Design exploration for network on chip based FPGAs: 2D and 3D tiles to router interface

Alaa Salaheldin a, Hassan Mostafa a,b,∗, Ahmed M. Soliman a

a Electronics and Communications Engineering Department, Cairo University, Giza 12613, Egypt
b University of Science and Technology, Nanotechnology and Nanoelectronics Program, Zewail City of Science and Technology, October Gardens, 6th of October, Giza 12578, Egypt

ABSTRACT

Nowadays, SoC uses Network on Chip (NoC) to connect its increasing number of building blocks. FPGAs, like SoCs, can use NoC to connect its increasing number of tiles, memories, DSP slices and embedded processors. However, one drawback of using NoC is that increasing its router ports affects the area, power and frequency of the system significantly. For FPGAs to benefit from the NoC approach, an efficient way has to be found to interface a large number of blocks without increasing NoC router ports. In this paper, a concentrator module or a Codec, is used to connect between routers and multiple Tiles (FPGA basic building block). Usage of Codec reduces the effect of increasing tile count on the area, power and frequency of the FPGA routing network. Different 2D and 3D network configurations are compared to investigate the effects of adding the Codec module.

I. Introduction

FPGAs (Field Programmable Gate Arrays) are used increasingly in most of today’s applications because of their low development cost, fast design cycle, configurability and short time to market. On the other hand, ASICs (Application Specific Integrated Circuits) have long design cycle, poor configurability and require high development effort. These strong points of the FPGA made it an appropriate candidate for most research and industry applications. However, these advantages come at a significant cost in delay, area and power consumption arising mainly from its programmable routing fabric.

An FPGA consists of three main components: Processing elements (PEs), storage elements (SEs) and a complex programmable routing fabric. PEs are programmable logic blocks that perform logic calculations, for example, look-up tables (LUTs) with a fixed configuration of logic gates. SEs are memory blocks placed across the chip area and used to store data or algorithm states. The programmable routing fabric is a huge network of wires, multiplexers and bus-based interconnections; all used to connect PEs, SEs and IPs (Intellectual Property cores).

Due to the continuous demand for more powerful and larger chips, new blocks are added to the FPGA architecture, such as Digital Signal Processing (DSP) blocks and embedded processors. As the system complexity increases, the negative impact of the routing fabric increases as well. Bus-based interconnections, such as ARM’s AMBA [1] and IBM’s CoreConnect [2], become bottlenecks since they are unable to meet systems requirements. In general, they are not suitable for large systems as their performance degrades when used to connect many blocks. In addition, these interconnects normally include very long wires (global wires) to be able to connect all parts of the chip, these global wires contribute heavily to the increased area and power consumption of the routing fabric.

Network on Chip (NoC) comes as a promising solution for the conventional interconnects problems. NoC has the benefits of independent implementation and optimization of nodes, simplified customization per application, support for multiple topologies and options, reduced area and power consumption, scalability and increased operating frequency.

Using the NoC approach instead of depending on long interconnect wires solves the conventional interconnects problems because NoC uses fast optimized lanes to transfer packets between the routers, and these routers interface with the main application blocks (such as SEs and PEs) through a configurable number of input/output ports solving most of the problems introduced by long and medium-size routing wires.

Correspondingly, the NoC approach is the right choice as an interconnect fabric of the next generation FPGA. On the other hand, the problems of integrating NoC into the FPGA architecture should be investigated and solved which has been addressed in this research work.
By applying the NoC approach in FPGAs, a significant problem appears because of the large number of tiles that should be connected to the network. In Ref. [3], it is stated that the area, delay and power consumption of NoC routers are increasing significantly with increasing the port count. In order to overcome this problem and to make the NoC approach useful in designing the next generation of FPGAs, a tile to NoC router connector/coder-decoder (hereinafter referred to as Codec) is used to enable multiple tiles to share single router port, given that the required rate for all multiplexed tiles does not exceed the router port's maximum rate.

