Noname manuscript No.
(will be inserted by the editor)

Reliability-Aware Heterogeneous 3D Chip Multiprocessor Design

Ismail Akturk · Ozcan Ozturk

the date of receipt and acceptance should be inserted later

Abstract Ability to stack separate chips in a single package enables three-dimensional integrated circuits (3D ICs). Heterogeneous 3D ICs provide even better opportunities to reduce the power and increase the performance per unit area. An important issue in designing a heterogeneous 3D IC is reliability. To achieve this, one needs to select the data mapping and processor layout carefully. This paper addresses this problem using an integer linear programming (ILP) approach. Specifically, on a heterogeneous 3D CMP, it explores how applications can be mapped onto 3D ICs to maximize reliability. Preliminary experimental evaluation indicates that the proposed technique generates promising results in both reliability and performance.

Keywords Reliability · Multicore · 3D · Data Mapping

1 Introduction

As technology scales, the International Technology Roadmap for Semiconductors projects that the number of cores will drastically increase to satisfy performance requirements of future applications [5]. Once the number of cores passes some threshold (16 cores), conventional point-to-point buses will no longer be a sufficient interconnect structure. These future applications will therefore require a Network-on-Chip (NoC) [14], where a dedicated on-chip network (with switches and links) is used to perform the communication between cores. NoCs have shown to be able to handle the required communications between the cores in a scalable, flexible, programmable, and reliable fashion [14].

In addition to NoCs, three-dimensional integrated circuit (3D IC) [4] is an attractive option for overcoming the barriers in interconnect scaling. 3D

Bilkent University Computer Engineering Department
Bilkent, Ankara, Turkey
E-mail: {iakturk,ozturk}@cs.bilkent.edu.tr

ICs are built using multiple device layers stacked together with a direct tunnel between them, thereby allowing them to reduce the global interconnect. Moreover, 3D ICs provide higher performance and lower power consumption due to the reduced interconnect (wire) length. Other benefits include support for realization of mixed-technology chips, higher packing density, and smaller footprint.

3D NoCs [6,7,14] have been introduced to combine these two techniques (3D ICs and NoCs) in order to achieve better performance with higher scalability. 3D ICs reduce the global interconnect, thereby improve performance. On the other hand, NoCs provide scalable communication framework. While, homogeneous NoCs have been widely used for both 2D and 3D ICs, they are limited compared to their heterogeneous counterparts. This follows from the fact that every application has a different processing requirement and memory footprint. A powerful core will be a better match for an application with a high level of instruction-level parallelism, while a simpler core will be sufficient for applications with lower instruction-level parallelism. Therefore, it is more effective to use heterogeneous NoCs.

As the technology shrinks, one of the challenging problems in the context of 3D NoC systems is reliability. Reliability of 3D ICs is effected by both temperature and thermo-mechanical stress. This is especially caused by the limited cooling capability between the layers. Specifically, vias become more and more sensitive and when the via fails to make proper connection, unwanted loss in yield and decrease in reliability may occur. Reliability for 3D ICs have been explored from different angles [2,8–11,16]. Through Silicon Vias(TSVs) are the most recent medium in stacking [8] multiple dies on a 3D IC. However, these vias become more sensitive with higher temperatures that can be caused by more activity or traffic. Since TSVs are bridges between layers, they are potentially more prone to thermal stress. Therefore, reducing the TSV communication load has potential of improving reliability. This work aims at increasing the reliability of an application through effective mapping on 3D heterogeneous IC. Contribution of the approach is in two folds:

- An ILP formulation of the problem of maximizing the reliability of a given application. This is achieved through optimal placement of nodes in a 3D NoC.
- Minimization of the communication cost between the nodes, thereby improving both performance and energy consumption.

ILP-based approach presented here targets at reducing the amount of layer-to-layer communication on TSVs, while keeping the overall communication overheads minimum. The remainder of this paper is structured as follows. Section 2 gives the related work on heterogeneous 3D NoCs and reliability. Section 3 discusses the overview of proposed approach. The details of the ILP (integer linear programming) based formulation are given in Section 4, and an experimental evaluation is presented in Section 5. The paper is concluded in Section 6.

2 Related Work

Related work can be summarized in three parts, namely, 3D ICs, 3D NoCs and 3D reliability. 3D technologies and benefits of 3D ICs over 2D ICs have been presented in [4]. In [18], authors review the process steps and design aspects of 3D ICs. Andry et al [3] discuss a three-dimensional (3D) chip stacking technology using fine-pitched interconnects. Sakuma et al [15] reviews the 3D integration technologies, including process technology and reliability characterization. Akasaka [1] presents 3D IC technology for fabrication, total power consumption estimation and chip cooling. Zhao et al [19] studies DC current crowding and its impact on 3D power integrity.

