Networks-on-Chips: Theory and Practice

Editors: Fayez Gebali and Haytham Elmiligi
Contents

1 Three-Dimensional Network-on-Chip Architectures .......................... 1

1.1 Introduction .............................................................................. 1

1.2 Related Work ........................................................................... 3

1.3 Alternative Vertical Interconnection Topologies ......................... 6

1.4 Overview of the Exploration Methodology ............................... 9

1.5 Evaluation – Experimental Results ........................................... 12

1.5.1 Experimental Setup ............................................................ 12

1.5.2 Routing Procedure ............................................................... 14

1.5.3 Impact of Traffic Load ......................................................... 17

1.5.4 3D NoC Performance under Uniform Traffic ..................... 18

1.5.5 3D NoC Performance under Hotspot Traffic ...................... 22

1.5.6 3D NoC Performance under Transpose Traffic .................. 25

1.5.7 Energy Dissipation Breakdown ............................................ 27

1.5.8 Summary .............................................................................. 27
1.6 Conclusions .......................................................... 28
Chapter 1

Three-Dimensional Network-on-Chip Architectures

1.1 Introduction

Future integrated systems will contain billion of transistors [51], composing tens to hundreds of IP cores. These IP cores, implementing emerging complex multimedia and network applications, should be able to deliver rich multimedia and networking services. An efficient cooperation among these IP cores (e.g., efficient data transfers) can be achieved through utilization of the available resources.

The design of such complex systems includes several challenges to be addressed. Among others one challenge is to design an on-chip interconnection network that should be able to efficiently connect the IP cores. Another challenge is to derive such an application mapping that will make efficient usage of the available hardware resources [39, 21]. An architecture that is able to accommodate such a high number of cores, satisfying the need for communication and data transfers, is the Network-on-Chip (NoC) architecture [5, 25]. For these reasons Networks-on-Chip become a popular choice for designing the on-chip interconnect for Systems-on-Chip (MPSoCs), and are supported from the industry (such as the Æthereal
Chapter 1. Three-Dimensional Network-On-Chip Architectures

NoC [18] from Philips, the STNoC [55] from STMicroelectronics and an 80-core NoC from Intel [57]). As it is presented in [43], the key design challenges of emerging NoC design are a) the communication infrastructure, b) the communication paradigm selection and c) the application mapping optimization.

The type of the IP cores (their characteristics, capabilities) as well as the topology and interconnection scheme plays an important role on how efficiently an NoC will perform for a certain application or set of applications. Furthermore, the application features (e.g., data transfers, communication and computation needs) plays an equally important role in the overall performance of the NoC system. For this reason, in order to take full advantage of the hardware resources the NoC architecture should be able to accommodate efficiently the applications’ needs providing an application-specific (an application-domain specific) architecture. An overview of the cost considerations on the design of NoCs is given at [9].

Up to now NoC designs were limited to two dimensions. But the currently emerging 3D integration technology exhibits, among others, two major advantages, namely higher performance and smaller energy consumption [7]. A survey of existing 3D fabrication technologies is presented in [8], showing the available interconnection architectures among the layers of 3D integrated circuits and illustrating the main research issues in current and future 3D technologies. So, due to process / integration technology advancements, it is feasible to design and manufacture NoCs that will expand to the third dimension (3D NoCs). In order to satisfy the demands of emerging systems for scaling, performance and functionality 3D integration is a way to accommodate these demands. For example, a considerable reduction can be achieved in the number and length of global interconnect using three-dimensional integration [27].

In this chapter we present an architectural exploration methodology designing alternative 3D NoC architectures. We define as 3D NoCs these architectures composed of many layers, where each layer is a two-dimensional NoC grid, where the grids are the same for all the layers, composed of elements of the same types. The main objective of the methodology is to derive to heterogeneous 3D NoC topologies with a mix of 2D and 3D routers and vertical link interconnection patterns that performs best to the incoming traffic. The cost factors we
consider are: (i) energy consumption; (ii) average packet latency and (iii) total switch block area, and we perform comparisons against an NoC that all the routers are 3D ones. We have employed and extended the Worm_Sim NoC simulator [36], being able to model these heterogeneous architectures and simulate them, gathering information on how they perform. The NoC heterogeneity can be achieved using a mix of two- and three-dimensional routers for each layer of the NoC, which implies to a “reduced” presence of vertical interconnection links. The design methodology evaluates such heterogeneous topologies, targeting mesh and torus ones, for various inputs and shown which ones can handle best the corresponding types of traffic.

The rest of the chapter is organized as follows. In Section 1.2 the related work is described. In Section 1.3 we present the 3D NoC topologies under consideration, whereas in Section 1.4 the proposed methodology is introduced. In Section 1.5 the simulation process and the achieved results are presented. Finally, in Section 1.6 the conclusions are drawn and the future work is outlined.

