http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7297-7304, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7297.7304

Power-and Thermal-aware Mapping for 3D Network-on-chip

[1,2]Ge Fen, [2]Feng Gui, [2]Yu Shuang and [2]Wu Ning
[1]Science and Technology on Avionics Integration Laboratory, Shanghai, 200233, China
[2]College of Electronic and Information Engineering,
Nanjing University of Aero. and Astro., Nanjing, Jiangsu, 210016, China

Abstract: In this study, we present a 3D Network-on-Chip (NoC) comprehensive simulation framework which is integrated with 3D Nirgam, HotSpot and Orion. Based on this framework, a power- and thermal-aware mapping approach for 3D NoC is proposed. The proposed approach uses genetic algorithm to map IP cores onto 3D NoC architecture with the goal of minimizing the power consumption and temperature deviation. Various multimedia benchmark applications are evaluated to confirm the efficiency of the proposed approach. The power consumption in power-aware mapping is saved 27.07% on average compared to random mapping. The temperature deviation between nodes is reduced by 34.56% on average with our approach compared to random mapping while a little influence is caused on performance.

Key words: 3D Network-on-chip, genetic algorithm, power consumption, temperature

INTRODUCTION

Continuous development of semi-conductor technologies makes it possible to integrate more and more IP cores in a chip. The traditional System-on-Chip (SoC) design encounters more challenges in latency, power consumption and throughput. As a result, Network-on-Chip (NoC) emerged as a novel paradigm to provide an effective, reliable and flexible infrastructure for system modules based on data packet transmission scheme (Benini and De Micheli, 2002). Simultaneously, the advent and increasing viability of 3D Integrated circuits show a new field for IC designer. By combining these two techniques, 3D NoC is suggested to become an effective solution to overcome difficulties associated with inter- connection length and interconnect power which provides better communication mechanisms than traditional networks (Shan and Bill, 2008).

In NoC design, mapping determines the topological placement of IP cores under different design constraints which has great influence on overall system performance. Thus, finding an efficient and accurate mapping algorithm is extremely important. Usually, metrics as power consumption and latency are considered in mapping algorithm. However, due to smaller chip area and the existence of the vertical layers, thermal management issue comes to be more serious than ever in 3D IC design. In reality, many of electronic circuit failures are caused by or related to unevenly distributed heat dissipation which may produce hotspots (Chu and Wong, 1998). Hotspots can give rise to timing failures and reliability concerns. As a result of these effects, temperature should be considered as one metric in 3D mapping.

In this study, we take both the power consumption and thermal management into account and propose a power- and thermal- aware mapping approach based on genetic algorithm for 3D NoC to optimize both power consumption and temperature distribution of the chip. Parts of our work have been presented in (Feng *et al*., 2013) to minimize the peak temperature and the temperature deviation. This study expands the previous work with a further analysis of the impact of mapping on power consumption and performance variation with different thermal optimization strategies. Also, we present a comprehensive simulation framework for 3D NoC to evaluate the mapping results.

RELATED WORK

In 2D NoC design field, tremendous works has been done on mapping problem. As we know, PBB is proposed to minimize the energy consumption for mesh architecture (Hu and Marculescu, 2003). And heuristic search algorithm gets more attention in recent years. A NAMP algorithm maps IP cores onto NoC architecture under the bandwidth constraint with the aim of minimizing communication delay (Murali and de Micheli, 2004). Simulation Annealing (SA) is presented to map IP cores considering link bandwidth with latency (Hredzak and Diessel, 2011).

Corresponding Author: Ge Fen, Science and Technology on Avionics Integration Laboratory, Shanghai, 200233, China

Recently, there are some studies on 3D NoC mapping. According to the properties of 3D NoC, different optimization goals attached great attention. A rank-based multi-objective genetic algorithm has been proposed for latency-aware mapping to reduce latency under congestion and no congestion (Wang *et al*., 2011). Ying *et al*. (2012) present a genetic algorithm based optimization method for low vertical link density, considering the tradeoff of the number of TSVs and performance. Addo-Quaye (2005) first proposed to address the thermal-aware mapping for 3D NoC. Three ILP-based thermal-aware mapping for 3D NoC are proposed to explore the thermal constraints and their effects on temperature and performance (Hamedani *et al*., 2012). However, only the peak temperature of the chip is considered in their work. Another important fact in thermal issues is the distribution of temperature which is not included.

