Microprocessors and Microsystems 48 (2017) 36–47

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Area and power savings via asymmetric organization of buffers in 3D-NoCs for heterogeneous 3D-SoCs

Jan Moritz Joseph[a,*], Christopher Blochwitz[b], Alberto García-Ortiz[c], Thilo Pionteck[a]

[a] Otto-von-Guericke-Universität Magdeburg, Institut für Informations- und Kommunikationstechnik, 39106 Magdeburg, Germany
[b] Universität zu Lübeck, Institut für Technische Informatik, 23562 Lübeck, Germany
[c] University of Bremen, Institute of Electrodynamics and Microelectronics, 28359 Bremen, Germany

ARTICLE INFO

Article history:
Received 22 February 2016
Revised 14 June 2016
Accepted 28 September 2016
Available online 29 September 2016

Keywords:
Network-on-Chip
Heterogeneous 3D-System-on-Chip
Asymmetric 3D-Noc
Buffer reorganization
Buffer depths

ABSTRACT

In this paper we investigate the effects of asymmetric organization and depths of Network-on-Chip (NoC) router buffers among dies in heterogeneous 3D-System-on-Chips (SoCs). In our novel approach the properties of the routers are aligned with the characteristics of the technological nodes per layer. We call these designs Asymmetric 3D-NoCs (A-3D-NoCs). In this work we demonstrate potentials of A-3D-NoCs in comparison to a conventional, symmetric 3D-NoC: Applying asymmetric buffer reorganization we achieve area savings of 8.3% and power savings of 5.4% for link buffers while accepting a minor average system performance loss of 2.1%. With additional asymmetry in buffer depth up to 28% cost savings and 15% power reduction are given in combination with a 4.6% performance decline. Thus, the proposed buffer organization scheme is applicable for cost and power critical applications of NoCs in heterogeneous 3D-SoCs.

1. Introduction

The increasing requirements of high performance, low power consumption, reduced area footprint, and the need for higher degrees of re-usability are steadily increasing the complexity of chip design. Traditional 2D-architectures struggle to keep up with these demands. A promising approach is the stacking of multiple silicon dies resulting in 3D-structures. 3D-integration provides multiple advantages over 2D-architectures, e.g. reduced power consumption and costs and increased performance [1]. There are critical challenges as well, such as thermal issues, low yield, and design complexity [2]. As pointed out by [3], the final goal of 3D-technology is to allow heterogeneous integration. Dies with different characteristics and technologies, such as analog, mixed-signal, logic, and memory, are stacked on each other and are connected by Through-Silicon Vias (TSVs) [4]. Hence, the technology of single dies can be aligned with the requirements of their components.

To exploit the potential of 3D-SoCs, 3D-Network-on-Chips (NoCs) offer a powerful, flexible, and scalable interconnect architecture. In addition, heterogeneous 3D-SoCs are intrinsically asynchronous and NoCs facilitate communication through different clock domains. An important issue for 3D-NoC-design has not sufficiently been considered in the literature so far: dedicated, asymmetric router architectures for 3D-NoCs that exploit the heterogeneity of 3D-SoCs. Therefore, we introduce the novel term **asymmetric 3D-NoCs (A-3D-NoCs)**. This is in contrast to the concept of *heterogeneous 3D-NoC* [5,6], which denotes NoC designs with non-uniform properties at the architectural level. These do not consider the manufacturing technologies. A-3D-NoCs furthermore add an additional design dimension to *homogeneous 3D-NoCs*, in which 2D-designs are extended by a dimension, yet not by new manufacturing technologies (e.g. see [7]): Existing works tacitly assume a multilayer homogeneous 3D-SoC that disregards the technological asymmetry intrinsically present in heterogeneous 3D-SoCs. Because of this heterogeneity, the actual costs and constraints of the communication infrastructure in each die are different. However, most of the existing works on 3D-NoCs, as pointed out in [8], only exploit incremental advantages of 3D-technology without addressing its unique features. These features, however, provide a large optimization potential [9]. For instance, different degrees of heterogeneity in the architecture can be further exploited, e.g. the overall costs of the TSV-redundancy schemes required for yield improvement [10] are decreased. Moreover, 3D-technologies allow larger

* Corresponding author.
E-mail addresses: jan.joseph@iovgu.de (J.M. Joseph), blochwitz@iti.uni-luebeck.de (C. Blochwitz), agarcia@item.uni-bremen.de (A. García-Ortiz), thilo.pionteck@ovgu.de (T. Pionteck).
URL: http://www.iikt.ovgu.de (J.M. Joseph), http://www.iti.uni-luebeck.de (C. Blochwitz), http://www.item.uni-bremen.de (A. García-Ortiz), http://www.iikt.ovgu.de (T. Pionteck)

http://dx.doi.org/10.1016/j.micpro.2016.09.011
0141-9331/

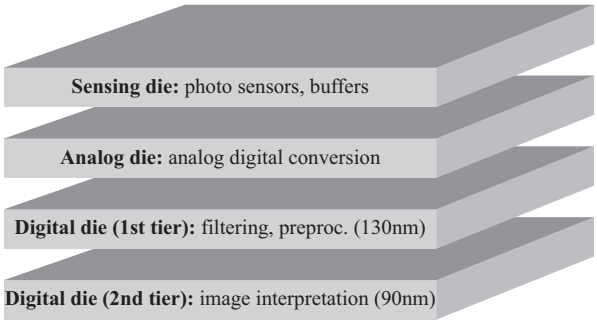

Fig. 1. 3D-VSoC design [16].

degrees of freedom in the system design. For example, previously discarded topologies for NoCs are feasible again, such as torus [11], since the wire-length constraints and layout complications of 2D-designs are reduced [12].