1.2 Related Work

The on-chip interconnection is a widely studied research field and good overviews are [16, 14], illustrating the various interconnection schemes available for present ICs and emerging Multiprocessor Systems-on-Chip (MPSoC) architectures. The use of an NoC-based interconnection is able to provide an efficient and scalable infrastructure that is going to be able to handle the increased communication needs. Lee et al. [31] presented a quantitative evaluation of 2D point-to-point, bus and NoC interconnection approaches. In this work an MPEG-2 implementation is studied, exploiting the aforementioned interconnection solutions, proved that the NoC-based solution scales very well in terms of area, performance and power consumption.

In order to support the NoC design, a number of simulators has been developed, such as the Nostrum [35], Polaris [53], XPipes [13] and Worm_Sim [36] using C++ and/or SystemC [44]. To provide adequate input / stimuli to an NoC design, usually synthetic traffic is
used. Several synthetic traffic generators have been proposed [47, 54, 20, 50] that are able to provide adequate inputs to NoC simulators for evaluation and exploring the communication infrastructure designs.

In [41] a methodology that is able to synthesize NoC architectures where long-range links are inserted on top of a mesh network is proposed. In this way the NoC is transformed to an application specific one, serving best the traffic, but it is limited to two dimensions. Li et al. [33] presented a mesh-based 3D network-in-memory architecture, using a hybrid NoC/bus interconnection fabric to accommodate efficiently processors and L2 cache memories in 3D NoCs. It is demonstrated that by using a 3D L2 memory architecture, better results are achieved comparing with the two-dimensional designs.

In the work of Koyanagi et al. [30] a 3D integration technique of vertical stacking and gluing together of several wafers is presented. Utilizing this technology the authors are able to increase the wiring connectivity while reducing the number of long interconnections. A fabricated three-dimensional shared memory is presented in [32]. The memory module has three layers and can perform wafer stacking using the following technologies: i) formation of buried interconnection; ii) micro-bumps; iii) wafer thinning; iv) wafer alignment and v) wafer bonding. Another 3D integration scheme is proposed in [24], where wireless interconnections are being employed in order to offer connectivity.

In [37] an overview of the available interconnect solutions for Systems-on-Chip (SoC) are presented. This study includes interconnects for 3D ICs and shows that 3D integration reduces the length of the longest global interconnects [28] and reduces the total required wire length and thus the dissipated energy [26].

In the work of Benkart et al. [6] a overview of the 3D chip stacking technology using through-chip interconnects is presented. In this work the trade-off between the high number of vertical interconnects versus the circuit density is highlighted, since the through-silicon vias occupy active chip area. Furthermore, Davis et al. in the work presented in [15] show an implementation of an FFT in a 3D IC achieving 33% reduction in maximum wire length, proving that the move to 3D ICs is beneficial. However, they highlight as limiting factors the heat dissipation and yield.
1.2. RELATED WORK

The placement and routing in 3D integrated circuits is being studied in [2] and a system on package solution for three-dimensional systems is presented in [34]. A big challenge remains the heat dissipation of 3D circuits [22]. In order to tackle this several analysis techniques have been proposed [23, 10, 48]. Another way to approach this issue is to perform thermal-aware placement and mapping for 3D NoCs, such as the work presented in [3]. Furthermore, the insertion of thermal vias can lower the chip temperature as illustrated in [19, 12].

In [42] a generalized NoC router model is presented, and then based on that the authors perform NoC performance analysis. Using the aforementioned router model it is feasible to perform NoC evaluation which is significantly faster than performing simulation. Additionally, Pande et al. [45] present an evaluation methodology in order to compare the performance and other metrics of a variety of NoC architectures. But, this comparison is made only amongst two dimensional NoC architectures. The work of Feero and Pande, presented in [17], extends the aforementioned work considering 3D NoCs and illustrates that the 3D NoCs are advantageous when compared to 2D ones (with both having the same number of components in total). It is demonstrated that besides reducing the footprint in a fabricated design, three-dimensional network structures provide a better performance compared to traditional, two-dimensional architectures. This works shows that despite the face of a small area penalty, 3D NoCs achieve significant gains in terms of energy, latency and throughput.

In [46] Pavlidis and Friedman presented and evaluated various 3D NoC topologies, proposing an analytic model for 3D NoCs. Mesh topologies are considered, modeling the zero-load latency of them. The authors assumed 100% vertical interconnection vias focused on the physical level study of these silicon vias. Kim et al. [29] presentd an exploration of communication architectures on 3D NoCs. A dimensionally-decomposed router and its comparison with a hop-by-hop router connection and hybrid NoC-bus architecture is presented. The aforementioned works, both from the physical level as well as adding more communication architectures, such as full 3D crossbar and bus-based communication, are complementary to the one presented here and can be used for extension of the methodology.