Pavlidis et al [14] compare 3D NoC over 2D NoC from a physical constraints perspective. These physical constraints include the maximum number of planes that can be vertically stacked and the asymmetry between the horizontal and vertical communication channels of the network. In [6], authors focus on the second level (L2) cache design for 3D architectures. Similarly, Ozturk et al [13] try to place processor cores and data blocks optimally in a 3D design. Mesochronous communication scheme for 3D NoCs have been explored in [7].

Minz et al. proposed a 3D module and decoupling capacitance(a.k.a decap) placement algorithm that tries to distribute the thermal profile on the circuit evenly and reduce the power noise [10]. The algorithm is trying to find the location of each block in the 3D placement layers without overlap and it tries to minimize the footprint area, total wire length, maximum block temperature, total amount of decap required to suppress simultaneous switching noise (SSN) under the given tolerance value. They showed that there is little correlation between thermal and decap objectives that allows them to optimize these objectives simultaneously. They extended existing 2D sequence pair scheme of Murata et al [11]. Specifically, each layer has its own sequence pair to represent the relative positions among the blocks in it. Then, they used simulated annealing to search through the solution space using various intra- inter-layer moves. Malta et al. discussed the characterization of thermo-mechanical stress and reliability issues for Cu-filled Through Silicon Vias (TSVs) [8]. An X-ray imaging method was used for fast nondestructive analysis of Cu TSV plating profiles. It was observed that TSVs exposed to increased temperature exhibited a substantial increase in grain size which was associated with the Cu protrusion effect. Alam et al. developed a framework to enable reliability analysis in 3D circuits called ERNI-3D [2]. It is a Reliability Computer Aided Design (RCAD) tool that is a capable of comparison of 2D and 3D circuit layouts. Similar to study of Alam et al., Shayan et al. proposed a framework to analyze the reliability of 3D power distribution network under local through silicon via failures [17]. The 3D power distribution network is extracted and modeled in frequency domain considering skin effect. The model is first solved in frequency domain to identify the behavior of the system. Then, the time domain voltage noise under worst case transistor switching currents is obtained with enhanced vector fitting algorithm. The objective of the optimization is to increase the

reliability of the 3D structure and reduce the voltage noise while minimizing the block out area from through silicon via design rules. They showed that the increase in the dimension and density reduce the routable area of the stacked dies. As the width in the tiers increases the power noise decreases. Selvanayagam et al. worked on thermo-mechanical reliability of through silicon via for different dimensions [16]. The increase in the through silicon via diameter will increase the thermo-mechanical strains and as a result the reliability is reduced. They showed the existence of the trade off between the reliability and the power noise reduction as the TSV diameters increases. Minas et al. presented the challenges of and some emerging solutions for 3D processing phases including TSV insertion and wafer thinning [9]. These processes have an impact on the functionality, performance and reliability of the circuit. Approach presented in this paper is different from these techniques such that it implements a reliability-aware node/task mapping.

3 Overview

High level view of the ILP-based approach is shown in Figure 1. After necessary parallelization and mapping steps, the input code is fed to a compiler analysis module. The compiler analysis module captures the communication characteristics of processors are captured and this is subsequently used in the ILP solver. Each processor can potentially have different characteristics in terms of performance, energy consumption, area requirement, and communication bandwidth. According to the processor characteristics and communication requirements, processors are laid out on the 3D NoC. Moreover, various constraints such as die area, temperature limit, number of layers, and performance are applied. Location of processors is selected based on a reliability objective while keeping the communication cost at reasonable levels. Note also that, the objective function can be replaced with a combination of reliability, performance, and energy using different weights.

An example 3D NoC architecture is given in Figure 2, where multiple layers of heterogeneous processors are connected using network switches/routers represented by R. This heterogeneous 3D NoC architecture is exposed to the compiler to enable accesses to the state of the processors, the network switches, and the data/code movements. Note that, heterogeneous processors are represented by CPU and memory hierarchies are represented by MH. Each layer of the 3D NoC architecture is considered as a grid, where processors are mapped according to their dimensions. Processors are considered to have widths and heights based on the same unit length as the grid. In-layer communication distances are captured using the coordinates of processors in 2D grid space with Manhattan distance. Moreover, inter-layer distances include the communication overhead caused by layer-to-layer transmissions.