This paper extends our work presented in [13]. There, we focused on buffer distribution in A-3D-NoCs as one aspect to demonstrate their advantages in heterogeneous 3D-SoCs. Placing as many buffers as possible in a layer, which is optimized for memory, was shown to be advantageous for power consumption and costs. Therefore, we introduced two optimization techniques for router architectures with asymmetric buffer reorganization among dies. Here, we provide the following novel aspects to extend the previous work: We determine the influence of differing buffer depths per layer for the proposed reorganization technique. Adding this new degree of asymmetry to 3D-NoC design, we analyze architectural metrics such as area, power consumption, and performance. We evaluate the influence of model simplifications by extending our performance evaluation with simulations of different clock speeds per layer. Furthermore, we conduct more detailed benchmarks including synthetic traffic patterns with multiple temporal and spatial distributions [7]. We relate costs to performance loss from a system-level point of view. NoC design paradigms, as presented here, consider technological characteristics of heterogeneous SoCs. The presented results show the potentials of asymmetries in communication infrastructures of heterogeneous SoCs, although we still consider only a subset of asymmetric NoC design options. To the best of our knowledge comparable approaches for exploiting the optimization potential for NoCs in heterogeneous 3D SoCs do not exist in literature so far.

2. Related work

2.1. Heterogeneous 3D-SoCs

Heterogeneous 3D-SoCs offer architectural advantages and the significant promises of the 3D-technology have generated a plethora of different architectures and paradigms. For instance, dedicated and interleaved dies with either memory or processing units offer a better performance than dies that implement both elements side by side [14]. Heterogeneity leads to higher energy efficiency in server multi-core processors as well [15]. As another example, as part of 3D-Vision-System-on-Chips (3D-VSoCs) [16], conservative mixed-signal technologyies (e.g. 130 nm) are connected with a high-speed digital technologies (e.g. 65 nm). In general, the layers in a VSoC are optimized for their tasks: VSoCs consists of an (analog) photo sensor array, a signal conversion layer with mixed-signal processor arrays and analog digital (AD) converters, a layer for frame buffers and switch matrices, and, on the fourth layer, a foveal processor array. Such an example is shown in Fig. 1. The CMOS sensor on top generates data which are converted into digital signals on the second layer. In the first tier digital chip, image filtering and pre-processing is computed using a slow digital technology. The second tier digital plane allows for complex operations such as image interpretation, using a high speed digital technology. In [17] the conceptual heterogeneous 3D-SoC “eCube” was introduced. It targets small, self-sustaining, and interconnected chips and, thus, implements power, application, and radio layers within a single SoC. Using heterogeneous integration, each layer is optimized for its function. Other examples are 3D-DRAM subsystems [18] and 3D-FPGAs [19]. 3D-technology is steadily gaining in maturity. It is already used in commercial applications such as the hybrid memory cube (HMC) by Mircon, which is a stack of four DRAM dies on a logic (SerDES) chip containing interfaces [20].

2.2. 3D-NoC designs and evaluation

In general, performance, power consumption, and area footprint are the three main design metrics that are optimized for 2D- and 3D-NoCs. Increased throughput and reduced communication latency are the fundamental measures for an increased NoC performance. Optimizations for homogeneous 2D-NoC routers are also applicable for (homogeneous) 3D-systems. For instance, in [21], a 3D-routing algorithm is presented, which reduces the system latency and increases the throughput by approximately 45% using look-ahead strategies. These results are similar to 2D-look-ahead routing algorithms [22]. However, measuring the performance of interconnects is far from trivial since significant and comparable results require standard benchmarks. To the best of our knowledge, there is no standard benchmark for communication infrastructures of heterogeneous 3D-SoCs. Generally, there are three options for NoC benchmarking: First, synthetic traffic patterns can be used as they provide basic comparability. However, they do not reflect properties of real-world applications. Second, there are benchmark suits such as PARSEC [23], which, in combination with full system simulation (e.g. co-simulation with Gem5 [24]) or traces with dependency tracking [25], can also be used for performance evaluation of heterogeneous 3D-SoCs. Third, abstract task graph modeling [26–28] is another benchmarking option and typical network loads for 2D-SoCs are available [26,27]. However, all of these benchmark methods are tailored for 2D-SoCs’ applications and the lack of standard 3D-SoC models limits the significance and comparability of the results.

Numerous simulators have been proposed for performance evaluation of NoCs. Common simulators for 2D-NoCs (e.g. Booksim [29] or Noxim [30]) have a wide variety of features. Many network parameters can be adjusted such as router buffer depths, topologies, and routing algorithms. Using standard benchmarks on the basis of synthetic traffic patterns, the network’s performance and energy consumption can be estimated. In general, these 2D-simulators allow for an extension to 3D-SoCs yet are not specifically designed with this focus. Particular 3D-NoC simulators are published as well, e.g. [31], which do not target manufacturing technology-specific features of individual dies. This is only included in the simulator, which is developed as part of [9]. Still, this NoC model is implemented for a non-standard router design and thus lacks of generality. However, for the evolution of asymmetric 3D-NoCs a general NoC simulator is important, which includes manufacturing technology specific features of chip layers. Thus, as part of this work, a simulator is implemented to cover asymmetric buffer distributions in NoCs for heterogeneous 3D-SoCs.

Performance and power consumption of interconnections are tightly coupled. Power consumption can be optimized through novel NoC designs mainly by the adaption of router architectures. As one example for 2D-designs, in [32] multiple low-power techniques are implemented such as partially activated crossbar, clock frequency scaling, and serial link coding. Similar strategies can be applied for 3D-SoCs as well. Another approach is to change the