PROBLEM FORMULATION AND DEFINITIONS

In this section, we first analyze the power and thermal model and then present the mapping problem formulation.

Power model: Power dissipation is a critical issue in 3D circuits. Although the total power consumption of 3D systems is expected to be lower than that of mainstream 2D circuits (since the global interconnects are shorter), the increased power density constitutes a new challenge for this novel design paradigm.

Pavlidis and Friedman (2007) show that the power consumption per bit between a source destination node pair in 3D NoC is given by:

$$P_{bit} = hopsP_{stotal} + hops_{2-D}P_{htotal} + hops_{3-D}P_{vtotal} \tag{1}$$

where, P_{stotal} is the power consumed on the crossbar switch in a router. P_{htotal} and P_{vtotal} are the power consumed on the horizontal and the vertical link respectively. Hops is the average number of routers that a packet traverses to reach the destination node. $hops_{2-D}$ is the average number of hops within the 2D layers and $hops_{3-D}$ is the average number of hops on the third dimension

For 3D Mesh NoC with minimal routing, hops is determined by the Manhattan distance between node $i(x_i, y_i, z_i)$ and node $j(x_j, y_j, z_i)$:

$$hops = |x_i - x_j| + |y_i - y_j| + |z_i - z_j| \tag{2}$$

$$hops_{2-D} = |x_i - x_j| + |y_i - y_j| \tag{3}$$

$$hops_{3-D} = |z_i - z_j| \tag{4}$$

Since Eq. 1 is a linear combination of the variable hops and the constants P_{stotal}, P_{htotal} and P_{vtotal}, we assert that minimizing the average hop distance is equivalent to minimizing the power consumption, regardless of the constant values.

Temperature model: Thermal management issues in 3D NoC draw more and more attention. In [11] researchers show steady state temperature of each IP core in 3D NoC is given by:

$$Th_{i,j,k} = T_{Amb} + \sum_{m=1}^{k} \frac{R_{i,j,m}}{A} \times (\sum_{s=m}^{n} P_{i,j,s} + PR_{i,j,s}) \tag{5}$$

where $Th_{i,j,k}$ is the temperature of IP core at the position (i, j, k) in 3D NoC. T_{Amb} is the ambient temperature. $R_{i,j,m}$ is the thermal resistance of the IP core at the position (i, j, m). A is the area of IP core. n is the total number of layers. $P_{i,j,s}$ and $PR_{i,j,s}$ are the average power consumption of IP core and router at the position (i, j, s) in 3D NoC. $P_{i,j,s}$ can be assigned a random power based on the average power consumption of the core (at the range of 10~60 W/cm^2) at the specific technology (Tsai and Kang, 2000). $PR_{i,j,s}$ can be calculated as follows:

$$PR_{i,j,s} = \sum_{\forall L \rightarrow R_{i,j,s}} \lambda_L \times P_{rbit} \tag{6}$$

where λ_L is the amount of data routed from this router. P_{rbit} is the bit power consumption of the router.

Problem formulation: The mapping is to determine the topological placement of IP cores onto different resource nodes, such that the temperature distribution of the chip, power consumption and performance metrics are optimized. Fig. 1 shows an instance for the mapping problem in 3D NoC architecture.

For the convenience of analysis and discussion, the mapping problem can be formulated as follows.

Given a core communication graph denoted by CCG(C, A), where each vertex $c_i \in C$ represents an IP core and each directed edge $a_{i,j} \in A$ represents the communication trace from IP c_i to IP c_j. The weight of the edge, denoted by $b(a_{i,j})$ represents the total volume of the communication.