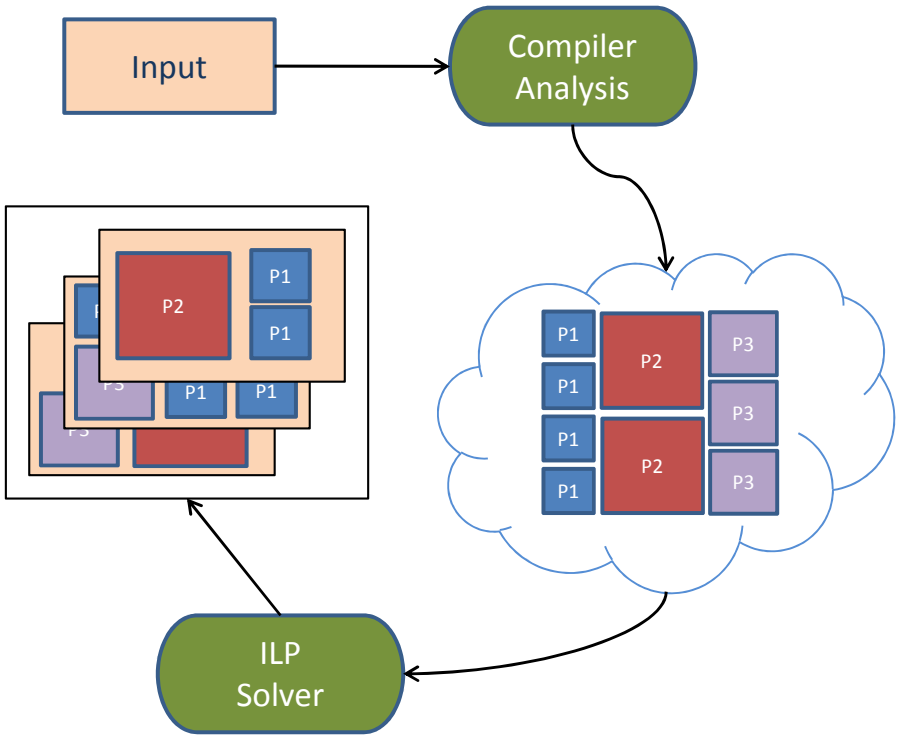

Fig. 1 *High level view of the ILP-based approach.*

4 ILP Formulation

Integer linear programming (ILP) is an optimization technique which targets optimization of a linear objective function subject to linear function constraints and integer solution variables. A special case of ILP is the 0-1 ILP, where solution variables are required to be either 0 or 1. In this context, integer linear programming (ILP) is used to formulate the reliability problem on a 3D NoC by finding the optimal location of each processor. There are two important goals in selecting the location of processors:

- Reduce communication overheads by placing the frequently communicating nodes as close as possible.
- Improve reliability by minimizing the inter-layer communications.

These two goals can potentially contradict with each other when the layer-to-layer communication is considered. This is due to the fact that third dimension provides a lot of opportunities in terms of improving the connectivity of processors. Processors can potentially communicate much faster through TSVs compared to in-layer communications. However, this also results in with

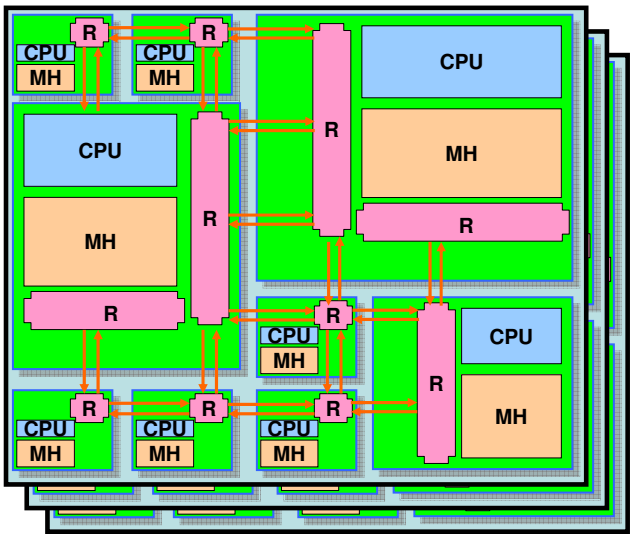

Fig. 2 *3D NoC-based CMP architecture.*

increased levels of heat density around vias, making them more and more sensitive. Therefore, ILP aims at reducing the communication cost, simultaneously tries not to map high communicating nodes onto separate layers as this increases the use of Through Silicon Vias(TSVs) which are less reliable compared to the in-layer communication.