The main differentiator with the related work is that we do not assume full vertical
interconnection (as it is shown in Figure 1.1), but an heterogeneous interconnection fabric, composed of a mix of 3D and 2D routers. Additional motivation for this heterogeneous design is not only the reduced total interconnection network length but the reduced size of the 2D routers have when compared to the 3D ones [17]. In this way, reducing the number of vertical interconnection links the fabrication of the design is easier and more active chip area is being used by the available logic / memory blocks. Two-dimensional routers are the routers that have connections with the neighboring ones of the same grid. Whereas, when we say that a router is a 3D one, it means that it has direct, hop-by-hop, connections not only with the neighboring routers belonging to the same grid but also to the ones belonging to the adjacent layers. This difference between two- and three-dimensional routers for a 3D mesh NoC is illustrated in Figure 1.1, where it is shown a grid that belongs to a 3D NoC and in that grid are present 2D and 3D routers.

1.3 Alternative Vertical Interconnection Topologies

We assume four different groups of interconnection patterns, as well as the ten vertical interconnection patterns used in the context of this work. Considering a 3D NoC, where each layer has dimensions $x \times y$, and only $K\%$ of the routers can have connections to the vertical direction as well (called 3D routers). The available scenarios of how these 3D routers can be placed on a layer - grid are:

1. **Uniform**: distribution of the 3D routers over the different layers. Using this scheme we “spread” the 3D routers along every layer of the 3D NoC. In order to find the place of each router we work like this:

   - first place the first 3D router of the $(0,0)$ position of each layer ($z = 0,\ldots$, number of grids),
   - then the four neighboring 2D routers are placed in the positions $(x+r+1, y, z)$, $(x-r-1, y, z)$, $(x, y+r+1, z)$ and $(x, y-r-1, z)$. The $r$ parameter is defined as:

   $$ r = \left\lfloor \frac{1}{K\%} - 1 \right\rfloor $$


1.3. ALTERNATIVE VERTICAL INTERCONNECTION TOPOLOGIES

and it represents the number of 2D routers among consecutive 3D ones. In Figure 1.1(b) is illustrated this scheme, showing one layer of a 3D NoC, with $K = 25\%$, meaning that $r = 3$.

2. **Center**: All the “3D routers” are positioned at the center of each layer, as it is shown in Figure 1.1(c). Since vertical interconnection links exist only in the center of the layer, in the outer region of the NoC grid the routers are 2D ones, connecting only to the neighboring routers of the same grid.

3. **Periphery**: The 3D routers are positioned at the periphery of each layer (as it is shown in Figure 1.1(d)), in a sense the opposite vertical interconnection link to the scheme presented earlier. In this case, the NoC is focused in serving best the communication needs of the outer cores.

4. **Full Custom**: The position of the 3D routers is fully customized matching perfectly the needs of the application with the NoC architecture. This solution fits best the needs of the application, while it minimizes the occupied area by the switching blocks, by “reducing” the number of vertical links and thus the number of 3D routers. However, derivation of a full custom solution requires high design-time cost, since this exploration is going to be performed for every application. Furthermore, this will create a non-regular design that will not adjust well in a potential change of the functionality, the number of applications that are going to be executed, etc.

The aforementioned patterns were based on the work on 3D FPGAs by [52]. In order to perform exploration towards full custom interconnection schemes real applications and/or application traces are needed. In this chapter we have adopted various types of synthetic traffic, so the exploration for full custom interconnections schemes is out of the scope. More specifically we perform exploration towards pattern based vertical interconnection topologies (categories 1-3). We have considered ten different vertical link interconnection topologies. For each of these topologies the number of 3D routers is given and inside parenthesis the corresponding $K$ percentage, considering a $4 \times 4 \times 4$ NoC architecture.

- **Full**: where all the routers of the NoC are 3D ones (number of 3D routers: 64 (100%)).
CHAPTER 1. THREE-DIMENSIONAL NETWORK-ON-CHIP ARCHITECTURES

(a) Full vertical interconnection (100%) for a 3D NoC.

(b) Uniform distribution of vertical links.

(c) Positioning of vertical links at the center of the NoC.

(d) Positioning of the vertical links at the periphery of the NoC.

(e) Legend:

- 3D Router
- 2D Router
- Interconnection Link
- Processing Node

Figure 1.1: Positioning of the vertical interconnection links, for each layer of the 3D NoC (each layer is a 6 × 6) grid.
• **Uniform based**: pattern based topologies with \( r \) value equals to three (by\_three pattern, as shown in Figure 1.1(b)), four (by\_four) and five (by\_five). Correspondingly the number of 3D routers is: 44 (68.75%), 48 (75%) and 52 (81.25%).

• **Odd**: In this pattern all the routers belonging to the same row are of the same type. Two adjacent rows never have the same type of router (number of 3D routers: 32 (50%)).

• **Edges**: Where the center (dimensions \( x \times y \)) of the 3D NoC has only 2D routers (number of 3D routers: 48 (75%)).

• **Center**: Where only the center (dimensions \( x \times y \)) of the 3D NoC has 3D routers (number of 3D routers: 16 (25%)).

• **Side based**: Where a side (e.g., outer row) of each layer has 2D routers. Patterns evaluated had one (one\_side), two (two\_side), or three (three\_side) sides as “2D only”. The number of 3D routers for each pattern is 48 (75%), 36 (56.25%) and 24 (37.5%) correspondingly.

