Comparison Between Optimal Interconnection Network in Different 2D and 3D NoC Structures

Farzad Radfar  
Department of Electrical Engineering,  
Sharif University of Technology,  
Tehran, Iran,  
farzad_radfar@ee.sharif.edu

Masoud Zabih  
Department of Electrical Engineering,  
Sharif University of Technology,  
Tehran, Iran,  
masoud_zabih@ee.sharif.edu

Reza Sarvari  
Department of Electrical Engineering,  
Sharif University of Technology,  
Tehran, Iran,  
sarvari@sharif.edu

Abstract—The current article studies optimal intercore interconnect network in a NoC structure for 2D and 3D mesh, torus and hypercube topologies. Optimal wire width spacing is calculated by numerically maximizing bandwidth times the reciprocal delay, which depends on the technology node and hop length. Through 3D integration and increasing tiers, optimal interconnect width and spacing in torus and hypercube topologies will decrease. The core-to-core channel width in all topologies will be obtained by assigning 20% of the power consumption to the routers. By increasing number of cores, channel width will decrease due to reduced power consumption of each core. This is more in hypercube topology, due to the fact that the number of router ports will increase along with an increase in the number of cores. In terms of the worst case delay, mesh topology is worse than the two other topologies. Also it is not scalable due to the increase in the number of cores. In all topologies, power consumption of the chip and the worst case delay will decrease by 3D integration and utilizing more tiers. Mesh and torus topologies make the least and the most use of wiring area, respectively. Bisection bandwidth increases in all topologies by 3D integration.

Keywords—3D NoC; interconnect; optimal interconnection network; channel width

I. INTRODUCTION

For the past decade, increasing the number of cores replaced increasing clock frequency due to power challenges facing Moore’s Law [1]. As the number of processing elements increases, interconnection network plays an important role in the system design. Based on the system’s conditions, one of the interconnection networks will be selected from crossbar, ring, bus or Network-on-Chip (NoC) [2]. Scalability makes NoC a better option than the others for on-chip structures [3].

NoC includes links and routers (within each core). In 2D architectures, links are global copper interconnects while in 3D structures, which consist of both vertical Trough Silicon Via (TSV) links and copper wires. Scaling reduces gate delay and at the same the delay per unit length of wires increases [4]. 3D integration and stacking silicon layers will reduce the length of global wires. This will lead to the lower power consumption and higher performance of 3D ICs compared to 2D counterparts [5].

Results of the current study will give an optimal design for the global and core-to-core interconnects which can be used by the designer at the early stages of the design process. First, based on both delay and bandwidth, we optimized width and spacing of the global wires for several technology nodes and topologies. Then, according to the power consumption of routers, we obtained core-to-core channel width. Finally, we compared power dissipation and performance of several optimal designs in different topologies. In the present article, 2D and 3D mesh, torus and hypercube topologies have been investigated. Same methodology can be applied to the other topologies as well.

In section II optimal width and spacing of global wires are investigated. In section III, the core-to-core channel width is optimized and the number of bits of communication between cores for several topologies, are investigated. In section IV, the power dissipation and performance of several topologies is compared and finally, in section V, a conclusion will be presented.

II. WIDTH AND SPACING OPTIMIZATION

A. Global interconnect width and spacing in NoC structures

In this section, optimal width and spacing of the global copper interconnects in a NoC structure despite an optimal number of repeaters will be addressed. Cross section of global interconnects is shown in Fig.1, where $W$ stands for interconnect width, $S$ for spacing between two adjacent interconnects, $H$ for thickness of dielectric, and $T$ for thickness of interconnect, moreover, $T = H$ is assumed, and according to the aspect ratio of interconnect we have $T = H = AR \times W$. Considering Elmore delay model [6] and optimal repeaters, the delay of interconnects is as follows [7]:

$$D = 2 \log 2 \times L \sqrt{R_c C_{arc}} \left(1 + \frac{1}{2} \left(1 + \frac{V_d}{V_0}\right)\right)$$

Fig. 1. Cross section view and parameters of global copper interconnects.
Fig. 2. Global copper interconnects with optimal repeaters

Where \( r \) and \( c \) stand for interconnect resistance and capacitance per unit length, respectively; \( L \) for interconnect length, and \( R_s, C_p \) and \( C_0 \) stand for output resistance, output capacitance and input capacitance of minimum size of repeater, respectively.

