Journal of Computer and System Sciences 79 (2013) 475–491

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Cluster-based topologies for 3D Networks-on-Chip using advanced inter-layer bus architecture

Masoumeh Ebrahimi, Masoud Daneshtalab *, Pasi Liljeberg, Juha Plosila, Hannu Tenhunen

Department of Information Technology, University of Turku, Finland

ARTICLE INFO

Article history:
Received 30 December 2010
Received in revised form 30 April 2011
Accepted 26 September 2012
Available online 9 October 2012

Keywords:
Three-dimensional Networks-on-Chip
Bus
Network topology
Inter-layer communication

ABSTRACT

Three-dimensional integrated circuits (3D ICs) have emerged as a viable candidate to achieve better performance and packaging density as compared to traditional two-dimensional (2D) ICs. In addition, combining the benefits of 3D ICs and Networks-on-Chip (NoCs) schemes provides a significant performance gain for 3D architectures. In recent years, through-silicon-via (TSV), employed for inter-layer connectivity (vertical channel), has attracted a lot of interest since it enables faster and more power efficient inter-layer communication across multiple stacked layers. The router-based and bus-based organizations are the two dominant architectures for utilizing TSVs as inter-layer communication channel in 3D architectures. Both approaches have some disadvantages. The former suffers from poor scalability and deteriorates the performance at high injection rates, and the latter consumes more area and power. The area overhead of TSVs reduces wafer utilization and yield, which can impact designing 3D architectures with a large number of TSVs. In this paper, two mesh-based topologies for 3D architectures are introduced to mitigate TSV footprint and power dissipation on each layer with a small performance penalty. On top of that, we propose a novel pipeline bus structure for inter-layer communication to improve the performance by reducing the delay and complexity of traditional bus arbitration.

1. Introduction

As technology geometries have shrunk to the deep submicron, the communication delay and power consumption of global interconnections of high performance Multi-Processor Systems-on-Chip (MPSoCs) are becoming a major bottleneck [1,2]. The Network-on-Chip (NoC) architecture paradigm, based on a modular packet-switched mechanism, can address many of the on-chip communication design issues such as performance limitations of long interconnects, and integration of a large number of intellectual property (IP) cores in a chip [3–5]. However, two-dimensional (2D) chip fabrication technology is facing several challenges in the deep submicron regime even when utilizing NoC architectures [6,7], e.g. designing the clock-tree network for a large chip, having limited floor-planning choices, increasing the wire delay and power consumption, integrating diversity components that are digital, analog, MEMS RF, etc.

The three-dimensional (3D) integration has emerged as a potent solution to address these problems and the design complexity of MPSoC in 2D Integrated Circuits (IC). 3D ICs reduce the interconnect delay by stacking vertically active silicon layers as well as offering a number of advantages over the traditional 2D chip [7–10]: (1) shorter global interconnects; (2) higher performance; (3) lower interconnect power consumption due to wire-length reduction; (4) higher packing density

* Corresponding author.
E-mail addresses: masebr@utu.fi (M. Ebrahimi), masdan@utu.fi (M. Daneshtalab), pakrli@utu.fi (P. Liljeberg), juplos@utu.fi (J. Plosila), hanten@utu.fi (H. Tenhunen).

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2012.09.005

and smaller footprint; and (5) support for the implementation of mixed-technology chips. In this paper we focused on wafer stacking technology. In wafer-to-wafer bonding technology, which is one of the popular options for 3D integrations, dies are vertically stacked. Short, fat, and vertical Through Silicon Vias (TSVs) are exploited for inter-layer communication. The distance between wafers can range from 5 μm to 50 μm [10,12], which is much shorter than the wire length between cores on a tier, and the pitches of TSVs can range from 1 μm to 10 μm square [10,12]. That is, the wire delay, power consumption and chip forming factor are significantly reduced [13,14,16].