Given a NoC architecture graph which is denoted by NAG(R, P). Each vertex $r_i \in R$ represents a resource node in the architecture and each edge denoted as $p_{i,j} \in P$ represents the communication path from resource node r_i

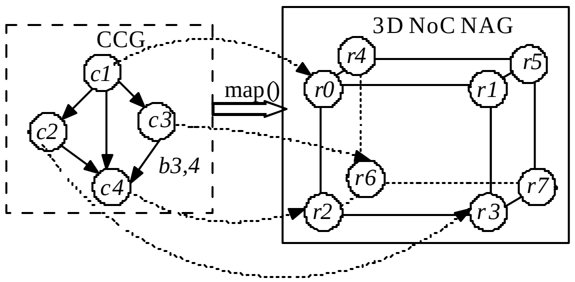


Fig. 1: Mapping problem for 3D NoC architecture

to r_j. The weight of the edge, denoted by $pw_{i,j}$, represents the average power consumption of sending one bit of data from r_i to r_j, i.e. P_{bit}.

Find a map function:

$$\text{map}: C \to R \Rightarrow r_i = \text{map}(c_i), \forall c_i \in C, \exists r_i \in R, |C| \le |R|$$

Such that:

- The total communication power consumption is minimized:

$$\min E(A) = \sum_{\forall a_{i,j} \in A} b(a_{i,j}) \times hops_{i,j}) \tag{7}$$

- The chip temperature is evenly distributed

According to thermal model, the temperature of the node mainly depends on the router power. When the technology of the chip is chosen, the other parameters are fixed. As stated in Eq. 6, the router power is calculated based on the router power consumption per bit and the amount of data which is routed from this router. Hence, the goal can be simplified as:

$$\min\{\lambda_L\} \tag{8}$$

Where:

$$\lambda_L = \sum_{\forall ai, aj \in A} b_{i,j} \times M$$

and:

$$M = \begin{cases} 1 & \text{if } L \in \text{Link}(r_i, r_j) \\ 0 & \text{else} \end{cases}$$

Link (r_i, r_j) is the set of links between node i and node j based on the NoC routing algorithm.

Our proposed algorithm takes temperature deviation into account, so as to balance the temperature distribution of each layer. We use the deviation of temperature on chip to represent whether the heat dissipation is evenly distributed and it is given by:

$$\sigma_n = \frac{1}{n}\sum_{i=0}^{n}(Th_i - T_{avg})^2 \tag{9}$$

where, σ_n is the deviation of temperature of nodes on chip (the smaller values means evenly distributed). n is the total number of nodes. Th_i is temperature of the ith node. And T_{avg} is the average temperature of the total nodes.

In order to balance the performance and temperature deviation between nodes, we sort the temperature of nodes in descending order and take top m values to calculate deviation. The value of m (m = n) can be adjusted according to the requirements of performance and temperature optimization. Thus, the temperature deviation of nodes on chip is modified by:

$$\sigma_m = \frac{1}{m}\sum_{i=0}^{m}(Th_i - T_{avg})^2 \tag{10}$$

Finally, the goal of optimizing the chip temperature is to find:

$$\min\{\sigma m\} \tag{11}$$

s.t.:

$$(Th_i)_{max} < T_{max}$$

where, T_{max} is the maximum temperature that a chip can bear.

3D NoC SIMULATION FRAMEWORK

The goal of the proposed mapping approach is to optimize power consumption and temperature distribution of the chip. However, the existing NoC simulators, such as Nirgam, Noxim and NNSE, only can be used to evaluate performance metrics of 2D NoC system. Therefore, we establish a comprehensive simulation framework for 3D NoC to evaluate various configured 3D NoC architectures with different topologies, routing algorithms and traffic patterns, as shown in Fig. 2. We first expand existing Nirgam simulator for 2D NoC to be suitable for the simulation of 3D NoC and combine the power simulator Orion and thermal model Hotspot with it to form the comprehensive simulation framework. This platform can evaluate the 3D NoC architectures with performance, power consumption and chip temperature which provide a fundamental platform for 3D NoC design.

