

A CASE STUDY OF NoC ARCHITECTURE DESIGN

Santosh Krishnan, Adithya Suresh (skkrishnan, ksadithya at ucsd dot edu)

University of California, San Diego

Abstract—

Obtaining high throughput and reducing power for an NoC architecture has become the principal aim over the past few years. In this paper, we survey some unique and recent techniques that have been implemented to accomplish the same. Having given an introduction that highlights the need of NoC architectures and some basic network topologies that are generally deployed in NoCs, we provide two cutting edge architectures that are used as NoCs namely the Optical NoCs and the wireless NoCs. The motivation behind the birth of the architectures, their method of operation and a critique of both architectures has been given. The parameters that are to be considered are the power loss and the latency of the architecture.

Index Terms—NoC architecture, network on chip, wireless NoC, optical NoC, network topology, 3D.

I. INTRODUCTION

WITH the proliferation of compact mobile devices, increased use of cloud storage and the advent of the IoT phenomenon, applications have gotten increasingly computationally intensive. This trend is expected to continue in the future too. To meet this increase in the computational requirements, the number of computing resources on a chip is expected to increase substantially [1]. Current on chip communication technique revolve around the bus architecture. While this architecture continues to serve us well, its performance is expected to fall woefully short of that of future requirements due to the following reasons. Firstly bus architectures are not scalable [2]. As the technology moves into the sub-nanometer region, the wire delay is expected to increase exponentially leading to large latencies for signals. Moreover since only one master can have control of the bus at a given time, this type of communication technique is not suitable for future multiprocessor architectures. Current commercial bus architectures like the AMBA bus from ARM and CoreConnect from IBM employ many smaller busses that are connected in a hierarchical manner instead of a single shared bus, facilitating multiple master communications at a time [3]. However long distance inter-bus communication involves multiple hops leading to poor latency performance. In data communication space LAN and WAN have successfully dealt with this problem by employing a layered

architecture. Following the same trends, networks have started to replace busses in much smaller systems for example PCI-Express is a network-on-a board, replacing the PCI board-level bus.[4]. If we can use the same concept for modern SoC's it can greatly improve inter resource communication performance. Networks-on-Chip (NoCs) have been recently proposed as a promising solution to complex on-chip communication problems. The following paper dives into various NoC topologies, emerging interconnect technologies and a brief comparison of some upcoming NoC architectures. This paper is organized as follows: Section II discusses the related work done in this field. Section III discusses different possible network topologies. Section IV discusses briefly NoCs for Section VI and V discusses the wireless NoC architecture and the optical ring NoC architecture. Finally section VII and VIII discusses concluding remarks and future scope.

II. RELATED WORK

The concept of network on chip was known long ago it is only now that it is finding its way into present day commercial SoC's and chips. Hemani et al.[6] Wingrad [9] and Dally[7] were among the first to propose the concept of NoC's. Dally [7] further goes on to elaborate the challenges in the architecture and design of these networks. Kumar et al[8] proposed a packet switched platform for a single chip system that scales well with a number of processor like resources. One of the first instances where the communication issues for SoC's were studied was carried out by Cesar [5] where they carried out a comparison between bus and NoC architectures by using mathematical modeling the same.

Just as with data communication, where there are many architectures possible, there are many possible NoC architectures each having its own advantages and disadvantages. Pande et al[10] evaluated the different NoC architectures with regard to latency, throughput, energy dissipation, and silicon area requirements for multiprocessor SoC platform. They were also the first ones to characterize different NoC architectures with respect to their performance and design trade-offs.

III. NETWORK TOPOLOGIES

Some of the basic topologies that are used in data communication networks are mesh, ring, bus and star. Some of these topologies like mesh, ring and bus can be used for on chip communication too. For example the mesh topology is very popular since it is simple to implement and is very

compact. However there topologies including mesh topologies do not scale very well of large NoC's. For future NoC's what is needed is a clustered communication approach wherein most of the communication can be limited to small set of nodes.[11]. Some of the suggested topologies are concentrated mesh[11][13], fat tree[12][13][14], flattened butterfly[11][14] and torus[12][13]. They have been implemented for a 2D NoC with wired interconnects.

A. Concentrated Mesh

A concentrated mesh requires less number of routers needed by having each router serve its four adjacent resources. This reduces the number of hops needed thus reducing latency compared to that of mesh. [13].