This section presents an ILP formulation of the problem of maximizing reliability while minimizing the data communication cost of a given application. This is achieved through optimal placement of nodes in a 3D NoC. While overall ILP formulation has more details, for clarity, we only give the important parts of it. A commercial tool, *Xpress-MP* [12], is used to formulate and test the ILP-based approach. *Xpress-MP* takes the problem as a Mosel description which is a plain text file with descriptions of binary variables, constraints, and objective function. Solver (*Xpress-MP*) generates the output as a plain text file which lists the values of decision variables. Table 1 gives the constant terms and binary variables used in the ILP formulation.

Assuming that 3D NoC chip has dimensions of D_X and D_Y in the 2D grid space and L number of layers, the ILP problem can be formulated to map the P number of processors on this 3D NoC. Note that, each processor has its own dimensions expressed as DPX_p and DPY_p. Communication intensity of two processors, namely p_1 and p_2, is given with I_{p_1,p_2}. As mentioned before, using TSVs has contradicting effects. A weighted objective function is considered to capture the potential effects on reliability and overall communication. This is achieved by the ϕ constant which is used as a knob for choosing in-layer versus layer-to-layer communication.

In the ILP formulation, location of processor p is captured by $Loc(p)^l_{x,y}$, where,

Constant	Definition
P	Number of processors
D_X, D_Y	Dimensions of the 2D grid
L	Number of layers in 3D NoC
DPX_p, DPY_p	Dimensions of processor p
$I_{p1,p2}$	Communication intensity of processors p_1 and p_2
ϕ	In-layer vs. layer-to-layer communication cost ratio
Variable	**Definition**
$Loc(p)^l_{x,y}$	Processor p is in (x, y) coordinates on layer l
$Occ(p)^l_{x,y}$	Processor p occupies (x, y) coordinates on layer l
$In-layer(p_1, p_2)_d$	Manhattan distance between processors p_1 and p_2 is d
$Inter-layer(p_1, p_2)_l$	Layer-to-layer distance between processors p_1 and p_2 is l
$Comm_{In-layer}$	Total in-layer communication
$Comm_{Inter-layer}$	Total layer-to-layer communication
$Comm$	Total communication

Table 1 The constant terms and binary variables used in the ILP formulation. These are either architecture specific or program specific. L indicates the number of layers in the 3D chip.

– $Loc(p)^l_{x,y}$: indicates whether processor p is in (x, y) coordinates in the 2D grid space and on the l layer.

Since a processor can potentially occupy multiple unit spaces in the 2D grid space, a 0-1 variable named as $Occ(p)^l_{x,y}$ is introduced. This binary variable will depend on the dimensions of the processor given with DPX_p, DPY_p.

– $Occ(p)^l_{x,y}$: indicates whether processor p occupies (x, y) coordinates of the l layer.

Two binary variables have been introduced to capture the distances between two processors; $In-layer$ and $Inter-layer$. Specifically,

– $In-layer(p_1, p_2)_d$: indicates whether the Manhattan distance in 2D grid space between processors p_1 and p_2 is equal to d.

– $Inter-layer(p_1, p_2)_l$: indicates whether the layer-to-layer distance between processors p_1 and p_2 is equal to l.

In addition to the specified binary variables, there are also non-binary variables to capture different values in the optimization problem. However, these variables are not given here for simplicity. These binary and non-binary variables are used in satisfying various constraints, first of which is one-to-one mapping between processor and 2D-grid coordinate system at the specified layer the processor is in.

$$\sum_{x=1}^{D_X}\sum_{y=1}^{D_Y}\sum_{l=1}^{L} Loc(p)^l_{x,y} = 1, \quad \forall p \in (1, P). \tag{1}$$

To ensure one-to-one mapping, processor needs to be assigned a single coordinate, where x and y indicate the 2D-grid coordinates, whereas l indicates the

layer. Similarly, a specific coordinate on every layer can only be mapped to a single processor which is captured by:

$$\sum_{p=1}^{P} Occ(p)^{l}_{x,y} = 1, \forall x \in (1, D_X), \forall y \in (1, D_Y), \forall l \in (1, L). \tag{2}$$

As mentioned earlier, total data communication requirement at a certain layer is estimated by using the Manhattan distance on a 2D-grid space.

