-wide temperature profile. This section describes the flexibility of the proposed framework to deal with design-time thermal/performance trade-off, exploiting the per-core duty-cycle control knob that the framework provides. One key observation is that thermal hotspots are generally localized at the centre of a 2D-mesh architecture, independently of the size of the mesh. This fact is due to thermal coupling: cores surrounded by other cores are influenced by the core-to-core heat exchange. Moreover, temperature gradually decreases toward the edges: the higher the number of the cores, the higher the maximum operating temperature. To deal with this space-related property, we propose an insight of a suitable methodology to deal with temperature profile optimization under performance constraints. The proposed methodology constructs a ring-based view of the chip, in which rings are concentric each other, and in each ring a duty-cycle level is synthesized through the appropriate clock-toggling specification. Each core belongs to one and only one ring, and cores belonging to the same ring share similar temperature dissipation properties. We developed several

Table IV
SIMULATED TEMPERATURE, AGAINST CLOCK-TOGGLING LEVEL (ONE MINUS DUTY-CYCLE), AND RELATIVE SLOPE.

Observed temperature [K]	**Clock-toggling**	**Ratio**
332.06	0.64	–
333.11	0.58	17.5
333.92	0.53	16.9
334.96	0.47	17.06
337.13	0.35	17.48
337.97	0.30	17.38
339.01	0.24	17.38
340.07	0.18	17.41
340.86	0.13	17.26
341.92	0.07	17.3



(a) (b)

Figure 4. Normalized degradation for a floating-point intensive benchmark.

experiments, by setting an appropriate duty-cycle level to each core belonging to the same ring, and checked the output temperature. We experienced a strong linear relation between the two, meaning that the duty-cycle is a suitable control knob to control heat dissipation and chip-level operating temperature, although the results are guaranteed only for in-order processor cores in 2D-mesh architectures. Even though the applications running on each core are slightly different, they have comparable power consumption profiles, such that we can affirm that the effects of different workload are negligible to the thermal coupling effects. Table IV reports the simulated temperature values as a function of the applied clock-toggling synthesizing the desired duty-cycle, and the relative ratio: assuming all values are taken relative to the lowest (332.06K), the ratio is computed as the fraction of the observed temperature value to the applied clock-toggling level. *Ratio* shows that a strong linear relation between the observed temperature and the applied control knob exists, under a valid broad range of temperatures, and with very little variance.

C. NBTI degradation estimation

Negative Bias Temperature Instability (NBTI) is one of the most pressing failure mechanisms in scaled technologies. The stress/recovery phases typical of the NBTI process [18] makes it suitable to control the degradation through appropriate duty-cycle usage optimization, at microarchitecture and architecture levels [19]. Thus, it is of paramount importance being able to provide designers with early estimates of NBTI degradation. The framework presented in this paper is able to capture the impact of functional unit usage and instruction allocation on the unit degradation.

The impact of the application class on the degradation of the circuit can be seen in Fig. 4. The degradation is shown normalized with respect to a specified time frame, and considering three

different functional units: integer ALU, integer multiply/divide and floating-point unit. The results are shown for a floating-point intensive application executed as bare-metal on the reference processor. As evident from Fig. 4, a floating-point intensive benchmark will have greater degradation in the floating-point unit, since the duty-cycle usage in that unit is higher than in the other ones.

Fig. 4(b) shows the normalized threshold voltage degradation due to NBTI of the integer multiply/divide unit, considering six different applications taken from `WCET`, `MiBench` and `SPLASH-2` benchmark suites. As done before, the degradation is shown along a predefined time interval, after a warm-up period of 3×10^6 instructions executed. The different starting point at 3×10^6 is due to process variation, i.e. our framework is able to capture process variability along the microarchitecture design of the reference processor. Different applications provide different access patterns to the functional units according to the instruction flow, and this translates to different degradation induced by the NBTI mechanism, as it is clearly shown in the trend curves.

IV. Conclusions

This paper proposes a novel simulation framework for thermal and reliability analysis in modern multi-core architectures, with full support for Network-on-Chip interconnect. Differently from state-of-the-art simulators, the proposed framework allows for accurate and detailed joint analysis of power, performance, thermal and reliability metrics that are relevant for current computer architecture and microarchitecture research. The simulation flow is composed of state-of-the-art tools, as well as tools developed from scratch, to provide the designers with a comprehensive design and estimation framework. We proposed and described a set of experiments showing the flexibility of the framework to support any kind of estimation objectives and methodologies, ranging from temperature-aware floorplan design to temperature/performance trade-off estimation, up to reliability projections.