B. Fat Tree

In telecommunication, a Clos network is a multistage switching network where the outputs of the stage under consideration are connected to all the crossbar switches in the next stage. By combining the output stage and the input stage together, we get a fat tree network. However this network has double the cost of a flattened butterfly network while providing the same capacity [14]. Another variation of a fat tree topology is a tapered fat tree topology where the bandwidth decreases towards the root. By doing so it reduces the amount of routers needed. [13]

C. Flattened Butterfly

The flattened butterfly is a topology that exploits high radix routers to realize lower cost than a Clos on load balanced traffic, and provide better performance and path diversity than a conventional butterfly. Fig.1 shows how by collapsing the row routers onto one route we can convert a standard butterfly to a flattened butterfly. Compared to a fat tree network, a flattened butterfly gives better performance in terms of throughput and latency for benign traffic and equal performance for worst case traffic.[14]

D. Torus

The Torus is constructed from the Mesh by adding end-around channels at the periphery.[13] The performance of torus topology is very similar to mesh topology with regards to area and power[13][12]. However the torus topology has very good latency performance since it requires lesser number of hops. This yields superior area delay product compared to mesh.[13]

Comparing the above topologies, it is observed that the concentrated mesh has superior energy delay product and area delay product compared to torus and fat tree. According to Kim [20], flattened butterfly provides lower latency and 2.5 times reduction in area when compared to concentrated mesh. Table 1 mentions the comparative results of concentrated mesh, torus and fat tree topologies.

IV. NOC FOR 3D IC.

A. Motivation for 3D NoC

Current 3D integration technologies like TSV provide a fascinating solution to the future problems of interconnect scaling. [15]. While 3D integration and NoC are proposed

solution to solve the problem of interconnect scaling, it is only in the recent past that work has been done to combine the two. Combining the concept of network on chip along with 3D offers a potential solution for building the communication infrastructure of future multi-core systems-on-chip [15].

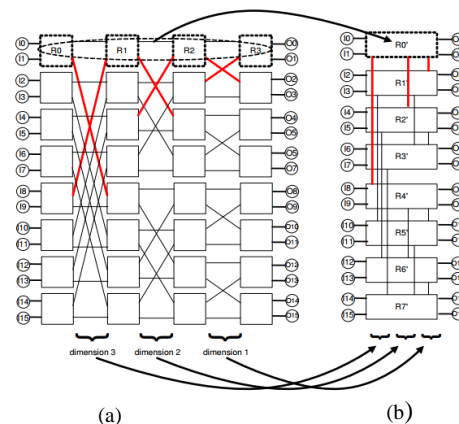


Fig. 1. (a) 2-ary 4-fly butterfly and (b) 2-ary 4-flat – the corresponding flattened butterfly with three dimensions. Lines denote unidirectional links in the butterfly and bidirectional links (i.e. two unidirectional links) in the flattened butterfly.

Parameter	Concentrated Mesh	Fat Tree	Torus
Rel. Area Delay prod (w.r.t. mesh)	0.543	0.927	0.709
Rel Energy delay prod (w.r.t. mesh)	0.331	0.807	0.781
Area	191mm ²	284mm ²	217mm ²
Rel Latency (w.r.t. mesh)	0.59	0.68	0.68

Table. 1 Comparison of results

B. Architecture

Some of the architectural solution is given below. Note that the topology used is a mesh topology.

- 1) *Symmetric NoC Router Design:* This is a natural extension to the natural NoC router for 2D integration. It is achieved by having two additional links for adjacent layer communication. The drawback is that the routers become bulky.[15]
- 2) *3D NoC-Bus Hybrid Router Design:* This design has a vertical bus for inter layer communication. The number of ports now reduce by one to 6. The drawback of this technique is that only one device can communicate vertically at a time.[15]
- 3) *True 3D Router Design.* We can develop a true 3D router that can ensure connection in 3D space in multiple hops to save area(Fig. 2). The drawback of this method is that now there will be multiple paths in the router to reach the destination. Managing these multiple paths will need extra overhead.[15]

A mesh topology is preferred in because mesh topology ensures that the routers are compact for 3D NoC. For example if we choose concentrated mesh topology, the 3D NoC-bus hybrid approach would result in a 9-port router design. [15]

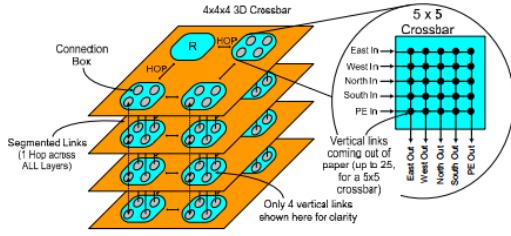


Fig. 2. A true 3D router design.

