Scalability of Network-on-Chip Communication Architecture for 3-D Meshes

Awet Yemane Weldezion*, Matt Grange+, Dinesh Pamunuwa+, Zhonghai Lu*,
Axel Jantsch*, Roshan Weerasekera+, and Hannu Tenhunen*

*School of Information and Communication Technologies,
Department of Electronics, Computer, and Software Systems,
KTH Royal Institute of Technology, Electrum 229, Kista SE 16440, Sweden
{aywe,zhonghai, axel, hannu}@kth.se

+Centre for Microsystems Engineering, Department of Engineering
Lancaster University, Lancaster LA1 4YW, UK
{m.grange, d.pamunuwa, r.weerasekera}@lancaster.ac.uk

Abstract

Design Constraints imposed by global interconnect delays as well as limitations in integration of disparate technologies make 3-D chip stacks an enticing technology solution for massively integrated electronic systems. The scarcity of vertical interconnects however imposes special constraints on the design of the communication architecture. This article examines the performance and scalability of different communication topologies for 3-D Network-on-Chips (NoC) using Through-Silicon-Vias (TSV) for inter-die connectivity. Cycle accurate RTL-level simulations are conducted for two communication schemes based on a 7-port switch and a centrally arbitrated vertical bus using different traffic patterns. The scalability of the 3-D NoC is examined under both communication architectures and compared to 2-D NoC structures in terms of throughput and latency in order to quantify the variation of network performance with the number of nodes and derive key design guidelines.

1. Introduction

The performance bottleneck imposed by on-chip interconnects in the deep submicron regime of process technology has been widely documented [15], as have the benefits of standardization of on-chip communication [3]. Implementing a standardized communication architecture such as a packet-switched NoC for massively integrated multiprocessor systems provides an abstraction of the global interconnection link and can greatly reduce design effort, potentially at the cost of some area and possibly power and performance penalties. The suitability of the NoC as a communication architecture depends on the overall system; i.e. the number of autonomous functioning blocks, the degree of parallelism, and area and performance requirements dictate its usefulness. The precise quantification of the performance and overhead of the network is of paramount importance in making this decision. The potential of the 2-D NoC for interconnecting multi-processor systems has been widely researched, and most recently an 80 tile, 100M transistor system with 1.28 TFLOP peak performance has been demonstrated in a $65nm$, 1V technology [17]. However a major new paradigm for continued Moore's law integration is 3-D chip stacks based on a variety of vertical interconnection techniques [1], [16]. 3-D integration provides opportunities for cost reduction and yield improvement in integration of different technologies such as CMOS, DRAM and MEMS circuits through the ability to implement them over multiple die layers on the same chip. It can also reduce form factor in applications where size is critical, while effective heat dissipation and temperature control can be a challenge. To get the most benefit out of 3-D chip stacks in multiprocessor systems, the communication architecture has to support efficient and high throughput vertical communication. In this article we examine the scalability of the NoC for such systems.

We build on previous work published in the literature and investigate the scalability of the three NoC architectures of 2-D mesh, 3-D mesh (with switch connectivity between layers) and 3-D bus (with bus connectivity between layers) with respect to performance. This is accomplished by quantifying latency and throughput of the different architectures for various network sizes, under two representative traffic

978-1-4244-4143-3/09/$25.00

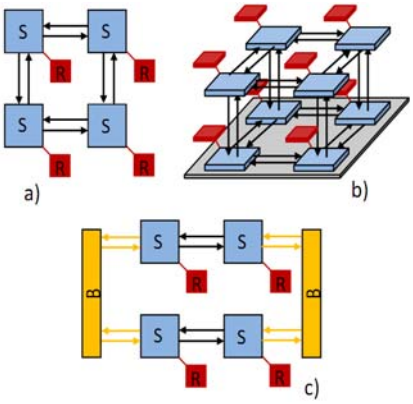


Figure 1. Each switch, S, is connected to a resource, R. a) a 2x2, 2-D mesh b) a 2x2x2, 3-D mesh structures, c) a 1x2x2, 3-D centrally arbitrated vertical bus for layer-to-layer network communication

patterns, uniform random and local. The Nostrum network communication protocol used in this study is based on a bufferless hot-potato routing algorithm [12], while the bus protocol utilizes a centrally arbitrated least-served-first priority scheme.

Cycle accurate RTL simulations are carried out to capture latency and throughput in terms of hops between nodes and hops per node per cycle respectively, for symmetric networks up to 1000 nodes. Various injection rates are used to study saturation points for the two traffic patterns. The main contribution of this paper is in providing a novel and extensive analysis into the scalability of 3-D NoC architectures as the number of nodes and layers grows. This study quantifies performance to aid the design of the communication architecture of future massively integrated 3-D systems.The rest of the paper is organized in the following manner. Section 2 discusses related work while the following section discusses network implementation issues including the protocols, network topology and architecture of the different switches and bus arbiter. Sections 4 and 5 describe the exploration space, the simulation and calculation methodology and the results. We end with a discussion of the results and our conclusions.

