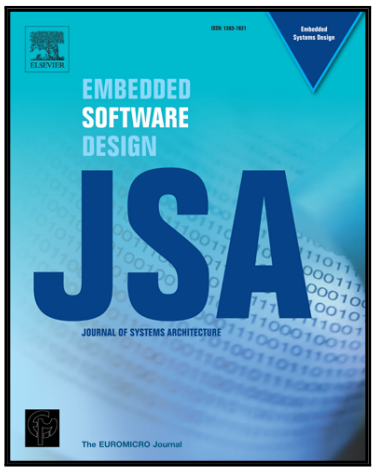

Energy and Performance-Aware Application Mapping for Inhomogeneous 3D Networks-on-Chip

Michael Opoku Agyeman, Ali Ahmadinia, Nader Bagherzadeh

PII: S1383-7621(17)30156-X
DOI: https://doi.org/10.1016/j.sysarc.2018.08.002
Reference: SYSARC 1516

To appear in: *Journal of Systems Architecture*

Received date: 22 March 2017
Revised date: 2 August 2018
Accepted date: 5 August 2018

Please cite this article as: Michael Opoku Agyeman, Ali Ahmadinia, Nader Bagherzadeh, Energy and Performance-Aware Application Mapping for Inhomogeneous 3D Networks-on-Chip, *Journal of Systems Architecture* (2018), doi: https://doi.org/10.1016/j.sysarc.2018.08.002

Energy and Performance-Aware Application Mapping for Inhomogeneous 3D Networks-on-Chip

Michael Opoku Agyeman, Ali Ahmadinia, Nader Bagherzadeh

***Abstract*—Three dimensional Networks-on-Chip (3D NoCs) have evolved as an ideal solution to the communication demands and complexity of future high density many core architectures. However, the design practicality of 3D NoCs faces several challenges such as thermal issues, high power consumption and area overhead of 3D routers as well as high complexity and cost of vertical link implementation. To mitigate the performance and manufacturing cost of 3D NoCs, inhomogeneous architectures have emerged to combine 2D and 3D routers in 3D NoCs producing lower area and energy consumption while maintaining the performance of homogeneous 3D NoCs. Due to the limited number of vertical links, application mapping on inhomogeneous 3D NoCs can be complex. However, application mapping has a great impact on the performance and energy consumption of NoCs. This paper presents an energy and performance aware application mapping algorithm for inhomogeneous 3D NoCs. The algorithm has been evaluated with various realistic traffic patterns and compared with existing mapping algorithms. Experimental results show NoCs mapped with the proposed algorithm have lower energy consumption and significant reduction in packet delays compared to the existing algorithms and comparable average packet latency with Branch-and-Bound.**

***Index Terms*—Multi-core Architectures; Network-on-Chip; 3D Integration**

I. Introduction

Current advances in semiconductor integration and processing capability of System-on-Chip (SoC) enable a large number of high performance cores to be integrated into a single chip. Networks-on-Chip (NoC) have been proposed to adopt the idea of networking in the macro-world for current and future SoCs but with limited resources and tighter constraints. 3D IC fabrication has also emerged to stack several layers of 2D ICs vertically, providing shorter vertical links with enhanced connectivity and lower delays. Combining 3D IC fabrication and Network-on-Chip technique, 3D NoC creates new design and research opportunities and challenges for high density SoCs. High integration and performance opportunity of 3D NoCs encourage enhanced heterogeneity of application design with implementation of multiple applications on a SoC.

Conventionally, homogeneous 3D NoCs have been employed for 3D-Integration where 3D routers are employed for inter-layer and planar communication. However, the 3D routers have a larger area and power consumptions than a 2D router with a similar architecture. Consequently, the homogeneous 3D router distribution may lead to significant area and power overheads if applied to applications whose communication patterns vary significantly among embedded cores. Moreover, Through Silicon Via (TSV) which has been accepted as a viable inter-layer wiring technique has a complex and expensive manufacturing process [1], [2]. To optimize the performance and manufacturing cost of 3D NoCs with minimal distortion to the modularity, inhomogeneous architectures have been proposed to combine 2D and 3D routers in 3D NoCs [3], [4], [5], [2], [6]. Several inhomogeneous 3D architectures focusing on different NoC router architectures, minimal hop-count between 2D and 3D routers in each layer, and uniform distribution of 2D and 3D routers have been proposed [7], [2], [8]. However, due to the limited number of 3D routers and vertical links, mapping of applications to an inhomogeneous 3D NoC can be challenging. Specifically, different applications have variable characteristics at design time and run-time with possible link, node or task failures due to current technology limitations.

In this paper, an energy-aware application mapping technique for inhomogeneous 3D NoCs is proposed. The proposed algorithm efficiently maps a given application with minimized communication energy while maintaining the performance. The application must be known and its communication task graph much be constructed for this approach, similar to the applied benchmarks [34][56-58] as has been used in literature [3-9][13-18]. Consequently, the mapping algorithm involves three main stages:

1) Initial NoC size determination and clustering which assign regions along the 3D NoC vertices to reduce the cost of 3D routers.
2) Architecture matching stage which allocates the application graph to the clusters in the 3D NoC such that optimized numbers of 2D and 3D routers are assigned by enumerating the communication bandwidth and energy constraints in each cluster. Thus, this stage re-evaluates and assigns a suitable NoC size and automatically determines the suitable number of 3D and 2D routers and their assigned tiles.
3) Task to NoC region mapping which assigns the IP cores to the tiles in different clusters with minimized total energy consumption and maximized performance. This stage efficiently exploits area and power efficiencies of 2D routers as well as the higher bandwidth and lower latency characteristics of the vertical links associated with 3D routers.

Corresponding Author: aahmadinia@csusm.edu

M. O. Agyeman is with Department of Computing and Immersive Technologies, University of Northampton, UK

A. Ahmadinia is with Department of Computer Science, California State University San Marcos, San Marcos, CA 92078, USA

N. Bagherzadeh is with Dept. of Electrical Engineering and Computer Science, University of California, Irvine, Irvine, CA 92697, USA