V. OPTICAL RING NOC

A. Motivation for Optical Interconnects

Future increases in the computational resources needed and off chip communication bandwidth and limitations imposed on the overall power budget and maximum on chip temperature will make power consumption of the communication network a critical factor. Photonic networks-on-chip (NoC) have been proposed as a solution to for low power communication [15]. According to Le Beux et al. [16], the main advantages of optical NoC's are increased bandwidth, immunity to electromagnetic noise, decreased latency, decreased power and current CMOS technologies are mature enough to build on chip modulators, light sources, waveguides and detectors.

B. Architecture

Most optical NoC's that have been proposed thus far all have a hybrid architecture where communication is handled by an electrical network and an optical network[16][17]. However people are now trying to develop architectures that only use an optical network to achieve communication. [18][19]. The hybrid architecture is achieved by grouping several resources and processors together into a cluster. Each cluster has an electrical to optical interface that connects the cluster to optical network. Two types of communication are distinguished:

1. *Inter cluster communication* is used for data transfer between different clusters. The optical network is used to accomplish inter cluster communication. The optical network also consists of micro resonators that are characterized to a particular wavelength. The principle of operation of these microresonators is explained in Fig 3. If the incoming wavelength is equal to the characteristic wavelength of the resonator, the wavelength is coupled into the resonator else it simply passes through. [16]

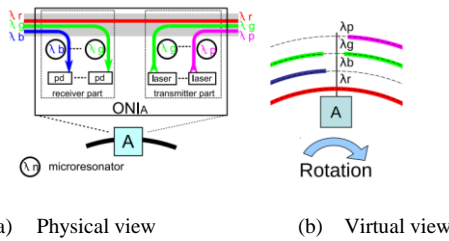


Fig. 3. Microresonators in optical portion of the ONI. From (a) we can see that wavelengths corresponding to color green, violet and blue are being coupled into the resonators.

2. *Intra cluster communication* is used for data transfer between the resources inside a cluster. The electrical network is used to accomplish intra cluster communication. Apart from intra cluster communication, the electrical network also performs serializing of data, selecting appropriate wavelength-waveguide pair for inter cluster communication and modulation.[16]

C. Principle of Operation.

The topology used is that of a ring topology. While the ring topology, which like bus allows only one master to take control at a time, is not suitable for a NoC with only electrical interconnects, it is a very good topology for optical NoC's for the following reasons. Firstly the latency for long paths is much smaller for optical paths when compared to electrical paths since the former does not suffer the problems that the latter will have with regards to large delay on long wires with future technology scaling. Secondly using concepts like WDM and wavelength reuse, one can achieve multiple master communications at a given time. Lastly the ring architecture is not only very simple to implement but also using concept like WDM and wavelength reuse, this architecture becomes more scalable for multiprocessor systems. This architecture is also contention free thus ensuring that we do not need an arbitration network for higher throughput and efficiency [16].

The principle mode of operation is explained below with the help of Fig 4.

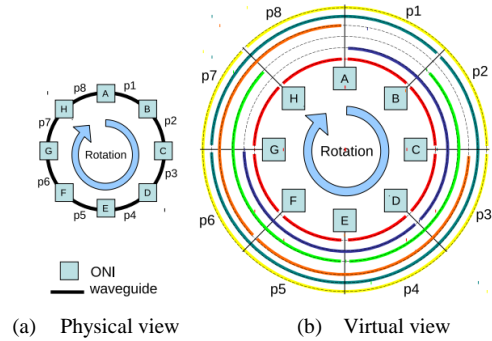


Fig. 4. The ORNoC highlighting multiplexed wavelengths and wavelength reuse to implement connections

In Fig 4b a single waveguide has been represented as 6 separate virtual rings each corresponding to a particular wavelength. The ring is divided into 8 sections here with each section connecting two optical network interfaces. Each optical network interface represents the cluster it is a part of. Notice that wavelength λ_1 corresponding to red color is reused for communication between B to C, C to D, D to E and so on. The number of wavelengths that can be used for WDM is limited and it depends on the optical properties of waveguides. In practice up to 24 wavelengths can be used. [16] Hence in order to provide multiple communication channels, multiple waveguides are used.