2. Related Work

A comprehensive overview of processing and technological issues in 3-D integration can be found in [16] while a discussion of the physical level cost and performance tradeoffs associated with the different techniques is provided in [18]. The different NoC-based system architectures for 3-D systems have been exhaustively enumerated in [14], depending on whether a die housed in a tile has one layer or several layers, and whether the NoC itself is 2-D or 3-D. The average unloaded latency per bit has also been derived, showing the potential improvements in latency and power consumption obtainable through a 3-D architecture, but the performance of the network for different traffic patterns under a packet-switched protocol is necessary to evaluate its scalability. An investigation into different router structures was conducted in [6] which proposes a 3-D crossbar-style NoC and performs cycle accurate simulations to extract energy and latency metrics. However the study fixes on a 64 node network. In [13] a multi-layer NoC router architecture is explored for a number of traffic patterns. In this paper, the number of nodes is fixed at 36 and the number of layers in the 3-D stack is kept constant at four. A well known paper that describes the performance of communication networks of varying dimension for wormhole routing is [2], which generalizes the interconnection network as being a k-ary n-cube torus, with n being the dimension of the cube, and k being the radix, or number of switches in a given dimension. In it Dally points out that VLSI circuits are wire-limited, and any growth in dimension has to be accompanied by a related reduction in the parallelism of each link and therefore the appropriate analysis of performance of interconnection networks for VLSI circuits has to be under the constraint of constant bisection bandwidth. It is shown that under various wire delay models, including transmission-line like and RC like behavior, for relatively large networks with 1M nodes, the best performance is delivered by switches of relatively low dimension, around 5 or 6. Due to layout restrictions, the most common implementation of the NoC for both 2-D and 3-D is a mesh, which is not directly comparable to the tori considered in [2]. In a 2-D mesh a switch has links to four other switches and its resource, while the straight-

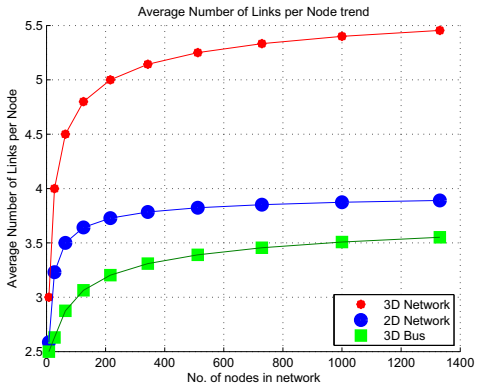


Figure 2. Number of links per node for a 2-D mesh, a 3-D mesh and a 3-D bus NoC

forward extension for 3-D is to consider a switch with two additional inter-layer links (with the seventh port being connected to the resource). Therefore it is of interest to investigate the scalability (i.e. variation of key metrics related to latency and throughput with size of network) of the NoC for this topology. Further, two key assumptions in [2] are likely to have an effect; firstly the study is carried out under the assumption of uniform links, but as explained below the vertical links have a significantly different delay model than the horizontal links. Secondly, the assumption of constant bisection bandwidth needs to be revisited as the footprint of the vertical interconnects is different in general to the horizontal links.

The electrical behavior of the relatively short and wide TSV is much better than that of long on-chip interconnects, primarily due to the low resistance, and can support much higher signaling speeds, but a relatively high area penalty due to via blockage may impose constraints on the number of TSVs that can be used for communication. This has led the authors of [7] to propose a dynamic time-division-multiple-access (dTDMA) bus with a centralized arbiter for the vertical communication link, which allows single hop latency for packets between any number of vertical layers. Most recently [5] describes a detailed study using different traffic patterns under a realistic protocol for both architectures to derive figures of merit related to throughput, latency, energy and area overhead to characterize various topologies. However the question of the general scalability of the different architectures with network size has not been addressed. Finally [11] describes a working 3-layer 27 node prototype that provides proof of concept of the 3-D NoC, but identify the need for a 3-D network simulator and system-level explorations of the kind discussed here.

3. Link and Network layer Protocols

The NoC protocol employed in this study is a hot-potato implementation with switch architecture as described in [12]. The routing strategy is based on non-minimal and load dependent deflection type packet switching, with adaptive per hop routing. A relative addressing scheme is implemented which simplifies the duplication of identical switches when network structures of varying sizes are designed.

The switches employed in all of the network configurations in this study are bufferless, and directly connected with their associated resource with a 1:1 resource to switch ratio. A packet cannot be stored in a switch, and thus in each cycle the packets must be moved from switch to switch or deflected back to a resource. To reduce the complexity of the switches, deterministic routing is favored over adaptive routing in buffered networks [10],[4],[1]-[3]. In bufferless networks, deflection routing is advocated [16] because it is

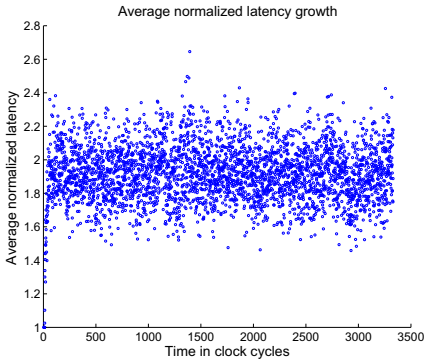


Figure 3. Variation of average normalised latency for a 5×5×5 mesh with 0.5 injection rate under the local traffic model, illustrating the warming up phase

possible to design fast and small switches with simplified control circuitry. The implemented switch uses 128-bit input and output channels for each switch-to-switch Physical Channel (PC) and resource to switch connection. For 2-D structures a 5-port switch is used as in Figure (1a), while for 3-D structures a 7-port switch, as shown in Figure (1b), is used where the inter layer connectivity is implemented with two simplex channels having the same switch-to-switch bitwidth as the horizontal channels.