Acknowledgments

This research work is supported by European Community Seventh Framework Programme (FP7/2007-2013), under agreements no. 248716 (`2PARMA` project www.2parma.eu).

References

[1] S. Borkar, "Thousand core chips: a technology perspective," in *Annual ACM IEEE Design Automation Conference*, 2007.

[2] A. Banerjee, R. Mullins, and S. Moore, "A Power and Energy Exploration of Network-on-Chip Architectures," in *NOCS '07*. IEEE Computer Society, 2007, pp. 163–172.

[3] C. J. M. Lasance, "Thermally driven reliability issues in microelectronic systems: status-quo and challenges," *Microelectronics Reliability*, vol. 43, no. 12, pp. 1969–1974, 2003.

[4] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: a framework for architectural-level power analysis and optimizations," in *Proceedings of the 27th annual international symposium on Computer architecture*, pp. 83–94.

[5] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture: Modeling and implementation," *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 1, no. 1, 2004.

[6] J. Renau, B. Fraguela, J. Tuck, W. Liu, M. Prvulovic, L. Ceze, S. Sarangi, P. Sack, K. Strauss, and P. Montesinos, "SESC simulator," January 2005, http://sesc.sourceforge.net.

[7] L. Benini and G. De Micheli, "Networks on chips: a new soc paradigm," *Computer*, vol. 35, no. 1, pp. 70 –78, 2002.

[8] V. Soteriou, N. Eisley, H. Wang, B. Li, and L.-S. Peh, "Polaris: A system-level roadmap for on-chip interconnection networks," in *ICCD 2006.*, pp. 134 –141.

[9] M.-y. Hsieh, A. Rodrigues, R. Riesen, K. Thompson, and W. Song, "A framework for architecture-level power, area, and thermal simulation and its application to network-on-chip design exploration," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, pp. 63–68.

[10] A. Bartolini, M. Cacciari, A. Tilli, L. Benini, and M. Gries, "A virtual platform environment for exploring power, thermal and reliability management control strategies in high-performance multicores," in *GLSVLSI'10*, pp. 311–316.

[11] M. Lis, P. Ren, M. H. Cho, K. S. Shim, C. Fletcher, O. Khan, and S. Devadas, "Scalable, accurate multicore simulation in the 1000-core era," in *Performance Analysis of Systems and Software, IEEE International Symposium on*, 2011, pp. 175 –185.

[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 *Microarchitecture IEEE/ACM International Symposium on*, 2009, pp. 469 –480.

[13] A. Kahng, B. Li, L.-S. Peh, and K. Samadi, "Orion 2.0: A fast and accurate noc power and area model for early-stage design space exploration," in *DATE '09.*, 2009, pp. 423 –428.

[14] A. Ajami, K. Banerjee, and M. Pedram, "Modeling and analysis of nonuniform substrate temperature effects on global ulsi interconnects," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, no. 6, pp. 849–861, 2005.

[15] S. S. Mukherjee, P. Bannon, S. Lang, A. Spink, and D. Webb, "The alpha 21364 network architecture," in *Proceedings of the The Ninth Symposium on High Performance Interconnects*.

[16] L.-S. Peh and W. Dally, "A delay model and speculative architecture for pipelined routers," in *High-Performance Computer Architecture, 2001. The Seventh International Symposium on*, pp. 255–266.

[17] N. Agarwal, T. Krishna, L.-S. Peh, and N. Jha, "Garnet: A detailed on-chip network model inside a full-system simulator," in *Performance Analysis of Systems and Software. IEEE International Symposium on*, 2009, pp. 33–42.

[18] M. Alam and S. Mahapatra, "A comprehensive model of pmos nbti degradation," *Microelectronics Reliability*, vol. 45, no. 1, pp. 71–81, 2005.

[19] L. Li, Y. Zhang, J. Yang, and J. Zhao, "Proactive nbti mitigation for busy functional units in out-of-order microprocessors," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2010*, pp. 411–416.

[20] Corbetta. S and Fornaciari. W, "NBTI Mitigation in Microprocessor Designs", in *GLSVLSI'12: Proceedings of the great lakes symposium on VLSI*, pp. 33–38.