DEVELOPMENT OF HIERARCHICAL MANAGEMENT OF DATA CENTER SERVERS' HARDWARE

Babkin O.V.[1], Varlamov A.A.[2], Gorshunov R.A.[3], Dos E.V.[4], Kropachev A.V.[5], Zuev D.O.[6]

[1]Babkin Oleg Vyacheslavovich - Strategy Consultant, IBM;
[2]Varlamov Aleksandr Aleksandrovich - CTO, SHARXDC LLC, MOSCOW;
[3]Gorshunov Roman Aleksandrovich - Solution Architect, AT&T, BRATISLAVA, SLOVAKIA;
[4]Dos Evgenii Vladimirovich - Lead DevOps Architect, EPAM, MINSK, REPUBLIC OF BELARUS;
[5]Kropachev Artemii Vasilyevich - Principal Architect, LI9 TECHNOLOGY SOLUTIONS, NORTH CAROLINA;
[6]Zuev Denis Olegovich - Independent Consultant, NEW JERSEY, USA

Abstract: *methods of data center performance estimation based on mathematical simulation of consumption system were analyzed. Multi-processor system on chip performance enhancement was proved to be optimal instrument of modeling could be. In order to optimize the model centralized control concept, inter-tier liquid cooling and proactive management scheme that rely on model predictive controller were discussed. It was demonstrated that modern thermal management techniques have to be studied. To develop the methodology operating power supply of the platform to near-threshold values, multiple supply voltages utilization optimization method for the voltage islands distribution and microarchitectural techniques to control the thermal hotspots were analyzed. Multi-processor system on chip performance enhancement was demonstrated as application for minimization of the global thermal impact, specifically temperature-aware floorplanning and simulated annealing utilization. While power consumption is generated by two sources it was decided that cost function was defined as a sum of the power input vector and required workload. Developed control system is based on interval steps, which starts at current time. The result of the optimization is proved to be an optimal sequence of control actions. To evaluate developed model were compared application of thermal management load balancing, look up table, fuzzy logic and proactive liquid cooling techniques. Unified thermal modeling methodology based on the finite difference method helps to make a proper analysis of the problem and to build proper applications up to the particular properties. The methodology uses paradigm of search optimal control criteria to find the optimal microchannel width. The main problem of optimization is to minimize the peak temperature and thermal gradients of the model, which allows to reduce the cooling system consumption.*
Keywords: *data center, power consumption, liquid cooling technique, load balancing, look up table, fuzzy logic, thermal modeling.*

1. Introduction

Requirements for data center servers' room power and temperature facilities have significantly grown for the last decades. Development of proper model of data center performance estimation could be done by mathematical simulation of servers' room power consumption system. 3D integrated circuits (IC) were proved to be optimal instruments of multi-processor system-on-chip (MPSoC) performance enhancement. Modern hardware resources and advanced 3D architecture stays a serious challenge in thermal dissipation and power management. Unified thermal modeling methodology based on the finite difference method for evaluation of 3D MPSoCs proposed in this work will help to make a proper analysis of the problem and to build proper applications up to the particular properties of the data center infrastructure. The integration of this methodology in bounds of the virtual platform will enable server chip's dynamic thermal evaluation procedure.

For development of the unified thermal modeling methodology meta-analysis of recent studies was done. There were analyzed key aspects of thermal management of liquid-cooled 3D MPSoCs [1-4]. In order to optimize the model centralized control concept [2], inter-tier liquid cooling [3] and proactive management scheme that rely on model predictive controller were discussed. Modern thermal management techniques were studied, particularly: load balancing liquid cooling policy [5], fuzzy logic thermal management mechanism [6] LUT-based flow rate control load balancing [7]. To develop the methodology operating power supply of the platform to near-threshold values [8], multiple supply voltages utilization optimization method for the voltage islands distribution in 3D MPSoCs [9] and microarchitectural techniques to control the thermal hotspots in 3D MPSoCs [10] were analyzed. MPSoC was demonstrated as application for minimization of the global thermal impact, specifically temperature-aware floorplanning [11], simulated annealing utilization [12] and temperature-

aware floorplanning genetic algorithms [13]. In the context of 3D MPSoCs floorplanning has been studied to analyze interlayer thermal dissipation [12, 14-17].

Meta-analysis shows possibility development of efficient thermal modeling methodology based on the finite difference method.