A TDMA vertical bus to provide connectivity between network layers is also designed as seen in Figure (1c). It is a centrally arbitrated bus employing a least-served-first priority scheme for packet arbitration. The matrix arbiter circuit in [4] has been extended to deal with up to 10 layers. The bus has a 10 packet deep FIFO buffer at each input to prevent packet loss for the maximum network size. The bus runs on a clock 16 times faster than the network clock to provide serialization and de-serialization time, potentially circumventing area limitations for vertical TSV connectivity between layers. Hence the bus can accommodate a reduction in the bit width of the vertical channel up to a ratio of 1/16 and still receive and deliver a packet to a destination, on any layer, within one cycle of the network clock. It is connected to 6-port switches through a 128-bit Input/Output PC. In this study, every switch on a given layer is connected to a separate bus, and a single bus connects all switches that are vertically in-line. Thus, a 5x5x5 network will have 25 buses in total.

The packets are generated by each resource and injected into the network with an injection rate of up to 0.9 packets per node per cycle. For this study, a packet is considered to be one flit long. Each resource has a FIFO buffer to temporarily store packets if they cannot enter the network due

to congestion. These packets are queued and re-injected into the network with a higher priority than newly generated packets. The implemented network does not drop packets. The packet headers generated by the resource contain final destination addresses and the switches make routing decisions on the fly based on this information.

Any NoC structure is comprised of switches located in general at the corners, edges, surface and center of the physical structure. For example a 2-D switch has 4 links to other switches in 4 directions, namely, North, South, East and West. A 3-D switch has two additional, Up and Down links giving a total of 6 bi-directional links, whereas a bus architecture adds only one extra link to the switch, to give a total of 5 bi-directional links. There is also an additional bi-directional port from switch to resource in each case. For a given NoC structure, not all of the links can be used for routing packets. For example, only half of the switch-to-switch links in the corner of the network are connected, halving the available switch bandwidth due to the unconnected edge ports. The total number of connected links, in any NoC structure depends on the topology, dimension and size of the network. Figure 2 depicts the number of links as a function of network size in a mesh topology to highlight the differences between a 2-D 5-port switch, a 3-D 7-port switch, and a 3-D bus architecture as defined in equations (1), (2) and (3) respectively.

For a 2-D $n \times n$ network:

$$L_{2D} = 4n(n-1) \tag{1}$$

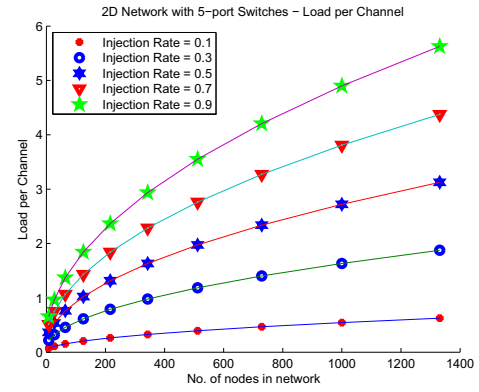
For a 3-D $n \times n \times n$ network:

$$L_{3DN} = 6n^2(n-1) \tag{2}$$

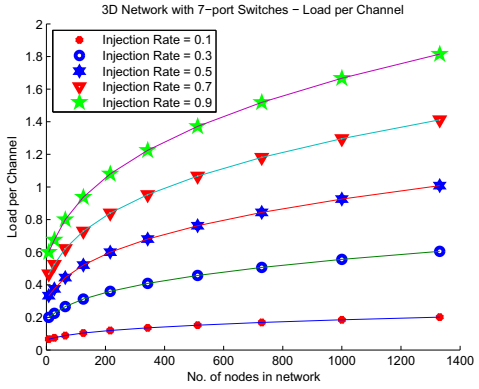
For a 3-D Bus $n \times n \times n$ network:

$$L_{3DB} = 4n^2(n-5) + 32 \tag{3}$$

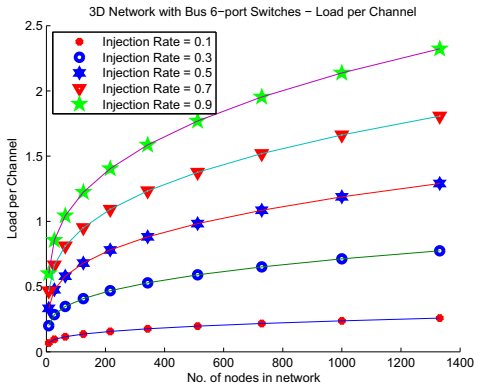
All three curves for the different architectures show a sharp rise as network size increases and then begin to level off as the number of corner and edge nodes (i.e. nodes with some ports unconnected) become outweighed by the fully connected nodes. This trend of growth of links per node allows for higher throughput as the network size increases due to the increased bandwidth available to route a packet in the network. Additionally, it is clear from this figure that the 7-port switch has an advantage over the other topologies due to the increased number of switch-to-switch links per node available to route the traffic in the network for a given network size. The 3-D bus architecture has the least number of links due to the fact that all switches in a given vertical line share a common uni-directional link. The 7-port design benefits from separate PCs between every layer to handle multiple packets simultaneously within one cycle on any given vertical line from the bottom to the top layer.



(a)



(b)



(c)

Figure 4. Load per channel for 2-D, 3-D mesh and 3-D Bus architectures

The bus in contrast can only deal with a single incoming or outgoing packet within a cycle of the network. This trend is evident in the simulation results, as the bus' available band-

width is necessarily limited by its design.

