Bio-Inspired Thermal Management Techniques for Three Dimensional Heterogeneous Stacked Network-on-Chip Systems

Ranjita Kumari Dash
National Institute of Technology
Rourkela, India, 769008
Email: ranjita.rakhi@gmail.com

José L. Risco-Martín
Complutense University
of Madrid
Madrid, Spain
Email: jlrisco@ucm.es

Ashok Kumar Turuk
National Institute of Technology
Rourkela, India, 769008
Email: akturuk@gmail.com

José L. Ayala
Complutense University
of Madrid
Madrid, Spain
Email: jlayala@ucm.es

With 3D NoCs help improve circuit performance, fault tolerance and energy efficiency through the reduction of average wire-length and the increase in communication bandwidth of on-chip wiring, the soaring increase of on-chip temperature remains one of the most challenging obstacles to their commercialization. We present a physical design flow that integrates thermal driven floor-planning with MOEA. The thermal aware floor-planning help reduce the magnitude of hotspots in each layer , in turn, alleviate the negative impact of heat dissipation on chip performance and reliability. The essence of the flow is to analyze the layered thermal map of the chip stack and then apply MOEA operators, which helps to put the power hungry functional units far from each other.3D stacked heterogeneous mesh architecture with 3 layers is used as the baseline for our experimental work. Further another two layers has been added to check the impact of increasing number of layers on the peak temperature of individual layers. The experimental results show the suitability of our algorithm for significantly reducing maximum on-chip temperature. As our approach is independent of any topology, it paves the way for thermal driven design methods consisting of 3D layouts made up of several layers.

***Index Terms*—MOEA, floorplanning, 3D NoC, thermal map, heterogeneous mesh architecture.**

I. Introduction and Related Work

The complexity of System on Chip (SoC) arises with continuous scaling in semiconductor technology. In the future, hundreds or even thousands of processors will be integrated in a single chip. At such high level of integration, a major role is performed by communication channels to properly communicate among all the functional units. NoC has emerged as a promising candidate for the on-chip communication system because of scalability, reduced power consumption, fault tolerance, reusability and better throughput [1], [2], [3].

3D NoC is the combination of NoC and 3D IC. The NoC concept replaces design-specific global on-chip wires with a generic on-chip interconnection network realized by specialized routers that connect generic processing elements (PE) such as processors, ASICs, FPGAs , memories and so forth to the network and facilitate communications or links between them. Vertical placement of dies reduces interconnection length notably [4] . The reduced interconnection length results in low latency and increase in performance. As 3D NoC supports heterogeneous integration, the dies placed one above the other need not to be similar ones leading to optimization of chip components according to their diverse function [5], [6], [7].

Thermal and communication aware mapping using genetic algorithm is used to decrease the peak temperature in 2D and 3D NoCs [8]. Skadron *et al.* [9] proposed to use equivalent electrical RC-circuits in order to model the thermal behavior of a chip [9]. Several other works focused on 3D floor-planning, placement and thermal related issues [10], [11]. Coskum *et al.* [12] proposed dynamic thermal management methods for 3D NoCs.

In order to balance thermal distribution, floorplan optimization and wire routing are two other methods proposed in [13], [14], [15]. Jason Cong *et al.* proposed a thermal aware floorplan algorithm, where mapping is done from a 2D floorplan to a 3D floorplan, and placement of TSVs is taken in to account for improvement of thermal state and on-chip performance. Wire routing was proposed by Tianpei Zhang *et al.* , which lowers the temperature by adding thermal vias and wires for heat dissipation.

The paper is organized as follows: Section 2 covers mathematical formulation of the problem, design methodology and floorplanning algorithm used to achieve the result. Section 3 illustrates the experimental analysis and Section 4 concludes this paper.