2. Proposed method

Thermal management procedures for 3D MPSoCs are usually meant to use a variable-flow liquid cooling with experimentally estimated sets of rules to control the temperature profile check performance requirements. In this case has to be developed a centralized control concept, which must be scalable for the controlled parameters increase scenario. It was proposed a cyber-physical approach 3D MPSoCs thermal management with inter-tier liquid cooling [93]. Control mechanism has to be developed with software-based thermal estimation and prediction which includes application of non-uniform liquid flow model. Non-uniform liquid flow and different microchannels requirements are used to conform to the specifications of all modules. Thereby control decisions has to be done on software-based thermal estimation and prediction platform and simulation non-uniform liquid flow in different microchannels meets all cooling demands. Thus, effective model has to demonstrate the overhead of software-based thermal estimation realization of non-uniform flow process in different channels. Proactive thermal management scheme relies on model predictive controller (MPC). It has to be developed a thermal management algorithm that controls task scheduling and the data center server room cooling infrastructure. It is main target is building of the cooling infrastructure of interlayer liquid cooled 3D MPSoC with dynamical change of the liquid flow rate. While at each time moment or interval system get a new set of tasks the management scheme should allocate schedule to various cores and change the flow rate up to the predicted peak temperature to reduce the 3D MPSoC power consumption for cooling and computation needs.

While power consumption is generated by two sources cost function J could be defined (Figure 1) as a sum of the power input vector $p(\tau)$ weighted by matrix R and required workload $u(\tau)$ weighted by matrix T:

$$J(p,u) = \sum_{\tau=0}^{\tau_H}(|R \cdot p(\tau)| + |T \cdot u(\tau)|), \qquad (1)$$

where τ is a time range limited by τ_H value of predictive policy horizon. Matrix R estimates maximum values of the tiers and the cooling system power consumption, while matrix T estimates optimization of required workload from the scheduler. To estimate structure of vector $p(\tau)$ formally it should be defined:

$$\begin{cases} p(\tau) = [l(\tau), m(\tau)] \\ \tau \in [0; \tau_H] \end{cases}, \qquad (2)$$

where $l(\tau)$ is the power input vector and $m(\tau)$ is liquid cooling management value. Thereby target value of the cost function has to be defined as min (J).

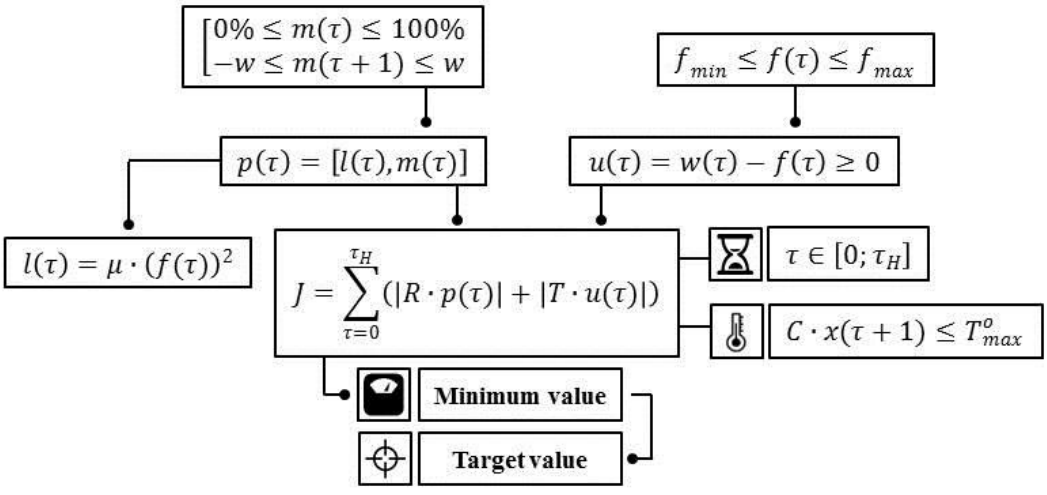


Fig. 1. Sources cost function estimation algorithm

To find a range of operating frequencies values of f_{min} and f_{max} have to be founded:

$$\begin{cases} f_{min} \le f(\tau) \le f_{max} \\ \tau \in [0; \tau_H] \end{cases}, \qquad (3)$$

which adds to the optimization problem solving a limitation on the number of allowed frequency values estimatiom.