D. Scenarios

There are two possible scenarios where one can use ORNoC. They are given below

1. 2D or planar

This is the typical use case scenario for ORNoC. The resources are grouped into clusters and 9 clusters are connected in a 3x3 matrix form. Each cluster has an optical network interface. We define a $N \times N$ connectivity matrix for a N cluster architecture as one where the entry $c_{i,j}$ is unity if there exist a connection between optical interface i and j . Fig. 5 shows the 2d planar architecture along with its connectivity matrix.

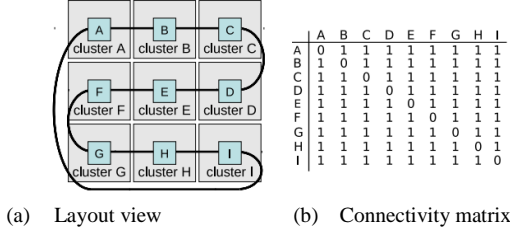


Fig. 5. ORNoC 2D planar architecture

2. 3D

The 3D architecture has an optical network layer sandwiched between two electrical layers. The electrical layers have many clusters with each cluster consisting of an optical interface.[16] The optical layer is used for inter layer communication. Fig. 6 shows the proposed 3D architecture and the corresponding connection matrix.

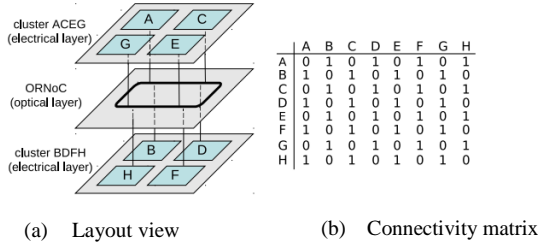


Fig. 6. ORNoC 3D architecture

E. Case Studies

Le Beux [16] implemented both 2D and 3D architectures for $N \times N$ clusters with $N = 2, 3, 4, 5$ and 6 . The maximum number of wavelengths per waveguide was chosen as $P = 8, 10, 16$ and 24 .

For both the 2D and 3D scenarios for a particular value of P , the minimum number of waveguides needed for effective communication increased sharply with N pointing to potential scaling problems for large multi-processor SoC's.

For both the 2D and 3D scenarios for a particular value of N , the minimum number of waveguides needed for effective communication increased reduced with N pointing to the need of developing waveguides that will allow maximum number of wavelengths through it.

Le Beux [16] also compared this architecture with other existing architectures like a 64 cluster cross and the Corona architecture connecting 256 (4×64) cores. For both the architecture, the above architecture requires much lesser number of waveguides and lesser number of wavelengths to multiplex.

VI. WIRELESS NOC

A. Motivation for wireless NoCs

To alleviate the long wire delays and high power consumption of future multicore computer ICs, many prototypes and commercial designs are using network-on-chip (NoC) packet switching architectures. Wireless interconnects can improve NoCs by reducing the power dissipation of long “global” wires while providing high-bandwidth and low-latency communication [21],[22]. Wireless interconnects can provide some unique benefits including:

- Reduced power dissipation by avoiding multi-hop communication as in traditional metallic interconnects
- Reduced IC area overhead (fewer wires, waveguides) and lower parasitics.
- Reuse of complementary metal oxide semiconductor (CMOS) wireless transceiver device designs

Wireless technologies have an advantage of being a mature form of communication with many well-known applications implemented in wireless local area networks, cell phones, and so on. This existing knowledge in the wireless/radio frequency (RF) field will facilitate the integration of wireless interconnects for NoCs, or WINoCs. Yet even with the relative maturity of wireless communication technologies, scaling these to very small sizes while concurrently scaling data rates to multiples of tens of Gb/s presents significant challenges in multiple areas, including network architecture, wireless propagation modeling and antennas, and low-power circuit and device design.

B. WINoC challenges

In order to ensure that wireless links truly enhance NoC performance, they must:

- Provide high throughputs (e.g., tens of gigabits per second)
 - Employ power- and area-efficient transceivers
 - Employ efficient MA across the shared spatial channel.
- Providing tens of gigabits per second among multiple cores is nontrivial; this is particularly true when frequency spectrum is limited. Although link distances are very short, wireless transceiver power dissipation must be minimized, and in the low millimeter wave frequency range, antennas will be inefficient due to their small electrical size. These large data rates also challenge circuit design, as most digital circuits cannot currently operate at these rates, and required serial-parallel conversions may introduce unacceptable overhead in power and complexity, so very simple modulation/demodulation schemes may be required. When spectrum is limited, time and frequency division must be used to allow sharing of the wireless medium. Spatial-division multiplexing (SDM) could provide welcome spatial reuse of time-frequency resources, but this is extraordinarily challenging at millimeter wave frequencies at present.

