3-D Topologies for Networks-on-Chip

Vasilis F. Pavlidis and Eby G. Friedman
Department of Electrical and Computer Engineering,
University of Rochester
Rochester, New York 14627

***Abstract*—Several interesting topologies emerge by incorporating the third dimension in the design of Networks-on-Chip (NoC). An analytic model for the zero-load latency of each network that considers the effect of the topology on the performance of a 3-D NoC is developed. A tradeoff between the number of nodes utilized in the third dimension of the network, which reduces the average number of hops traversed by a packet, and the number of physical planes used to integrate the processing elements (PE) of the network, which decreases the wire delay of the communication channel, is evaluated. A performance improvement of up to 33% is demonstrated for 3-D NoC as compared to a traditional 2-D NoC topology for a network size of N = 128 nodes.**

I. INTRODUCTION

Interconnect related problems, emerging from technology scaling and the integration limitations of Systems-on-Chip (SoC), originate from the functional diversity demanded by the electronics market. These issues have triggered a quest for non-conventional IC design paradigms. 3-D integration [1], and networks-on-chip [2] have been proposed as potent solutions to address these interconnect problems and the design complexity of SoC. Each of these design paradigms offers unique opportunities.

The major advantage of 3-D ICs is the considerable reduction in the length and number of global interconnects, resulting in an increase in the performance of wire-limited circuits. Another important advantage of 3-D ICs is that this paradigm enables the integration of disparate technologies which can be non-silicon or even electro-mechanical [1]. Despite the significant advantages of three-dimensional integration, important challenges remain such as crosstalk noise analysis and reduction, thermal mitigation, and interconnect modeling.

NoC offer high flexibility and the regularity of a network structure, supporting simpler interconnect modeling and more robust circuits. The canonical interconnect backbone of the network combined with appropriate communication protocols enhance the flexibility of such systems [2]. The intra-PE delay, however, cannot be reduced by the network. Furthermore, the length of the communication channel is primarily determined by the area of the PE which is unaffected by the network structure. In addition, each PE is limited by the CMOS process. By merging these two approaches, many of the individual limitations of 3-D ICs and NoC are circumvented, yielding a robust design paradigm with unprecented capabilities.

Research in 3-D NoC is only now emerging [3]. Multi-dimensional interconnection networks have been studied under various constraints, such as constant bisection-width and pin-out constraints [4]. NoC differ from generic interconnection networks, however, in that NoC are not limited by the channel width or pin-out. Alternatively, physical constraints, such as the number of nodes that can be implemented in the third dimension and the asymmetry in the length of the channels of the network, have to be considered. In this work, various possible topologies for 3-D NoC are presented. Additionally, simple analytic models for the zero-load latency of these networks that capture the effects of the topology on the performance of the 3-D NoC are described. An optimum topology is shown to exist that minimizes the zero-load latency of a network.

The paper is organized as follows. In the following section, various topological choices for 3-D NoC are reviewed. In Section III, an analytic model of the zero-load latency of traditional interconnection networks is adapted for each of the proposed 3-D NoC topologies. In Section IV, the proposed 3-D NoC topologies are compared in terms of the zero-load network latency, and guidelines for the optimum design of NoC structures are provided. Finally, some conclusions are offered in Section V.

II. 3-D NOC TOPOLOGIES

Various topologies for 3-D networks are presented in this section, and related terminology is introduced. Mesh structures have been a popular network topology for conventional 2-D NoC [5]. A mesh network is illustrated in Fig. 1a, where each PE is connected to the network through a router. Each router is connected to

This research is supported in part by the Semiconductor Research Corporation under Contract No. 2003-TJ-1068 and 2004-TJ-1207, the National Science Foundation under Contract Nos. CCR-0304574 and CCF-0541206, grants from the New York State Office of Science, Technology & Academic Research to the Center for Advanced Technology in Electronic Imaging Systems, and by grants from Intel Corporation, Eastman Kodak Company, Manhattan Routing, and Intrinsix Corporation.