Each of the aforementioned vertical interconnection schemes has advantages and disadvantages and how these schemes perform is based on the behavior of the applications that are implemented on the NoC. As it is explained in the experimental results (Section 1.5) a wrong choice may diminish the gains of using a 3D architecture.

1.4 Overview of the Exploration Methodology

An overview of the proposed methodology is shown in Figure 1.2. In order to perform the exploration for alternative topologies of 3D NoC architectures, we have used as a basis the Worm\_Sim NoC Simulator [36] that utilizes wormhole switching [40] (this is the center block in Figure 1.2).

In order to support 3D architectures / topologies, we have extended this simulator, adapting the provided routing schemes, and assuming compatibility with the Trident traffic format [54]. As it is shown in Figure 1.2 now the simulator supports 3D NoC architectures (3D Mesh and 3D Torus – as shown in Figure 1.3) and vertical link interconnection patterns.
Figure 1.2: An overview of the exploration methodology of alternative topologies for 3D Networks-on-Chip.
1.4. OVERVIEW OF THE EXPLORATION METHODOLOGY

Figure 1.3: 3D NoC architectures: (a) Mesh and (b) Torus.

Each of these 3D architectures is composed of many grids, with each grid in turn from tiles that are connected to each other using mesh and torus interconnection networks. Each tile is composed of a processing core and a router. Since we are considering 3D architectures the router is connected to the six neighboring tiles and its local processing core via channels, consisting of two one-directional point-to-point links.

The NoC simulator can be configured using these parameters (as it is shown in Figure 1.2):

1. The NoC architecture (two- or three-dimensional mesh and torus architectures, as well as defining the specific grid size (x and y parameters) and number of layers (z parameter).
2. The type of input traffic (uniform, transpose or hotspot) as well as how heavy the traffic load will be.
3. The routing scheme.
4. The vertical link configuration file, which defines where vertical links are present or not.
5. The router model as well as the models used in order to calculate the energy and delay figures.

The output of the simulation is a log file containing the relevant cost factors we evaluate, such as overall latency, average latency per packet and the energy breakdown of the NoC, providing numbers for link energy consumption, crossbar and router energy consumption.
etc. From these energy figures we calculate the total energy consumption of the 3D NoCs.

The 3D architectures to be explored may have a mix of two- and three-dimensional routers, ranging from very few 3D routers to only 3D routers (100% vertical interconnection link presence). In order to steer the exploration we are based on different patterns (as they were presented in Section 1.3. The proposed 3D NoCs can be constructed by placing a number of identical two-dimensional NoCs on individual layers, providing communication by inter-layer vias among vertically adjacent routers. This means that the position of silicon vias is exactly the same for each layer. Hence, the router configuration is extended to the third dimension, while the structure of the individual logic blocks (IP cores) remains unchanged.

1.5 Evaluation – Experimental Results

The main objective of the methodology and the exploration process is to find alternative irregular 3D Network-on-Chip topologies with a mix of two- and three-dimensional routers, exhibiting vertical link interconnection patterns that perform best to the incoming traffic. Our primary cost function is the energy consumption, with the other cost factors being the average packet latency and total switch block area. We compare these patterns against the fully vertically interconnected 3D NoC as well the 2D one (with all having the same number of nodes).

1.5.1 Experimental Setup

The three-dimensional router uses as a switching fabric a $7 \times 7$ crossbar switch, whereas the two-dimensional one uses as a switching fabric a $5 \times 5$ crossbar switch. Additionally, each router has a routing table and based on the source/destination address, the routing table decides which output link the packet should deliver to. The routing table is being built using the algorithm described in Figure 1.4.

As an energy model the NoC simulator is using the Ebit model, proposed in [58]. We
make the assumption (based on the work presented in [49]) that the vertical communication links between the layers are electrically equivalent to horizontal routing tracks with the same length. In this way we consider that the energy consumption of a vertical link between two routers is the same one as the consumption of a link between two neighboring routers of the same layer (if they have the same length).

More specifically and based on the fact that the 3D integration technology, which provides communication among layers using through silicon vias (TSVs), has not been explored sufficiently yet, careful design of systems that employ such interconnection is required. Due to the large variation of the 3D TSV parameters, such as diameter, length, dielectric thickness, and fill material among alternative process technologies, a wide range of measured resistances, capacitances, and inductances have been reported in the literature. Typical values for the size (diameter) of TSVs is about $4 \times 4 \mu m$, with a minimum pitch around $8 - 10 \mu m$, while their total length starting from plane T1 and terminating on plane T3 is $17.94 \mu m$, implying wafer thinning of planes T2 and T3 to approximately $10 - 15 \mu m$ [38, 56, 1].