Next stage of 3D MPSoC simulation includes equations that defines the evolution of the system and temperature limit T_{max}^{o}:

$$\begin{cases} \begin{bmatrix} x_{\tau+1} = A \cdot x(\tau) + B \cdot p(\tau) \\ C \cdot x(\tau+1) \le T_{max}^{o} \end{bmatrix}, \\ \tau \in [0; \tau_H] \end{cases} \quad (4)$$

where matrices A, B and C refer to 3D MPSoC system description and represent the the system using a coarse granularity of the thermal cells.

Next equations define required workload $u(\tau)$ as undone work at moment τ. While certain part of work will be always undone till the last moment this value should be equal or more than zero:

$$\begin{cases} u(\tau) = w(\tau) - f(\tau) \ge 0 \\ \tau \in [0; \tau_H] \end{cases}. \quad (5)$$

It should be noticed that operational frequency can also define power vector l:

$$\begin{cases} l(\tau) = \mu \cdot (f(\tau))^2 \\ \tau \in [0; \tau_H] \end{cases}, \quad (6)$$

where μ is a 3D MPSoC simulation technology-dependent constant.

To define limits of the liquid cooling management value $m(\tau)$ is to be used. The normalized pumping power value scales from 0% which refers to no liquid injection to 100% which refers to power at the maximum pressure. Though maximum change in the pumping power value is limited by normalized value w which models dynamics of the pump:

$$\begin{cases} \begin{bmatrix} 0\% \le m(\tau) \le 100\% \\ -w \le m(\tau+1) \le w. \end{bmatrix} \\ \tau \in [0; \tau_H] \end{cases} \quad (7)$$

Control problem is based on interval steps, which starts at current timeτ. The result of the optimization is an optimal sequence of control actions, such as amount of tasks to be executed for each tier. First samples of the sequence have to be applied to the target 3D MPSoC, while the remaining ones have to be discarded. At each time moment, a new optimal control problem based on new temperature measurements and required frequencies is solved over a shifted prediction horizon [87], which refers to transforming open-loop design method into a feedback method. Therefore at every time moment the input value is applied to the process parameters up to the real time process measurements.

3. Experimental results and analysis

To evaluate developed model we have to compare application of different thermal management liquid cooling techniques:

- LB (load balancing);
- LUT (look up table);
- FL (fuzzy logic);
- PRA (proactive).

Liquid cooling implies maximum cooling flow rate and application of load balancing policy. LB balances the workload by moving threads from a core's queue up to the queue lengths difference and threshold value. LUT-based flow rate control dynamically changes the flow rate up to the predicted maximum temperature, while the tasks have to be scheduled with standard LB procedure. Fuzzy-logic control is basedon fuzzy logic mechanisms which forms thermal management algorithms that controls the liquid flow rate.

Liquid cooling techniques are comparision is based on maximum and average temperatures values as computational and cooling power consumption regimes. Thermal impact of the techniques is shown at Figure 2. It has to be noticed that LB reduces the peak temperature more than LUT and FL, but still avoids hot-spots. It is same to PRA technique, which the peak temperature reaches 84°C. While each technique has a different management policy and control elements, it afects the peak and average temperatures values.

Figure 3 demonstrates comparison of the total consumed power rate for different rechniques based on on the four-tier MPSoC with the average workload [94]. Energy consumption values were normalized up to the 3D-MPSoC liquid cooling technique load balancing policy. It has to be metioned that PRA policy manages reducing

of the cooling power and thereby overall system power by 23 % with respect to LC policy, by 40% with respect to LUT policy and by 22% with respect to LUT policy.