4. Traffic Patterns and Simulation Setup

The simulations were performed for a 2-D network with size $n \times n$, for n = 3, 5, 8, 12, 15, 19, 23, 27 and 32 nodes, with packet injection rates, r, varying from 0.1 to 0.9 packets per node per cycle in step increments of 0.1. For the 3-D 7-port switch and 3-D bus architectures, the network size was $n \times n \times n$ for n = 2, 3, 4, 5, 6, 7, 8, 9 and 10 with the same injection rate as in the 2-D case. The number of simulated nodes in the 2-D mesh follows the 3-D setup to provide a fair comparison. The data sample used is extracted following the warm-up phase of the network and preceding the cool-down phase to ensure reliable results. For example, Figure 3 shows the average latency growth for a 5×5×5 3-D mesh network with an injection rate of 0.5. The warm-up phase occupies the first 200 clock cycles. After 500 clock cycles the network is fully stable until the simulation ends. In our experiments samples were obtained after the network reached stability but for certain combinations of network size and injection rate, the network never becomes stable. These cases are identified in the discussion.

The metrics used to quantify the performance of the network in each case are raw latency, normalized latency, and throughput, defined in (4), (5) and (7) respectively.

$$T_{Raw} = \frac{(C_{Final} - C_{Init})}{HC} \tag{4}$$

The raw latency, T_{Raw}, is the distance traveled by a packet from the source to the destination address in terms of hop counts, denoted by HC. C_{Final} and C_{Init} represent the final and initial clock cycles respectively. When the network is at zero-load, the raw latency is equivalent to the minimum distance. The normalized latency, T_{Norm} is the ratio of the raw latency (i.e. the actual distance traveled by a packet) to the minimum distance with zero-load, T_{Zero_load}, defined as:

$$T_{Norm} = \frac{T_{Raw}}{T_{Zero\_load}} \tag{5}$$

The average normalized latency is the mean of the normalized latencies for the collected samples defined as:

$$T_{Norm\_Avg} = mean(T_{Norm}) \tag{6}$$

In each cycle, packets arrive at all nodes at a rate based on the injection rate and the congestion level of the network. The throughput per node per cycle, λ, is defined in (7), where P_{Total} is the total number of packets received over the simulated range, N is the number of nodes in the network and C is the number of cycles in the sampling region.

$$\lambda = \frac{P_{Total}}{N \times C} \tag{7}$$

The generated traffic patterns attempt to load the network in a realistic manner as would be encountered on a multi-processor SoC. A hybrid of two traffic patterns are employed for all cases, uniform random traffic (URT) and local traffic models. The URT model stipulates that each node in the network is equally likely to become the destination address of any packet emitted from a resource. From a design perspective, it is customary to place frequently communicating resources close to each other to maximize efficiency. This increases the performance of the communication in terms of power, timing and resource management. The local traffic pattern generated for each resource replicates this localized communication behavior. The destination address is assigned randomly and then a localized probability formula [8] is applied to the address, as shown in Equation (8) and (9). This formula increases the probability that the final destination of the packet will be local to the sending resource rather than farther away. This follows the principle that the resources will be arranged within the network such that their nearest neighbors are devices with which they communicate with most frequently.

The localized probability for local traffic patterns is defined in (8), where D is the maximum distance in the network and (9) is a normalizing factor guaranteeing that the sum of all probabilities is 1 [8].

$$P(d) = \frac{1}{(A(D)2^d)} \tag{8}$$

$$A(D) = \sum \frac{1}{(2^d)} \tag{9}$$

In order to compare traffic patterns, a fully URT pattern without localization was also implemented. This pattern means that any address within a given network is equally likely to be assigned to each packet being generated. To maintain stability in the network, the injection rate is lowered as the number of nodes is increased. The injection rate as a function of size is determined by Equation (15) and shown in Table 1.

The load, γ, that each node puts on the network is shown in (10), where r is injection rate from 0.1 to 0.9 and HC is the average hop count in the network [9].

$$\gamma_{Ntwrk} = r \times HC \tag{10}$$

We assume that a packet loads the network with each hop by 1, because it occupies 1 link per hop. Equation (10) shows that the load depends on the injection rate of each node and on the average distance of each packet. HC grows with the size of the network. In k-dimensional meshes when