The different TSV fabrication processes lead to a high variation in the corresponding electrical characteristics. More specifically, from the state-of-the-art solutions that can be found in relevant literature, the resistance of a single 3D via varies from $20m\Omega$ to as high as $600m\Omega$ [56, 1], with a feasible value (in terms of fabrication) around $30m\Omega$. Regarding the capacitances of these vias, their value in literature vary from $40f\Phi$ to over $1p\Phi$ [4], with feasible value for fabrication to be around $180f\Phi$. In the context of this work we assume a resistance of $350m\Omega$ and capacitance of $2.5f\Phi$.

Using our extended version of the NoC simulator, we have performed simulations involving a 64-node and a 144-node architecture with 3D mesh and torus topologies with synthetic traffic patterns. The configuration files describing the corresponding link patterns are supplied to the simulator as an input. The sizes of the 3D NoCs we simulated were $4 \times 4 \times 4$ and $6 \times 6 \times 4$, whereas the equivalent 2D ones were the $8 \times 8$ and $12 \times 12$. We have used three types of input (synthetic traffic) and three traffic loads (heavy, normal and low):

- **Uniform**: Where we have uniform distribution of the traffic across the NoC with the nodes receiving approximately the same number of packets.
• **Transpose:** In this traffic scheme packets originating from node \(a, b, c\) have as destination the node \((X - a, Y - b, Z - c)\), where \(X, Y, Z\) are the dimensions of the NoC.

• **Hotspot:** Where some nodes (a minority) receive increased number of packets (in our case it was at least 100% increased) than the majority of the rest of the nodes (which they receive packets in a uniform manner). The hotspot nodes in the 2D grids are positioned in the middle of every quadrant, where the size of the quadrant is specified by the dimensions of each layer in the 3D NoC architecture under simulation. Whereas, in the 3D NoC, a hotspot is located in the middle of each layer.

We have used three routing schemes present in Worm_Sim [36], and extended them in order to function in a 3D NoC:

• **XYZ-OLD:** Where it is a extended version on XY routing.

• **XYZ:** It is based on XY routing but this variation checks which direction has lower delay and takes the one with the lower delay.

• **ODD-EVEN:** This is the odd-even routing scheme presented in [11]. In this scheme the packets take some turns is order to avoid deadlock situations.

From the simulations performed we have extracted figures regarding the energy consumption (in J) and the average packet latency (in cycles). Additionally, for each vertical interconnection pattern, as well as for the 2D NoC we have the occupied area of the switching block, based in the gate equivalent of the switching fabric presented in [17]. A good design is one that exhibits lower values in the aforementioned metrics when compared to the 2D NoC as well to the 3D NoC which has full vertical connectivity (all the routers are 3D ones). Furthermore, all the simulation measurements were taken for the same number of cycles the network was operational (200,000 cycles).

### 1.5.2 Routing Procedure

Furthermore, we have modified the routing procedure, as shown in Figure 1.4 (valid for all routing schemes) in order to be able to route packets over the 3D topologies. This
1.5. EVALUATION – EXPERIMENTAL RESULTS

1: function ROUTINGXYZ
2:   src : type Node;  // this is the source node
3:   dst : type Node;  // this is the destination node (final)
4:   findCoordinates(); // returns src.x, src.y, src.z, dst.x, dst.y and dst.z
5:   for all layer ∈ NoC do
6:     if packet passes from layer then
7:       findTmpDestination(); // find a temporary destination of the packet for each
8:       end if
9:     end for
10:    while tmpDestination NOT dst do // if we have not reached the final destination...
11:      packet.header = tmpDestination;
12:     end while
13:  end function

14: function findTmpDestination // for each layer that the packet is going to traverse
15:   tmpDestination.x = dst.x
16:   tmpDestination.y = dst.y
17:   tmpDestination.z = src.z  // for xyz routing
18:   for all validNodes ∈ layer do
19:     if link NOT valid then // if vertical link does not exist. This information is
20:       newLink = computeManhattanDistance(); // returns the position of a vertical
21:       tmpDestination = newLink;
22:     else
23:       tmpDestination = link;
24:     end if
25:   end for
26: end function

Figure 1.4: Routing algorithm modifications. (The // denote a comment in the algorithm)

modification allows the customization of the routing scheme in order to efficiently cope with
the heterogeneous topologies, based on vertical link connectivity patterns.

The steps of the routing algorithm are:

1. For each packet we know the source and destination nodes (lines 2 and 3) and we can
   find the positions of these nodes in the topology. The on-chip “coordinates” of the
   nodes are: for the destination one are dst.x, dst.y, dst.z and for the source one are
   src.x, src.y, src.z (line 5).

2. By doing so we can formulate the temporary destinations, that is one temporary des-
CHAPTER 1. THREE-DIMENSIONAL NETWORK-ON-CHIP ARCHITECTURES

tination per layer. More specifically, for the number of layers a packet has to traverse in order to arrive to its final destination, the algorithm sets the route to a temporary destination located at position \( \text{dst.x, dst.y, src.z} \) initially (lines 17-19). The algorithm takes under consideration the “direction” of the packet is going to follow across the layers (i.e., if it is going to an upper or lower layer according to its “source” layer) and finds the nearest valid link at each layer. This process has as an outcome to update properly the \( z \) coefficient of the temporary destination’s position (lines ). Valid link is every vertical interconnection link available in the layer that the packet traverses. This information is obtained through the vertical interconnection patterns file. A link is uniquely identified by the node that is connected to and its direction. So, for all the specified valid links that are located at the same layer with the header flit of the packet check if it matches with the desired for the route to the destination up/down link.

