Exploration of Temperature Constraints for Thermal-Aware Mapping of 3D Networks-on-Chip

Parisa Khadem Hamedani, The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

Natalie Enright Jerger, The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

Shaahin Hessabi, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Hamid Sarbazi-Azad, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, & Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

ABSTRACT

This paper proposes three ILP-based static thermal-aware mapping algorithms for 3D Networks-on-Chip (NoC). With these three mapping algorithms, the authors explore the thermal constraints and their effects on temperature and performance. Through complexity analysis, the authors show that the first algorithm, an optimal one, is not suitable for 3D NoCs. Therefore, the authors develop two approximation algorithms and analyze their algorithmic complexities to show their proficiency. According to simulation results, mapping algorithms that employ direct thermal calculation to minimize the temperature reduce the peak temperature by up to 24% and 22%, for the benchmarks that have the highest communication rate and largest number of tasks, respectively. This peak temperature reduction comes at the price of a higher power-delay product. The authors' exploration shows that considering power balancing early in the mapping algorithm does not affect chip temperature. Moreover, the authors show that considering explicit performance constraints in the thermal mapping has no major effect on performance.

Keywords: *3D Integrated Circuits (3D ICs), Integer Linear Programming, Mapping, Networks-on-Chip (NoC), Temperature*

INTRODUCTION

With rapid advances in integrated circuit manufacturing technology, the number of processing cores in systems-on-chip is continuously increasing. Due to technology scaling, interconnect has become a bottleneck for performance; wires have not scaled at the same rate as logic with each technology generation. Furthermore, the rapid increase in the capability and complexity of applications demand greater interconnect bandwidth which leads to

DOI: 10.4018/jaras.2013070103

an increase in the density and complexity of on-chip interconnects. Two solutions to alleviate the interconnect problems are NoCs (Jantsch, 2003) and 3D integrated circuits (3D ICs) (De Micheli, 2009; Pavlidisi, 2009). By combining these two techniques, we take advantage of their different benefits (De Micheli, 2009). In a 3D IC, various layers of components are stacked vertically on each other; communication across layers is through vertical interconnects. Advantages of 3D ICs include: abatement of overall interconnection length, scaling of chip dimensions, decrease of interconnect power and the possibility of mixing different technologies (e.g. logic, analog, DRAM) on one chip. Although a 3D IC has many advantages, it also has some disadvantages (De Micheli, 2009; Pavlidisi, 2009). For example, thermal management is a much more significant issue than for 2D ICs. Although the power consumption of a 3D IC is less than that of the equivalent 2D chip due to the decrease in average interconnection length, power density of a 3D IC is higher due to smaller chip area and the existence of the vertical layers. Also, the decrease in overall interconnect length can improve the operating frequency of a 3D chip, which can also contribute to the growth of chip power density (De Micheli, 2009; Pavlidisi, 2009).

In addition, the increase in transistor count and the higher operating frequency increase chip power consumption, which results in a higher power density on chip, which in turn increases the chip temperature. High temperature has several undesirable effects on nano-scale designs, such as performance degradation of transistors, increased static power consumption, increased RC delay in interconnects, and reliability issues such as reduced Mean Time To Failure (MTTF) and thermo-migration (Tu, 2011; Pedram, 2006). As a result of these effects, designing and utilizing temperature control algorithms for SoCs is essential.

There are several methods to control chip temperature. These methods can be classified into three categories. The first group of methods is devised and implemented before the chip is built, such as thermal aware placement (Cong, 2007), and insertion of through silicon vias (TSVs) (Cong, 2005). The second group contains methods, such as DVFS (Murali, 2007), that are devised before the chip is built, but whether or not they are used (and how) depends on the runtime conditions of the system. Finally, the third group represents post-silicon techniques that do not require extra effort during hardware design Methods such as task mapping (Addo-Quaye, 2005) and migration (Donald, 2006) algorithms manage chip temperature while the system is running.

A mapping algorithm decides how various tasks are assigned to cores based on optimization criteria (Marculescu, 2009). These criteria can include minimizing the total network communication rate (Murali, 2004), minimizing the network communication energy (Hu, 2005), decreasing network congestion (Chou, 2008) and network thermal balancing (Hung, 2004; Zhou, 2006; Addo-Quaye, 2005). Genetic algorithms have been proposed for thermal-aware mapping to reduce the peak temperature in 2D NoCs (Hung, 2004; Zhou, 2006); thermal- and communication-aware mapping for 3D NoCs can be performed using genetic algorithms (Addo-Quaye, 2005).

The criteria used by the mapping algorithm impacts the quality of the obtained results and the complexity of the algorithm. Therefore, we need to explore temperature constraints for thermal-aware mapping. We explore answers to the following questions: Is it important for chip temperature that the mapping algorithm consider thermal constraints, such as power or temperature? Which thermal constraints are sufficient to achieve an acceptable chip temperature: implicit constraints such as power estimation, or explicit constraints such as temperature? To what extent do these constraints affect the chip temperature? Do these constraints negatively affect performance and power consumption? Is an implicit constraint such as router power consumption, which is used to estimate the temperature of a tile, sufficient to obtain the optimal power and performance?