The rest of the paper is organized as follows: In Section 2, a brief background of CONNECT router is introduced followed by a comparison between CONNECT-only and CONNECT with Codec networks. In Section 3, the effects of adding the Codec module to 3D-NoC systems are investigated and finally, a conclusion is drawn in Section 4.

2. 2D comparison using CONNECT and codec

2.1. CONNECT

CONNECT is a soft router designed for FPGAs [4]. CONNECT adds new features such as virtual links and peak flow control. It maximizes routing resources utilization by using wider buses between routers. CONNECT has a configurable number of input/output ports, packet length and buffer depth. It is implemented using Bluespec System Verilog (BSV) and provides a flexible design. Fig. 1 shows its architecture that uses a single stage pipeline leading to a lower cost and latency. CONNECT supports four variations of separable input-output allocators [4].

2.2. Codec modeling

In Ref. [5], a Simulink system-level model is built using the SimEvents toolbox to measure the throughput difference between two networks, Network A which includes routers only, and Network B which includes routers and Codecs. SimEvents provides a discrete-event simulation engine and component library for analyzing event-driven system models and optimizing performance characteristics such as latency, throughput, and packet loss.

A $2 \times 2$ mesh topology with sixteen tiles is used for both networks; each router is connected to four tiles either by direct port connections or through a Codec. Network A does not use Codec, so each router interfaces with four tiles and two neighbor routers. The 6-port router is shown in Fig. 2.

A router consists of routing core and input/output queues. The routing core consists of routing logic and output switch; its routing logic is implemented as delay server to model the packet processing latency and a routing table to determine which output port the packet goes to. Packet processing time in a 6-port router is longer than in a 3-port router.

In network B, each router uses a Codec module to interface with four tiles, so the used routers are 3-port routers like the one shown in Fig. 3. The packet length is increased by two bits in this case to handle the switching required from Codec to tiles.

After running the simulation, the number of received packets at each tile is counted. Table 1 shows the comparison between the number of received packets at some tiles in network A compared to network B. The total number of received packets is 281 packets in network A and 677 packets in network B.

2.3. Codec implementation and comparison setup

The sending path of Codec (from tiles to router) acts as a path combiner that rotates across all attached tiles and checks for available payloads to send.
As shown in the lower part of Fig. 4, the receiving path is similar to a router; a packet is examined once received from a router and according to the destination address, the packet is sent to one tile.

Codec design is made modular to facilitate generating any network configuration with different tile widths and counts. Also, Codec inherits some parameters from CONNECT design for compatibility given that CONNECT router is dedicated to the embedded NoC in the next generation FPGA [6,7].

In Ref. [5], and similar to the Simulink-based comparison mentioned previously in this section, an RTL comparison is held between two network models both have four routers in a 2 × 2 mesh topology. The first model is called network A; it uses CONNECT routers only, three configurations of this model are built to interface with 16-tiles, 32-tiles and 54-tiles. A 64-tiles network could not be built as CONNECT generation tool in Ref. [8] is limited to 16-port per router; two ports out of the 16 are used to connect with neighbor routers and 14 are left to interface with the tiles, which gives a total number of 54 ports for the four routers inside the network. The second model is called network B; it uses Codecs to interface with 16-tiles, 32-tiles and 64-tiles.

For each model, three configurations are built for a different number of Codecs per router. Table 2 illustrates the Codec network configurations. CpR is the Codecs per router and TpC is the tiles per Codec. For example, to build a 2 × 2 16-tiles configuration of network B, each router connects with one Codec (1 CpR) and each Codec connects to four tiles (4 TpC), or each router connects to two Codecs (2 CpR) and each Codec connects to two tiles (2 TpC).

