Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation

Brett Stanley Feero, *Member*, *IEEE*, and Partha Pratim Pande, *Member*, *IEEE*

Abstract—The Network-on-Chip (NoC) paradigm has emerged as a revolutionary methodology for integrating a very high number of intellectual property (IP) blocks in a single die. The achievable performance benefit arising out of adopting NoCs is constrained by the performance limitation imposed by the metal wire, which is the physical realization of communication channels. With technology scaling, only depending on the material innovation will extend the lifetime of conventional interconnect systems a few technology generations. According to the International Technology Roadmap for Semiconductors (ITRS) for the longer term, new interconnect paradigms are in need. The conventional 2D integrated circuit (IC) has limited floorplanning choices, and consequently, it limits the performance enhancements arising out of NoC architectures. Three-dimensional ICs are capable of achieving better performance, functionality, and packaging density compared to more traditional planar ICs. On the other hand, NoC is an enabling solution for integrating large numbers of embedded cores in a single die. Three-dimensional NoC architectures combine the benefits of these two new domains to offer an unprecedented performance gain. In this paper, we evaluate the performance of 3D NoC architectures and demonstrate their superior functionality in terms of throughput, latency, energy dissipation, and wiring area overhead compared to traditional 2D implementations.

Index Terms—Network-on-chip, 3D integrated circuits, throughput, latency, energy, area, wormhole routing.

1 Introduction

With shrinking geometries, global interconnects are becoming the principal performance bottleneck for high-performance Systems-on-Chip (SoCs) [1], [2], [3]. These long interconnects are quickly becoming a performance impediment in terms of communication latency and power [4]. The Network-on-Chip (NoC) model is emerging as a revolutionary methodology in solving the performance limitations arising out of long interconnects, outperforming more mainstream bus architectures [5], [6]. In addition to providing a solution for the global wire delay problem, the NoC paradigm also eases integration of high numbers of intellectual property (IP) cores in a single SoC. However, restricted floorplanning choices of the 2D integrated circuits (ICs) limit the performance enhancements arising out of NoC architectures. As shown in Fig. 1, 3D ICs, which contain multiple layers of active devices, have the potential for enhancing system performance [7], [8], [9], [10]. According to [7], 3D ICs allow for performance enhancements even in the absence of scaling. This is the result of the reduction in interconnect length. Besides this clear benefit, package density is increased significantly, power is reduced from shorter wires, and circuitry is more immune to noise [7]. The performance improvement arising from the architectural advantages of NoCs will be significantly enhanced if 3D ICs are adopted as the basic fabrication methodology. The amalgamation of two emerging paradigms, NoC and 3D IC, allows for the creation of new structures that enable significant performance enhancements over more traditional solutions. With freedom in the third dimension, architectures that were impossible or prohibitive due to wiring constraints in planar ICs are now possible, and many 3D implementations can outperform their 2D counterparts.

In this paper, we characterize the performance of multiple 3D NoC architectures in the presence of realistic traffic patterns through cycle-accurate simulation and establish the performance benchmark and related design trade-offs.

This paper is organized as follows: Section 2 covers previous research and developments pertinent to this study. Section 3 is an introduction to the different 3D NoC architectures evaluated in this paper. Section 4 reviews the performance metrics used to evaluate the network architectures. Section 5 details the experimental analysis, and finally, Section 6 concludes this paper.

2 Related Work

Current SoCs are implemented predominantly following 2D architectures. However, the emergence of 3D ICs will present a fundamental change. Topol et al. [7] describe, in detail, the challenges of manufacturing in a 3D IC process. It shows that 3D ICs are capable of improvements in power, noise, logical span, density, performance, and functionality. One major advantage of the 3D IC paradigm is that it allows for the integration of "dissimilar technologies," e.g., memory, analog, MEMS, and so forth, in a single die. This paper describes the benefits and drawbacks of different fabrication methods including face-to-face bonding and face-to-back bonding. With an SOI face-to-back process, the via pitch is minimized, at 0.4 μm with a separation of 2 μm between layers of SOI devices [7].

Jacob et al. [11] propose using 3D ICs to improve the performance of microprocessors by forming a processor-memory stack. They show that the integration of processor

- B.S. Feero is with ARM Ltd., 110 Fulbourn Road, Cambridge CB2 0NU, UK. E-mail: brett.feero@arm.com.
- P.P. Pande is with the School of Electrical Engineering and Computer Science, Washington State University, PO Box 642752, Pullman, WA 99164-2752. E-mail: pande@eecs.wsu.edu.

Manuscript received 26 June 2007; revised 21 Feb. 2008; accepted 20 May 2008; published online 6 Aug. 2008.
Recommended for acceptance by F. Lombardi.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TC-2007-06-0260.

0018-9340/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

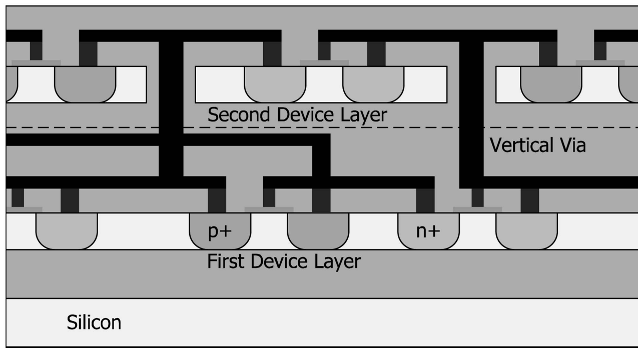


Fig. 1. Three-dimensional IC from an SOI process.

and memory in a stack enables a large increase in performance. In particular, 3D integration enables the use of very wide buses (> 1,024 bits) for vertical communication. In addition to ultrawide buses, a stack provides a very short distance between the processor and memory, decreasing memory access times considerably.

In [12], 3D ICs were proposed to improve performance of chip multiprocessors. Drawing upon 3D IC research, they chose a hybridization of buses and networks to provide the interconnect fabric between CPUs and L2 caches. The performance of this fusion of NoC and bus architectures was evaluated using standard CPU benchmarks. However, this analysis pertains only to chip multiprocessors and does not consider the use of 3D network structures for application-specific SoCs. Three-dimensional NoCs are analyzed in terms of temperature in [13]. Pavlidis and Friedman [14] compared 2D MESH structures with their 3D counterparts by analyzing the zero-load latency and power consumption of each network. This is an evaluation that shows some of the advantages of 3D NoCs, but it neither applies any real traffic pattern nor does it measure other relevant performance metrics. We aim to address these concerns by applying real traffic patterns in a cycle-accurate simulation and by measuring performance through established metrics for 3D NoC structures.

3 3D NoC Architectures

Enabling design in the vertical dimension permits a large degree of freedom in choosing an on-chip network topology. Due to wire-length constraints and layout complications, the more conventional 2D ICs have placed limitations on the types of network structures that are possible. With the advent of 3D ICs, a wide range of on-chip network structures that were not explored earlier are being considered [12], [14]. This paper investigates five different topologies in 3D space and compares them with three well-known NoC architectures from 2D implementations. We consider a SoC with a 400-mm^2 floorplan and 64 functional IP blocks. We selected this system size to reflect the state of the art of emerging SoCs. At ISSCC 2007, the design of an 80-core processor arranged in an 8×10 regular grid built on fundamental NoC concepts was demonstrated [15]. Therefore, we believe that the system size assumed in this work is representative of the current trends. IP blocks for 3D SoC are mapped onto four 10 mm $\times$ 10 mm layers, in order to occupy the same total area as a single-layer, 20 mm $\times$ 20 mm layout.

3.1 Mesh-Based Networks

One of the well-known 2D NoC architectures is the 2D Mesh, as shown in Fig. 2a. This architecture consists of an $m \times n$ mesh of switches interconnecting IP blocks placed along with them. It is known for its regular structure and short interswitch wires. From this structure, a variety of 3D topologies can be derived. The straightforward extension of this popular planar structure is the 3D Mesh. Fig. 2b shows an example of 3D Mesh-based NoC. It employs seven-port