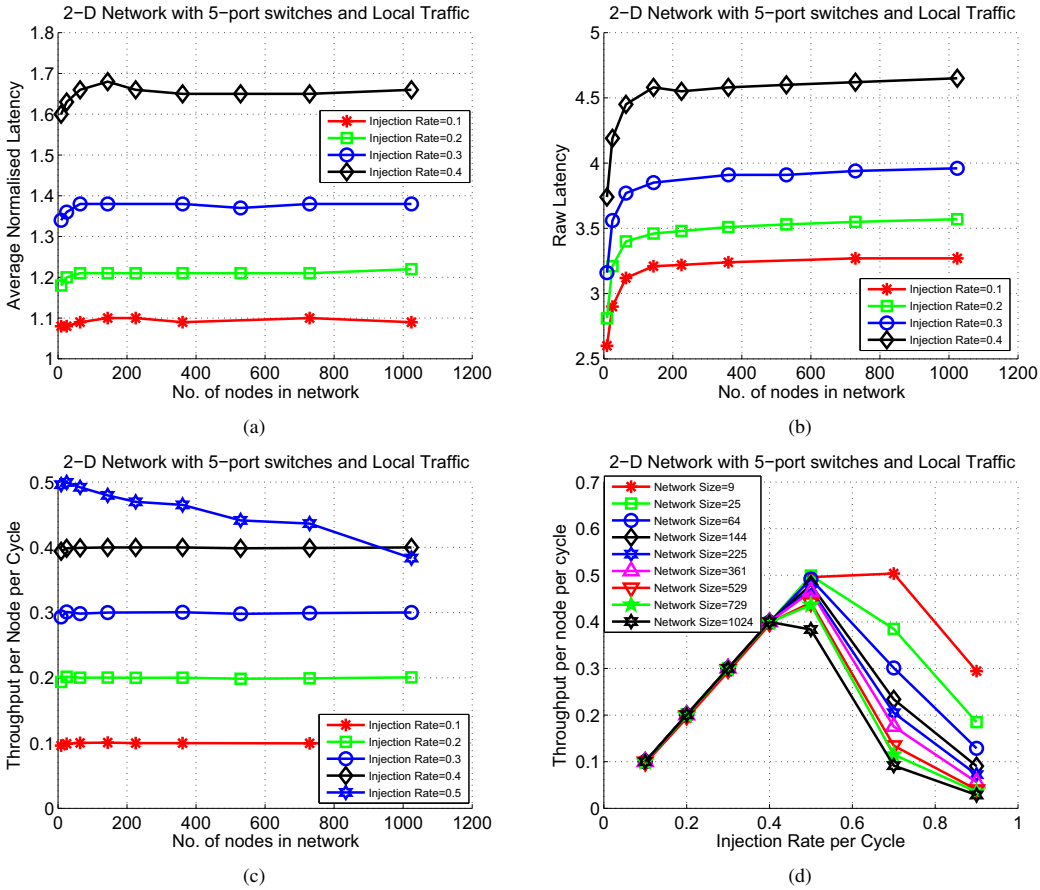


Figure 5. 2-D Mesh Performance

n is even, we have

$$HC = \frac{k \times n}{3} \tag{11}$$

In 3-D meshes HC is coincidentally $HC = \frac{3\times n}{3} = n$, whereas for a 2-D mesh $HC = \frac{2\times n}{3}$. Hence the total load in the network is

$$\gamma_{Ntwrk,2D} = \frac{2r \times n^3}{3} \tag{12}$$

$$\gamma_{Ntwrk,3D} = r \times n^4 \tag{13}$$

The network capacity can be expressed as the number of links for 2-D and 3-D meshes as given in Equation (1) and (2) respectively. As the network grows, the network capacity grows and the network load grows. However, the load under uniform random traffic grows faster than the network capacity. Consequently, the injection rate has to decrease as the network grows. The load per channel, γ_{Chnl}, is defined as:

$$\gamma_{Chnl,3D} = \frac{\gamma_{Ntwrk,3D}}{L_{3d}} \tag{14}$$

n	N	r
2	8	0.75
3	27	0.67
4	64	0.56
5	125	0.48
6	216	0.42
7	343	0.37
8	512	0.33
9	729	0.30
10	1024	0.27

Table 1. Injection rate for a channel load of 0.5 in a 3-D $n\times n\times n$ mesh

The load per channel for 2-D-mesh and 3-D mesh and 3-D bus architectures for varying injection rates is shown in Figures 4(a), (b) and (c) respectively. Substituting equations