C. Architecture

The WINoC architecture is a scalable, wireless hybrid NoC. The architecture is separated into hierarchical subsections that define the communication protocol as shown in Fig. 7. Four cores ($N = 4$) are concentrated into (wired to) one cluster, and each cluster has its own data packet router. Routers are the

nodes that are connected by the links (wired or wireless). The function of the router is to move packets from source to destination. Routers consist of buffers that store data, crossbars that switch or move data, and wireless transceivers. The four-core cluster has been shown [22] to be an effective design to reduce the router area overhead as well as serialization latency. The WINoC MA scheme uses both time and frequency division to enable wireless transmission from any core in any cluster to any other core in any other non-adjacent cluster (adjacent cluster communication is wired). Clusters are grouped into sets. Set-to-set duplexing is via frequency division, and transmit multiplexing employs time division; this ensures single-carrier wireless transmission by any wireless modulator. The use of both wired and wireless communications provides efficiency and flexibility at the expense of a slight increase in MA complexity.

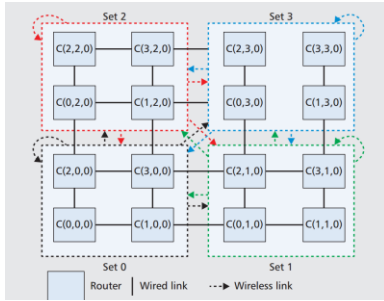


Fig 7. WINoC with C=4 and S=4 showing wireless communications between sets

D. Channel design

The frequency selection for this NoC had to be done carefully since a trade-off is involved. It is important to note that it is not possible to obtain any truly accurate channel models without precise specification of the physical “landscape” of the WiNoC. The landscape is defined by the dimensions and electrical properties (conductivity σ , permittivity ϵ , and permeability μ) of all objects/surfaces in the environment through which the electromagnetic waves from transmitter to receiver propagate. This landscape could ultimately be quite complex, rendering accurate analysis of the channel impossible without sophisticated computer computations. At present we do not define the landscape in detail, and our initial channel model simply employs estimates. Channel modeling is ongoing work. By “channel model”, the primary meaning is the attenuation and delay dispersion characteristics, over frequency band(s) of interest, over expected link distances. Antennas are often excluded from many channel models, but in the tens-to-hundreds of gigahertz or terahertz frequency ranges, we will ultimately need to incorporate antenna characteristics as well. Novel approaches such as carbon nanotubes, semiconductor nanowires, or even meta materials may be required to keep antenna sizes small with acceptable radiation efficiencies. Electromagnetic compatibility is also an issue for future study.

The table 2 shows the tradeoff for different frequency bands that are used when being used as multiplexed channels.

Technology/ Design Area	Frequency Band		
	50–150 GHz	150–500 GHz	500 GHz–3 THz
Circuits, Devices	Status: currently feasible Technology: RF-CMOS, substrate SOI	Status: encouraging Technology: SiGe-BiCMOS, substrate SOI	Status: immature Technology: III-V/Si hybrid, substrate alumina
Antennas, Propagation	Status: challenging Issues: electrically-small (inefficient) antennas, near field coupling	Status: challenging Issues: nearing conventional antennas, far-field conditions	Status: reasonable Issues: at highest f_s , propagation analysis conventional, antennas immature
System, Architecture	Issues: throughput low to-moderate, SDM very difficult Area: low-Q inductors, antenna size Power: manageable	Issues: sufficient throughput, SDM challenging Area: very lossy substrates, ultra-low Q Power: challenging	Issues: ample throughput, SDM possible Area: limited by waveguides and-sources Power: very challenging

Table 2 WINoC frequency trade-offs for 3 potential bands

E. Case studies

An initial working model of a WINoC known as an iWISE was implemented by DiTomaso et al[23] in which a hybrid communication network was used to reduce area overhead with smaller routers and shared buffers, and also improved performance by increasing the hop count. 2X power and 2X area was saved by implementing their design.

Also, by scaling it to 256 cores, the performance could be increased by 2.5X and the power was saved by 2X.