the neighboring routers in four directions. The architecture of the router is considered here to be a canonical router with input and output buffering [6]. Each PE and router are called a network node. For a 2-D mesh network, the total number of nodes N is $N = n_1 \times n_2$, where n_i is the number of nodes included in the i^{th} physical dimension. For a 3-D NoC as shown in Fig. 1b, the total number of nodes is $N = n_1 \times n_2 \times n_3$, where n_3 is the number of nodes in the third dimension. In this topology, each PE is on a single physical plane (2-D IC – 3-D NoC). In Fig. 1c, a 3-D NoC topology is illustrated, where the interconnect network is contained within one physical plane, while the PEs are integrated in multiple planes (3-D IC – 2-D NoC). Finally, a hybrid 3-D NoC based on the two previous topologies is depicted in Fig. 1d. In such an NoC, both the interconnect network and the PEs can span more than one physical plane of the stack (3-D IC – 3-D NoC). In the following section, zero-load latency expressions for each of the NoC topologies are described, assuming a zero-latency model.

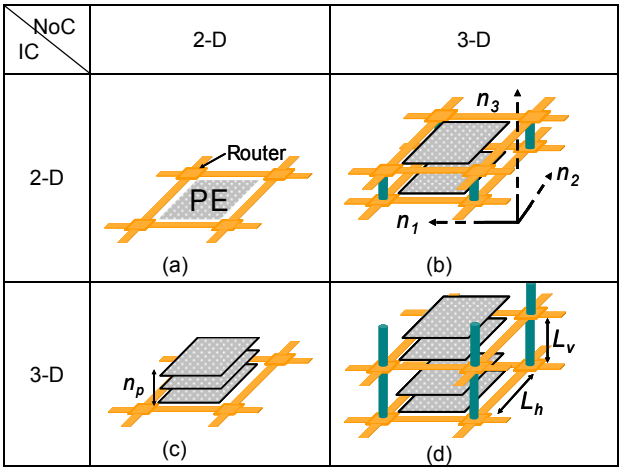

Figure 1: Various NoC topologies (not to scale), (a) 2-D IC – 2-D NoC, (b) 2-D IC – 3-D NoC, (c) 3-D IC – 2-D NoC, and (d) 3-D IC – 3-D NoC.

III. ZERO-LOAD LATENCY FOR 3-D NOC

In this section, analytic models of the zero-load latency of each of the 3-D NoC topologies are described. The zero-load network latency has been widely used as a performance metric in traditional interconnection networks [6]. The zero-load latency of a network is the latency of a network where only one packet traverses the network. Although such a model does not consider contention among packets, the zero-load latency can be used to characterize the effect of a topology on the performance of a network. The zero-load latency of an NoC with wormhole switching is [6]

$$T = hops \cdot t_r + t_c + \frac{L_p}{b}. \quad (1)$$

where the first term is the routing delay, t_c is the delay of the network channel, and the third term is the serialization delay of the packet. *hops* is the average number of hops that a packet traverses to reach the destination node, t_r is the router delay, L_p is the length of the packets in bits, and b is the bandwidth of the channel defined as $b \equiv wf_c$, where w is the width of the channel in bits and f_c is the inverse of the propagation delay of a bit along the longest communication channel.

Since the number of planes that can be stacked in a 3-D NoC is constrained by the target technology, n_3 is also constrained. The average number of hops in a 3-D NoC is

$$hops = \frac{n_1 n_2 n_3 (n_1 + n_2 + n_3) - n_3 (n_1 + n_2) - n_1 n_2}{3(n_1 n_2 n_3 - 1)}. \quad (2)$$

assuming dimension-order routing. The number of hops in (2) can be divided into two components, the average number of hops within the two dimensions n_1 and n_2, and the average number of hops within the third dimension n_3.