II. Problem Formulation

A. Mathematical Model

In this paper, the fitness function is organized as a weighted sum of two objectives. The first objective is defined by means of the topological relations among placed blocks. It represents the number of topological constraints violated (no overlapping between placed blocks as current volume $\leq$ maximum volume). This objective ensures that current area created by functional units is less or equal than maximum area and also ensures no overlapping between placed units. Without this objective , our floorplanner can not find feasible solutions. Every block in the model i $(i = 1, 2, 3, ..., n)$ is characterized by a height h_i, length l_i and width w_i. The entire chip can be considered as a design volume having maximum height H, maximum length L and maximum width W. Our aim is to find a feasible floorplan by maximizing the distance between hottest elements (N_e). We have defined the geometrical location of block i, where $0 \leq x_i \leq L - l_i$, $0 \leq y_i \leq W - w_i$ and $0 \leq z_i \leq H - h_i$. We take (x_i, y_i, z_i) to denote the left-bottom-back coordinate of block i.

The second objective is a measure of the thermal impact, computed as follows. Given a block b, it's temperature can be modeled as follows:

$$T_b = \sum_{i}^{N} A_{bi}^{-1}.P_i \approx A_{bb}^{-1}.P_b \propto P_b \tag{1}$$

where P_b is the nomalized power density of block b.

Therefore, our objective can be stated as follows,

$$Min.(T) = \sum_{i<j} \frac{B_i.B_j}{\sqrt{dx_{ij}^2 + dy_{ij}^2 + dz_{ij}^2}} \tag{2}$$

where B=$\{PE_1, PE_2, ..., PE_n, SW_1, SW_2, ..., SW_n\}$;

where, PE is a processing element and SW represents a switch.

Our proposed algorithm will try to place the hottest processing elements and switches as far as possible. As described in [16] , this method can reduce maximum temperature on the chip. For our 3D NoC stacked mesh scenario, we have used two objectives. The Performance and temperature objectives have an opposite nature: the better the performance, higher the temperature. In our proposed algorithm, our objective function is stated as follows:

$$Min.(F) = (n + r) \times \frac{T}{T^*} \tag{3}$$

where **r** is the set of topological constraints violated, T^* is the evaluation of Equation 4, in a stacked mesh topology. We have used the mesh topology as a baseline because of its high performance. For our future work, we are currently analyzing mechanisms to include communication as a third objective, making the mesh topology an excellent baseline for the fitness function.

The use of MOEA is proposed in this work as it is an efficient method to solve NP-hard problems. MOEAs are stochastic optimization heuristics in nature. These approaches have been successfully applied to many NP-hard combinatorial optimization problems. In order to apply MOEAs to a problem, a genetic representation of each individual has to be found out.

In each generation, two chromosomes are selected in tournament selection. Tournament selection runs several tournaments among few individuals chosen at random from the population. The winner of the tournament is the chromosome having the best fitness function. The winner chromosome is selected for crossover. Selection pressure is adjusted by changing the tournament size. Bigger tournament size discards the chance of weak individuals to be selected. The selection operator in our floor-planner selects two random chromosomes from the entire population and then as described above , the best of these are selected. This task is repeated twice in order to obtain two chromosomes or parents. Advantages of tournament selections are it is efficient to code, works on parallel architectures and it allows the selection pressure to be easily adjusted.

Our floor-planner helps to keep the heat sources as far as possible and it generally places them at the border of the chip. This technique helps in reducing the on chip temperature due to diffusion. Vertical heat spread is taken into account by not placing the heat sources one above the other. From the optimized floor-plans , it is clearly visible that most of the IPs are placed either in the first layer or in the last layer. Routers are placed at the border lines of the chip which yields to less heat.