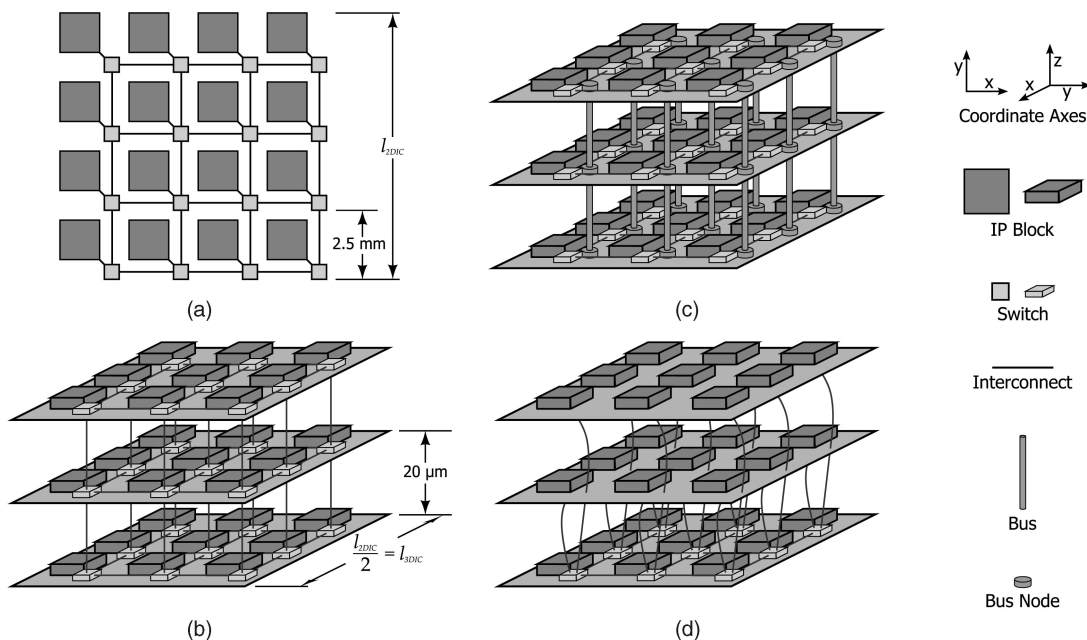


Fig. 2. Mesh-based NoC architectures. (a) Two-dimensional mesh. (b) Three-dimensional mesh. (c) Stacked mesh. (d) Ciliated 3D mesh.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

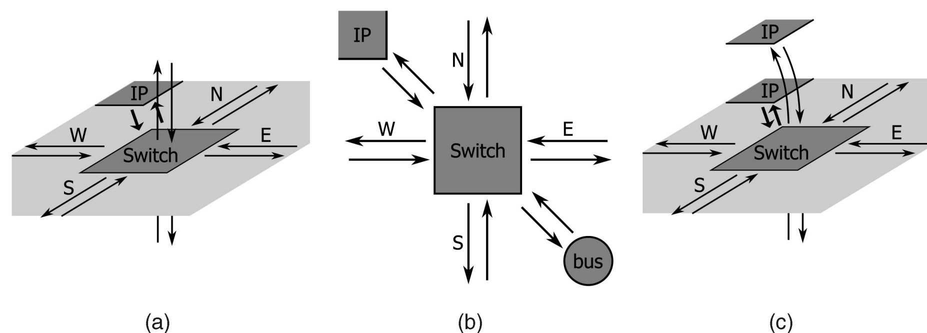


Fig. 3. Switches for mesh-based NoCs. (a) Three-dimensional mesh. (b) Stacked mesh. (c) Ciliated 3D mesh.

switches: one port to the IP block, one each to switches above and below, and one in each cardinal direction (North, South, East, and West), as shown in Fig. 3a.

A second derivation, Stacked Mesh (Fig. 2c), takes advantage of the short interlayer distances that are characteristics of a 3D IC, which can be around 20 $\mu\mathrm{m}$ [7]. Stacked Mesh is a hybrid between a packet-switched network and a bus. It integrates multiple layers of 2D Mesh networks by connecting them with a bus spanning the entire vertical distance of the chip. As the distance between the individual 2D layers in 3D IC is extremely small, the overall length of the bus is also small, making it a suitable choice for communicating in the z-dimension [12]. Furthermore, each bus has only a small number of nodes (i.e., equal to the number of layers of silicon), keeping the overall capacitance on the bus small and greatly simplifying bus arbitration. For consistency with [12], our analysis considers the use of a dynamic, time-division multiple-access (dTDMA) bus, although any other type of bus may be used as well. A switch in a Stacked Mesh network has, at most, six ports: one to the IP, one to the bus, and four for the cardinal directions (Fig. 3b). Additionally, it is possible to utilize ultrawide buses similar to the approach introduced in [11] to implement cost-effective, high-bandwidth communication between layers.

A third method of constructing a 3D NoC is by adding layers of functional IP blocks and restricting the switches to one layer or a small number of layers. With this in mind, we introduce a new architecture, ciliated 3D Mesh. This structure is essentially a 3D Mesh network with multiple IP blocks per switch. For the ciliated 3D Mesh, we consider a $4 \times 4 \times 2$ 3D Mesh network with two IPs per switch, where the two functional IP blocks occupy, more or less, the same footprint but reside at different layers. This is shown in Fig. 2d. In a ciliated 3D Mesh network, each switch contains seven ports (one for each cardinal direction, one either up or down, and one to each of the two IP blocks), as shown in Fig. 3c. This architecture will clearly exhibit lower overall bandwidth than a complete 3D Mesh due to multiple IP blocks per switch and reduced connectivity; however, we will show that this type of network offers an advantage in terms of energy dissipation, especially in the presence of specific traffic patterns.

It is important to note that each mesh-based network introduced in this section can be easily translated into a toroidal structure.

3.2 Tree-Based Networks

Two types of tree-based interconnection networks that have been considered for NoC applications are butterfly fat tree (BFT) [16], [17] and the generic fat tree, or SPIN [18]. This paper endeavors to quantify the enhancements achieved when these networks are instantiated in a 3D IC environment. Unlike the work with mesh-based NoCs, we do not propose any new topologies for tree-based systems. Instead, we investigate the achievable performance benefits by instantiating already-existing tree-based NoC topologies in a 3D environment.

The BFT topology we consider is shown in Fig. 4a. For a 64-IP SoC, a BFT network will contain 28 switches. Each switch (Fig. 5a) in a BFT network consists of six ports, one to each of four child nodes and two to parent nodes, with the exception of the switches at the topmost layer. When mapped to a 2D structure, the longest interswitch wire length for a BFT-based NoC is $l_{2DIC}/2$, where l_{2DIC} is the die length on one side [17], [5]. If the NoC is spread over a 20 mm $\times$ 20 mm die, then the longest interswitch wire is 10 mm [5], as shown in Fig. 4c. Yet, when the same BFT network is mapped onto a four-layer 3D SoC, wire routing becomes simpler, and the longest interswitch wire length is reduced by at least a factor of two, as can be seen in Fig. 4d. This will lead to reduced energy dissipation as well as less area overhead. The fat tree topology in Fig. 4b will have the same advantages when mapped on to a 3D IC as the BFT.

4 Performance Analysis and Design Cost

In order to properly analyze the various 3D NoC topologies, a standard set of metrics must be used [19]. We assume wormhole routing [20] as the data transport mechanism where the packet is divided into fixed length flow control units or flits. The header flit holds the routing and control information. It establishes a path, and subsequent payload or body flits follow that path. Our comparative analysis focuses on the four established benchmarks [19] of throughput, latency, energy, and area overhead.

Throughput is a metric that quantifies the rate in which message traffic can be sent across a communication fabric. It is defined as the average number of flits arriving per IP block per clock cycle, so the maximum throughput of a system is directly related to the peak data rate that a system can sustain. Accordingly, it is important to note that a throughput of 1 corresponds to every node accepting one flit in every clock cycle.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

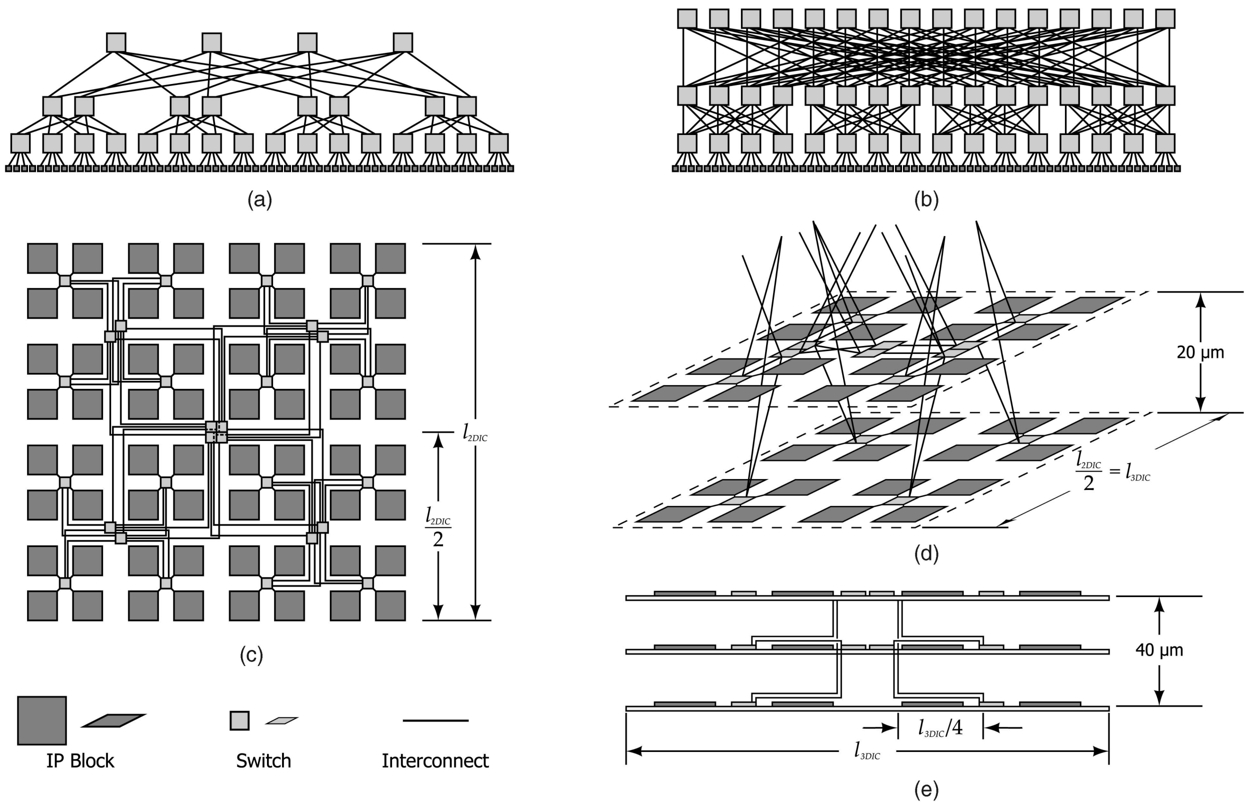


Fig. 4. Tree architectures. (a) BFT. (b) SPIN. (c) Two-dimensional BFT floorplan. (d) Three-dimensional BFT floorplan for the first two layers. (e) The first three layers of a 3D BFT floorplan as seen in elevation view.