$$hops_{2D} = \frac{n_3 (n_1 + n_2)(n_1 n_2 - 1)}{3(n_1 n_2 n_3 - 1)}. \quad (3)$$

$$hops_{3D} = \frac{(n_3^2 - 1) n_1 n_2 \sim}{3(n_1 n_2 n_3 - 1)} \quad (4)$$

The delay of the router t_r can be determined from the models described in [7]. The delay of the channel t_c is

$$t_c = t_v hops_{3D} + t_h hops_{2D}. \quad (5)$$

where t_v and t_h are the 50% propagation delay of the vertical and horizontal channel, respectively (see Fig. 1d). Note that if n_3 = 1, (5) describes the propagation delay of a 2-D NoC. The delays t_v and t_h can be described by the delay expression in [8],

$$t_{v(h)} = 0.377 r_{v(h)} c_{v(h)} L_{v(h)}^2 + 0.693(R_S C_L + R_S c_{v(h)} L_{v(h)} + r_{v(h)} L_{v(h)} C_L), \quad (6)$$

where $r_{v(h)}$ and $c_{v(h)}$ are the resistance and capacitance per unit length of the vertical and horizontal channel, respectively, R_s is the driver resistance, C_L is the input capacitance of the neighboring router, and $L_{v(h)}$ is the length of the vertical (horizontal) channel. The length of the vertical communication channel for the 3-D NoC shown in Fig. 1d is

$$L_v = \begin{cases} l_v, & \text{for 2DIC} - \text{3DNoC} \quad (7a) \\ (n_p - 1) l_v, & \text{for 3DIC} - \text{3DNoC} \quad (7b) \\ 0, & \text{for 2DIC} - \text{2DNoC and 3DIC} - \text{2DNoC}, \quad (7c) \end{cases}$$

where l_v is the length of a through-silicon via connecting two routers on adjacent physical planes. n_p is the number of physical planes used to implement the PEs. The length of the horizontal communication channel is assumed to be equal to the side length of the PE and is

$$L_h = \begin{cases} \sqrt{A_{PE}}, & \text{for 2DIC}-\text{2DNoC and 2DIC}-\text{3DNoC} \quad (8a) \\ \sqrt{A_{PE}/n_p}, & \text{for 3DIC}-\text{2DNoC and 3DIC}-\text{3DNoC}, \quad (8b) \end{cases}$$

where A_{PE} is the area of the processing element. The area of all of the PEs and, consequently, the length of each horizontal channel are assumed to be equal. In the following section, these expressions are used to determine the 3-D NoC topology that minimizes the zero-load latency for various network sizes.

IV. PERFORMANCE TRADEOFFS FOR 3-D NOC

In this section, the zero-load latency is determined for various network sizes. The 3-D NoC topology that provides the minimum zero-latency is also determined. Different constraints apply for each topology. For example, n_3 and n_p are constrained by the maximum number of physical planes n_{max} that can be vertically stacked. A maximum 16 planes is assumed in this study. The constraints that apply for each of the 3-D NoC topologies shown in Fig. 1 are

$$n_3 \le n_{\max}, \quad \textit{for 2D IC} - \textit{3D NoC}, \quad (9a)$$

$$n_p \le n_{\max}, \quad \textit{for 3D IC} - \textit{2D NoC}, \quad (9b)$$

$$n_3 n_p \le n_{\max}, \quad \textit{for 3D IC} - \textit{3D NoC}. \quad (9c)$$

In Fig. 2, the number of hops for a 2-D IC – 3-D NoC is compared to a 2-D IC – 2-D NoC. A significant reduction in the number of hops is achieved by introducing the third dimension. This reduction is greater for larger networks. Alternatively, the router delay slightly increases due to the increase in the number of ports required for a 3-D router (from five ports in a 2-D NoC to seven ports in a 3-D NoC). This increase, however, has a logarithmic dependence on the number of ports and is independent of the network size, while from (1), the latency of the network depends linearly on the number of nodes. Decreasing the number of hops reduces the routing and channel latency components in (1), while the serialization latency is independent of the number of hops. The interconnect and network parameters are listed in Table 1 for each of the topologies. The interconnect parameters are extracted using a commercial impedance extraction tool [9]. The zero-load latency of a 3-D IC – 2-D NoC is compared in Fig. 3 to that of a 2-D NoC. The average number of hops between these two network topologies is the same, as the network structure is in both cases two dimensional. By substituting (8) into (6), the latency of the network is shown to be inversely proportional to n_p. Thus, by integrating the PEs in multiple planes, the length of the communication channel and, consequently, the latency of the network are reduced. The decrease in the length of the communication channel reduces the channel latency and the serialization latency in (1). The routing latency, however, remains the same as the number of hops does not change. Finally, note that the decrease in the network latency for 3-D IC – 2-D NoC becomes smaller for larger network sizes as the increase in the latency caused by the greater number of hops cannot be compensated by the reduction in the communication channel length.