Algorithm 1 Thermal-aware Floorplanning Algorithm for 3D stacked heterogeneous Network-on-Chip

```
Require: Floorplan with number of layers, power values and
    layer dimensions
Ensure: Best fitness function
Input number of layers, layer dimensions and power values
    of each IP and Switches;
Produce n arrays of random integer permutations;
Generate chip floorplan;
Initialize algorithm with physical dimensions of each unit
    and power data;
Arrange the blocks according to descending order of
    power density;
For i=1 to max-generation
generate placement;
Calculate peak temperature of each layer;
Evaluate fitness function;
crossover();
mutation();
i=i+1;
EndFor
Display best function;
```

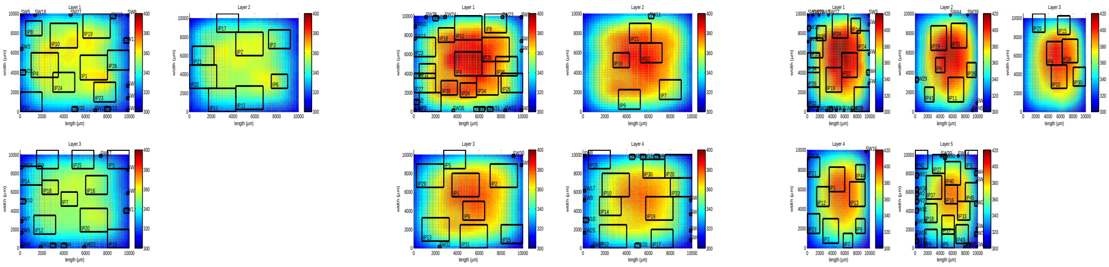
The proposed algorithm is configured with a maximum population of 100 individuals, and a maximum number of 250 generations. The probability of mutation depends on the number of variables. In this case, it is the inverse of the number of blocks. Then, we set a cycle crossover with a probability of 0.90 and the tournament selection method, as mentioned in [17].

III. EXPERIMENTAL RESULTS

This section presents the thermal results obtained in the scenarios described in Section 3. The benchmarks used here is divided in to 3 categories , they are 3 layer, 4 layer, and 5 layer stacked 3D NoC heterogeneous stacked mesh architecture.

Architecture Used	Layer Number	Original Temperature (in °K)	Resultant Temperature (in °K)	Temperature Reduction (in °K)
3layer	layer1	372.0910	359.8613	14.2297
	layer2	377.8307	361.9631	15.8676
	layer3	384.2310	365.9055	18.3255
4layer	layer1	387.0078	377.0070	10.0008
	layer2	393.8061	382.2046	11.6015
	layer3	402.4738	388.2689	14.2049
	layer4	410.2078	394.7030	15.5048
5layer	layer1	399.3518	392.8972	6.4546
	layer2	406.8852	399.2916	7.5936
	layer3	416.4039	406.5383	9.8656
	layer4	428.3669	414.7866	13.5803
	layer5	437.2575	419.3000	17.9575

TABLE I: Peak and optimized temperature per Layer in Heterogeneous 3-D Stacked Network-on-Chip



(a) Thermal map of 3 layer heterogeneous floorplan after applying MOEA

(b) Thermal map of 4 layer heterogeneous floorplan after applying MOEA

(c) Thermal map of 5 layer heterogeneous floorplan after applying MOEA

Fig. 1: Original and optimized thermal diagram of 3 layer, 4 layer and 5 layer heterogeneous stacked 3D mesh based NoC

The floor-plans are modified in order to include an increased number of functional units in every layer. The inter-layer communication is carried out with a set of buses that route the communication signals from one layer to another.

As can be seen in Table 1, we achieve a reduction of 32.6588 °K in peak temperature when compared with the baseline scenario for 3 layer scenario. However, as Table 2 and Table 3 shows the peak temperature of 4 layer and 5 layer benchmarks have gone up to 410.2078°K and 437.2575°K respectively. The high rise in temperature is because of the fact that the inner layers are not able to dissipate the heat to the ambient. After applying the optimization algorithm, the corresponding maximum reduction in temperature are 15.5048°K and 17.9575°K respectively, for the layers showing highest temperature. The peak temperature of other layers can be referred from Table 1, Table 2 and Table 3.