$$\begin{aligned} In-layer(p_1, p_2)_d \geq Loc(p_1)^{l_1}_{x_1,y_1} + Loc(p_2)^{l_2}_{x_2,y_2} - 1, \\ d = |x_1 - x_2| + |y_1 - y_2|. \end{aligned} \tag{3}$$

On the other hand, inter-layer communication distance can be captured using the layers the two processors are in:

$$\begin{aligned} Inter-layer(p_1, p_2)_l \geq Loc(p_1)^{l_1}_{x_1,y_1} + Loc(p_2)^{l_2}_{x_2,y_2}, -1, \\ l = |l_1 - l_2|. \end{aligned} \tag{4}$$

Total communication load within 2D layers can be obtained through:

$$Comm_{In-layer} = \sum_{p_1=1}^{P} \sum_{p_2=1}^{P} \sum_{d=1}^{D_X+D_Y} I_{p_1,p_2} \times In-layer(p_1, p_2)_d \times d. \tag{5}$$

Similarly, layer-to-layer communication overhead can be expressed as a multiplication of communicating processors' communication intensity and layer-to-layer distances:

$$Comm_{Inter-layer} = \sum_{p_1=1}^{P} \sum_{p_2=1}^{P} \sum_{l=1}^{L} I_{p_1,p_2} \times Inter-layer(p_1, p_2)_l \times l. \tag{6}$$

Both $Comm_{In-layer}$ and $Comm_{Inter-layer}$ uses I_{p_1,p_2} to express the affinity between two processors, which is multiplied with the distance given by d or l.

Based on the above constraints, the objective function can be defined as:

$$\min \quad Comm = Comm_{In-layer} + \phi \ Comm_{Inter-layer}. \tag{7}$$

As expressed before, ϕ can be used as a knob to evaluate communication reduction versus reliability. In the experimental results section, ϕ constant's value and its effects are evaluated. From a pure communication reduction perspective this value will probably be much higher. However, if TSV usage is not preferred due to reliability concerns, ϕ parameter can be adjusted to reflect this. In the baseline implementation, ϕ parameter is conservatively set to 0.1.

While the objective function does not consider performance specifically, it will indirectly optimize the performance by reducing the overall communication overheads. Note that, this performance improvement will also be limited with the ϕ constant. Moreover, additional constraints will be required for performance evaluation; for example, a constraint that captures simultaneous communication. Similarly, energy results can also be obtained with necessary constraints.

Benchmark	Source	Description	Number of Data Accesses
3step-log	DSPstone	Motion Estimation	91×10^6
adi	Livermore	Alternate Direction Integration	71×10^6
ammp	Spec	Computational Chemistry	87×10^6
equake	Spec	Seismic Wave Propagation Sim.	84×10^6
mcf	Spec	Combinatorial Optimization	115×10^6
mesa	Spec	3D Graphics Library	135×10^6
vortex	Spec	Object-oriented Database	164×10^6
vpr	Spec	FPGA Circuit Placement	117×10^6

Table 2 Benchmark codes used in this study.

5 Experimental Evaluation

Experimental evaluation is performed on parallelized array-based applications. Parallelizations and code optimizations are implemented through Stanford University Intermediate Format (SUIF). Benchmarks used in experiments are shown in Table 2. Experiments are conducted by first fast-forwarding the first 1 billion instructions, and simulating the next 300 million instructions. Number of data accesses are shown in the fourth column of Table 2. As shown in Table 3, the default number of device layers is set two and a single layer is composed of 48 unit areas which can be assigned to NoC nodes. As explained before, in the base configuration, ϕ parameter is set to 0.1, conservatively . The ILP solution times varied between 3 minutes and 7 hours, averaging on about 42 minutes. Overall complexity of the proposed scheme is NP-complete since it is based on ILP. However, when the offline nature of the proposed scheme is considered, the solution times are within tolerable ranges. Moreover, it is possible to generate a sub-otpimal solution in cases of longer solution times, which usually tends to be very close to the optimal solution.

Experiments are conducted on four different execution models, namely, 2D-HM, 2D-HT, 3D-HM, and 3D-HT:

- *2D-HM:* A single layer of 2D conventional NoC topology with homogeneous processors.
- *2D-HT:* Optimal placement of heterogeneous processors on a 2D grid using an integer linear programming based strategy. This uses the same optimization framework proposed so far, except it only considers a single layer, that is, it is the optimal placement scheme for 2D.
- *3D-HM:* Homogeneous processors are distributed among a 3D stacked chip based on the communication requirements. Note that, this scheme also applies an ILP-based approach and finds optimal placement.
- *3D-HT:* Heterogeneous processor cores are placed on several layers optimally using the proposed integer linear programming based placement strategy. This scheme represents the optimal placement for 3D depending on the communication frequencies of nodes and the level of reliability.

