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

Reliability-Aware 3D Chip Multiprocessor Design

Ismail Akturk · Ozcan Ozturk

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. In this paper, we try to perform this mapping and processor layout effectively. Specifically, on a heterogeneous 3D CMP, we explore how applications can be mapped onto 3D ICs to maximize reliability. Our preliminary experimental evaluation indicates that the proposed technique generates promising results in both reliability and performance.

Keywords Reliability · Multicore · 3D · Data Mapping

1 Introduction

Three-dimensional integrated circuit (3D IC) [2] is an attractive option for overcoming the barriers in interconnect scaling. 3D 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.

As the technology shrinks, one of the challenging problems in the context of 3D Network-on-Chip (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 [5,6,3,1,7,4]. Our goal is to increase the reliability of an application through effective mapping on 3D heterogeneous IC. Our contribution is in two folds:

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

The remainder of this paper is structured as follows. The next section gives the details of our approach, and the paper is concluded in Section 3.

This research is supported in part by a Marie Curie International Reintegration Grant within the 7th European Community Framework Programme.

Ismail Akturk
Bilkent University Computer Engineering Department, Bilkent, Ankara, Turkey
Tel.: +90-312-290 1945
Fax: +90-312-266 4047
E-mail: iakturk@cs.bilkent.edu.tr

Ozcan Ozturk
Bilkent University Computer Engineering Department, Bilkent, Ankara, Turkey
Tel.: +90-312-290 3444
Fax: +90-312-266 4047
E-mail: ozturk@cs.bilkent.edu.tr

2 Our Approach

2.1 Overview

In our framework, information about the set of processor nodes that communicate with each other is passed to the solver/heuristic which determines the location of each node within the 3D NoC. Our goal in selecting the location of each node is to maximize the reliability while keeping the communication cost at reasonable levels.

2.2 Formulation

Our goal in this section is to present a 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. More specifically, we try to map frequently communicating nodes as close as possible to reduce the communication cost. At the same time, we try 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. While overall formulation has more details, for clarity, we only give the important parts of it.

Assume that we are given N number of nodes with dimensions (X_i, Y_i), where $1 \leq i \leq N$. Our approach uses 0-1 variables to place these nodes on the 3D grid with (C_X, C_Y, C_Z) dimensions, and at the end, returns the coordinates of each NoC based CMP node. Note that, C_Z indicates the number of layers in the 3D chip. Communication load between two nodes is expressed by $A_{i,j}$, which indicates the affinity between two nodes.

We define our cost function as the sum of the data communication loads in both vertical and horizontal dimensions. More specifically, we denote the total data communication using $Comm_H$ and $Comm_V$ for horizontal, and vertical communication costs, respectively. Note that, for both communication costs, we use $A_{i,j}$ to express the affinity between two nodes. Consequently, our objective function can be expressed as:

$$\min \quad Comm = Comm_H + \alpha \ Comm_V. \tag{1}$$

Note that, in the objective function given in Expression 1, the difference between horizontal and vertical communication costs is captured by the α parameter which is conservatively set to 2 in our baseline implementation. More specifically, accessing a data from a neighboring node on a different layer is two times costlier than accessing a neighbor on the same layer. This way we are able to penalize the inter-layer transfers. The α parameter can be exercised and the most suitable value can be used, however we do not discuss this any further.

Note that, in our formulation, we employ area and temperature as two main constraints, whereas performance, energy, and communication bandwidth and other possible constraints are left out. For example, depending on the switch present in a node, bandwidth available to the connected links will be limited. Our formulation, in its current form, does not cover this constraint. However, our formulation can easily be modified to include such constraints. In addition to additional constraints, our formulation can also be modified to optimize for a different objective function instead of data communication cost. We do not discuss the details of additional constraints and different objective functions in this paper.

3 Conclusion

Reliability problem has become more important for 3D ICs with the shrinking technologies. This paper proposes an optimal 3D node mapping to maximize reliability while minimizing the communication costs.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. Minz, J., Wong, E., Lim, S.K.: Reliability-aware floorplanning for 3d circuits. In: SOC Conference, 2005. Proceedings. IEEE International, pp. 81 – 82 (2005)
6. 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)
7. 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)