In Network-on-Chip (NoC) structure, the router latches data on the rising or falling edges of the clock. Therefore, the delay of intercore interconnects in a NoC structure will be obtained from the following relationship:

\[
\tau_{\text{interconnect}} = \left( \left\lfloor \frac{D}{\tau_{\text{clk}}} \right\rfloor + 1 \right) \tau_{\text{clk}} \tag{2}
\]

Where \( \lfloor \rfloor \) stands for floor function. In addition to delay of interconnect, the delay of two routers connected together by a hop (connection of one core with the adjacent core) should also be taken into consideration. Delay of router depends on the network data traffic and it equals to 5 clock cycle in no-load mode [11]. For optimization, we assume network at no-load mode. Therefore, the overall delay of one hop in no-load mode is as follows:

\[
\tau = \tau_{\text{interconnect}} + (2 \times 5\tau_{\text{clk}}) \tag{3}
\]

As in Fig.2, bandwidth \( B \) of global interconnect equals number of global interconnects times reciprocal delay. In NoC structure, flits get pipelined by routers, hence for any number of hops the bandwidth remains constant [12] and is:

\[
B = \frac{D_{\text{chip}}}{W+S} \frac{1}{\tau_{\text{interconnect}}} \tag{4}
\]

With increasing width and spacing of interconnects, delay will decrease and bandwidth will also decrease, unfortunately. Regarding the importance of the delay and bandwidth, the optimal interconnect width and spacing is selected so that a trade-off occurs between delay and bandwidth. Thus, multiplying bandwidth by reciprocal delay is introduced as a figure-of-merit, \( F \) [7, 13, 14]:

\[
F = \frac{B}{\tau} \tag{5}
\]

Optimal interconnect width and spacing are values that maximizes \( F \). Figure-of-merit versus interconnect width and spacing for a hop of 10mm is shown in Fig.3. All technology parameters are extracted from predictions presented in International Technology Roadmap for Semiconductors (ITRS) [15] for the year 2016. Discontinuities of \( F \) is due to the existence of floor function in (2) and maximum figure-of-merit occurs at a point where \( D \) equals \( \tau_{\text{clk}} \). Therefore, optimal interconnect width and spacing are values for which the wire delay equals one clock cycle. In Fig. 4, optimal global copper interconnect width and spacing versus hop length for some technology generations is depicted. Optimization is made numerically. As can be seen in the figure, optimal width and spacing of wires are functions of both interconnect length (direct relation) and technology node.

Formula (1) has been obtained considering RC model for transmission line and interconnect inductance is neglected. It can be inferred that RC model is a good model for analyzing transmission line, when the following condition is met [16].

\[
\left( 0.377 \frac{\tau_{\text{clk}}}{Z_{0}} + 0.693 \frac{R_s}{Z_{0}} \right)^{-1} \leq 1 \tag{6}
\]

Where \( Z_0 \) stands for the characteristic impedance. Moreover, (1) is obtained considering an optimal number of repeaters. Hence, the following condition should be met before using this relationship [7]:

\[
\frac{h_{\text{opt}}}{L} \leq 1 \tag{7}
\]

Where \( h_{\text{opt}} \) is optimal spacing between two repeaters (Fig.2) which is obtained from the following relationship [7]:

\[
h_{\text{opt}} = \sqrt{\frac{2R_s(C_0+C_p)}{rc}} \tag{8}
\]
The two prerequisites namely (6) and (7) with presence of $\tau_{th}$ and $\tau_{th}$, versus hop length for some technology generations are depicted in Fig. 5. As is clear in this figure and according to the range of existent hop lengths in NoC structures, the two prerequisites are met.