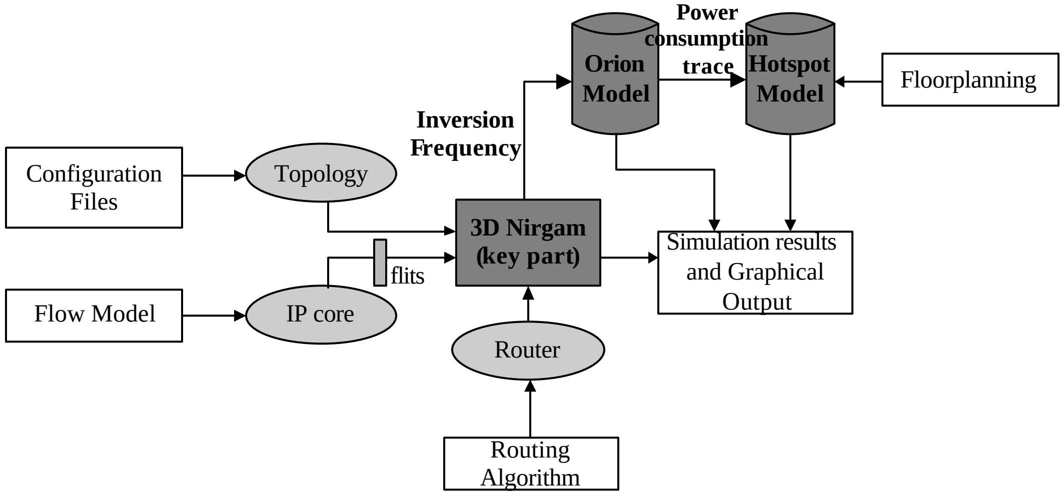


Fig. 2: 3D NoC comprehensive simulation framework

Nirgam simulator which is the key part of the simulation platform loads the configuration information from the configuration files and component libraries to initialize the NoC architecture with the network topology, the routing algorithm used in the router and the flow model used in communication IP cores and then starts the simulation process. Nirgam analyzes the network latency and throughput and invokes the Orion model to compute the power consumption of the whole system and each component. Then, the power consumption traces are input to the Hotspot model and the temperature distribution of NoC system can be computed. All the simulation results can be output graphically.

PROPOSED MAPPING APPROACH

Based on the above simulation framework, we propose a power-and thermal-aware mapping approach for 3D NoC using genetic algorithm. The approach consists of four phases and the algorithm flow is shown in Fig. 3.

First phase: An optimal mapping set is produced by utilizing genetic algorithm. The following steps are used to decide the position of each IP core in the 3D NoC and the pseudo code is shown in Fig. 4.

- Generate an initial population of n chromosome which consists of many randomly generated IP core placements. Each chromosome is encoded into integer strings, with its length equal to the number of vertices in a core communication graph. As shown in Fig. 5, each gene (vertex in CCG) in the chromosome contains an integer indicating a randomly selected vertex of the 3D NoC architecture graph

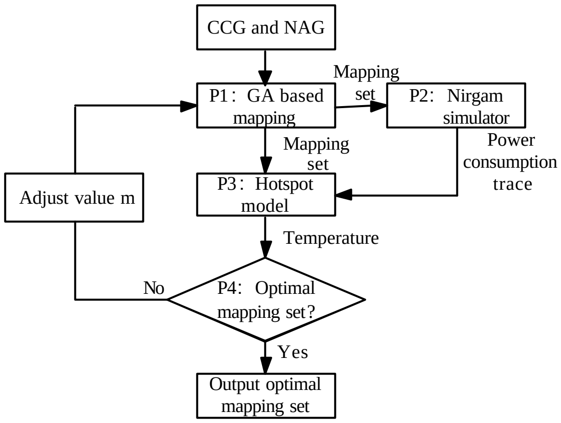


Fig. 3: Algorithm flow

```
Initalize population
Set the parameters
For i=1 to max_generation
    For each solution
        Attain IP core communication graph
        Generate placement
    Compute the fitness(the communication cost in hops, temperature
    for each IP core)
    Sort temperature
    Calculate the temperature deviation
    Evaluate cost function
End for
    select
    crossover
    mutate
    upgrade the population
    I=i+1
end for
output the optimal solution
```