Parameter	Value
Types of processor cores	4
Number of blocks	48
Number of layers	2
Total storage capacity	128KB
Set associativity	2 way
Line size	32 Bytes
Number of lines per block	90
Temperature bound	110°C
Reliability (ϕ)	0.1

Table 3 The default simulation parameters.

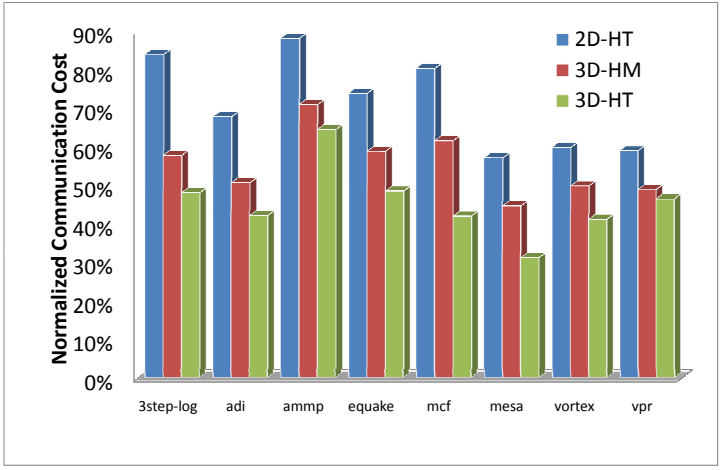

Fig. 3 Reliability-oriented data communication costs of 2D-HT, 3D-HM, and 3D-HT normalized with respect to 2D-HM.

Reliability-oriented data communication results normalized with respect to 2D-HM scheme based on two layers is given in Figure 3. Using the default values given in Table 3, average reduction in reliability-oriented data access costs for 2D-HT and 3D-HM are around 30% and 44%, respectively. 3D-HT reduces the communication further by about 54% on average. 3D reduces the global interconnect length and improves overall communication while maintaining reliability. This is more pronounced with heterogeneous processors as there are more opportunities.

Recall that the original number of 3D layers used were two. The bar-chart in Figure 4 shows the normalized costs (with respect to those of the 2D-HM scheme) for the benchmark ammp with the different number of layers (the results with the original number of layers are also shown for convenience), ranging from 1 to 4. Note that, the total storage capacity is kept constant for all these experiments and the only difference between two experiments is the number of layers and size of each layer. The number of layers and the corresponding number of blocks per layer for each topology tested are given in Table 4. One can see from these results that the effectiveness of the ILP-based

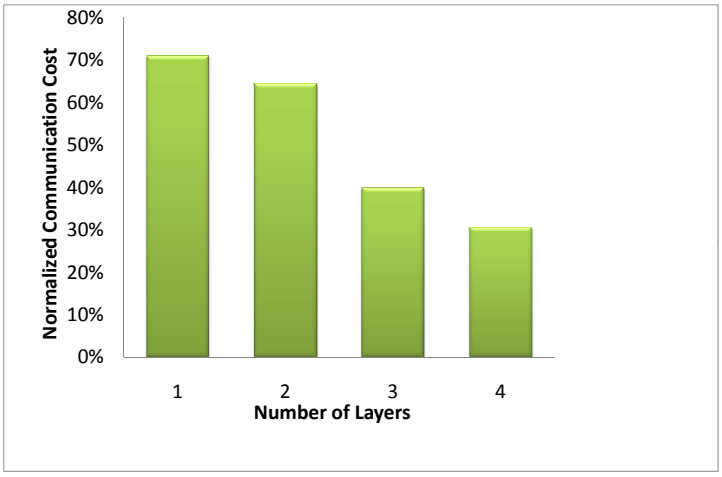

Fig. 4 Normalized reliability-communication costs with the different number of layers (ammp).

Number of Layers	Number of Blocks per Layer
1	48
2	24
3	16
4	12

Table 4 Different topologies.

approach increases with increasing number of layers. The main reason for this behavior is that adding more layers gives more flexibility to the proposed approach in placement.