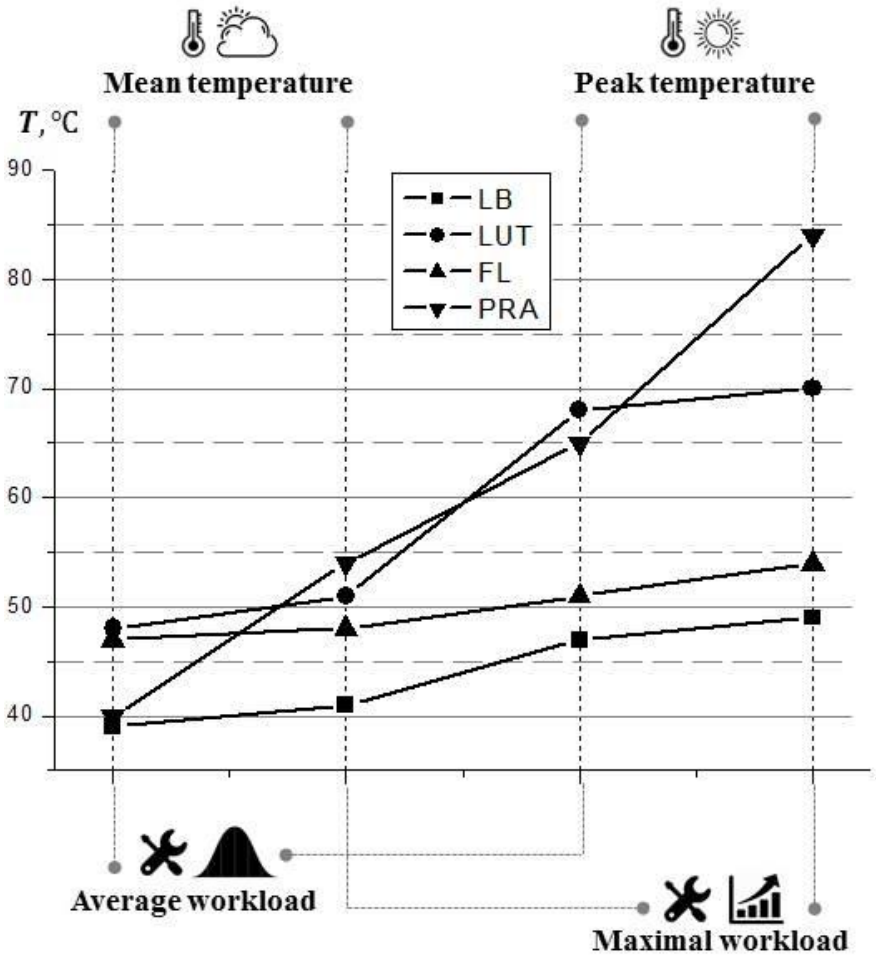


Fig. 2. Temperature values observed using all the policies for different workloads regime on four-tier 3D MPSoC

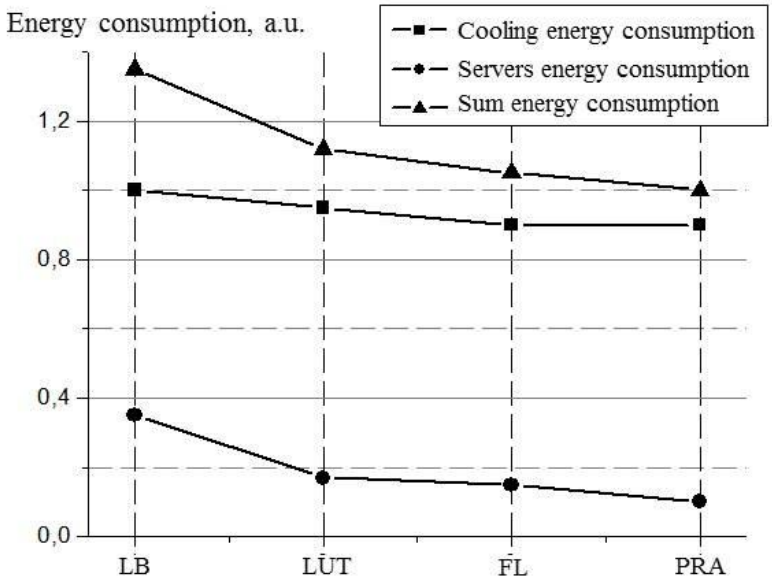


Fig. 3. The normalized energy consumption in the whole 3D MPSoC system

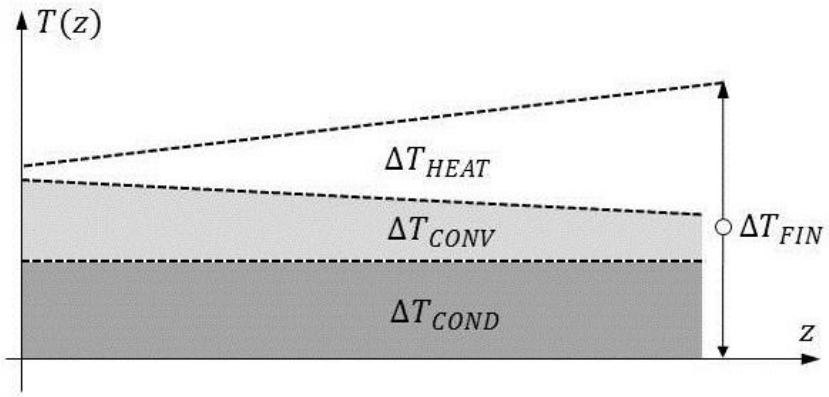
Microchannel width significantly affects the change in temperature due to convection ΔT_{CONV}. With Nusselt number NU value, hydraulic diameter of channel d and thermal conductivity h_{COOL} the heat transfer coefficient H can be obtained as:

$$H = \frac{h_{COOL} \cdot NU}{d}. \qquad (8)$$

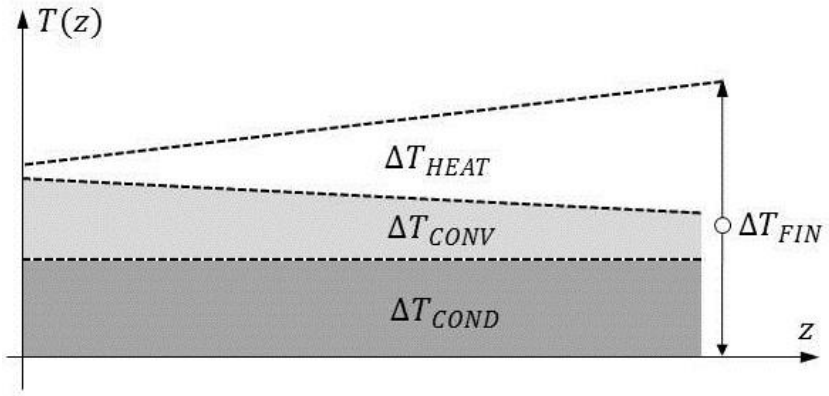
The effective heat transfer coefficient can be evaluated by projecting the heat transfer coefficient above from the side wall surfaces by channel height h_C and channel width w_c:

$$H_{eff} = H \frac{2h_C + w_c}{W}. \qquad (8)$$

The convective resistance R_{CONV} for the system can be obtained as a reciprocal value of the H_{eff}. Figure 4 and 5 shows that the convective temperature ΔT_{CONV} as well as convective resistance R_{CONV} decreases up to the channel width is reduction. The main problem is modification of the convective resistance to compensate T_{HEAT}. Thereby the channel width should be a function of the distance along the channel $w_C(z)$. The maximal width is at the inlet where the fluid temperature is low and minimal width is near the outlet where the fluid temperature is high. Thereby, for the case of uniform heat flux, it should be modulated the channel width from inlet to outlet.



(a)



(b)

Fig. 4. Microchannel temperature distribution for the structure with (a) uniform constant channel width and (b) modulated channel width

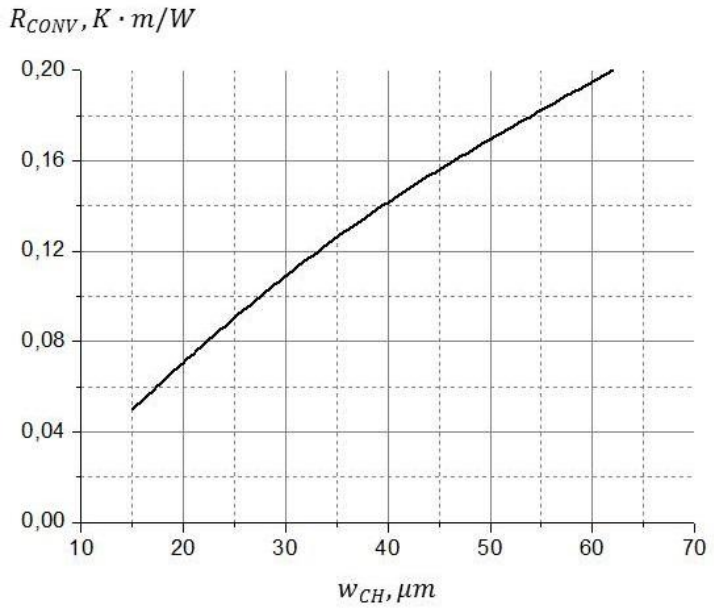


Fig. 5. Dependence of the convective resistance on the channel width

Developed methodology uses this paradigm of formulating an optimal control criteria to find the optimal microchannel width, from the fluid inlet to outlet. The main problem of optimization is to minimize the peak temperature and thermal gradients of the 3D MPSoC model, which allows to reduce the energy needed by cooling system of servers' room.