A 2 × 2 16-tiles configuration using (4 CpR) is not applicable because in this configuration each Codec module would connect a single tile to the network. In this case it is more reasonable to connect this tile directly to a router port without using Codec.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Simulink comparison results.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tile number</td>
<td>1</td>
</tr>
<tr>
<td>Network A</td>
<td>18</td>
</tr>
<tr>
<td>Network B</td>
<td>54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Network B CpR and TpC configurations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration\Tiles</td>
<td>16 Tiles</td>
</tr>
<tr>
<td>1CpR</td>
<td>4TpC</td>
</tr>
<tr>
<td>2CpR</td>
<td>2TpC</td>
</tr>
<tr>
<td>4CpR</td>
<td>n/a</td>
</tr>
</tbody>
</table>
2.4. 2D comparison results

Altera Arria II GX EP2AGX260 FPGA is used as target to compare synthesis results. It has 205200 combinational ALUTs, 102600 memory ALUTs and 692 IO pins. Quartus II 12.0 is used with ModelSim Altera Starter Edition for synthesis and RTL simulation.

The logic utilization values shown in the comparison figures are the sum of both combinational and memory resources consumed on the FPGA. Quartus PowerPlay Analyzer tool is used to estimate the consumed power.

2.4.1. Frequency

As shown in Fig. 5, the maximum operating frequency of network A decreases with increasing the number of tiles. This is due to the increased delay of allocators and crossbars; a 6-port router in a 16-tile network runs on a higher frequency than a 10-port router in a 32-tile network. On the other hand, network B starts at a higher frequency and decreases slightly as the number of tiles increases.

The reason for the difference between network B configurations shown in Fig. 6, is not the change in the Codec circuit. The reason for this difference is mostly the increased size of CONNECT routers. A 3-port CONNECT router used in 1CpR and a 4-port router is used in 2CpR network; the 4-port router occupies more area and operates with a lower frequency compared to the 3-port router.

2.4.2. Logic utilization

As displayed in Fig. 7, a 56-tiles network A uses 30% of the FPGA resources. In this network, each of the four routers has sixteen ports in order to be able to interface with 14 tiles. However, a 64-tiles network B with 1CpR uses at maximum 2% of the resources as each of the four routers has only three ports; two ports to interface with the adjacent routers and one to interface with the Codec which connects to sixteen tiles.

The reason for this large logic utilization difference is that a 16-port CONNECT router consumes larger area than a 16-port Codec.

In Fig. 8, a comparison between 1CpR, 2CpR and 4CpR configurations is shown. The 4CpR configuration consumes largest area because each router has six ports; two to connect with the adjacent routers and four to connect with four Codecs. 2CpR network configuration has a 4-port router and 1CpR has a 3-port router.

The number of Codec ports increases as the number of connected tiles increases; this explains the slight increase of logic utilization between different network B configurations.

2.4.3. Power consumption

As shown in Figs. 9 and 10, network A consumes more power than all network B configurations because of the large area consumed by CONNECT 6-port routers. For example, a 16-tile network B using 1CpR consumes less than 50% of the power compared to 16-tile network B.

Also, it is shown that increasing the input/output port count affects the power consumption of network A more significantly than network B.

2.5. Summary

It is clear from the simulation results that comparing two 2 × 2 networks, one uses routers only and the other uses routers and Codecs, the routers and Codecs network is found to take less than 15% area, consume less than 50% power of the routers only network and operates with 2.5X frequency.
3. 3D comparison using NOCET and codec

