High-Performance TSV Architecture for 3-D ICs

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

Department of Information Technology, University of Turku, Finland
{ masdan, masebr, pakrli, juplos, hatenhu}@utu.fi

Abstract- *Three-dimensional integrated circuits (3-D ICs) outperform traditional planar ICs in terms of performance, packaging density, interconnection power consumption, and functionality. Since the performance of 3-D ICs employing Through Silicon Vias (TSVs) depends on vertical interlayer interconnects, in this paper we present a high-performance bus architecture for TSVs.*

I. INTRODUCTION

3D integration has emerged to reduce the interconnect delay problem by stacking vertically active silicon layers [1][2][3]. 3D ICs offer a number of advantages over the traditional 2D design: (1) shorter global interconnects; (2) higher performance; (3) lower interconnect power consumption due to wire-length reduction; (4) higher packing density and smaller footprint; and (5) support for the implementation of mixed-technology chips. The power-performance and efficiency of 3D-NoCs essentially depend on the underlying topology formation. On-chip network topology is a crucial factor of the 3D chips in terms of performance, cost, and energy consumption [2][3]. Various network topologies have been studied for 3D-NoCs. 3D Symmetric NoC and 3D NoC-Bus Hybrid (stacked mesh) are popularly used in NoCs, because their grid-based regular structure is intuitively considered to be matched to the two-dimensional VLSI layout for each stack layer [1][2][3]. By adding two additional physical ports to each baseline-router (one for up and one for down) in the popular 2D-mesh structure the 3D symmetric NoC is formed (Fig. 1(a)). This structure suffers from: 1) adding two additional ports requires larger crossbar incurring significant area and power overhead. That is, the more input ports employed the more blocking probability occurred inside the router. 2) Since the distance between layers is very short and the vertical interlayer links are very fast, placing a router as a hop on the vertical path is not desirable because each flit must pass buffering and arbitration at every hop. This increases the latency of moving flits through the upward and downward paths. Thereby, a single-hop communication (bus) among layers is a feasible solution. A router in this structure has at most 6 ports, one to the IP-core, one to the bus, and four for cardinal directions. This structure is called 3D NoC-Bus Hybrid (Fig. 1(b)). Due to the fact that the 3D IC is emerging as a promising solution to continue the growth of the number of transistors on a chip, in this research work we explore the topology formation and communication routing protocols in three dimensional ICs.

Our contribution of this work is to propose a new 3D topology based on stacked mesh architecture. In this topology, a high performance pipeline bus structure is introduced to overcome the drawbacks of the previously presented bus structures such as segmented bus [5], SAMBA bus [5], and dTDMA bus [6] to integrate multiple layers in 3D ICs. The benefits of the proposed bus structure are as follows:

First, this novel bus architecture, improve the performance, by reducing the delay and complexity of previously proposed bus' arbitration module which is the foremost impediment in bus communications.

Second, if in some cases, each layer is built by different vendors and with completely different processes, each layer should have its own clock tree with associated clock buffers [7]. Furthermore, because there is no clear solution of modular and skew-free clock distribution in 3D ICs, a clock synchronization mechanism between active layers through vertical connections (interlayer vias) is required [7]. In this work, we introduced a programmable synchronous/asynchronous pipeline-bus structure to cope with the mesochronous communication. If communications among layers suffer from clock skewing, the asynchronous mode of the bus will be programmed. This structure can also be employed as an interlayer via to handle the communication among layers with different clock frequencies in 3D chips.

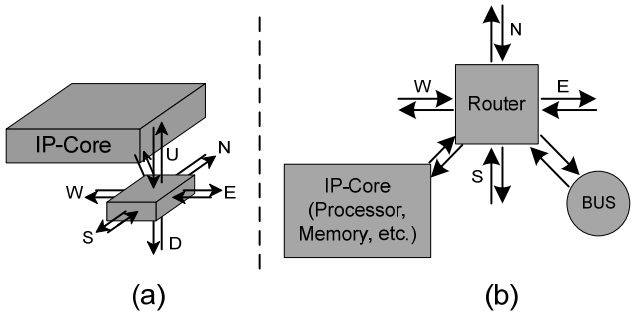

Fig. 1. Mesh-based NoC architectures: (a) 3D symmetric NoC (b) 3D NoC-Bus Hybrid**.**

II. PROPOSED INTERLAYER BUS ARCHITECTURE

Traditionally, a bus is described as a shared link which can be owned by one attached subsystem at the time, i.e., when one module is transmitting via the bus, the others can only be receiving. Parallelism can be added to the structure by partitioning the bus into segments with bridges and allowing these segments to operate concurrently [2]. However, on one side, the overall system performance in such designs is still limited by the lack of parallel bus transactions. And on the other hand, because of using many control wires for the central arbitration in such segmented buses, it is not a suitable approach to use them as vertical bus (TSVs) in 3D ICs. Our solution for these bottlenecks for vertical buses is to consider the system bus with a bidirectional pipeline which is capable of transferring data concurrently from one or more sources to several destinations. The proposed architecture is illustrated in Fig. 2. The system is partitioned into a set of modules each of which forms its own timing domain. In fact, each module is used to connect the corresponding layer to the pipeline bus. As the system is based on GALS (Globally Asynchronous Locally Synchronous) design paradigm, the layers can internally operate at different clock frequencies.