Next, latency refers to the length of time elapsed between the injection of a message header at the source node and the reception of the tail flit at the destination.

The transport of messages across a network leads to a quantifiable amount of energy dissipation. Activity in the logic gates of the network switches as well as the charging and discharging of interconnection wires lead to the consumption of energy.

Last, the amount of silicon area used by an interconnection network is a necessary consideration. Network switches, layer-to-layer vias, interswitch wires, and buffers incurred by relatively longer wires all use a certain amount of area of the chip.

4.1 Performance Analysis in 3D Mesh-Based NoCs

In this section, we analyze the performance of the 3D Mesh-based NoC architectures in terms of the parameters mentioned above.

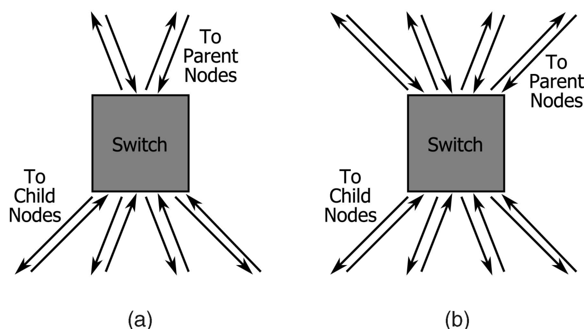


Fig. 5. Switches for tree networks. (a) BFT. (b) SPIN.

Throughput is given in the number of accepted flits per IP per cycle. This metric, therefore, is closely related to the maximum amount of sustainable traffic in a certain network type. Any improvements in throughput in 3D networks are principally related to two factors: the number of physical links and the average number of hops.

In general, for a Mesh-based NoC, the number of links is given as follows:

$$links = N_1N_2(N_3 - 1) + N_1N_3(N_2 - 1) + N_2N_3(N_1 - 1), \quad (1)$$

where N_i represents the number of switches in the ith dimension. For instance, in an 8×8 2D Mesh-based NoC, this yield 112 links. In a $4 \times 4 \times 4$ 3D Mesh-based NoC, the number of links turns out to be 144. With a greater number of links, a 3D Mesh network, for example, is able to contain a greater number of flits and therefore transmit a greater number of messages.

However, only considering the number of links will not characterize the overall throughput of a network. The average hop count also has a definitive effect on throughput. Following [14], the average number of hops in a mesh-based NoC is given by

$$hops_{NoC} = \frac{n_1n_2n_3(n_1 + n_2 + n_3) - n_3(n_1 + n_2) - n_1n_2}{3(n_1n_2n_3 - 1)}, \quad (2)$$

where n_i is the number of nodes in the ith dimension. This equation applies both to the $4 \times 4 \times 4$ 3D Mesh and

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

$4 \times 4 \times 2$ ciliated 3D Mesh networks. The number of hops for the Stacked Mesh is equal to

$$hops_{Stacked} = \frac{n_1 + n_2}{3} + \frac{n_3 - 1}{n_3}. \quad (3)$$

For our $4 \times 4 \times 4$ 3D Mesh and 8×8 2D Mesh, average hop counts are 3.81 and 5.33, respectively. There are 40 percent more hops in the 2D Mesh compared to that in the 3D Mesh. Consequently, flits in 3D Mesh needs to traverse less number of stages between a pair of source and destination than the 2D counterpart. As a result of this, we expect a corresponding increase in throughput. A lower average hop count will also allow more flits to be transmitted through the network. With a lower hop count, a wormhole-routed packet will utilize fewer links, thus leaving more room to increase the maximum sustainable traffic.

Transport latency, like throughput, is also affected by average hop count. It is also affected heavily by the number of links and the injection load. In 3D architectures, a decrease in latency is expected due to a lower hop count and an increased number of links.

In the SoC realm, energy dissipation characteristics of the interconnect structures are crucial, as the interconnect fabric can consume a significant portion of the overall energy budget [13]. The energy dissipation in an NoC depends on the energy dissipated by the switch blocks and the interswitch wire segments. Both of these factors depend on the network architecture. Additionally, the injection load has a significant contribution as it is the cause for any activity in the switches and interswitch wires. Intuitively, it is clear that with more packets traversing the network, power will increase. This is why packet energy is an important attribute for characterizing NoC structures. The energy dissipated per flit per hop is given by

$$E_{hop} = E_{switch} + E_{wire}, \quad (4)$$

where E_{switch} and E_{wire} are the energy dissipated by each switch and interswitch wire segments, respectively. The energy of a packet of length n flits that completes h hops is given by

$$E_{packet} = n \sum_{j=1}^{h} E_{hop,j}. \quad (5)$$

From this, a formula for packet energy can be realized. If P packets are transmitted, then the average energy dissipated per packet is given as

$$\overline{E_{packet}} = \frac{\sum_{i=1}^{P} E_{packet,i}}{P} = \frac{\sum_{i=1}^{P} \left(n_i \sum_{j=1}^{h_i} E_{hop,j} \right)}{P}. \quad (6)$$

Now, it is clear that a strong correlation exists between packet energy and the number of hops from source to destination. Consequently, a network topology that exhibits smaller hop counts will also exhibit correspondingly lower packet energy. As all 3D mesh-based NoC architectures exhibit a lower hop count, they should also dissipate less energy per packet.

Last, the area overhead must be analyzed for mesh-based NoCs. Area overhead for an NoC includes switch overhead and wiring overhead. Switch area is affected by the overall number of switches and the area per switch, which is highly correlated to the number of ports. Since all 3D mesh-based NoCs have more ports, the area per switch will increase. However, the ciliated structure has a reduced number of switches, which should significantly reduce the overall switch area. For 3D NoCs in general, wiring overhead includes the interlayer via footprint in addition to the area incurred by horizontal and vertical wirings. The addition of interlayer vias and their corresponding area overhead is a characteristic that is unique to 3D ICs, and it is included in the area overhead calculations presented in Section 5.

TABLE 1
Interswitch Wire Lengths in 3D Tree-Based NoCs

	2D NoC	4-layer 3D NoC
1st Level	≤ $l/8$ = 2.5 mm	≤ $l/4$ = 2.5 mm
2nd Level	$l/4$ = 5 mm	$l/4$ = 2.5 mm
3rd Level	$l/2$ = 10 mm	$l/4$ = 2.5 mm

Wire overhead is reduced when moving to a 3D IC. However, this is not due to reductions in the length of most interswitch wires. Horizontal wire length is given by l_{IC}/n_{side}, where n_{side} represents the number of IPs in one dimension of the IC and l_{IC} is the die length on one side, as shown in Figs. 2a and 2b. For the 8×8 2D Mesh, this evaluates to 20 mm/8 or 2.5 mm, and for all 3D mesh-based architectures, the expression evaluates to 10 mm/4, also 2.5 mm. With this in mind, reductions in wire overhead come from the interlayer wires. The 3D structures have a reduced number of horizontal links due to the presence of interlayer wires. These interlayer wires are very small, and hence, they are the source of wire overhead savings in mesh-based 3D NoCs.

4.2 Performance Analysis in 3D Tree-Based NoCs

Unlike the previous discussion pertaining to mesh-based NoCs, the tree-based networks we consider for 3D implementations have identical topologies to their 2D counterparts. The only variable is the interswitch wire length. As a result, there are significant improvements both in terms of energy and area overhead.