3D-NoC has been investigated and explored intensively during the last years. However, 3D-NoC is still considered a new and emerging technology. Most research is targeted to the development of 3D-NoC modeling and simulation tools. In Ref. [9], the authors developed an open-source and generic NoC simulator using SystemC called Noxim, the tool does not support 3D-NoC natively but can be modified to mimic the behavior of 3D-NoC systems. In Ref. [10], the authors implemented a 16-processor NoC-based 3D system that consists of two tiers in a mesh topology. In Refs. [11,12], the authors explored the performance improvements and constraints for different 3D topologies. In Ref. [13], the authors designed a NoC router that exploits the vertical nature of 3D-NoCs. In Ref. [14], the authors introduced two look-up table based routing algorithms for 3D-NoC. In Ref. [15], a virtual channel based routing algorithm is introduced to avoid deadlock in irregular networks and it also uses a compact form of routing tables in order to minimize their overhead. In Ref. [16], the authors proposed a routing algorithm that depends on splitting the network into layers in order to provide deadlock and live-lock free operation. In Ref. [17], a power consumption and latency optimized routing algorithm is introduced. Since mesh topology is the most used topology for 3D-NoCs, in Refs. [18,19], the authors provide routing mechanisms designed specifically for mesh topologies. In Ref. [20], the authors developed a simulation tool called 3D-NOCET, the tool offers a generic and flexible solution to generate different 3D-NoC configurations. This tool could be used to do different performance evaluations according to the main network factors which are number of tiers, number of routers per tier and the planar topology for the tiers.

3.1. 3D-NOCET as an exploration tool

3D-NOCET tool supports full-mesh and ring as 2D topologies. The tool supports a maximum number of 16 tiers and 256 routers per tier. As shown in Fig. 11, the tool provides a simple GUI (Graphical User
Interface), by choosing the “Mesh”, “Tier #1” and “Tier #2” boxes, the tool is easily configured to generate a 3D-NoC configuration with two tiers and full-mesh topology as 2D topology for each tier. Behind its GUI, lays the automation infrastructure which consists of few scripts that generate the synthesizable SystemVerilog RTL code.

The authors of 3D-NOCET tool have made it possible to extend the tool further to include more planar topologies. To study the effect of adding Codec to 3D-NoC networks; it is required to add the Codec block to the auto-generated RTL code. The updated tool, as shown in Fig. 11, supports the integration of Codec to full-mesh and ring topologies, the number of tiles per Codec is set to four.

3.2. Comparison setup and results for full-mesh topology

A comparison setup is created to study the performance differences between 3D-NoCs with and without the Codec. The comparison methodology is done similar to the comparison in Ref. [20]. First, with respect to vertical complexity in which the impact of increasing the number of tiers is investigated. Second, with respect to network complexity in which the impact of changing the 2D topology is investigated.

In this comparison, a full-mesh topology is used as the 2D topology for all tiers. In addition, one Codec per router (1CpR) and four tiles per Codec (4TpC) are used for all configurations. For investigating vertical complexity, a constant number of four tiles per tier is used (4TpT). For network complexity, a constant number of two tiers is used, the limitation of using only two tiers comes from the long compilation time required to do synthesis for full-mesh topology.

Altera Arria II GX EP2AGX260 FPGA is used as a target chip. It is the same target chip used in the 2D comparison discussed in Section 2.

3.2.1. Logic utilization

As shown in Figs. 12 and 13, increasing the number of tiers in a 3D network increases the consumed FPGA resources significantly. This is due to the increased number of instantiated routers; without using Codec, one router is required per PE, SE or tile. In addition to that, the increased amount of wiring and routing resources used to connect routers to the network and to other entities.

On the other hand, for 3D network configurations using Codec, every four tiles are sharing only one router port through a Codec module and this reduces the total number of instantiated routers to a one-fourth.
fourth. This leads to slightly increased FPGA resources due to extra wiring between tiles and Codecs and due to the logic consumption of the Codec modules themselves.

Increasing network complexity means adding more routers or nodes in a single tier with maintaining the number of tiers constant. As shown in Figs. 14 and 15, increasing the number of tiles per tier increases the consumed FPGA resources significantly. This is due to the increased number of router ports required to connect all routers in the full-mesh topology.

3.2.2. Frequency

Increasing the network size in the vertical dimension (increasing number of tiers) with fixing the number of tiles per tier constant does not affect the complexity of arbiters and switching logic. Hence, the maximum operation frequency is not affected significantly. As shown in Fig. 16, $F_{\text{max}}$ is not affected heavily by increasing the number of tiers.