In the next set of experiments, the effect of the ϕ parameter in savings is tested. As one can expect, savings increase with lower ϕ values. The main reason for this behavior is the reduction in the relative layer-to-layer communication cost, thereby increasing the flexibility on the vertical placement. On the other hand, from a reliability point of view, it is preferable to minimize the vertical communication on TSVs. Figure 5 shows the performance and reliability effects of the ϕ parameter. As mentioned before, the default value of ϕ is 0.1. Hence, all reliability and performance values are normalized with respect to $\phi = 0.1$. As can be seen from the figure, when ϕ is increased, the normalized communication cost increases since the cost of transfer between layers is higher. Similarly, reliability of the communications also increases due to reduced usage of TSVs. Note that, reliability is measured using the amount of vertical communication cost which is measured by $Comm_V$ variable discussed in ILP formulation.

In the last set of experiments, the impact of the temperature constraint on the savings is measured. Recall from Table 3 that the default temperature bound used in the experiments so far was 110°C. The bar-chart in Figure 6

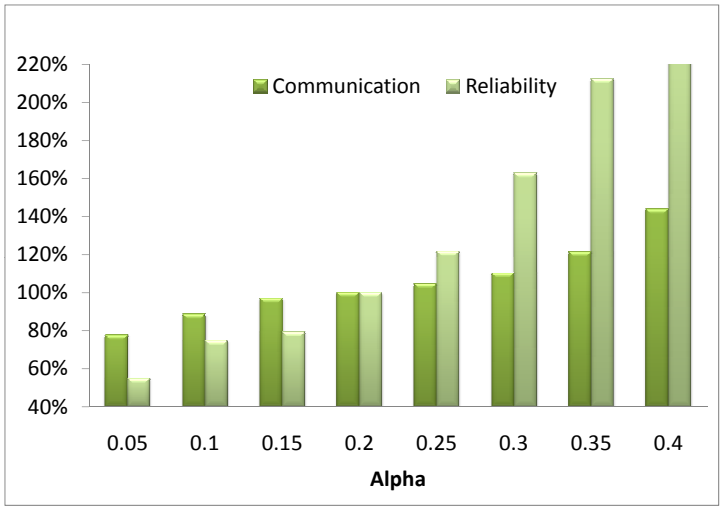

Fig. 5 Normalized reliability-communication costs under the different ϕ values (ammp).

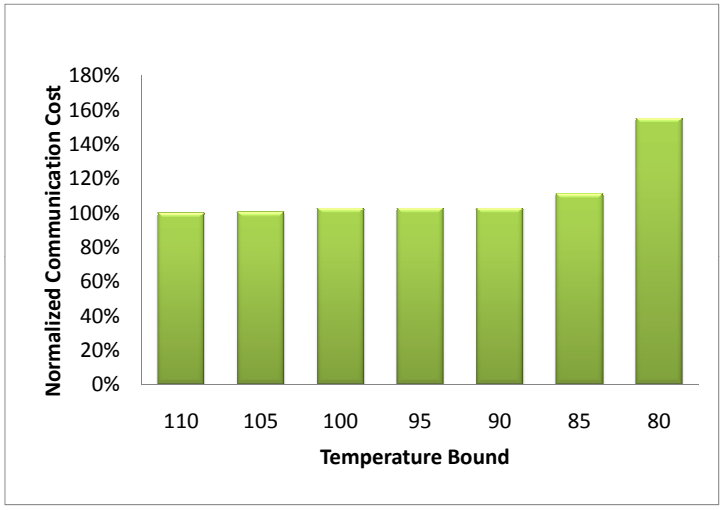

Fig. 6 Normalized reliability-communication costs under the different temperature bounds (ammp).

shows the normalized costs for the benchmark ammp with the different temperature bounds, ranging from 80°C to 110°C. Note that, the values given in this graph are normalized with respect to the default 3D-HT case, where the best results are obtained. As can be seen from this graph, having a tighter temperature bound reduces savings beyond a point. The reason for this behavior is that decreasing the temperature bound also decreases the flexibility in processor core assignment. For this particular example, reducing the temperature bound below 80°C did not return any feasible solution.

6 Conclusion

3D NoCs have been proposed to provide higher performance and lower power consumption by reducing the global interconnect length. However, reliability problem has become more important for 3D ICs with the shrinking technologies. This paper proposes an ILP-based optimal 3D node mapping to maximize reliability while minimizing the communication costs. Experiments indicate that, through effective mapping, it is possible to achieve performance benefits while improving reliability. Although initial experiments are limited to few layers of 3D stacking, it is planned to increase the layers of 3D stacking and test with more complex structures.