Fig. 4: Pseudo code of the genetic algorithm based mapping

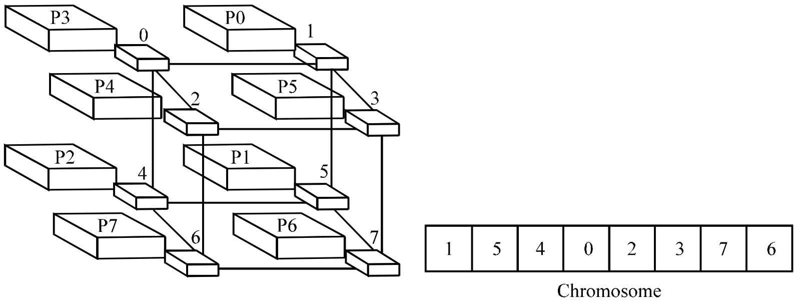


Fig. 5: Chromosome encoding for 3D NoC mapping

- Evaluate the fitness of each chromosome in the population. In this step, the fitness function given by Eq. 7 or 11 is used to minimize the communication power consumption
- Create a new population by applying three operators similar to the natural selection operators-selection, crossover and mutation. Selection selects two parent chromosomes from the population according to their fitness (the better fitness, the larger chance to be selected). For crossover, two cross points are randomly selected, the genes between two cross points of the first parent are copied to the second offspring, then the other parent is scanned and if the number is not yet in the offspring, it is added to the offspring. For mutation, we exchange two members in chromosome that are randomly selected
- Repeat the step 2~3 until the lowest average hop distance has not changed for a sufficient number of iterations, then stop and report the best mapping as the generated optimal IP core placement in the 3D NoC

Second phase: According to the mapping results obtained in last phase, we input the determined communication relationship between IP cores into Nirgam and set the configuration files. Then, the performance (network latency and throughput) and power consumption of the optimal mapping can be obtained.

Third phase: The mapping set and power consumption traces which are got in phase1 and phase 2 respectively are taken as inputs to Hotspot in the simulation framework. Temperature of each node is to be calculated.

The fourth phase: We have a mapping set and its corresponding performance, power consumption and temperature information. Then we need to judge whether the mapping set is the optimal one. If it is, the optimal mapping set can be output. If not, the value of m should be adjusted and the algorithm returns to the first phase.