In 2D space, the longest interswitch wire length in a BFT or SPIN network is equal to $l_{2DIC}/2$ [17], [5], where l_{2DIC} is the die length on one side. This interswitch wire length corresponds to the topmost level of the tree. In a 3D IC, however, this changes significantly. For instance, as shown in Figs. 4d and 4e, the longest wire length for 3D, tree-based NoC is equal to the length of the horizontal travel in addition to the length of the vertical via. Considering a 20 mm $\times$ 20 mm 2D die, the longest interswitch wire length is equal to 10 mm, whereas with a 10 mm $\times$ 10 mm stack of four layers, the maximum wire length is equal to the sum of $l_{3DIC}/4$, or 2.5 mm, and the span of two layers, 40 μm. This is almost a factor-of-four reduction compared to 2D implementations. Similarly, mid-level wire lengths are reduced by a factor of 2. As a result, this reduction in wire length, shown in Table 1, causes a significant reduction in energy.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

TABLE 2
Wire Delays

Wire Type	Wire Length	Delay (ps)	Architectures Used
Interlayer	20 μm	16	all 3D mesh-based
Vertical Bus	60 μm	110/450**	3D Stacked Mesh
Horizontal	2.5 mm	219	mesh-based, 2D tree-based
Horizontal + Interlayer	2.54 mm	231	all 3D tree-based
Horizontal	5 mm	436*	Mid-level in all 2D tree-based
Horizontal	10 mm	550*	Top-level in all 2D tree-based
		*Repeaters Necessary	**Bus Arbitration Included

TABLE 3
Architectural Parameters

Topology	Port Count	Switch Area (mm^2)	Switch Static Energy (pJ)	Longest Wire Delay (ps)
2D Mesh	5	0.0924	65.3	219
3D Mesh	7	0.1385	91.4	219
3D Stacked Mesh	6 (+ bus arbitration)	0.1225	81.3	219
Ciliated 3D Mesh	7	0.1346	91.2	219
2D BFT	6	0.1155	78.3	550
3D BFT	6	0.1155	78.3	231
2D Fat Tree	8	0.1616	104.5	550
3D Fat Tree	8	0.1616	104.5	231

In addition to the benefits in terms of energy, 3D ICs effect area improvements for tree-based NoCs. Again, as with energy, area gains pertain only to the interswitch wire segments; there is neither a change in the number of switches nor in the design of the switch.

As with the 3D mesh-based NoCs, wire overhead in a 3D tree-based NoC consists of the horizontal wiring in addition to the area incurred by the vertical wires and vias. Also, the longer interswitch wires, which are characteristics of 2D tree-based NoCs, require repeaters, and this is taken into account. For a BFT, the number of wires in an arbitrary tree level l as defined in [16] is

$$wires_{\text{layer}l} = w_{link} \cdot \left\lceil \frac{N}{2^{l-1}} \right\rceil, \quad (7)$$

where N is the number of IP blocks, and w_{link} is the link width in bits. For a generic fat tree, the number of wires in a tree level l is given by

$$wires_{\text{layer}l} = w_{link} \cdot N. \quad (8)$$

For instance, in a 64-IP BFT network with 32-bit-wide bidirectional interswitch links, there are 2,048 wires in the first level, 1,024 wires in the second level, and 512 wires in the third. Similarly, a 64-IP fat tree will have 2,048 wires in every level.

5 EXPERIMENTAL RESULTS

To model the performance of different NoC structures, a cycle-accurate network simulator is employed that can also simulate dTDMA buses. The simulator is flit-driven and uses wormhole routing. In this work, we assume a self-similar injection process [19], [21], [22], [23]. This type of traffic has been observed in the bursty traffic typical of on-chip modules in MPEG-2 video applications [23], as well as various other networking applications [22]. It has been shown to closely model real traffic [23]. In terms of spatial distribution, it is capable of producing both uniform and localized traffic patterns for injected packets. In order to acquire energy and area characteristics, the network switches, dTDMA arbiter, and FIFO buffers were modeled in VHDL. The network switches were designed in such a way that their delay can be constrained within the limit of one clock cycle. We assume the clock cycle to be equal to 15 fan-out-of-four (FO4) delay units. With the 90-nm standard cell library from CMP [24], this corresponds to a clock frequency of 1.67 GHz. As the switches were designed with differing numbers of ports, their delays vary with one another. However, it was important to ensure that all the delay numbers were kept within the 15 FO4 timing constraint. Consistent with [5], the longest delays were in the 2D/3D Fat Tree switches as they had the highest number of ports. Yet, even it can be run with a clock frequency of 11 FO4, well within the 15 FO4 limit. To have a consistent comparison, all the switches were run with a 15-FO4 clock.

Similarly, all interswitch wire delays must hold within the same constraints. As shown in Table 2, wire RC delays remain within the clock period of 600 ps [24]. For Stacked Mesh, even considering the bus arbitration, the delay is constrained within one clock cycle. For the vertical wires, the via resistance and capacitance are included in the analysis. Thus, all network architectures are able to run at the same clock frequency of 1.67 GHz. Additional architectural parameters for each topology is shown in Table 3.

Although the simulator is capable of running with an arbitrary specification, each switch was designed with four virtual channels per port and two-flit-deep virtual channel buffers, as discussed in [19]. Synopsys Design Analyzer was used to synthesize the hardware description using a 90-nm standard cell library from CMP [24], and

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

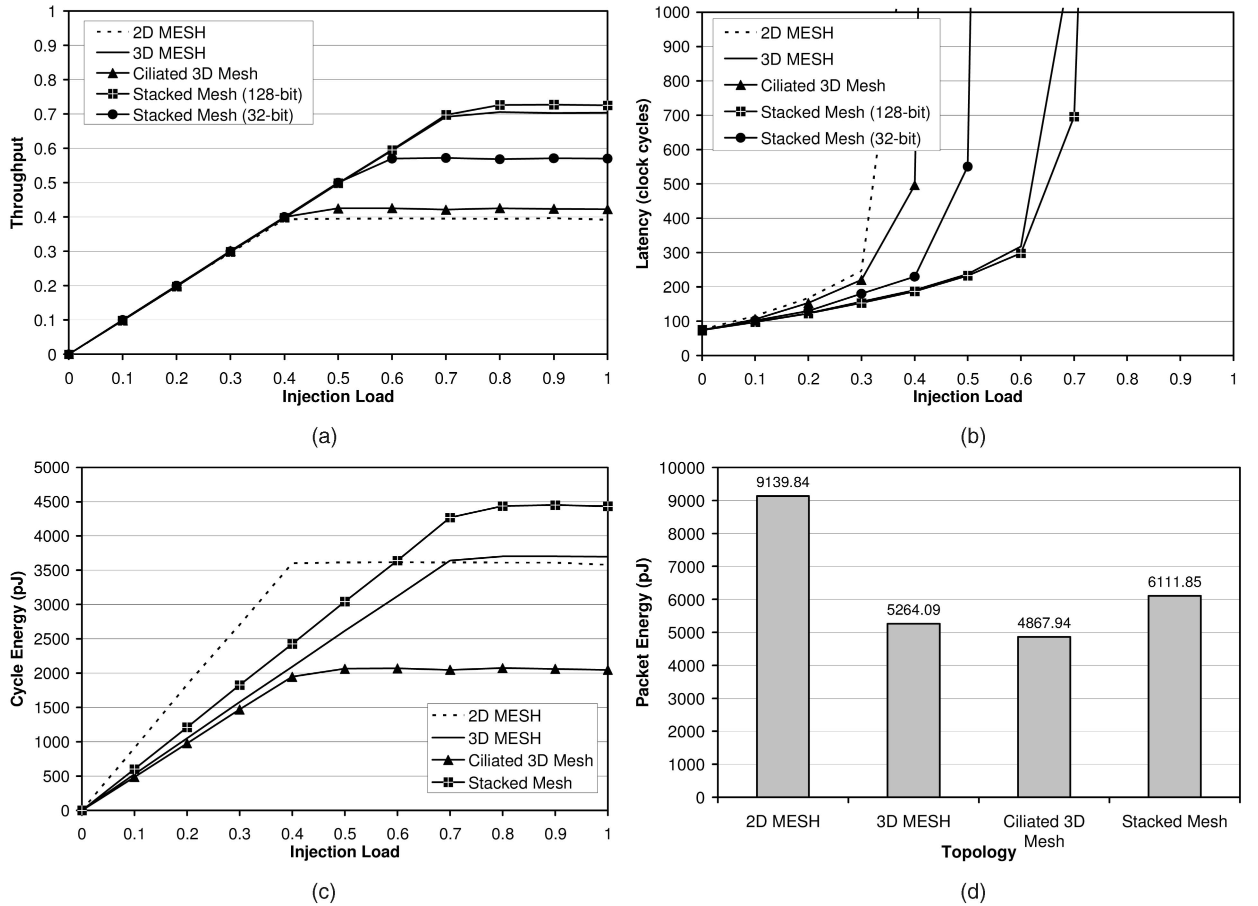


Fig. 6. Experimental results for mesh-based NoCs. (a) Throughput versus injection load. (b) Latency versus injection load. (c) Cycle energy versus injection load. (d) Packet energy.

Synopsys PrimePower was used to gather energy dissipation statistics. To calculate E_{switch} and $E_{interconnect}$ from (4), we follow the methodology discussed in [19]. We determine the energy dissipated by each switch, E_{switch}, by running its gate-level netlist through Synopsys PrimePower using large sets of input data patterns. In order to determine the interconnect energy, $E_{interconnect}$, we estimate the interconnects' capacitance, taking into account each interswitch wire's specific layout, by the following expression [19]:

$$C_{interconnect} = C_{wire} \cdot w_{a+1;a} + n \cdot m \cdot (C_G + C_J), \qquad (9)$$

where C_{wire} represents the capacitance per unit length of the wire, $w_{a+1;a}$ is the wire length between two consecutive switches, n is the number of repeaters, m represents the size of those repeaters with respect to minimum size devices, and last, C_G and C_J represent the gate and junction capacitances, respectively, of a minimum size inverter. We consider a worst case scenario where adjacent wires switch in opposite directions while determining C_{wire} [25].