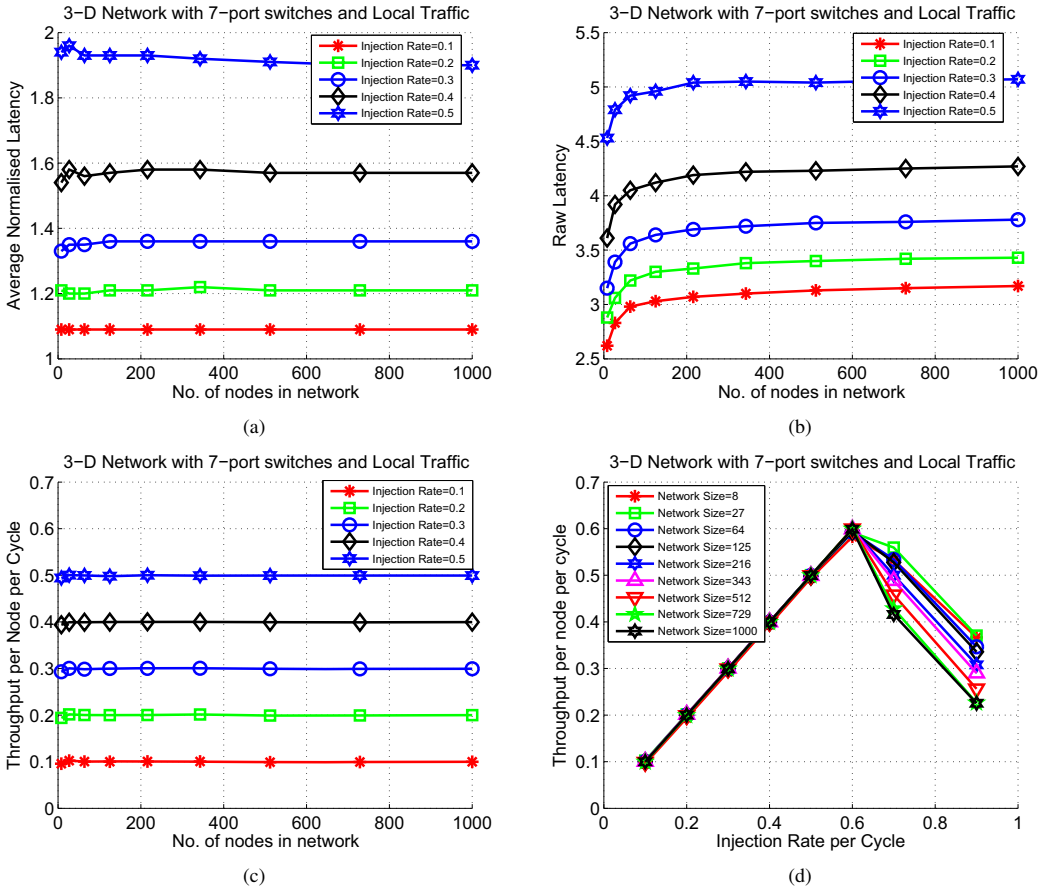


Figure 6. 3-D Mesh Performance

(2), (13), into (14) leads to

$$r = \frac{6 \times \gamma_{Chnl,3D} \times (n-1)}{n^2} \tag{15}$$

Now in order to maintain stability, the load per channel, $\gamma_{Chnl,3D}$, is set constant. For example if γ_{Chnl} is set to 0.5, i.e. a constant load of 0.5 packets per channel, the injection rates, r, corresponding to various network sizes are given in Table 1.

5. Results

The simulation results produced in this study highlight the effect of increasing the number of nodes and injection rates for different NoC topologies on network performance. The results quantify the normalised and raw latency, and throughput versus injection rate for a 2-D mesh, 3-D 7-port mesh, and a 3-D 6-port mesh with vertical bus connectivity for localized and uniform traffic. Figures 5(a), (b) and (c) are plots of normalized latency, throughput and raw latency, respectively, versus network size for injection rates ranging from 0.1 to 0.9 in a 2-D mesh. Figure 5(d) is a plot of throughput versus injection rate for different network sizes. These localized 2-D simulations show that as network size increases, the normalized latency quickly reaches a saturation point whereas the throughput matches the injection rate for stable input conditions. For injection rates of 0.1 to 0.4, once the size has increased beyond 225 nodes, scaling the network size up to 1024 nodes incurs no significant performance penalty in terms of latency or throughput due to the localized traffic pattern. Figure 5(d) shows that the 2-D network investigated in this study fails to match injection rate with throughput above 0.5 packets per node per cycle; i.e. the network becomes congested and unstable. The instability of injection rates above 0.5 was confirmed by an examination of the transmitter output buffers from each resource, where the network becomes unstable when the buffer fills to capacity. Figure 5(d) clearly shows the effect of an overly congested network on the throughput as the number of nodes grows. These 2-D simulations show