TABLE 1: Interconnect and Network Parameters

Parameter	*Value*
A_{pe}	0.01 cm^2
l_v	20 μm
r_v	506 Ω/cm
c_v	6 pF/cm
r_h	220 Ω/cm
c_h	2.5 pF/cm
R_S	550 Ω
C_L	10 fF
L_p	640 *bits*
w	64 *bits*

A topology that reduces all of the latency components in (1) produces the minimum zero-load network latency. The 3-D IC – 3-D NoC topology shown in Fig. 1d provides the greatest reduction in the network latency as n_{max} is optimally distributed among the number of nodes in the third dimension n_3 and the number of physical planes used for the PEs n_p. In the 3-D IC – 3-D NoC topology, both the number of hops and the length of the long horizontal communication channel are reduced. Alternatively, the router delay t_r increases. The delay of the vertical channel also increases, according to (7). The increase in the router delay, however, is small and the increased length of the vertical channel is insignificant as compared to the decreased length of the horizontal channel, as described in (7) and (8).

The minimum zero-load network latency achieved by each of the 3-D NoC topologies is plotted in Fig. 4 as a function of the router to communication channel delay ratio, t_r / t_c, for N = 128 nodes. The 3-D IC – 3-D NoC achieves the minimum latency. A latency improvement of up to 33% is achieved by the 3-D IC – 3-D NoC topology as compared to a traditional 2-D NoC. If $t_r / t_c << 1$, the network latency is primarily reduced by decreasing the communication channel length. The latency of the 3-D IC – 2-D NoC is therefore the same as the 3-D IC – 3-D NoC and the minimum latency is achieved for $n_3 = 1$ and $n_p = n_{max}$. If $t_r / t_c >> 1$, the network latency is primarily reduced by decreasing the number of hops that a packet experiences in the network. The performance of the 2-D IC – 3-D NoC therefore approaches that of the 3-D IC – 3-D NoC; and $n_3 = n_{max}$ and $n_p = 1$ achieve the minimum latency. For other values of t_r/t_c, the minimum latency is achieved for different values of n_3 and n_p such that (9c) is satisfied.

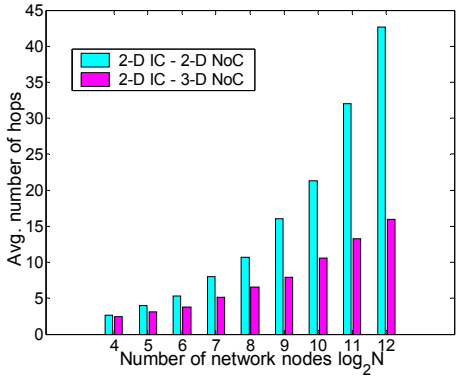

Figure 2: Average number of hops in 2-D IC – 2-D NoC and 2-D IC – 3-D NoC for different network sizes.

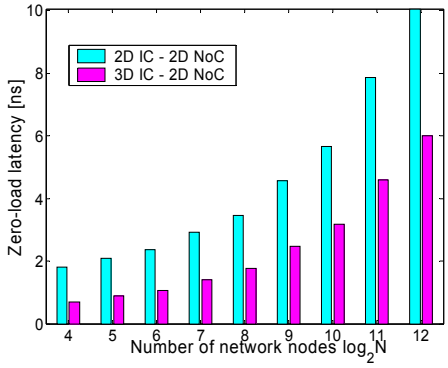

Figure 3: Zero-load network latency in 2-D IC – 2-D NoC and 3-D IC – 2-D NoC for different network sizes.