The simulation is initially run for 10,000 cycles to allow the 64-IP network to stabilize, and it is subsequently run for 100,000 more cycles. The simulator provides statistics for energy, throughput, and latency.

5.1 Mesh-Based Networks

We first consider the performance of 3D mesh-based NoC architectures. Fig. 6a shows the variation of throughput as a function of the injection load. A network cannot accept more traffic than is supplied, and limitations in routing and collisions cause saturation before throughput reaches unity. From Fig. 6a, we see clearly that both the 3D Mesh and Stacked Mesh topologies exhibit throughput improvements over their 2D counterparts. It is also clear that the ciliated 3D Mesh network shows only a small throughput improvement. However, this is not where a ciliated structure exhibits the best performance. It will be shown later that this network topology has significant benefits both in terms of energy dissipation and silicon area.

These results coincide with the analysis of 3D mesh-based NoC provided in Section 4.1. We have shown in (1) that a 3D mesh will have 29 percent more interconnection links than a 2D version; hop count calculations have shown that a flit in a 2D mesh network will, on average, traverse 40 percent more hops than a flit navigating a 3D mesh (according to Table 4); and 3D mesh switches have higher connectivity with the increased number of ports. These all account for throughput improvements. In general, the lower hop count allows a wormhole-routed packet to occupy fewer resources, freeing up links for additional packets. Consequently, there is a corresponding increase in throughput.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

TABLE 4
Average Hop Count in Mesh-Based NoCs

2D Mesh	5.33
3D Mesh	3.81
Stacked Mesh	3.42
Ciliated 3D Mesh	3.10

Next, we consider the Stacked Mesh architecture. An increase in throughput is evident, as shown in Fig. 6a. However, with a 32-bit bus (corresponding to the flit width) connecting the layers of the NoC, throughput improvements are not as substantial as with the 3D mesh. Contention issues in the bus limit the attainable performance gains. Yet, since communication between layers is bus based, we can increase the size of the bus without modifying the switch architectures. As a result, the bus width is increased to 128 bits. Any further increase did not have any significant impact on throughput, except to increase the total capacitance on the bus. With this improvement, Stacked Mesh saturates at a slightly higher injection load than a 3D Mesh network. The Stacked Mesh topology also offers a lower hop count in comparison to a strict 3D mesh. From (3), we see that the average hop count is equal to 3.42. With the lower hop count in addition to the wide, 128-bit bus for vertical transmission, this architecture offers the highest throughput among all the 3D mesh-based networks.

Throughput characteristics of the ciliated 3D Mesh topology differ significantly from the other 3D networks. This network has a saturating throughput that is slightly higher than a 2D Mesh network and considerably less than both 3D Mesh and Stacked Mesh networks. This is true despite having the lowest hop count at an average of 3.10 hops. However, with only 64 interswitch links, compared to 144 in the 3D Mesh and 112 in the 2D Mesh, throughput improvements due to hop count are negated by the reduced number of links. The fact that there are multiple functional IP blocks for every switch is also responsible for considerably lower throughput due to contention issues in the switches.

Fig. 6b depicts the latencies for the architectures under consideration. Here, it is seen that 3D mesh-based NoCs have superior latency characteristics over the 2D versions. This is a product of the reduced hop count characteristic of 3D mesh-based topologies.

Energy dissipation characteristics for 3D mesh-based NoCs reveal a substantial improvement over planar NoCs. The energy dissipation profiles of the mesh-based NoC architectures under consideration are shown in Fig. 6c. Energy dissipation is largely dependent on two factors: architecture and injection load. We consider these two parameters as the independent factors in our analysis. As shown in (4), the energy dissipation in an NoC depends on the energy dissipated by the switch blocks and the interswitch wire segments. Both these factors depend on the architectures. The design of the switch varies with the architecture, and interswitch wire length is also architecture dependent [19]. Besides the network architecture, injection load has a clear effect on the total energy dissipation of a NoC, in accordance with Fig. 6c. Intuitively, it is clear that with more packets traversing the network, power will increase. This is why packet energy, in Fig. 6d, is an important attribute for characterizing NoC structures. Notice that, at saturation, a 2D Mesh network dissipates less power than both Stacked Mesh and 3D Mesh networks. This is the result of the lower 2D Mesh throughput, and the 3D networks consume more energy because they transmit more flits at saturation. Packet energy is a more accurate representation of the cost of data transmission. With packet energy in mind, it can be seen that every 3D topology provides a very substantial improvement over 2D mesh. Also, the energy dissipation of the ciliated mesh topology is less, still, than that of the 3D mesh network. These results follow closely the hop count calculations summarized in Table 4, with the exception of the packet energy for a Stacked Mesh network. Energy is heavily dependant on interconnect energy, and this is where the Stacked Mesh suffers. Since vertical communication takes place through wide buses, the capacitive loading on those buses results in a significant amount of energy. As a result, though 3D Stacked Mesh has a lower hop count compared to 3D Mesh, it dissipates more packet energy on average. Regardless, the profound energy savings possible in these 3D architectures provides serious motivation for a SoC designer to consider a 3D IC.

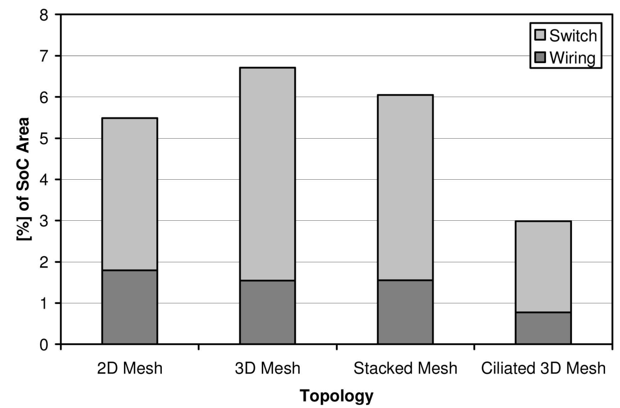


Fig. 7. Area overhead for mesh-based NoCs.

The final performance metric considered in this study is the overall area overhead incurred with the instantiation of the various networks. Fig. 7 shows the area penalty from each NoC design, both in terms of switch area and interconnect area. We see that while the 3D Mesh and Stacked Mesh NoCs reduce the amount of wiring area, switch overhead is increased. For both 3D Mesh and Stacked Mesh NoCs, the number of longer interswitch links in $x-y$ plane is reduced. There are 96 $x-y$ links for both topologies, for Stacked Mesh, 16 buses are present, and for the 3D Mesh, 48 vertical links are present. In comparison, the conventional 2D mesh-based NoC has 112 links in the horizontal plane. As the 3D NoCs have fewer long horizontal links, they incur less wiring area overhead. Although there are a large number of vertical links, the amount of area incurred by them is very small due to the $2\ \mu m \times 2\ \mu m$ interlayer vias. However, an increased number of ports per switch results in larger switch overhead for