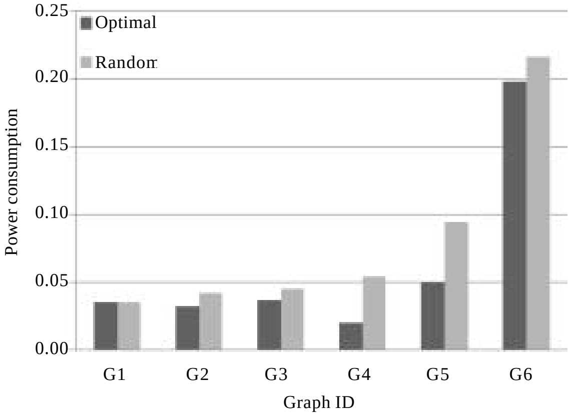


Fig. 6: Power consumption comparisons for multimedia benchmarks

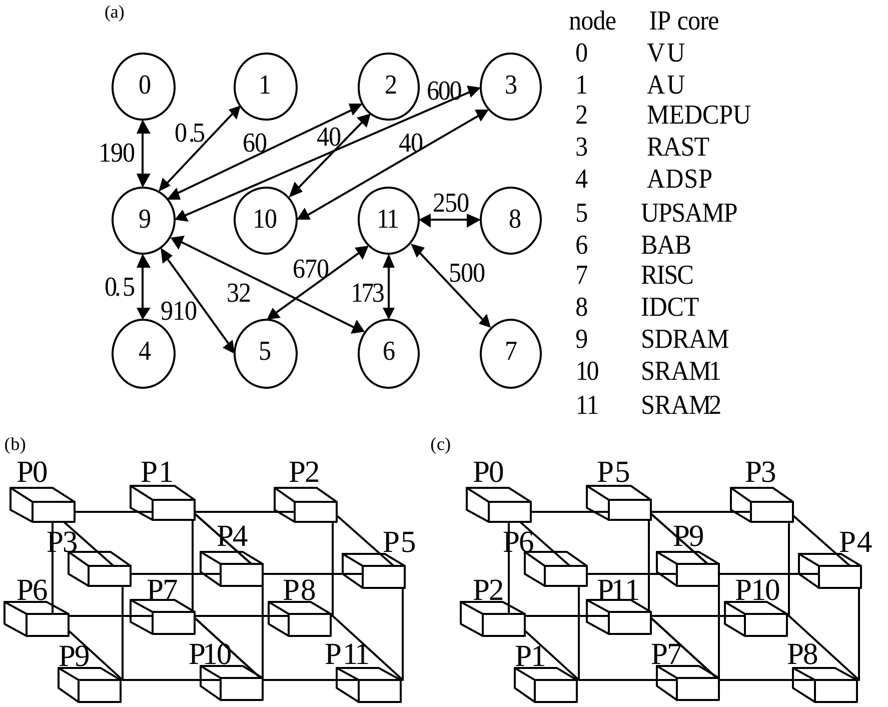
Table 1: Graph characteristics

Graph	Graph ID	Nodes	Edges	NoC X×Y×Z
MWD [19]	G1	12	13	2×3×2
VOPD [19]	G2	12	15	2×3×2
MPEG4 decoder [19]	G3	12	26	2×3×2
H.263 enc MP3 enc [20]	G4	14	19	2×4×2
MMS [21]	G5	25	33	3×4×3
D.38 tvopd [22]	G6	38	47	4×4×3

EXPERIMENTAL RESULTS

In this section, we present the experimental results obtained by executing the proposed approach on various multimedia benchmark applications. Table 1 lists the graph IDs and sizes of the *CCG* of the various benchmarks (Bertozzi *et al*., 2005; Murali *et al*., 2009). In order to evaluate the efficiency of the proposed approach, we compared the results produced by our proposed approach against random mapping.

Power consumption comparisons for multimedia benchmarks based on power-aware mapping using (7) as fitness function is shown in Fig. 6. The power consumption is reduced 27.07% on average compared with random mapping. Especially, with the number of cores increases the reduction gets more obvious.



node	IP core
0	VU
1	AU
2	MEDCPU
3	RAST
4	ADSP
5	UPSAMP
6	BAB
7	RISC
8	IDCT
9	SDRAM
10	SRAM1
11	SRAM2

Fig.7 (a-c): CCG and the mapping results of MPEG4 decoder, (a) The CCG of MPEG4 decoder, (b) A random mapping result and (c) Optimal mapping result

As an example, we give a core communication graph *CCG* of MPEG4 decoder in Fig. 7a, in which the labels of the edges in *CCG* denote the bandwidth requirement. The random mapping and optimal mapping solution are shown in Fig. 7b and c, respectively. From the figure, we can see that the communication amount between two cores is heavier, the more close they are attached to each other.

Mapping results comparisons for multimedia benchmarks based on power-aware mapping using Eq. 11 as fitness function is shown in Fig. 8. In Fig. 8a, we show a comparison between the optimal mapping and the random mapping on peak temperature using various benchmarks. From the figure, it can be seen that as the number of IP core increases, the reduction in peak temperature gets more obvious. 13% reduction is achieved on average. In particular, for the D.38 tvopd benchmark which has the largest number of IP cores, the proposed algorithm improves the peak temperature up to 17%. The average temperature is presented in Fig. 8b. It shows no much difference between the random mapping and optimal mapping, when the number of IP cores are relatively small. As the number of IP cores increase, the improvement is greater by utilizing optimal mapping. A comparison of the deviation of temperature for the different benchmarks is shown in Fig. 8c which shows 35% decrease on average. And for the MPEG4 benchmark which is the most popular application, the deviation is decreased 69%. We also present comparisons of power and latency in Fig. 8d and Fig. 8e respectively. From the figure, the optimal mapping save power consumption 14% on average and latency is reduced 8% on average. Therefore, the algorithm we proposed can reduce the temperature deviation efficiently while a little influence is caused on performance.

Table 2: Performance of MPEG4 with different values of m

Performances	m =3	m =6	m =12
Peak temperature (°C)	55.15	49.11	61.02
Average temperature (°C)	41.29	40.76	43.85
Average power consumption (mW)	39.69	38.64	44.95
Average delay (cycles)	20.30	21.70	20.48

In order to further discuss the impact of the value of m in Eq. 10, we take MPEG4 as an example again. Table 2 shows the performance of MPEG4 with different values m in our mapping approach. From the table, it is easy to see that, when m is taken as 6, the peak temperature is the lowest with higher latency. When m is 3, it has better performance than other values expect the peak temperature. For different applications, the number of hotspots varies. Besides, the main target of chip design is different, where the value of m plays a role of adjustment.