3D NoC topologies not only enable scalable networks to provide communication requirements in 3D ICs [7–10], but also are a crucial factor of 3D chips in terms of performance, cost, and energy consumption [7]. Various on-chip network topologies have been studied for 3D NoCs [7–11,13,15,17]. Mesh-based structures are popularly used in 3D systems, because their grid-based regular architecture is intuitively considered to be matched to the 2D VLSI layout for each stack layer [7–10,13]. Nevertheless, if the number of IP-cores and memories increases in each layer, more TSVs are necessitated to handle the inter-layer communication. Inasmuch as each TSV employs a pad for bonding, the area footprint of TSVs in each layer is augmented significantly [10,17]. The main contributions of this work are twofold.

First, we propose two novel stacked mesh topologies to reduce the area overhead of TSVs and power dissipation with a small performance penalty. The proposed stacked mesh topologies, named Clustered Mesh Inter-layer Topology (CMIT) and Concentrated Inter-layer Topology (CIT), benefit of clustering the mesh topology for each layer. Each cluster of the presented topologies has its dedicated vertical channel, composed of a set of TSVs. CMIT and CIT preserve the advantages of the clustered mesh topology and mitigates both power density and TSV area footprint on each layer. However, since in both CMIT and CIT topologies, multiple IPs/routers in each layer can be connected to the same vertical channel, conventional bus architectures are not efficient for inter-layer connections due to the following reasons. Conventional buses require a centralized arbiter to control access requests from different routers to the bus necessitating many control wires between each layer and the central arbitration module. Also, they suffer from the limited parallel accesses to a vertical bus which can degrade the inter-layer communication performance.

The second contribution of the paper is to introduce a novel pipeline bus structure to solve the problems imposed by traditional bus architectures for inter-layer communication in 3D ICs. The proposed bus structure allows simultaneous transmissions over the bus without using centralized arbitration unit and overcomes the drawbacks of previously presented bus structures for vertical channels. This distributed arbitration mechanism improves the performance by reducing the delay and complexity of traditional bus arbitration while reducing the wiring overhead. Moreover, the presented bus architecture takes advantage of a fair forwarding policy along with a non-blocking scheme to share the inter-layer bandwidth among different layers efficiently.

Finally, the presented pipeline bus structure can utilize bi-synchronous FIFO for synchronization between stacked layers, if each layer is fabricated by different technologies.

The rest of the paper is organized as follows. In Section 2, the background is discussed and a brief review of related works is presented in Section 3. The proposed cluster architectures and the pipeline bus architecture are presented in Sections 4 and 5, respectively. The experimental results are discussed in Section 6 while the summary and conclusion are given in the last section.

2. Background

2.1. 3D IC technology overview

There are many technologies for die stacking being pursued by industry and academia. Wafer-Bonding [19,20] and Multi-Layer Buried Structures (MLBS) [21,22] are the most promising ones. The details of these processes are described in [6]. Wafer-to-wafer bonding appears to be the leading contender in industry and many recent academic studies have assumed this type of 3D stacking technology [6–11,23].

Wafers can be stacked either Face-to-Face (F2F) or Face-to-Back (F2B) and both have pros and cons. While the former provides the greatest layer-to-layer via density, it is suitable only for two-layers; and additional layers would have to employ back-to-back placement using larger and longer vias. On the other hand, Face-to-Back provides uniform scalability to an arbitrary number of layers, despite a reduced inter-layer via density [15–24]. Layers, stacked on top of each other, are connected via vertical interconnects tunneling through them. Wire bonding, micro-bump, contactless, and TSV are some of the vertical interconnect technologies that have been used in stacked structures [21]. The TSV interconnection has the potential to offer the greatest vertical interconnect density and therefore is the most promising one among these vertical interconnect technologies [15–24]. In this work, we assumed the F2B method with TSV interconnects to provide more scalability when more than two layers are employed.

2.2. 3D NoC architecture

3D-Symmetric NoC and 3D NoC-Bus Hybrid (stacked mesh) structures are popularly used in 3D systems, because their grid-based regular structure is intuitively considered to match the 2D VLSI layout for each layer [7–10,13]. The 3D-symmetric NoC structure, shown in Fig. 1(a), is an extension of 2D mesh by adding two additional physical ports to each baseline-router (one for up and one for down) in the popular 2D mesh-based system [7,10]. Adding two additional ports requires