The methodology presented in this section - which not only interconnect width but also wire spacing is optimized - will result in a larger figure-of-merit compared to the previous works [11, 13, 14]. In Fig. 6, figure-of-merit is plotted versus hop length, for some technology nodes. The solid lines are related to the methodology presented in this article and the dotted lines indicate the methodology presented in [11] for which both width and spacing of the wires are equal to $W_{opt}$.

B. Optimal global interconnect width and spacing in NoC topologies

In this article, three widely used NoC topologies, mesh, torus and hypercube, are studied in 2D and 3D structures. It is assumed that intercore connection in 3D structure and between several tiers happens by Through Silicon Vias (TSVs) and copper wires in each tier. For intercore wiring in each tier, two orthogonal copper layers are assigned. TSVs selected between tiers are via-last [17], connecting global interconnects in each tier to the global interconnects of the next tier. In fact, 3D topologies are obtained through folding 2D structures. For example, one 3D structure with 2tier is obtained by folding down the middle of a 2D structure, and a 3D structure with 4tier is in fact the same 2D structure being folded down at the middle twice. In Fig. 7, the schematic of different topologies with 32 cores has been shown where red lines indicate TSVs and black lines indicate copper interconnects. For hypercube topology, the method presented in [18] has been used.

To obtain optimal width and spacing of the global wires, the longest length in each tier is selected as the basis and $F$ is maximized for that wire. Fig. 8 illustrates optimal wire width and spacing in various 2D and 3D topologies for the node 2016 of ITRS. The optimal interconnect width and spacing is directly proportional to its length, while the length of the longest interconnect decreases with increasing number of cores in mesh topology, increases in torus topology and remains constant in hypercube topology. This justifies the dependence of the optimal width and spacing versus number of cores in Fig. 8.
III. CHANNEL WIDTH

Most important limiting factor in chip-performance is power dissipation which is related to the channel width in a NoC structure. In this section we obtained channel width by assigning a percentage of power consumption to the routers. We used IntSim simulator to estimate total power consumption of chip, based on data available for technology node ITRS [19]. In addition, we used ORION 3.0 simulator to estimate power consumption of routers [20]. Assuming uniform distribution of gates on die level, one can divide power consumption evaluated by IntSim equally to the number of cores on a chip. In this work 20% of the power consumption inside each core is allocated to the routers.

In Fig. 9, channel width versus number of cores is plotted for node 2016. In this Figure, features of router are as follows: #Buffers=8, #Virtual Channels=4. Channel width in both mesh and torus topologies are the same, as routers in these two topologies have the same numbers of ports (input and output). Number of input/output ports for these two topologies is 5 for 2D structure, 6 for 3D-2tier structure, and 7 for 3D-4tier structure. Accordingly, routers consume more power for 3D integration, leading to a decrease in the channel width. For hypercube topology, power consumption of routers will not be affected by 3D integration, because number of input/output ports of routers is constant and equal to: (log2 #of cores) +1.

Hence, channel width will remain constant by 3D integration. By increasing number of cores, share of power consumption of each core will decrease leading to a decrease in the channel width. In hypercube topology by increasing number of cores, number of input/output ports of routers will also increase which further decreases channel width.

IV. A COMPARISON OF NOC TOPOLOGIES

In this section, we compare different NoC topologies based on the wire width-spacing and channel width, which has been optimized in two previous sections. The comparison parameters includes: energy dissipation, wiring area, worst case delay, and bisection bandwidth.

A. Energy Dissipation

In NoC structures, energy dissipation in a clock period $E$, consists of two main components: interconnects and routers. Interconnects include copper wires and TSVs. Their energy dissipation depends on the wire capacitance and number and size of repeaters. Energy dissipation in copper interconnects with optimal repeaters is obtained from the following relationship [7]:

$$E_{Copper\;Wire} = \text{hop length} \times C \times \left(1 + \frac{C_{TSV} + C_{E}}{2C_{0}}\right) V_{dd}^2$$ (9)

TSV capacitance is voltage dependent like any Metal-Insulator-Semiconductor (MIS) capacitance, in three different regions as: accumulation, depletion and inversion [21]. TSV capacitance in accumulation region is obtained as follows [22]:

$$C_{TSV} = 2\pi \varepsilon_{ox} AR \cdot D / \ln \left(1 - \frac{t}{2D}\right)$$ (10)

Where $\varepsilon_{ox}$ stands for the oxide permittivity, $D$ for diameter of TSV, and $t$ for the thickness of insulator. Having $C_{TSV}$ and $V_{dd}$, one can obtain energy dissipation in TSVs and routers using ORION 3.0.