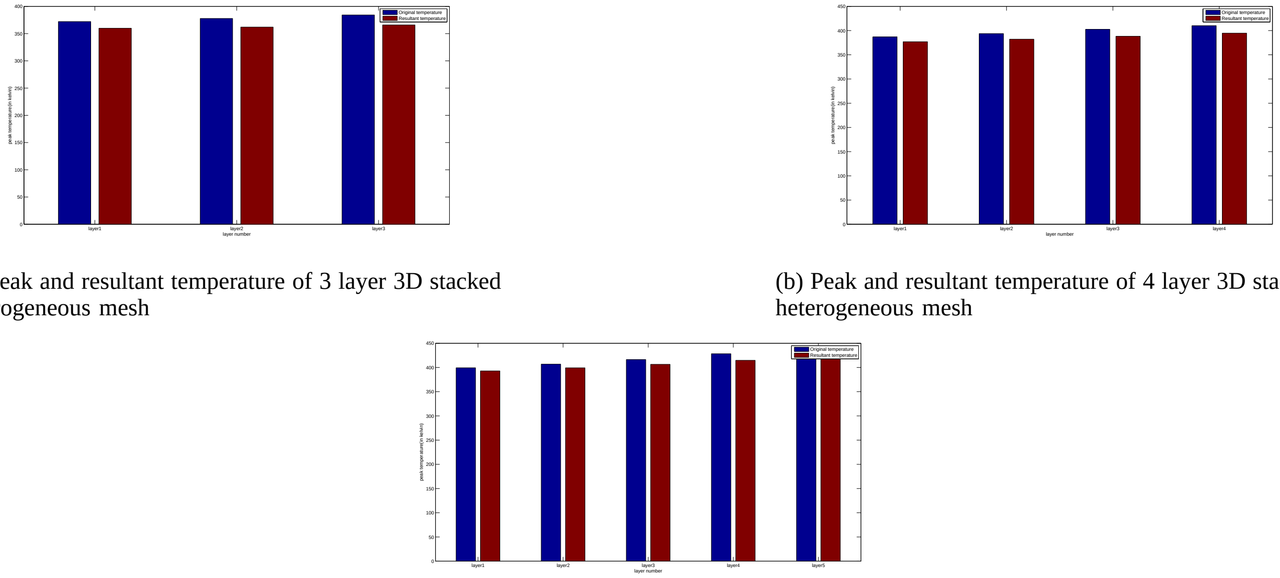
From the results reflected in the tables, it can be derived that our floor-planner algorithm decreases the maximum temperature in all the active layers for all the designs. This is due to the homogeneous thermal distribution due to replacement of functional units. The reduction of on-chip temperature translates in to a reduced reliability risk and diminished leakage currents.

From the visual analysis of the Figure 2, it is clearly noticed that there are high temperature regions , where hotspots are created. The problem is more exacerbated in first layers, where the functional units are not able to dissipate heat to the environment. The thermal problems affects the performance and reliability of the chip due to their high temperature. As the power dissipated in the die increases, the temperature increases exponentially.

From the results reflected in the Table 1, Table 2 and Table 3, it can be found out that the optimizer decreases the peak temperature in all active layers. This is due to the fact that, the replacement of functional units ensured a more homogeneous thermal distribution. It results into a minimized reliability risk and reduced leakage currents.

IV. Conclusion

We have developed an integrated CAD framework for thermal-aware 3D Network-on-Chip. This framework was applied to the 3D heterogeneous stacked NoC. Thermal simulation and floor-planning were conducted in the iterative loop seeking the maximum reduction in peak temperature of individual layers. Our algorithm first places the higher power consuming cores away from each other in different layers to avoid rise in peak temperature due to the diffusion phenomenon and then use thermal simulation to determine the thermal map of 3D heterogeneous stacked NoC. The peak temperature in the dies strongly increases with increase in number of layers. The effectiveness of the proposed algorithm regarding reduction in temperature has been demonstrated by the reduction of on-chip temperature with respect to its original floor-plan.