Increasing the number of tiles per tier in a full-mesh topology increases the router ports which leads to more complex arbiters and allocators, these two modules affects the max operating frequency significantly. As shown in Fig. 17, for all network configurations, the Codec networks operate with higher frequency that is at least $1.5 \times$ higher than the maximum frequency of regular networks.

3.2.3. Power consumption

As shown in Fig. 18, increasing the number of tiers in a 3D network increases the consumed power significantly. This is due to the increased number of instantiated routers.

As shown in Fig. 19, increasing the number of tiles per tier increases the size of arbiters, allocators and input/output buffers leading to more static and dynamic power dissipation. On the other hand, using Codec modules decreases the number of required routers compared to the regular network. The value of the dynamic power dissipated for 16pt-without-Codec configuration is interpolated using power values for 12pt and 8pt. The reason for the interpolation is that the synthesis tool is not able to fit this network on the target FPGA.

3.3. Comparison setup and results for ring topology

The comparison for ring topology extends the results obtained from the previous full-mesh comparison. Unlike full-mesh topology,
the number of ports per router in ring topologies does not depend on the number of routers per tier. This enables the exploration of larger 3D-NoC configurations without increasing compilation time significantly. The ring comparison setup is created to explore networks with a higher number of tiers. Networks with two to eight tiers are generated with a different number of tiles per tier spanning four to sixteen.

In this comparison, ring topology is used as the 2D topology for all tiers. One Codec per router (1CpR) and four tiles per Codec (4TpC) are used for all configurations. The maximum operating frequency is not considered in this comparison because it only changes slightly for ring topologies. The reason for that is that the port count of ring routers do not change with changing the network parameters, which are number of tiers and number of routers per tier.

3.3.1. Logic utilization

As shown in Figs. 20 and 21, the logic utilization of ring topologies is significantly larger than that of ring-with-Codec topologies. Approximately larger by order of magnitude.
3.3.2. Power consumption

As shown in Figs. 24 and 25, the power consumption of ring-only topology is higher than that of ring-with-Codec topology. For example, for the biggest network (#Tiers = 8, #Tiles per tier = 16), the power consumption of ring-only network is approximately twice that of the ring-with-Codec network.

3.4. Summary

3D-NoC comparison setups are created to study performance differences between 3D-NoCs with and without Codec. The comparison results for both full-mesh and ring topologies show that for the area, power and maximum operating frequency, 3D-NoC with Codec network outperforms the 3D-NoC only network.

4. Summary and conclusion

Given the importance of FPGA platforms in today’s market, this paper defines and tries to solve one issue that faces the integration of NoC within the FPGA architecture.

In Section 2, Codec is introduced to overcome one of the NoC problems, which is the performance degradation due to increasing input/output ports of NoC routers. Then a comparison between two 2×2 networks is made, one uses routers only and the other uses routers and Codecs. It is found that the Codec network takes less than 15% area, consumes less than 50% of a router only network and operates with 2.5X frequency.

In Section 3, the effects of using Codec in 3D-NoC are investigated. First, the 3D-NOET tool is updated to enable the exploration of different 3D configurations. Then a comparison is held between two 3D-NoCs, with and without Codec. The comparison results show that for the area, power and maximum operating frequency, 3D-NoC network with Codec outperforms 3D-NoC only network.

The contributions of this paper are briefly listed as follows:

- Introduce the Codec module as a solution to the increased router port count problem. It is used to interface between FPGA tiles and NoC routers. A comparison is done to investigate the effects of adding Codec to 2D networks.
- Investigate the impact of integrating Codec into 3D-NoC

Recommendations for future work:

- Investigate more topologies with 3D-NOET tool.
- Use a real NoC router in the generated RTL code instead of a simple one.
- Investigate latency and throughput performance using network simulators (such as Noxim).
- Investigate the effect of using TSVs in 3D-NoC based FPGAs.
- Better area analysis using semi-automatic layout generation.

Acknowledgement

This work was partially funded by ONE Lab at Zewail City of Science and Technology and Cairo University, NTRA, ASRT, and NSERC.

References