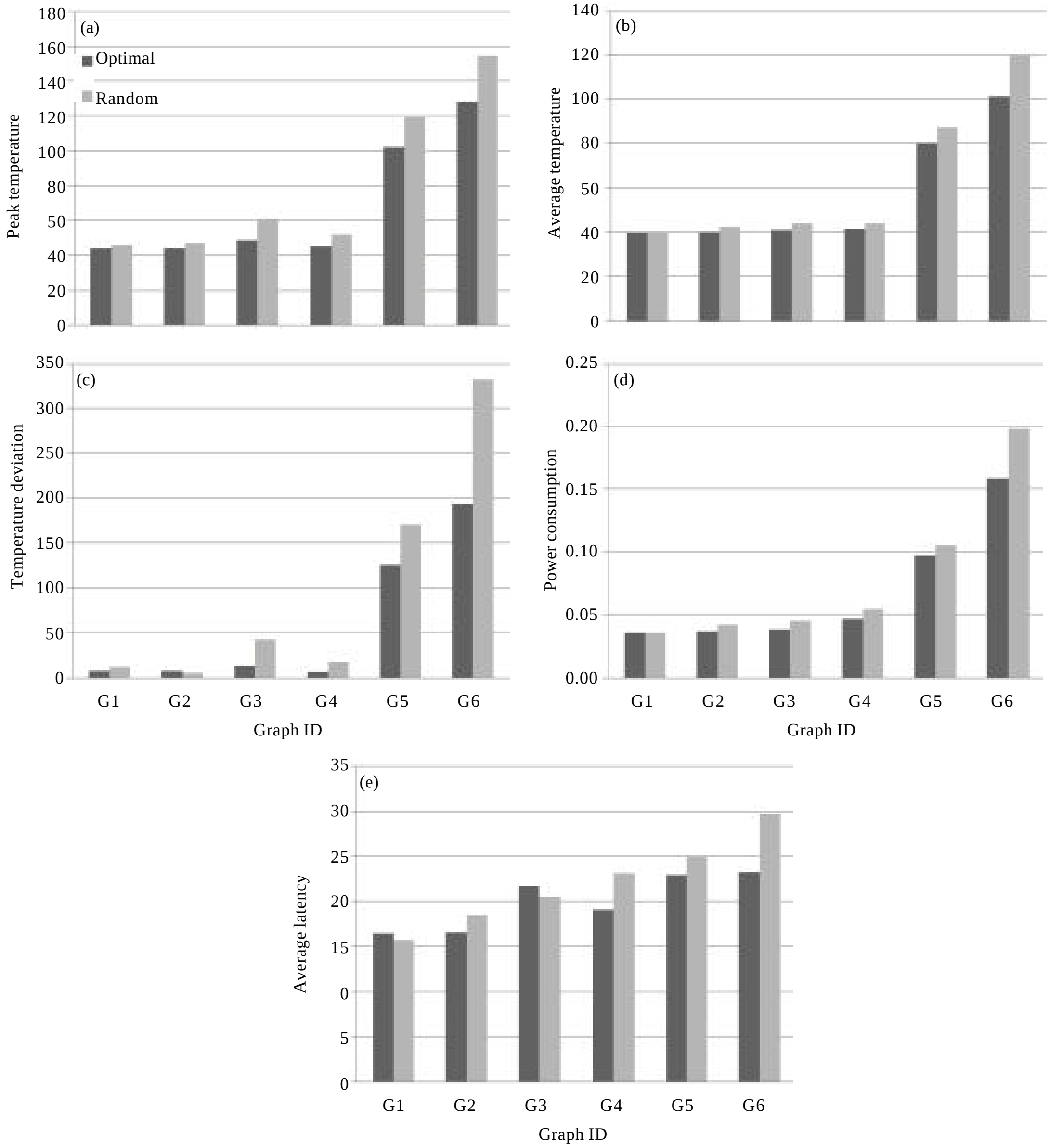


Fig. 8 (a-e): Temperature and performance comparisons of various benchmarks, (a) Peak temperature, (b) Average temperature, (c) Deviation of temperature, (d) Power consumption and (e) Average latency

CONCLUSION

This study presents a 3D NoC comprehensive simulation framework. And based on this framework, we propose a power- and thermal-aware mapping approach for 3D NoC. The aim is to optimize the communication power consumption and chip temperature distribution. Applying our approach on various multimedia benchmark applications gives experimental results showing significant achievements as compared to random mapping.