4. Conclusions

There were analyzed methods of data center performance estimation based on mathematical simulation of consumption system. It was shown that optimal instrument of modeling could be multi-processor system on chip performance enhancement. In order to optimize the model centralized control concept, inter-tier liquid cooling and proactive management scheme that rely on model predictive controller were discussed. To develop the methodology operating power supply of the platform to near-threshold values, multiple supply voltages utilization optimization method for the voltage islands distribution and microarchitectural techniques to control the thermal hotspots were analyzed. Multi-processor system on chip performance enhancement was demonstrated as application for minimization of the global thermal impact, specifically temperature-aware floorplanning and simulated annealing utilization. While power consumption is generated by two sources it was decided that cost function could be defined as a sum of the power input vector and required workload.

Developed control system is based on interval steps, which starts at current time. The result of the optimization is proved to be an optimal sequence of control actions. To evaluate developed model were compared application of thermal management load balancing, look up table, fuzzy logic and proactive liquid cooling techniques. Unified thermal modeling methodology based on the finite difference method helps to make a proper analysis of the problem and to build proper applications up to the particular properties. Developed methodology uses optimal control criteria to find the microchannel width. The main problem of optimization is to minimize the peak temperature and thermal gradients of the model, which allows to reduce the cooling system consumption.

References

1. *Zhang T., Cevrero A., Beanato G., Athanasopoulos P., Coskun, A.K. & Leblebici Y.,* 2013. 3D-MMC: A Modular 3D Multi-Core Architecture with Efficient Resource Pooling. Design. Automation & Test in Europe Conference & Exhibition (DATE), 2013.
2. *Emi T. et al.* Tape: Thermal-aware agent-based power economy for multi/many-core architectures. InICCAD. Pages 302 –309, 2009.
3. *Qian H. et al.* Cyber-physical thermal management of 3D multi-core cache-processor system with microfluidic cooling.ASP Journal of Low Power Electronics. 7(1):1–12, 2011.
4. *Zanini F., Sabry M.M., Atienza D. and De Micheli G.* Hierarchical thermal management policy for high-performance 3d systems with liquid cooling.IEEE JETCAS, 1(2):88–101, 2011.
5. *Aitken R., Flautner K. & Goodacre J.*, 2010. High-Performance Multiprocessor System on Chip: Trends in Chip Architecture for the Mass Market. Multiprocessor System-on-Chip. 223-239.

6. *Sabry M.M. et al.* Energy-Efficient Multi-Objective Thermal Control for Liquid-Cooled 3D Stacked Architectures.IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 30 (12):1883–1896, 2011.
7. ARM®CORTEX®-M4 Development Systems., 2015. Digital Signal Processing Using the ARM® CORTEX®-M4, 1-8.
8. *Dreslinski R.G. et al.* Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits. InProc. of the IEEE. 98(2), 2010.
9. *Xu N. et al.* Thermal-Aware Post Layout Voltage-Island Generation for 3D ICs. InJournal of Computer Science and Technology. 28(4):671–681, 2013.
10. *Puttaswamy K. and Loh G.H.* Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors. InHPCA. Pages 193–204, 2007.
11. *Han Y., Chakraborty K., Roy S. & Kuntamukkala V.*, 2011. A GPU Algorithm for IC Floorplanning: Specification, Analysis and Optimization. 2011 24th Internatioal Conference on VLSI Design.
12. *Sankaranarayanan K., Velusamy S., Stan M. and Skadron K.* A Case for Thermal-Aware Floorplanning at the Microarchitectural Level. InJournal of Instruction-Level Parallelism, 8:1–16, 2005.
13. *Hung W-L. et al.* Thermal-Aware Floorplanning Using Genetic Algorithms. InISQED, 2005.
14. *Liu W. & Nannarelli A.*, 2008. Net Balanced Floorplanning Based on Elastic Energy Model, 2008 Norchip.
15. *W.-L. Hung et al.* Interconnect and Thermal-Aware Floorplanning for 3D Microprocessors. InISQED, pages 98–104, 2006.
16. *Healy M. et al.* Multiobjective Microarchitectural Floorplanning for 2-D and 3-D ICs. InIEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. 26 (1), 2007.
17. Thermal-Aware Testing of Digital VLSI Circuits and Systems, 2018. doi:10.1201/9781351227780.