Fig. 10. Total energy dissipation and energy dissipation of copper global interconnects in NoC topologies versus # of cores for ITRS 2016

Fig. 10 illustrates total energy dissipation for optimal design at node 2016. It can be seen that, the highest power dissipation is due to the global Cu wires which decreases by 3D integration as the total length of wires decreases.
B. Wiring Area

Having optimal values for Cu wire width/space and channel width, one can determine area consumed by the wiring. This space refers to two metal layers prepared for vertical and horizontal core-to-core connections in each tier. Fig.11 illustrates the amount of wiring area used in various topologies for node 2016 of ITRS. The wiring area is more for torus compared to two other topologies.

Area consumed by TSVs in 3D topologies is plotted in Fig.12 for node 2016 of ITRS. TSV footprint area is negligible and is the same for 3D-2tier torus and mesh topologies as they have the same channel width.

C. The Worst Case Delay

The worst case delay is the time delay between two cores with the highest distance from each other. Fig.13 illustrates the worst case delay for 2D and 3D topologies for node 2016 of ITRS. In this Figure, delay of each hop and router are considered \( \tau_{clk} \) and \( 5\tau_{clk} \), respectively. In topologies where lengths of hops are not equal, designing copper interconnect was performed in a way that delay of the longest interconnect was equal to the clock cycle, \( \tau_{clk} \). Shorter interconnects will have a delay less than \( \tau_{clk} \); however, according to (2), delay of each interconnect cannot be less than \( \tau_{clk} \). In addition, delay of TSV for node 2016 of ITRS is far less than \( \tau_{clk} \) [23]. But, regarding the limitations created from data latching by routers on the rising or falling edges, the delay of TSVs is considered to be \( \tau_{clk} \).

According to Fig.13, 3D integration of mesh and torus topologies, will improve the worst case delay. Although worse case delay in hypercube topology will not benefit from 3D integration, as number of hops are constant. Mesh topology with large number of cores will result in a big worse case delay compared to the other two topologies as the largest distance between two cores will further increase due to the increase in the number of cores. However, hypercube topology is the best choice in terms of worse case delay as it bypasses many routers when transferring data between two cores with largest distance.

D. Bisection Bandwidth

Dividing the network into two equal parts, the bandwidth between two parts is called the bisection bandwidth. A system with higher bisection bandwidth handles the global traffic better. Bisection bandwidth of various topologies for node 2016 of ITRS is illustrated in Fig. 14. As can be seen, 3D integration will always improve bisection bandwidth and mesh topology has the least bisection bandwidth.

V. CONCLUSION

In this article, for the first time, optimal global interconnect width and spacing in a NoC structure is presented and applied to 2D and 3D mesh, torus and hypercube topologies. IntSim and ORION 3.0 simulators are used to find a core-to-core channel width which keeps the router’s power dissipation 20% of core’s power dissipation. We showed that in mesh and torus topologies by increasing tiers in 3D integration, channel width will decrease. However, by increasing number of cores in
improve.

the worst case delay of mesh and torus topologies will also consumption by reducing length of copper interconnects, and increasing number of tiers, one can decrease power also due to the large channel width, consumes more wiring area than other two topologies. By 3D integration of topologies and also hypercube topologies as the longest wire shortens. Torus topology, due to its large interconnect width and spacing and also due to the large channel width, consumes more wiring area than other two topologies. By 3D integration of topologies and increasing number of tiers, one can decrease power consumption by reducing length of copper interconnects, and the worst case delay of mesh and torus topologies will also improve.

REFERENCES