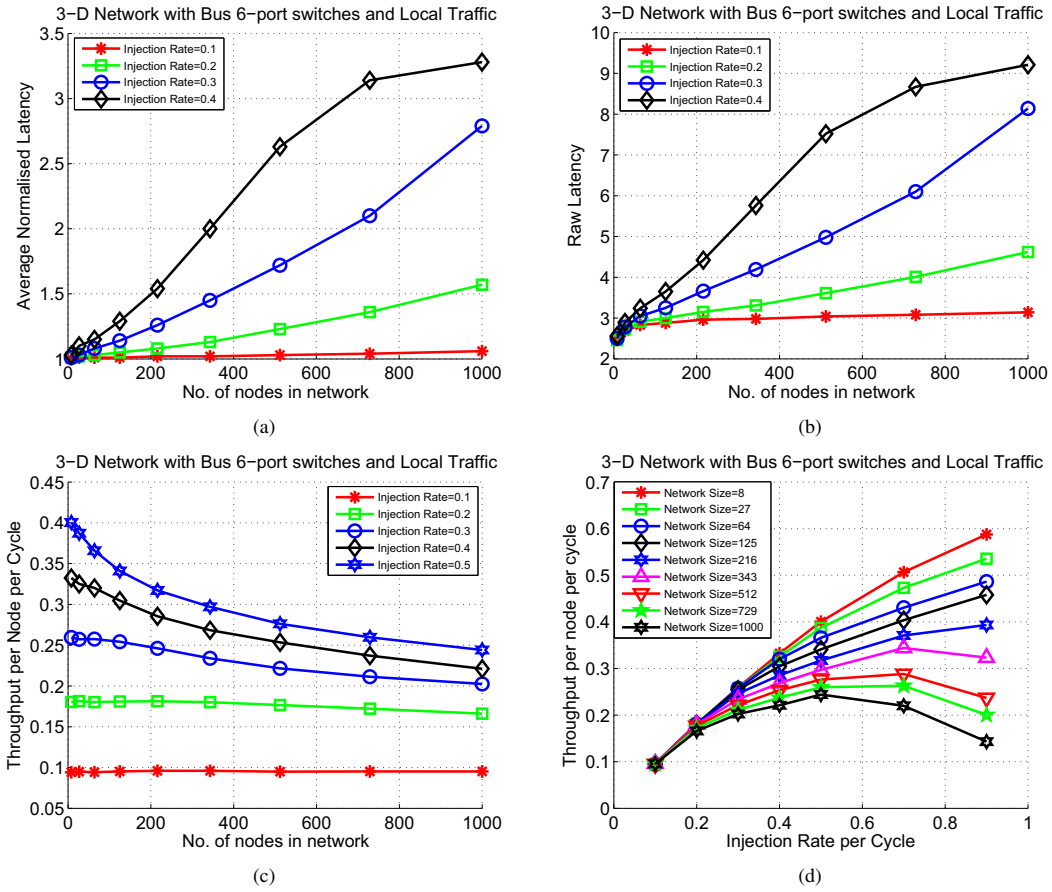


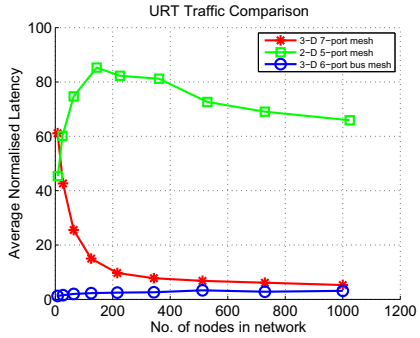
Figure 7. 3-D Bus Performance

that under a localized traffic model, a bufferless 2-D mesh network can maintain relative stability, and provide acceptable performance for up to 1000 nodes with injection rates below 0.5.

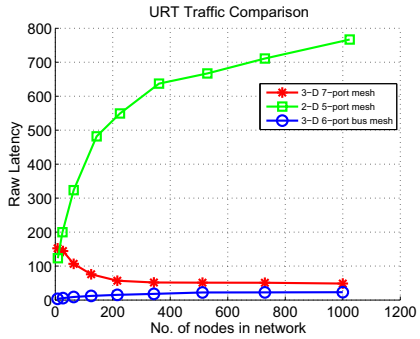
Figures 6(a), (b) and (c) show the normalized latency, throughput and raw latency against network size for varying injection rates in a 7-port 3-D NoC with local traffic. The 3-D 7-port mesh copes with a higher level of traffic better than the 2-D mesh. Where the 2-D network becomes congested at an injection rate of 0.5 as the number of nodes grows, the 3-D 7-port maintains throughput and a constant latency for all network sizes. The increased number of links per node in the 7-port 3-D design allows the network to remain stable for a wider range of network sizes and injection rates. This topology manages to match throughput to injection rate up to 0.6 packets per node per cycle. This result is interesting as it demonstrates the capability of a 3-D 7-port mesh to handle higher injection rates than a 2-D mesh for a given number of nodes. The plot in 6(a) however shows that the average normalized latency increases sharply with injection rates beyond 0.6 as the packets spend more time in output buffers at their origin resource due to congestion. Figure 6(d) is interesting because it correlates the injection rate to the throughput per node per cycle directly, so the saturation point for each network size can be determined. This understanding is paramount to designers as it determines the boundary conditions for maximum performance for different sized networks.

Figures 7(a), (b) and (c) plot the average latency, throughput, and raw latency against network size, respectively for the 3-D mesh network with the vertical bus. The 3-D bus has a markedly worse performance in comparison with the 7-port 3-D topology and the 2-D mesh. Figure 7(a), shows that the average normalized latency values are lower than the corresponding values for both the 2-D and 3-D cases when the network has a lower injection rate.

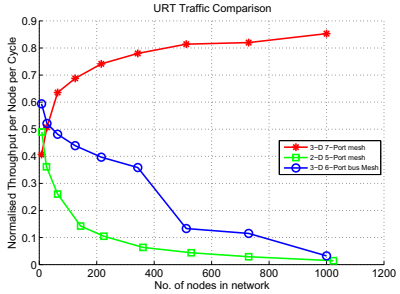
This is largely due to the bus only requiring 1 hop for a packet to reach its destination on any vertical layer. The bus performs well below 200 nodes for low injection rates in terms of latency; however, the throughput plotted in Figure



(a)



(b)



(c)

Figure 8. URT Performance

7(b) begins to drop for injection rates of 0.2 and above for any network size greater than 3x3x3. This is due to packets being deflected horizontally around the network, as contention for the vertical bus link increases with the number of layers. This effect increases as the network size grows as the number of users requesting access to the bus grows in proportion to the number of layers. Finally Figure 7(d) clearly shows that the bus struggles to match throughput to injection rates of over 0.3. The local traffic model is not friendly to the bus topology either, as the advantage the bus provides in sending a packet in one vertical hop to any layer is not fully utilized. Although the bus appears to be the worst of the three architectures under localized conditions, it may have advantages in other traffic conditions, such as when a stack of memory lies above a processor, where the bus will provide equal access time to any layer of memory. The results in Figure 8(a), (b) and (c) relate to the uniform random traffic model showing plots of the average normalized latency, raw latency and normalized throughput versus increasing network size respectively based on the injection rate given in Table 1 for each network size . Figure 8(a) shows that the bus has the lowest latency and that a conventional 2-D network has the highest, with the 3-D 7-port lying in-between. Figure 8(c) plots the ratio of the network throughput per node per cycle over the injection rates given in Table 1 versus the number of nodes. This shows that the 3-D 7-port switch behaves reasonably similar to the localized pattern in that it reaches a saturation point quickly for throughput and further increases in network size add little performance penalty. However the 2-D mesh and 3-D bus architectures are the worst for this traffic pattern. They both have decreasing throughput with increasing network size. In the 2-D case, the packet must travel over a long distance, and the network easily becomes congested. The bus again suffers from the reduced link bandwidth per node when it has to deal with traffic patterns with injection rates greater than 0.1. However, different from the local traffic simulations, the bus has outperformed the 2-D mesh in terms of throughput and latency. The uniform random traffic allows the bus to utilize its ability to transport a packet any distance in just one hop, and this shows through in these results. These figures appear to highlight that under uniform random patterns, the 7-port switch is the best option for large sized networks in terms of packet throughput, but trade offs such as area overhead and power may render the bus a viable alternative, especially when the vertical links are limited, and a scarce resource.