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

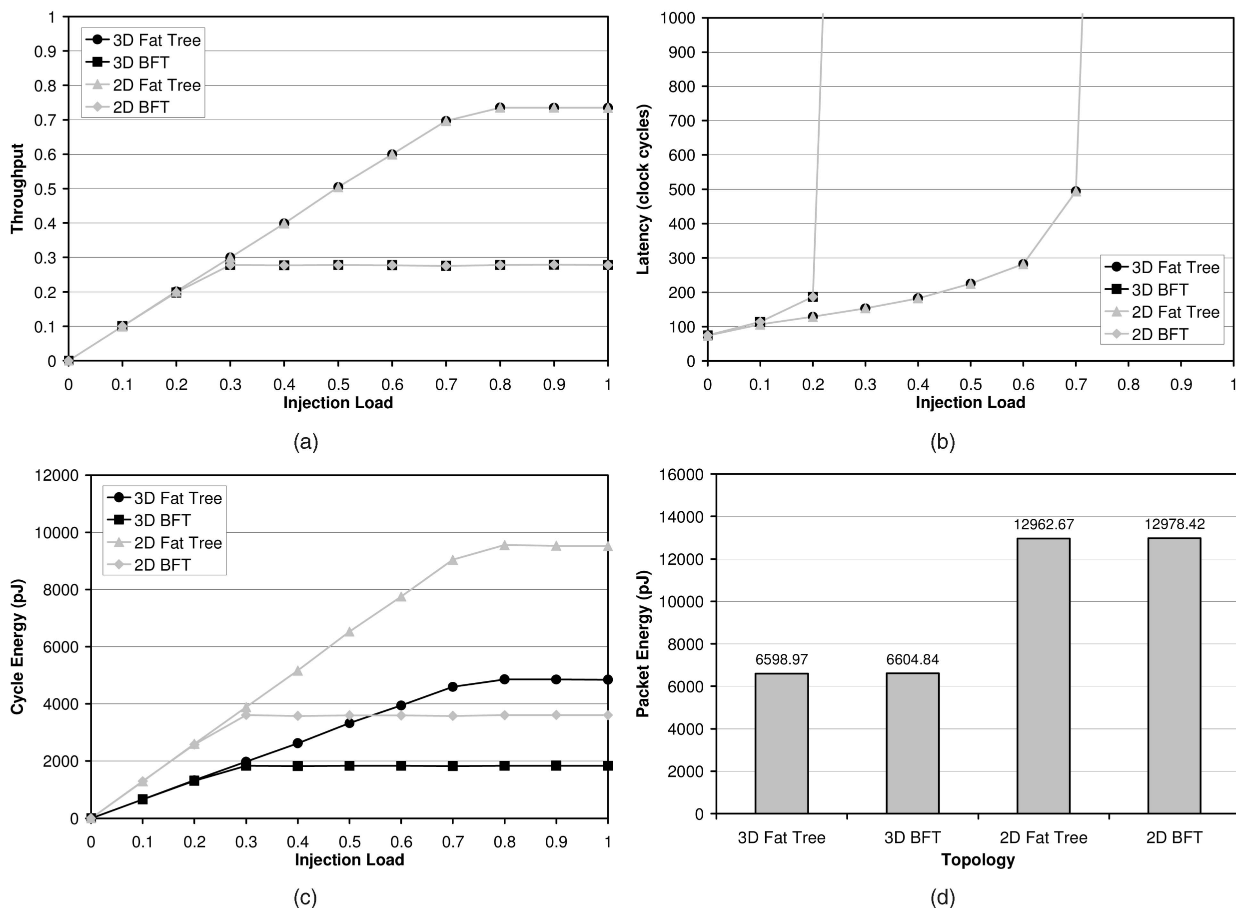


Fig. 8. Experimental results for tree-based NoCs. (a) Throughput versus injection load. (b) Latency versus injection load. (c) Cycle energy versus injection load. (d) Packet energy.

both of these NoC architectures, ultimately causing the 3D Mesh and Stacked Mesh topologies to incur more silicon area in spite of wiring improvements. On the other hand, ciliated 3D Mesh shows a significant improvement in terms of area. The $4 \times 4 \times 2$ ciliated 3D Mesh structure involves half the number of switches as the other mesh-based architectures in addition to only 64 links. As a result, the area overhead is accordingly smaller.

5.2 Tree-Based Networks

In this section, we evaluate the performance of the 3D tree-based NoCs. It has already been established that 2D and 3D versions of the tree topologies should have identical throughput and latency characteristics, and Figs. 8a and 8b support this. Consistent with the analysis of mesh-based NoCs, Fig. 8a shows the variation of throughput as a function of injection load, and Fig. 8b shows the effect of injection load on latency. The assumption here was that the switches and the interswitch wire segments are driven by the same clock as explained earlier. Consequently, under this assumption, in terms of throughput and latency there is no advantage to choosing a 3D IC over a traditional planar IC for a tree-based NoC. However, this is eclipsed by the superior performance achieved in terms of energy and area overhead. If the NoC switches can be designed to operate as fast as the interswitch wires, then we will show in Section 5.4 that the 3D tree-based architectures will have significant benefit in terms of even latency and bandwidth.

The energy profiles for 3D tree-based NoCs (Fig. 8c) reveal significant improvements over 2D implementations. Both BFT and fat tree (SPIN) networks show a very large reduction in energy when 3D ICs are used. Once again, we see that energy dissipation is largely dependant both on architecture and injection load. Each NoC shows that energy dissipation increases with injection load until the network becomes saturated, similar to the throughput curve shown in Fig. 8a. From the energy profiles, we also see that the fat tree networks cause higher energy dissipation than the BFT instantiations, but this is universally true only at high injection load.

Again, this is the motivation to consider packet energy of the networks as a relevant metric for comparison, as shown in Fig. 8d. Energy savings in excess of 45 percent are achievable by adopting 3D ICs as a manufacturing methodology, and both BFT and fat tree networks show similar improvements. In case of tree-based NoCs, where the basic network topology remains unchanged in 3D implementations, all improvements in energy dissipation are caused by the shorter wires. As we showed earlier in Table 1, a 3D structure greatly reduces the interswitch wire length. The overall energy dissipation in a NoC is heavily dependant on the interconnect energy, and this reduction in interswitch wire length effects very large savings.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

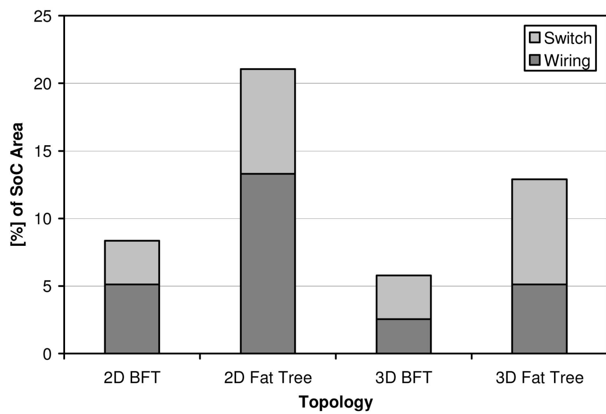


Fig. 9. Area overhead for tree-based NoCs.

Besides advantages in terms of energy, 3D ICs enable tree-based NoCs to reduce silicon area overhead by a sizable margin. Fig. 9 shows the overall area overhead of tree-based NoCs. We see that although no improvements are made in terms of switch area, the reductions in interswitch wire lengths and amount of repeaters are responsible for substantial reductions in wiring overhead. This is especially true for the fat tree network, which has more interconnects in the higher levels of the tree; wiring overhead is reduced more than 60 percent by instantiating the network into a 3D IC.

5.3 Effects of Traffic Localization

Until this point, we have assumed a uniform spatial distribution of traffic. In a SoC environment, different functions would map to different parts of the chip and the traffic patterns would be expected to be localized to different degrees [26]. We therefore studied the effect of traffic localization on the performance of the 3D NoCs and considered the illustrative case of spatial localization where local messages travel from a source to the set of the nearest destinations. In the case of BFT and fat tree, localized traffic is constrained to within a cluster consisting of a single subtree, while, in the case of 3D Mesh, it is constrained to within the destinations placed at the shortest Manhattan distance [19].

On the other hand, the 3D Stacked Mesh architecture is created simply to take advantage of the inexpensive vertical communication. The research pursued by Li et al. [12] suggested that in a 3D multiprocessor SoC, much of the communication should take place vertically, taking advantage of the short interlayer wire segments. This is a result of a large proportion of network traffic occurring between the processor and the closest cache memories, which are often placed along the z-dimension. Consequently, in these situations, the traffic will be highly localized, and we therefore consider localized traffic to be constrained to within a pillar for Stacked Mesh. Fig. 10 summarizes these

(a)

(b)

(c)

(d)

Fig. 10. Localization effects on mesh-based NoCs in terms of (a) throughput and (b) packet energy and localization effects on tree-based NoCs in terms of (c) throughput and (d) packet energy.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

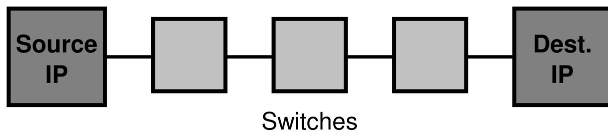


Fig. 11. The pipelined nature of NoCs.

effects, revealing the benefits of traffic localization. More packets can be injected into the network, improving the throughput characteristics of each topology, as shown in Figs. 10a and 10c, which also shows the throughput profile of the 2D topologies for reference. Analytically, increasing localization reduces the average number of hops that a flit must travel from source to destination.

Fig. 10a reveals that the stacked mesh network provides best performance in terms of throughput in the presence of localized traffic. However, this is achieved by using a wide bus for vertical communication. Let us consider what occurs when the bus size is equal to the flit width of 32 bits. With low localization, the achieved throughput is higher than that in a 2D Mesh network. However, when the fraction of localized traffic in the vertical pillars is increased, a huge performance degradation is seen. This is due to the contention in the bus. When the bus width is increased to 128 bits, throughput increases significantly with increase in localized traffic. This happens due to less contention in a wider communication channel.

Figs. 10b and 10d depict the effects of localization on packet energy, and unsurprisingly, there is a highly linear relationship between these two parameters. Packet energy is highly correlated with the number of hops from source to destination, and the resultant reduction of packet energy with localization supports this correlation. For the mesh-based networks, ciliated 3D Mesh exhibits the lowest packet energy due to its low hop count and very short vertical wires. In fact, at highest localization, the packet energy for a ciliated 3D Mesh topology is less than 50 percent of that of the next-best-performing topology: 3D Mesh. For the tree-based NoCs, both 3D networks have much-improved packet energy with traffic localization.

As can be seen from Fig. 10, there are trade-offs between packet energy and throughput. For instance, the best-performing topology in terms of energy, ciliated 3D Mesh, operates at the lowest throughput even when traffic is highly localized. On the other hand, although a 3D Stacked Mesh network with wider bus width achieves superior throughput without necessitating a highly local traffic distribution, it incurs more energy dissipation than other structures under local traffic due to the capacitive loading on the interlayer buses. However, the other topologies lie in some middle ground between these two extremes, and in general, it is clear that 3D ICs continue to effect improvements on NoCs under localized traffic.