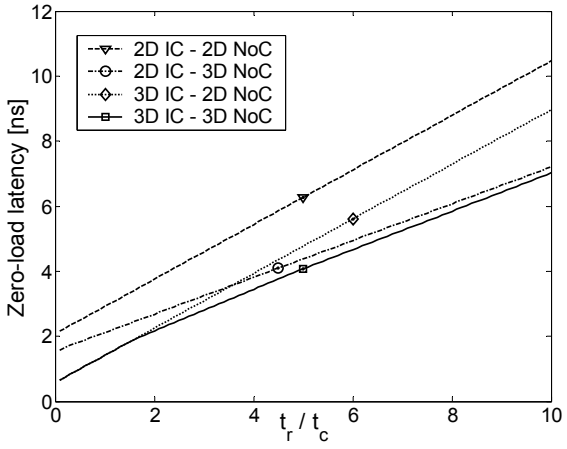

Figure 4: Minimum zero-load network latency for various 3-D NoC topologies as a function of the router to communication channel delay ratio (t_r / t_c) for N = 128.

Independent of the network size, the minimum latency occurs when n_1, n_2, and n_3 are equal to $N^{1/3}$. A similar result has been reported for generalized hypercubes in [10]. For example, consider a network with N = 64 nodes. The minimum zero-load latency for each of the 3-D NoC topologies is achieved for $n_1 = n_2 = n_3 = 4$. Due to the discrete nature of the variables, however, these optimum values are infeasible for certain network sizes (*e.g.*, N = 32). In this case, n_3 should be greater than or equal to either n_1 or n_2. By increasing the number of nodes in the third dimension, the average channel delay t_c decreases as the packets can utilize the short vertical channels.

In both of these cases, the remaining planes $n_p = n_{max} / n_3$ can be utilized to reduce the area of the PEs in the network and therefore, decrease the communication channel length. By reducing the number of hops and the communication channel length, the 3-D IC – 3-D NoC topology achieves the lowest zero-load network latency among the proposed 3-D NoC schemes.

V. CONCLUSIONS

3-D NoC is a natural evolution of 2-D NoC, exhibiting superior performance. Several novel 3-D NoC topologies are presented. The zero-load latency of the network is modeled for each of these topologies. The minimum latency can be achieved by reducing both the number of hops per packet and the length of the communication channel. The 3-D IC – 3-D NoC topology provides the optimum choice in terms of minimizing the zero-load network latency. Additionally, it is demonstrated that if the routing delay dominates, the performance of the 2-D IC – 3-D NoC topology approaches that of the 3-D IC – 3-D NoC topology, while if the communication channel delay dominates, the performance of the 3-D IC – 2-D NoC topology approaches that of the 3-D IC – 3-D NoC topology. Finally, design guidelines for determining the optimum number of nodes for any network size are provided.

REFERENCES

1 K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration," *Proceedings of the IEEE*, Vol. 89, No. 5, pp. 602-633, May 2001.

2 W. J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Netowrks," *Proceedings of the IEEE/ACM Deisgn Automation Conference*, pp. 684-689, June 2001.

3 C. Addo-Quaye, "Thermal-Aware Mapping and Placement for 3-D NoC Designs," *Proceedings of the IEEE International System-on-Chip Conference*, pp. 25-28, September 2005.

4 W. J. Dally, "Performance Analysis of *k*-ary *n*-cube Interconnection Networks," *IEEE Transaction on Computers,* Vol. 39, No. 6, pp. 775-785, June 1990.

5 A. Jantsch and H. Tenhunen, *Networks on Chip*, Kluwer Academic Publishers, 2003.

6 J. M. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks: An Engineering Approach*, Morgan Kaufmann, 2003.

7 L.-S. Peh and W. J. Dally, "A Delay Model for Router Microarchitectures," *IEEE Micro,* Vol. 21, No. 1, pp. 26-34, January/February 2001.

8 T. Sakurai, "Closed-Form Expressions for Interconnection Delay, Coupling, and Crosstalk in VLSI's," *IEEE Transactions on Electron Devices*, Vol. 40, No. 1, pp. 118-124, January 1993.

9 *Metal User's Guide,* OEA International Inc.

10 L. N. Bhuyan and D. P. Agrawal, "Generalized Hypercube and Hyperbus Structures for a Computer Network," *IEEE Transactions on Computers*, Vol. C-33, No. 4, pp. 323-333, April 1984.