6. Discussion and Conclusions

This paper has explored the scalability of two different bufferless 3-D NoC topologies in an effort to develop design guidelines for future development of massively integrated SoC devices. The cycle accurate simulation results presented clearly show the saturation points and performance in terms of latency and throughput of the different topologies as the number of nodes and layers in a 2-D and 3-D mesh network is increased. This is paramount for designers in determining the best balance between the number of features on a chip and the required communication performance. The results indicated that the 3-D 7-port switch is the best performer in terms of throughput and normalized latency as the number of nodes in a network is increased. It has the highest link per node ratio and thus the most bandwidth between the three designs. However, we have shown that a 2-D NoC with localized architecture can be scaled to a

very large dimension with no significant effect on throughput or latency. The TDMA vertical bus design was shown to perform the worst out of the three communication architectures in terms of scalability under local traffic, as it is physically limited by its raw bandwidth due to a smaller links per node ratio and contention issues as the number of layers increase. Although it can transfer packets through many layers in one hop, it struggles to handle the increased requests and congestion as the network grows. Although shown to be weak in this paper, the bus may be appropriate for hot spot traffic injection where many packets may need to be sent through several layers to a hot spot frequently. This may be akin to a processor on one layer, and a memory stack directly above it. The URT traffic pattern again highlights the 7-port switch's superior bandwidth capability. Of the three architectures, the 3-D 7-port switch is the only one that manages to gain throughput as network size increases.

We aim to build on these preliminary results and carry out further investigations into different traffic patterns, switch architectures, and communication protocols to quantify the performance differences in the various network topologies under more traffic patterns as well as the physical constraints imposed by the horizontal and vertical interconnections.

7. Acknowledgments

Funding support from the European Commission under grant $FP7 - ICT - 215030$ (ELITE) of the 7th framework program is gratefully acknowledged.

References

[1] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat. 3-d ics: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. *Proc. IEEE*, 89(5):602–633, May 2001.

[2] W. J. Dally. Performance analysis of k-ary n-cube interconnection networks. *IEEE Transactions on Computers*, 39(6):775–785, 1990.

[3] W. J. Dally and B. Towles. Route packets, not wires: on-chip interconnection networks. In *Proc. DAC*, pages 684–689, 2001.

[4] W.J. Dally. *Principles and practices of interconnection networks*. Morgan Kaufmann, 2004.

[5] B. Feero and P.P. Pande. Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation. *IEEE Transactions on Computers*, 6, 2008.

[6] J. Kim et al. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. 2007.

[7] F. Li et al. Design and Management of 3 D Chip Multiprocessors Using Network-in-Memory. *ACM SIGARCH Computer Architecture News*, 34(2):130–141, 2006.

[8] A. Salminen E. Lu, Z. Jantsch and C. Grecu. Network-on-Chip Benchmarking Specification, Part II: Micro-Benchmark Specification. In *OCP-IP*, pages 7–8. Springer London, Limited, 2008.

[9] Z. Lu and A. Jantsch. Flit admission in on-chip wormhole-switched networks with virtual channels. In *International Symposium on System-on-Chip Proceedings*, pages 21–24, 2004.

[10] Z. Lu and A. Jantsch. Admitting and ejecting flits in wormhole-switched networks on chip. *Computers and Digital Techniques, IET*, 1(5):546–556, Sept. 2007.

[11] C. Mineo, R. Jenkal, S. Melamed, and W.R. Davis. Inter-Die Signaling in Three Dimensional Integrated Circuits.

[12] E. Nilsson. Design and Implementation of a hot-potato Switch in a Network on Chip. *Mémoire, Departement of Microelectronics and Information Technology, Royal Institute of Technology*, 2002.

[13] D. Park, S. Eachempati, R. Das, A.K. Mishra, Y. Xie, N. Vijaykrishnan, and C.R. Das. MIRA: A Multi-layered On-Chip Interconnect Router Architecture. 2008.

[14] V.F. Pavlidis and E.G. Friedman. 3-D Topologies for Networks-on-Chip. *IEEE Transactions on Very Large Scale Integration Systems*, 15(10):1081, 2007.

[15] D. Sylvester and K. Keutzer. Getting to the bottom of deep submicron. In *Proc. ICCAD*, pages 203–211, 1998.

[16] A.W. Topol et al. Three-dimensional integrated circuits. *IBM Journal of Research and Development*, 50(4):491–506, 2006.

[17] S. Vangal et al. An 80-Tile 1.28 TFLOPS Network-on-Chip in 65nm CMOS. In *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pages 98–589, 2007.

[18] R. Weerasekera, L. Zheng, D. Pamunuwa, and H. Tenhunen. Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs. In *Proc. IEEE/ACM Int. Conf. on Comp.-Aided Design (ICCAD)*, pages 212–219, 2007.