5.4 Effects of Wire Delay on Latency and Bandwidth

In NoC architectures, the interswitch wire segments, along with the switch blocks, constitute a pipelined communication medium, as shown in Fig. 11. The overall latency (in nanoseconds) will be governed by the slowest pipelined stage. We had shown earlier, in Table 2, that the maximum wire delays for the network architectures are different. Though the vertical wire delays are very small, still the overall latency will be depended on the delay of the switch blocks. Though the delays of the switch blocks were constrained within the 15-FO4 limit, they were still the limiting stages in the pipeline, specifically when compared to the fast vertical links. Yet, if we consider a hypothetical case, which ignores the implications of switch design, where the clock period of the network is equal to the interswitch wire delay, then the clock frequency can be increased, and hence, the latency can be reduced significantly. With this in mind, latency in nanoseconds (instead of latency in clock cycles) and bandwidth (instead of throughput) are calculated. All other network parameters are kept consistent with our previous analysis.

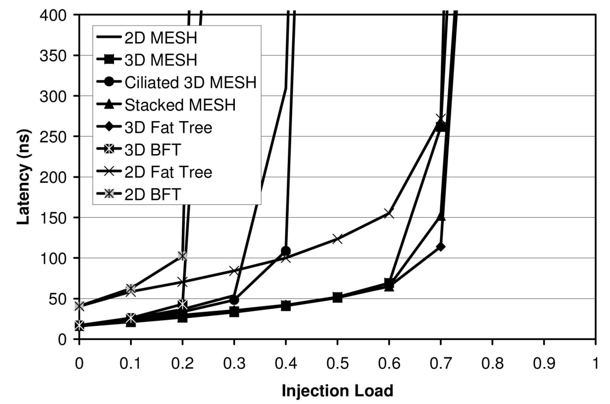


Fig. 12. Latency in nanoseconds at hypothetical clock frequencies.

A plot of latency for all network topologies is shown in Fig. 12, and Table 5 depicts the network bandwidth in units of terabits per second. To calculate bandwidth, we follow the following expression:

$$BW = TP_{\max} \cdot f \cdot w_{flit} \cdot N, \tag{10}$$

where $TP_{\max}$ represents the throughput at saturation, f represents the clock frequency, w_{flit} is the flit width, and N is the number of IP blocks. In Table 5, we show the performance difference achieved by running the NoC with a clock as fast as the interswitch wire, disregarding the

TABLE 5
Bandwidth of Network Architectures at Simulated and Hypothetical Frequencies (Terabits per Second)

	f = 1.67 GHz	f = 1/(max wire delay)	% increase
2D Mesh	1.357	3.711	173.5
3D Mesh	2.412	6.596	173.5
Ciliated 3D Mesh	1.457	3.983	173.5
3D Stacked Mesh	2.488	6.804	173.5
2D BFT	0.9543	1.039	8.9
2D Fat Tree	2.515	2.738	8.9
3D BFT	0.9543	2.474	159.2
3D Fat Tree	2.515	6.520	159.2

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

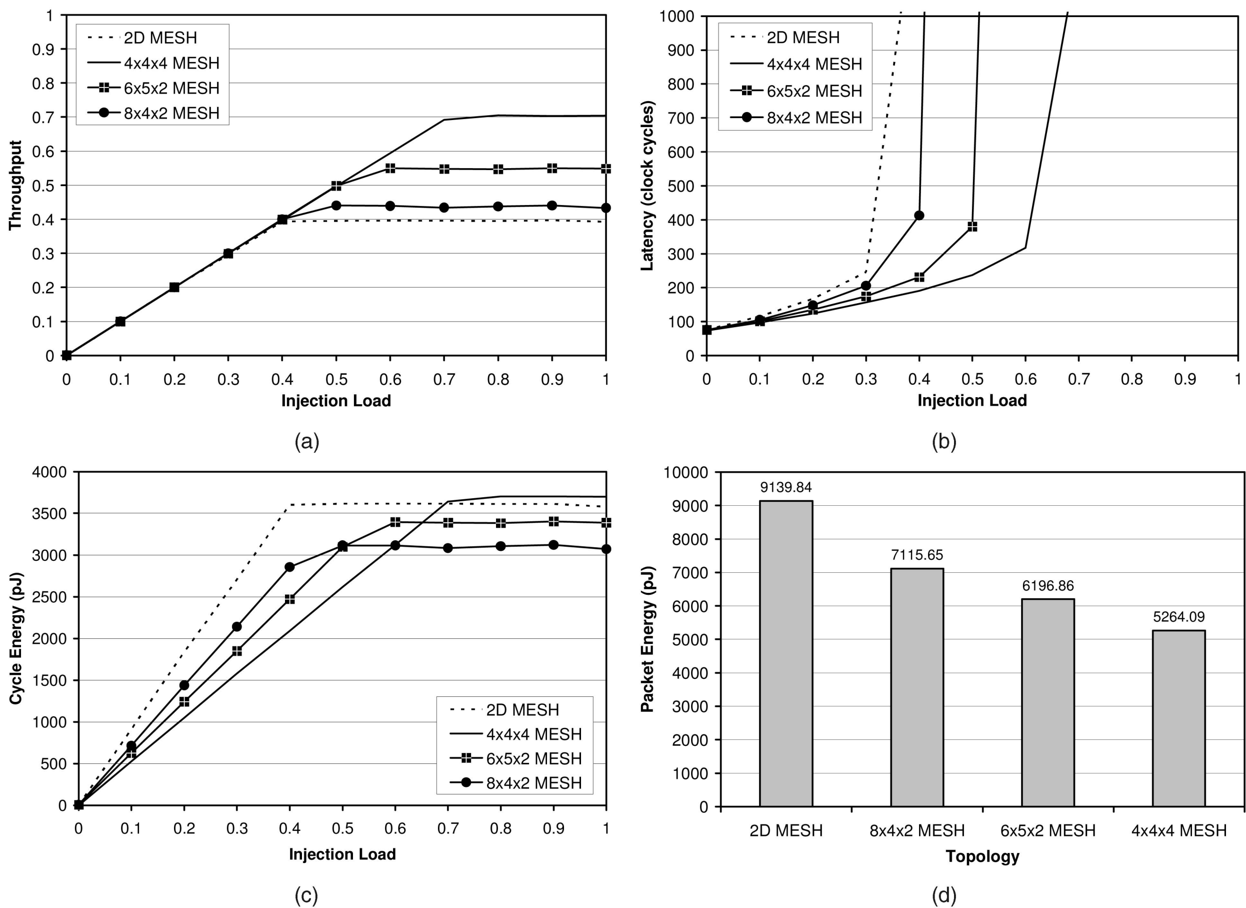


Fig. 13. Comparing two 2-layer NoCs. (a) Throughput versus injection load. (b) Latency versus injection load. (c) Cycle energy versus injection load. (d) Packet energy.

switch design constraints. It is evident that the tree-based architectures show the greatest performance improvement in this scenario going from 2D to 3D implementations, as the horizontal wire lengths are also reduced.

5.5 Network Aspect Ratio

The ability to stack layers of silicon is not without nuances. Upcoming 3D processes have a finite number of layers due to manufacturing difficulties and yield issues [7]. Furthermore, it is speculated [7] that the number of layers in a chip stack is not likely to scale with transistor geometries. This has a nontrivial effect on the performance of 3D NoCs. Consequently, future NoCs may have a greater number of IP blocks in the horizontal dimensions than vertically. The effect of this changing aspect ratio must be characterized.

For an illustration of these effects, we will examine, more in depth, the performance of a mesh-based NoC in a two-layer IC in comparison to the previously analyzed 3D $4 \times 4 \times 4$ Mesh and 2D 8×8 Mesh. Here, we consider a 64-IP $8 \times 4 \times 2$ Mesh to match the 64-IP network size, in order to make the comparison of latency and energy as fair as possible, and a 60-IP $6 \times 5 \times 2$ Mesh to show a network that is similar in size and that results in a more square overall footprint than the $8 \times 4 \times 2$ Mesh. Fig. 13 summarizes our analysis of these two-layer ICs. Throughput characteristics are seen in Fig. 13a. We see clearly that the $6 \times 5 \times 2$ Mesh achieves a significantly higher throughput than the 2D 8×8 Mesh and the $8 \times 4 \times 2$ Mesh, which suffers from a high average hop count (4.44 versus 4.11 for the $6 \times 5 \times 2$ Mesh), while achieving a lower maximum throughput than the four-layer mesh. Likewise, the two-layer mesh NoCs outperform the 2D Mesh in terms of latency, as shown in Fig. 13b, without exceeding the performance of the four-layer 3D instantiation. This trend continues when considering cycle energy (Fig. 13c) and packet energy (Fig. 13d). These results are as expected. With the first layer added, we see significant improvements in terms of each performance metric over the 2D case. Though the multilayer NoC exhibits superior performance characteristics compared to a 2D implementation, it will have to circumvent significant manufacturing challenges. Yet, even if we are limited to a two-layer 3D realization, that will still significantly outperform the planar NoCs.

5.6 Multilayer IPs

Throughout this paper, we have assumed each IP block to be instantiated in one layer of silicon. However, as discussed in [14], it is certainly possible for the IP blocks to be designed using multiple layers. So, each network architecture is analyzed with multilayer IPs. We assume the pipelined communication shown in Fig. 11, i.e., the NoCs are constrained by the switch delay and it cannot be driven

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.