3. If there is no match between them, compute the Manhattan distance (line 23) (in the case of 3D torus topology we have modified it in order to produce the correct Manhattan distance between the two nodes).

4. Finally, the valid link with the smallest Manhattan distance is chosen and its corresponding node is chosen to be the temporary destination at each layer the packet is going to traverse (lines 24-26).

5. After finding a set of temporary destinations (each one located at a different layer), they are stored into the header flit of the packet (line 13). The aforementioned temporary destinations may or may not be used, as the packet is being routed during the simulation, so they are “candidate” temporary destinations. The criterion of being just a candidate or the actual destination per layer is specified according to a set of vertical links that exhibited relatively high utilization during a previous simulation with the same network parameters and setting the desired minimum link communication volume or according to a given vertical link pattern as they were presented at Section 1.1.

Since the modification of the algorithm is composed of a check if a vertical link exists in the temporary destination of the packet, and if not find the closest router with such a link, we manage to keep the routing complexity low.
1.5. EVALUATION – EXPERIMENTAL RESULTS

1.5.3 Impact of Traffic Load

On top of the traffic schemes three different traffic loads were used (heavy, medium/normal, low). In this way, by altering the packet generation rate it is possible to test the performance of the NoC. The heavy load has 50% increased traffic, whereas the low one has 90% decreased traffic compared to the medium one respectively. The behavior of the NoCs in terms of the average packet latency is shown in Figure 1.5. In this Figure the latency is normalized using as basis the average packet latency of the full_connectivity 3D NoC under medium load and for each traffic scheme. It can be seen the impact of the traffic load (latency increases as the load increases) and that the NoCs can cope with the increased traffic as well as the differences between the different traffic schemes.

Mesh topologies exhibit similar behavior, though the latency figures are higher due to the decreased connectivity when compared to torus topologies. This is shown in Figure 1.6 where the latency of 64-node mesh and torus NoCs are compared (the basis for the latency...
Figure 1.6: Impact of traffic load on 2D and 3D mesh and torus NoCs (for uniform traffic). normalization is the average packet latency of the full_connectivity 3D torus. From this comparison it is shown that the mesh topologies have an increased packet latency of 34% compared to the torus ones (for the same traffic scheme, load and routing algorithm).

1.5.4 3D NoC Performance under Uniform Traffic

In Figure 1.7 the results of employing a non-fully vertical link connectivity to 3D mesh networks by using uniform traffic, medium load and xyz-old routing are presented. We make a comparison of the total energy consumption, average packet latency, total area of the switching blocks (routers) and the percentage of 2D routers (having 5 I/O ports instead of 7) under $4 \times 4 \times 4$ (Figure 1.7(a)) and $6 \times 6 \times 4$ (Figure 1.7(b)) mesh architecture. In the x-axis all the interconnection patterns are presented. In the y-axis, in a normalized manner (used as basis the figures of the full vertically interconnected 3D NoC), the cost factors for
1.5. EVALUATION – EXPERIMENTAL RESULTS

total energy consumption, average packet latency, total switching block area and percentage of vertical links are presented.

The advantages of 3D NoCs when compared to 2D ones are shown in Figure 1.7(a). In this case the $8 \times 8$ mesh dissipates 39% more energy and has 29% higher packet delivery latency. However, since its switching area is 71% of the area of the fully interconnected 3D NoC, since all its routers are 2D ones. Employing the by_five link pattern results in 3% reduction in energy and 5% increase in latency. In this pattern only 81% of the routers are 3D ones so we have a reduce area of the switching logic by 5%. Moving to bigger dimensions and as it can been seen from Figure 1.7(b) more patterns exhibit better results. It is worth noticing that the overall performance of the two-dimensional NoC significantly decrease, exhibiting around 50% increase in energy and latency.

When we increase the traffic load by increasing the packet generation rate by 50% we see that all patterns have a worst behavior than the one of the full_connectivity 3D NoC. The reason is that by using a pattern-based 3D NoC we decrease the number of 3D routers, decreasing the number of vertical links, thus reducing the connectivity within the NoC. As it is expected this reduced connectivity has a negative impact in cases where there is increased traffic.

In the case that there is a low traffic load in the NoC the patterns can become beneficial since there are not that high needs for communication resources. This effect is illustrated in Figure 1.8. In this Figure are presented the experimental results for a 64- and 144-node 2D and 3D NoCs under low uniform traffic and xyz routing. The exception is the edges pattern in the 64-node 3D NoC (Figure 1.8(a)), where all the 3D routers reside in the edges of each layer of the 3D NoC. This results in a 7% increase in the packet latency. Again it is worth noticing that as the NoC dimensions increase the performance of the 2D NoC decreases. This can be clearly seen in Figure 1.8(b) where the 2D NoC has 38% increased energy dissipation and 37%.