References

1. Akasaka, Y.: Three-dimensional ic trends. Proceedings of the IEEE **74**(12), 1703 – 1714 (1986)
2. Alam, S.M., Troxel, D.E., Thompson, C.V.: Circuit and system level tools for thermal-aware reliability assessments of ic designs. Tech. rep. (2004)
3. Andry, P., Sakuma, K., Dang, B., Maria, J., Tsang, C., Patel, C., Wright, S., Webb, B., Sprogis, E., Kang, S., Polastre, R., Horton, R., Knickerbocker, J.: 3d chip stacking technology with low-volume lead-free interconnections. In: Electronic Components and Technology Conference, 2007. ECTC '07. Proceedings. 57th, pp. 627–632 (2007)
4. Davis, W., Wilson, J., Mick, S., Xu, J., Hua, H., Mineo, C., Sule, A., Steer, M., Franzon, P.: Demystifying 3d ics: the pros and cons of going vertical. Design Test of Computers, IEEE **22**(6), 498–510 (2005)
5. ITRS: International technology roadmap for semiconductors
6. Li, F., Nicopoulos, C., Richardson, T., Xie, Y., Narayanan, V., Kandemir, M.: Design and management of 3d chip multiprocessors using network-in-memory. In: Computer Architecture, 2006. ISCA '06. 33rd International Symposium on, pp. 130–141 (2006)
7. Loi, I., Angiolini, F., Benini, L.: Developing mesochronous synchronizers to enable 3d nocs. In: Design, Automation and Test in Europe, 2008. DATE '08, pp. 1414–1419 (2008)
8. Malta, D., Gregory, C., Lueck, M., Temple, D., Krause, M., Altmann, F., Petzold, M., Weatherspoon, M., Miller, J.: Characterization of thermo-mechanical stress and reliability issues for cu-filled tsvs. In: Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st, pp. 1815 –1821 (2011)
9. Minas, N., De Wolf, I., Marinissen, E., Stucchi, M., Oprins, H., Mercha, A., Van der Plaas, G., Velenis, D., Marchal, P.: 3d integration: Circuit design, test, and reliability challenges. In: On-Line Testing Symposium (IOLTS), 2010 IEEE 16th International, p. 217 (2010)
10. Minz, J., Wong, E., Lim, S.K.: Reliability-aware floorplanning for 3d circuits. In: SOC Conference, 2005. Proceedings. IEEE International, pp. 81 – 82 (2005)
11. Murata, H., Fujiyoshi, K., Nakatake, S., Kajitani, Y.: Rectangle-packing-based module placement. In: Computer-Aided Design, 1995. ICCAD-95. Digest of Technical Papers., 1995 IEEE/ACM International Conference on, pp. 472 –479 (1995)
12. Optimization, D.: Xpressmp
13. Ozturk, O., Wang, F., Kandemir, M., Xie, Y.: Optimal topology exploration for application-specific 3d architectures. In: Design Automation, 2006. Asia and South Pacific Conference on (2006)
14. Pavlidis, V., Friedman, E.: 3-d topologies for networks-on-chip. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on **15**(10), 1081–1090 (2007)
15. Sakuma, K.: Development Trend of Three-Dimensional (3D) Integration Technology. IEEJ Transactions on Sensors and Micromachines **131**, 19–25 (2011)

16. Selvanayagam, C., Lau, J., Zhang, X., Seah, S., Vaidyanathan, K., Chai, T.: Nonlinear thermal stress/strain analyses of copper filled tsv (through silicon via) and their flip-chip microbumps. In: Electronic Components and Technology Conference, 2008. ECTC 2008. 58th, pp. 1073 –1081 (2008)
17. Shayan, A., Hu, X., Peng, H., Cheng, C.K., Yu, W., Popovich, M., Toms, T., Chen, X.: Reliability aware through silicon via planning for 3d stacked ics. In: Design, Automation Test in Europe Conference Exhibition, 2009. DATE '09., pp. 288 –291 (2009)
18. Topol, A.W., Tulipe, D.C.L., Shi, L., Frank, D.J., Bernstein, K., Steen, S.E., Kumar, A., Singco, G.U., Young, A.M., Guarini, K.W., Ieong, M.: Three-dimensional integrated circuits. IBM Journal of Research and Development **50**(4.5), 491–506 (2006)
19. Zhao, X., Scheuermann, M., Lim, S.K.: Analysis of dc current crowding in through-silicon-vias and its impact on power integrity in 3d ics. In: Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE, pp. 157–162 (2012)