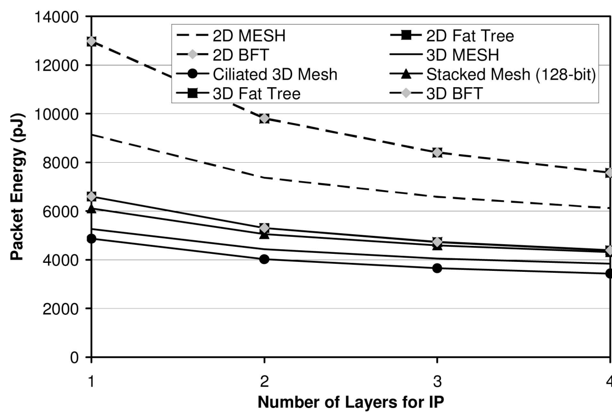


Fig. 14. The effect of multilayer IPs.

as fast as the interswitch wire. Considering this, multilayer IPs have no effect on either throughput or latency (assuming the same clock frequency for all networks), but there are nontrivial effects on the energy dissipation profile. This effect on packet energy is depicted in Fig. 14. The energy savings come from reduced horizontal wire lengths. For instance, if a 2.5 mm $\times$ 2.5 mm IP block is instantiated in two layers, the IP's circuitry is spread over two layers, and the footprint reduces by a factor of 1.414. Similarly, if instantiated in three layers, the footprint reduces by a factor of 1.732, and with four layers, the factor is 2. Although the vertical wire lengths are increased two, three, and four times, respectively, in order to span the entire multilayer IP, the negative effects on energy incurred by this are eclipsed by the significant reductions in horizontal wire lengths. However, multilayer IPs increase the number of layers in a 3D IC, placing an increased burden on manufacturability.

6 Conclusion

Three-dimensional NoCs are natural extensions of 2D designs. In this paper, we have demonstrated that besides reducing the footprint in a fabricated design, 3D network structures provide a better performance compared to traditional, 2D NoC architectures. We have demonstrated that both mesh and tree-based NoCs are capable of achieving better performance when instantiated in a 3D IC environment compared to more traditional 2D implementations. The mesh-based architectures show significant performance gains in terms of throughput, latency, and energy dissipation with a small area overhead. On the other hand, the 3D tree-based NoCs achieve significant gain in energy dissipation and area overhead without any change in throughput and latency. However, if the NoC switches are designed to be as fast as the interconnect, even the 3D tree-based NoCs will exhibit performance benefits in terms of latency and bandwidth. The NoC paradigm continues to attract significant research attention in both academia and industry. With the advent of 3D ICs, the achievable performance benefits from NoC methodology will be more pronounced as shown in this paper. Consequently, this will facilitate adoption of the NoC model as a mainstream design solution for larger multicore system chips.

References

[1] L. Benini and G. De Micheli, "Networks on Chips: A New SoC Paradigm," *Computer*, pp. 70-78, Jan. 2002.

[2] P. Magarshack and P.G. Paulin, "System-on-Chip beyond the Nanometer Wall," *Proc. 40th Design Automation Conf. (DAC '03)*, pp. 419-424, 2003.

[3] W.J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks," *Proc. 38th Design Automation Conf. (DAC '01)*, pp. 683-689, June 2001.

[4] M.A. Horowitz et al., "The Future of Wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490-504, Apr. 2001.

[5] C. Grecu, P.P. Pande, A. Ivanov, and R. Saleh, "Timing Analysis of Network on Chip Architectures for MP-SoC Platforms," *Microelectronics J.*, vol. 36, no. 9, pp. 833-845, Sept. 2005.

[6] H.G. Lee et al., "On-Chip Communication Architecture Exploration: A Quantitative Evaluation of Point-to-Point, Bus, and Network-on-Chip Approaches," *ACM Trans. Design Automation of Electronic Systems*, vol. 12, no. 3, pp. 1-20, Aug. 2007.

[7] A.W. Topol et al., "Three-Dimensional Integrated Circuits," *IBM J. Research and Development*, vol. 50, nos. 4/5, July-Sept. 2006.

[8] W.R. Davis et al., "Demystifying 3D ICS: The Pros and Cons of Going Vertical," *IEEE Design and Test of Computers*, vol. 22, no. 6, Nov./Dec. 2005.

[9] Y. Deng et al., "2.5D System Integration: A Design Driven System Implementation Schema," *Proc. Asia and South Pacific Design Automation Conf. (ASP-DAC)*, 2004.

[10] M. Ieong et al., "Three Dimensional CMOS Devices and Integrated Circuits," *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2003.

[11] P. Jacob et al., "Predicting the Performance of a 3D Processor-Memory Stack," *IEEE Design and Test of Computers*, vol. 22, no. 6, pp. 540-547, Nov./Dec. 2005.

[12] F. Li et al., "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," *Proc. 33rd Int'l Symp. Computer Architecture (ISCA '06)*, pp. 130-141, June 2005.

[13] C. Addo-Quaye, "Thermal-Aware Mapping and Placement for 3D NoC Designs," *Proc. IEEE Int'l System on Chip Conf. (SOCC '05)*, pp. 25-28, 2005.

[14] V.F. Pavlidis and E.G. Friedman, "3-D Topologies for Networks-on-Chip," *IEEE Trans. Very Large Scale Integration (VLSI '07)*, pp. 1081-1090, Oct. 2007.

[15] S. Vangal et al., "An 80-Tile 1.28TFLOPS Network-on-Chip in 65 nm CMOS," *Proc. IEEE Int'l Solid-State Circuits Conf. (ISSCC '07)*, pp. 98-99, 2007.

[16] R.I. Greenberg and L. Guan, "An Improved Analytical Model for Wormhole Routed Networks with Application to Butterfly Fat Trees," *Proc. Int'l Conf. Parallel Processing (ICPP '97)*, pp. 44-48, Aug. 1997.

[17] C. Grecu et al., "A Scalable Communication-Centric SoC Interconnect Architecture," *Proc. Fifth Int'l Symp. Quality Electronic Design (ISQED '04)*, pp. 343-348, 2004.

[18] P. Guerrier and A. Greiner, "A Generic Architecture for On-Chip Packet-Switched Interconnections," *Proc. Design, Automation and Test in Europe (DATE '00)*, pp. 250-256, Mar. 2000.

[19] P.P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance Evaluation and Design Trade-Offs for Network on Chip Interconnect Architectures," *IEEE Trans. Computers*, vol. 54, no. 8, pp. 1025-1040, Aug. 2005.

[20] J. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks—An Engineering Approach.* Morgan Kaufmann, 2002.

[21] K. Park and W. Willinger, *Self-Similar Network Traffic and Performance Evaluation.* John Wiley and Sons, 2000.

[22] D.R. Avresky, V. Shubranov, R. Horst, and P. Mehra, "Performance Evaluation of the ServerNet[R] SAN under Self-Similar Traffic," *Proc. 13th Int'l and 10th Symp. Parallel and Distributed Processing (IPPS/SPDP '99)*, pp. 143-147, Apr. 1999.

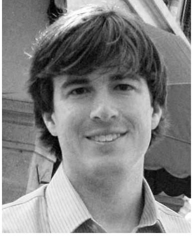
[23] G.V. Varatkar and R. Marculescu, "On-Chip Traffic Modeling and Synthesis for MPEG-2 Video Applications," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 335-339, June 2000.

[24] *Circuits Multi-Projects*, http://cmp.imag.fr/, 2008.

[25] K.C. Saraswat et al., "Technology and Reliability Constrained Future Copper Interconnects—Part II: Performance Implications," *IEEE Trans. Electron Devices*, vol. 49, no. 4, pp. 598-604, Apr. 2002.

[26] P.P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Effect of Traffic Localization on Energy Dissipation in NoC-Based Interconnect," *Proc. IEEE Int'l Symp. Circuits and Systems (ISCAS '05)*, pp. 1774-1777, May 2005.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.



Brett Stanley Feero received the BS degree in computer engineering (summa cum laude) and the MS degree in computer engineering from Washington State University, Pullman, in 2006 and 2008, respectively. His research interests include system-on-chip communication, networks-on-chip, and 3D integrated circuits. After his graduation, he became an engineer in Fabric IP at ARM, Cambridge, United Kingdom. He is a member of Tau Beta Pi and a member of the IEEE.

Partha Pratim Pande received the MS degree in computer science from the National University of Singapore in 2002 and the PhD degree in electrical and computer engineering from the University of British Columbia in 2005. He is an assistant professor in the School of Electrical Engineering and Computer Science, Washington State University, Pullman. His primary research interests lie in the areas of design and test of networks-on-chip, fault tolerance and reliability of multiprocessor SoC (MP-SoC) platforms, 3D integration, and on-chip wireless communication network. He serves in the program committees of various international conferences, including the IEEE International On-Line Testing Symposium (IOLTS), Asian Test Symposium (ATS), IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), and IEEE International Symposium on Electronic Design, Test and Applications (DELTA). He is a member of the IEEE.

▹ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Authorized licensed use limited to: University of Nevada Las Vegas. Downloaded on December 29, 2009 at 18:37 from IEEE Xplore. Restrictions apply.