We have also compared the performance of the proposed approach against that achievable with a torus network, which provides wrap around links added in a systematic manner. Note that the vertical links connecting the bottom with the upper layers are not removed,
Figure 1.7: Uniform traffic on a 3D NoC for alternative interconnection topologies.
1.5. **EVALUATION – EXPERIMENTAL RESULTS**

(a) Experimental results for a $4 \times 4 \times 4$ 3D Mesh.

(b) Experimental results for a $6 \times 6 \times 4$ 3D Mesh.

Figure 1.8: Uniform traffic on a 3D NoC for alternative interconnection topologies.
as this is the additional feature of the torus topology when compared to the mesh. Our simulations show that using the transpose traffic scheme, the vertical link patterns exhibit notable results, and this is goes better and better as the dimensions of the NoC get bigger. The explanation is that the flow of packets between a source and a destination is following a diagonal course among the nodes at each layer and this is also true the source-destination pair in 3D topologies, and this is where the wrap around links of the torus topology play a significant role in non reducing the performance even we remove some vertical links. And the results show that the bigger the dimensions of the NoC are, the energy savings also get bigger when the link patterns are applied. But, this is not true for the case of mesh topology. In particular, in the $6 \times 6 \times 4$ 3D torus architecture, using the by_five, by_four, by_three, one_side, two_side patterns show better results as long as the energy consumption is concerned, for instance, the two_side exhibit 7.5% energy savings and increased latency 32.84 cycles relatively to the 30 cycles of the fully vertical connected 3D torus topology.

1.5.5 3D NoC Performance under Hotspot Traffic

In the case of hotspot traffic (Figure 1.9), testing the $4 \times 4 \times 4$ 3D mesh architecture, seven out of nine link patterns perform better relatively to the fully vertical connected topology. For instance, the two_side pattern exhibits 2% decrease in network energy consumption whereas the increase in latency is 2.5 cycles, note that only 56.25% of the vertical links are present. The hotspot traffic in 3D mesh topologies favors of cube topologies (for example $6 \times 6 \times 6$), even so, in $6 \times 6 \times 4$ mesh architecture the center and two_side patterns exhibit similar performance regarding average cycles per packet compared to that of fully vertical connected architecture (that was expected due to the location where the hotspot nodes were positioned).

In Figure 1.10 the simulation results for the two 3D NoC architectures when triggered by a hotspot-type traffic are presented. In Figure 1.10(a) the results for the mesh architecture and in Figure 1.10(b) the results for the torus architecture are presented respectively, showing gains in energy consumption and area, with a negligible penalty in latency. Again the architectures where congestion is experienced are highlighted.
1.5. **EVALUATION – EXPERIMENTAL RESULTS**

![Graph](image)

(a) Experimental results for a $4 \times 4 \times 4$ 3D Mesh.

![Graph](image)

(b) Experimental results for a $6 \times 6 \times 4$ 3D Mesh.

Figure 1.9: Hotspot traffic on a 3D NoC for alternative interconnection topologies.
FIGURE 1.10: Hotspot traffic on a 3D NoC for alternative interconnection topologies.
1.5. EVALUATION – EXPERIMENTAL RESULTS

These results are also compared to their equivalent 2D architectures. For the $8 \times 8$ 2D NoC (same number of cores to the $4 \times 4 \times 4$ architecture) it shows 25% increased latency and 40% increased energy compared to one side link pattern, whereas the $12 \times 12$ (same number of cores to the $6 \times 6 \times 4$ architecture) mesh shows 46% increase in latency and 49% increase in energy consumption compared to the same pattern using uniform traffic. In addition, comparing the by_four pattern on 64-nodes architecture under transpose traffic shows 31% and 18% reduced latency and total network consumption, respectively. Whereas, in the case of hotspot traffic and employing the two_side link pattern, these numbers change to 24% reduced latency and 56% reduced energy consumption.

1.5.6 3D NoC Performance under Transpose Traffic

Under the transpose traffic scheme, when the by_four link pattern is adopted it shows 6.5% decrease in total network energy’s consumption at the expense of three cycles increased latency. In Figure 1.11 the simulation results for the 3D $4 \times 4 \times 4$ mesh and $6 \times 6 \times 4$ torus NoCs are presented for transpose type of traffic. From the Figure 1.11(a) we can see that we have a 4% gain in the energy consumption of the 3D NoCs with a 5% increase in the packet latency. Additionally we gain 6% in the area occupied by the switching blocks of the NoC. Comparing these patterns to the 2D NoC (having the same number of nodes) we can have on average a 14% decrease in energy consumption, a 33% decrease in total packet latency. But, on the area the cost of the 3D NoC is higher by 23%. From the Figure 1.11(b) we can see that the 2D NoC experiences traffic contention and not being able to cope with that amount of traffic (the actual value of the latency is close to 5000 cycles per packet). Additionally, 47% gains achieved in energy consumption. When this torus architecture is compared to the “full” 3D one, it shows 5% gains in energy consumption with 8% increase in the latency and 9% reduces switching block area.
64 node 2D and 3D NoCs (transpose traffic, medium load, xyz-old routing)

(a) Experimental results for a $4 \times 4 \times 4$ 3D Mesh.