In the future work, we will take 3D IC floorplanning information into consideration, such as add TSV constraint into our mapping approach. Moreover, in order to make better use of 3D NoC, layering technology will be applied in our work which will decrease the possibility of mixing different technologies (such as logic, analog, DRAM) on one chip.

ACKNOWLEDGMENTS

The authors would like to thank for the support by the Joint Fund from Science and Technology on Avionics

Integration Laboratory and Aeronautical Science Foundation of China under the Grant 20115552031.

REFERENCES

Addo-Quaye, C., 2005. Thermal-aware mapping and placement for 3-D NoC designs. Proceedings of the IEEE International SOC Conference, September 19-23, 2005, Herndon, VA., USA., pp: 25-28.

Benini, L. and G. De Micheli, 2002. Networks on chips: A new SoC paradigm. IEEE Comput., 35: 70-78.

Bertozzi, D., A. Jalabert, S. Murali, R. Tamhankar, S. Stergiou, L. Benini and G. De Micheli, 2005. NoC synthesis flow for customized domain specific multiprocessor systems-on-chip. IEEE Trans. Parallel Distrib. Syst., 16: 113-129.

Chu, C.C.N. and D.F. Wong, 1998. A matrix synthesis approach to thermal placement. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., 17: 1166-1174.

Feng, G., F. Ge, S. Yu and N. Wu, 2013. A thermal-aware mapping algorithm for 3D mesh network-on-chip architecture. Proceedings of the IEEE 10th International Conference on ASIC, October 28-31, 2013, Shenzhen, China, pp: 63-66.

Hamedani, P.K., S. Hessabi, H. Sarbazi-Azad and N.E. Jerger, 2012. Exploration of temperature constraints for thermal aware mapping of 3D networks on chip. Proceedings of the 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, February 15-17, 2012, Munich, Germany, pp: 499-506.

Hredzak, B. and O. Diessel, 2011. Optimization of placement of dynamic network-on-chip cores using simulated annealing. Proceedings of the 37th Annual Conference on IEEE Industrial Electronics Society, November 7-10, 2011, Melbourne, Australia, pp: 2400-2405.

Hu, J. and R. Marculescu, 2003. Energy-aware mapping for tile-based NoC architectures under performance constraints. Proceedings of the ASP-DAC Asia and South Pacific Design Automation Conference, January 21-24, 2003, Kitakyushu, Japan, pp: 233-239.

Murali, S. and G. de Micheli, 2004. Bandwidth-constrained mapping of cores onto NoC architectures. Proceedings of the Design, Automation and Test in Europe Conference and Exhibition, Volume 2, February 16-20, 2004, Paris, France, pp: 896-901.

Murali, S., C. Seiculescu, L. Benini and G. De Micheli, 2009. Synthesis of networks on chips for 3D systems on chips. Proceedings of the IEEE Asia and South Pacific Design Automation Conference, January 19-22, 2009, Yokohama, Japan, pp: 242-247.

Pavlidis, V.F. and E.G. Friedman, 2007. 3-D topologies for networks-on-chip. IEEE Trans. Very Large Scale Integr. Syst., 15: 1081-1090.

Shan, Y. and L. Bill, 2008. Design of application-specific 3D networks-on-chip architectures. Proceedings of the 26th International Conference on Computer Design, October 12-15, 2008, Lake Tahoe, CA., USA., pp: 142-149.

Tsai, C.H. and S.M. Kang, 2000. Cell-level placement for improving substrate thermal distribution. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., 19: 253-266.

Wang, J., L. Li, H. Pan, S. He and R. Zhang, 2011. Latency-aware mapping for 3D NoC using rank-based multi-objective genetic algorithm. Proceedings of the IEEE International Conference on ASIC, October 25-28, 2011, Xiamen, China, pp: 413-416.

Ying, H., K. Heid, T. Hollstein and K. Hofmann, 2012. A genetic algorithm based optimization method for low vertical link density 3-dimensional networks-on-chip many core systems. Proceedings of the IEEE NORCHIP Conference, November 12-13, 2012, Copenhagen, Denmark, pp: 1-4.