(a) Peak and resultant temperature of 3 layer 3D stacked heterogeneous mesh

(b) Peak and resultant temperature of 4 layer 3D stacked heterogeneous mesh

(c) Peak and resultant temperature of 5 layer 3D stacked heterogeneous mesh

Fig. 2: Comparison of original and final peak temperature heterogeneous 3D stacked mesh NoC architecture

References

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

[2] A. Jantsch, H. Tenhunen *et al.*, *Networks on chip*. Springer, 2003, vol. 38.

[3] T. Bjerregaard and S. Mahadevan, “A survey of research and practices of network-on-chip,” *ACM Computing Surveys (CSUR)*, vol. 38, no. 1, p. 1, 2006.

[4] L. Shang, L.-S. Peh, A. Kumar, and N. K. Jha, “Thermal modeling, characterization and management of on-chip networks,” in *Proceedings of the 37th annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2004, pp. 67–78.

[5] B. S. Feero and P. P. Pande, “Networks-on-chip in a three-dimensional environment: A performance evaluation,” *IEEE Transactions on Computers*, vol. 58, no. 1, pp. 32–45, 2009.

[6] P. K. Hamedani, S. Hessabi, H. Sarbazi-Azad, and N. E. Jerger, “Exploration of temperature constraints for thermal aware mapping of 3d networks on chip,” in *Parallel, Distributed and Network-Based Processing (PDP), 2012 20th Euromicro International Conference on*. IEEE, 2012, pp. 499–506.

[7] D. Chatterjee and T. Manikas, “Power-density aware floorplanning for reducing maximum on-chip temperature,” in *18th IASTED Int. Conf. on Modelling and Simulation (ICMS)*, 2007, pp. 319–324.

[8] W. Hung, C. Addo-Quaye, T. Theocharides, Y. Xie, N. Vijakrishnan, and M. J. Irwin, “Thermal-aware ip virtualization and placement for networks-on-chip architecture,” in *Computer Design: VLSI in Computers and Processors, 2004. ICCD 2004. Proceedings. IEEE International Conference on*. IEEE, 2004, pp. 430–437.

[9] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, “Temperature-aware computer systems: Opportunities and challenges,” *IEEE Micro*, no. 6, pp. 52–61, 2003.

[10] J. Hu and R. Marculescu, “Exploiting the routing flexibility for energy/performance aware mapping of regular noc architectures,” in *Design, Automation and Test in Europe Conference and Exhibition, 2003*. IEEE, 2003, pp. 688–693.

[11] S. Murali and G. De Micheli, “Sunmap: a tool for automatic topology selection and generation for nocs,” in *Proceedings of the 41st annual Design Automation Conference*. ACM, 2004, pp. 914–919.

[12] A. K. Coskun, J. L. Ayala, D. Atienza, T. S. Rosing, and Y. Leblebici, “Dynamic thermal management in 3d multicore architectures,” in *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE’09.* IEEE, 2009, pp. 1410–1415.

[13] J. Cong and Y. Zhang, “Thermal via planning for 3-d ics,” in *Computer-Aided Design, 2005. ICCAD-2005. IEEE/ACM International Conference on*. IEEE, 2005, pp. 745–752.

[14] J. Cong, G. Luo, J. Wei, and Y. Zhang, “Thermal-aware 3d ic placement via transformation,” in *Design Automation Conference, 2007. ASP-DAC’07. Asia and South Pacific*. IEEE, 2007, pp. 780–785.

[15] T. Zhang, Y. Zhan, and S. S. Sapatnekar, “Temperature-aware routing in 3d ics,” in *Design Automation, 2006. Asia and South Pacific Conference on*. IEEE, 2006, pp. 6–pp.

[16] D. Cuesta, J. L. Risco-Martin, and J. L. Ayala, “3d thermal-aware floorplanner using a milp approximation,” *Microprocessors and Microsystems*, vol. 36, no. 5, pp. 344–354, 2012.

[17] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: Nsga-ii,” *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002.