144 node 2D and 3D NoCs (transpose traffic, medium load, xyz-old routing)

(b) Experimental results for a $6 \times 6 \times 4$ 3D Torus.

Figure 1.11: Transpose traffic on a 3D NoC for alternative interconnection topologies.
1.5. EVALUATION – EXPERIMENTAL RESULTS

Figure 1.12: An overview of the energy breakdown in a 3D NoC ($4 \times 4 \times 4$ 3D mesh, uniform traffic, xyz-old routing).

1.5.7 Energy Dissipation Breakdown

What it can be seen from studying the analytical results derived from Ebit [58] energy model, is that the link’s, crossbar’s, arbiter’s, buffer’s read energy consumption gets smaller in exchange with an increase in the energy consumed when writing to the buffer and by the router’s routing engine. On average the link energy consumption accounts for the 8%, the crossbar’s energy for the 6%, the buffer’s read energy for the 23% and the buffer’s write energy for the 62% of the total energy respectively. The normalized results about the energy consumption for a uniform traffic on a $4 \times 4 \times 4$ NoC are presented in Figure 1.12.

1.5.8 Summary

A summary of the experimental results is presented in Table 1.1. There the energy and latency values that were obtained are compared to the ones of the 3D mesh full vertically interconnected NoC. The three types of traffic are shown in the first column. The next two
Table 1.1: Experimental results: min-max impact on costs (energy and latency) with medium traffic load.

<table>
<thead>
<tr>
<th>Traffic Patterns</th>
<th>Min-Max Impact on Costs (energy and latency)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normalized Energy</td>
</tr>
<tr>
<td>Uniform</td>
<td>92% 108%</td>
</tr>
<tr>
<td>Transpose</td>
<td>88% 116%</td>
</tr>
<tr>
<td>Hotspot</td>
<td>71% 116%</td>
</tr>
</tbody>
</table>

columns present the gains (min. values to max. values – in %) for the energy dissipation. The forth and fifth columns show the min-max values for the average packet latency respectively. As it can been seen energy reduction can be achieved up to 29%. But gains in energy dissipation cannot be reached without paying a penalty in average packet latency. It is the responsibility of the designer, utilizing this exploration methodology, to choose such a 3D NoC topology and vertical interconnection patterns that meets best the requirements of the system.

1.6 Conclusions

Networks-on-Chip are becoming more and more popular as a solution able to accommodate large numbers of IP cores, offering an efficient and scalable interconnection network. Three-dimensional NoCs are taking advantage of the progress of integration and packaging technologies offering advantages when compared to 2D ones. Existing 3D NoCs assume that every router of a grid can communicate directly with the neighboring routers of the same grid and with the ones of the adjacent layers. This communication can be achieved by employing wire bonding, microbumb or through-silicon vias [15].

All of these technologies have their advantages and disadvantages. Reducing the number of vertical connections make the design and final fabrication of 3D systems easier. The goal of the proposed methodology is to find heterogeneous 3D NoC topologies with a mix of 2D
and 3D routers and vertical link interconnection patterns that performs best to the incoming traffic. In this way the exploration process evaluates the incoming traffic and the interconnection network, proposing an incoming traffic-specific alternative 3D NoC. Aiming at that direction we have presented a methodology that shows that by employing an alternative 3D NoC vertical link interconnection network, in essence proposing a NoC with less vertical links, we can achieve gains in energy consumption (up to 29%), in the average packet latency (up to 2%) and in the area occupied by the routers of the NoC (up to 18%).

Extensions of this work could include not only more heterogeneous 3D architectures but also different router architectures, providing better adaptive routing algorithms and performing further customizations targeting heterogeneous NoC architectures. In this way it would be able to create even more heterogeneous 3D NoCs. For providing stimuli to the NoCs a move towards using real applications would be useful, apart from using even more types of synthetic traffic. By doing so, it would become feasible to propose application-domain-specific 3D NoC architectures.

Acknowledgments

The authors would like to thank Dr. Antonis Papanikolaou (IMEC vzw., Belgium) for his helpful comments and suggestions. This research is supported by the 03ED593 research project, implemented within the framework of the “Reinforcement Program of Human Research Manpower” (PENED) and co-financed by National and Community Funds (75% from E.U.-European Social Fund and 25% from the Greek Ministry of Development - General Secretariat of Research and Technology).

References


1.6. CONCLUSIONS


1.6. CONCLUSIONS

539, 2005.


CHAPTER 1. THREE-DIMENSIONAL NETWORK-ON-CHIP ARCHITECTURES