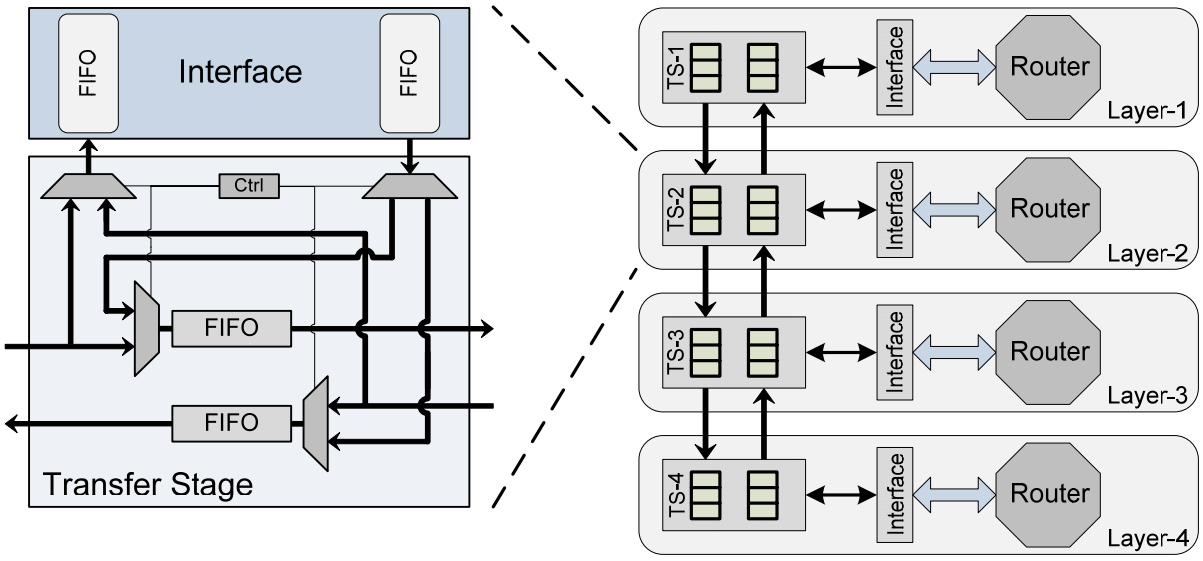

Fig. 2. Proposed Bus Architecture.

The interface between the layers can be either self-timed, i.e., based on asynchronous handshake signaling, or synchronous in which case the interface forms a clock domain of its own. The layers are independent of each other, in case there are some interlayer transactions, the layers exchange data asynchronously through the pipelined system bus, a self-timed segmented communication link which allows simultaneous transfer in both directions. The layers can concurrently access the bus without waiting for any grant signals, because of the pipelined structure of the proposed bus. The interface module role is to act as a dedicated adapter between the internal, possibly synchronous timing domain of the layer and the self-timed domain of the bus. Indeed, it acts as a synchronizer between the chip layer and the pipeline bus. To form the pipelined bus, the physical wires that implement the bus are divided into a set of segments separated from each other by *Transfer Stages (TS)*, one attached to each layer (Fig. 2). Each transfer stage contains internal FIFO queues for pipelining the data flow, and a bus segment between adjacent stages consists of two separate unidirectional point-to-point interconnects which transfer data asynchronously (or synchronously) between the stages in opposite directions. These two links of a segment can operate in parallel, and due to pipelining, all segments of the system bus can transfer data simultaneously. Each layer has a unique address for inter-layer communication. Furthermore, each IP-core in a layer has its own address which makes addressing of a specific module in a given cluster possible. Hence, a datagram propagating along the bus has a header containing both the layer address and the IP-core address. The former is analyzed at each transfer stage, and the latter is decoded by routers in each layer.

A transfer stage has three different functions:

1. It forwards incoming data from the preceding stage to the next stage through a buffer, in both directions.
2. If incoming data from an adjacent transfer stage is intended to be processed by the layer, the transfer stage transfers data to the interface module of the layer through a FIFO queue.
3. When the layer decides to send data to another layer, the transfer stage operates as an output buffer. This means that it takes care of first receiving data from the interface module of the attached layer through a FIFO queue and then sending this data to one of the two adjacent transfer stages, depending on the direction in which, the target layer is located.

III. RESULTS

We have developed a general propose cycle-accurate 3D network on chip simulator. We used an AXI-based NI [4], to bring backward compatibility with existing IP cores. Also, the 3D Symmetric and Hybrid (stacked mesh) are implemented. The Hybrid architecture has been implemented by three different bus structures which are segmented bus, dTDMA bus, and the proposed pipeline bus structure. The 3-D configuration is 4x4x4. Under uniform traffic, the proposed pipeline bus structure reduces the latency by 29% in compare with the segmented bus, and by 24% in compare with the dTDMA bus structure, respectively.

REFERENCES:

[1] D. Park, et al., “MIRA: A Multi-Layered On-Chip Interconnect Router Architecture”, ISCA 2008, pp. 251-261, Pennsylvania State, USA.

[2] B. S. Feero, et al., “Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation”, IEEE Transactions on Computers, vol. 58, no. 1, pp. 32-45, Jan. 2009.

[3] J. Kim, et al., “A novel dimensionally-decomposed router for on chip communication in 3D architectures,” in Proc. of the ISCA, pp. 138-149, Boston, USA, 2007.

[4] M. Ebrahimi, et al., “A High-Performance Network Interface Architecture for NoCs Using Reorder Buffer Sharing,” in Proceedings of 18th IEEE PDP, pp. 547-550, February 2010, Italy.

[5] R. Lu, et al., “SAMBA-Bus: A High Performance Bus Architecture for System- on-Chips”, IEEE Transactions on VLSI Systems, Vol. 15, Issue 1, pp. 69-79, Jan 2007.

[6] T.D. Richardson, et al., “A hybrid SoC interconnect with dynamic TDMA-based transaction-less buses and on-chip networks”, In Proc. VLSID, pp. 8-15, 2006.

[7] I. Loi, et al., “Developing mesochronous synchronizers to enable 3D NoCs”, Proc. of DATE, Germany, pp. 1414-1419, 2008.