Based on this work, Motalak et al[24] modified this design to further decrease the power by 34 per cent by making the architecture a shared interconnect architecture that consists of both wired interconnects and wireless interconnects. Also the performance could also be increased by 2.5X. This architecture is the one that is explained in detail.

VII. CONCLUSION

The unique requirements of NoCs have resulted in many different topologies. Comparisons have been made and it is found that the flattened butterfly topology provides the best results. However the choice of topology also depends on the type of channel. For example, the most preferred topology for optical NoCs is ring.

Wireless NoCs and optical NoCs represent two exciting solutions to solve the current problem of low power on chip communication. While both have their own advantages and challenges, both have been shown to low power on chip communication viz. wired NoCs

VIII. FUTURE WORK

If a grant were given, it would be invested to find ways to multiplex the data on channels at a faster rate than the hybrid design for WiNoC, make a trade-off for antenna size (good for high frequencies) and the device materials (good for low frequencies) for WiNoC. We also propose to compare of performance of wireless NoCs and optical NoCs with respect to different kinds of on chip communication traffic.

IX. REFERENCES

- [1] <http://gram.eng.uci.edu/comp.arch/lab/NoCOverview.htm>
- [2] Pierre Guerrier, Alain Greiner. (2001). “A Generic Architecture for On-Chip Packet-Switched Interconnections.” Presented at DATE 2000, Paris, France.
- [3] Milica Mitić, Mile Stojčev. Title “A Survey of Three System-on-Chip Buses: AMBA, CoreConnect and Wishbone.”
- [4] Arteris.inc “A comparison of Network-on-Chip and Busses.” White paper
- [5] Cesar A. Zeferino, Márcio E. Kreutz, Luigi Carro, Altamiro A. Susin (2002). “A Study on Communication Issues for Systems-on-Chip.” Presented at Symposium on Integrated Circuits and Systems Design

- [6] Ahmed Hemani, Axel Jantsch, Shashi Kumar, Adam Postula, Johnny Öberg, Mikael Millberg, Dan Lindqvist. "Network on a Chip: An architecture for billion transistor era."
- [7] William J. Dally, Brian Towles (2001, June). "Route Packets, Not Wires: On-Chip Interconnection Networks." Presented at DAC 2001.
- [8] Shashi Kumar, Axel Jantsch, Juha-Pekka Soininen, Martti Forsell, Mikael Millberg, Johnny Öberg, Kari Tiensyrjä, Ahmed Heman. (2002). "A Network on Chip Architecture and Design Methodology. Presented at IEEE Computer Society Annual Symposium on VLSI".
- [9] Drew Wingard. (2001, June). "Micronetwork based integration of NOC's." Presented at DAC 2001
- [10] Partha Pratim Pande, Cristian Grecu, Michael Jones, Andre' Ivanov, Resve Saleh. (2005, June). "Performance Evaluation and Design Trade-Offs for Network-on-Chip Interconnect Architectures." Published
- [11] Reetuparna Das, Soumya Eachempati, Asit K. Mishra, Vijaykrishnan Narayanan, Chita R. Das. (2008). "Design and Evaluation of a Hierarchical On-Chip Interconnect for Next-Generation CMPs." Published
- [12] Prabhakar Mishra, Nidhi A, J K Kishore, Chetan Kotturshettar. (2014). "Design of Configurable, Network Based Interconnection Modules for Communication Centric System-On-Chip Application." Presented at ISCAIE 2014
- [13] James Balfour, William J. Dally (2006). "Design Tradeoffs for Tiled CMP On-Chip Networks." Presented at ICS 2006
- [14] John Kim, William J. Dally, Dennis Abts (2007, June). "Flattened Butterfly : A Cost-Efficient Topology for High-Radix Networks." Presented at ICSCA 2007
- [15] Luca P. Carloni, Partha Pande, Yuan Xie (2009). "Networks-on-Chip in Emerging Interconnect Paradigms: Advantages and Challenges" Published.
- [16] S'ebastien Le Beux, Jelena Trajkovic, Ian O'Connor, Gabriela Nicolescu, Guy Bois, Pierre Paulin (year, month). "Optical Ring Network-on-Chip (ORNoC): Architecture and Design Methodology." Presented at DATE 2011
- [17] Jason Miller, James Psota, George Kurian, Nathan Beckmann, Jonathan Eastep, Jifeng Liu, Mark Beals, Jurgen Michel, Lionel Kimerling, and Anant Agarwal. (2009, April). "ATAC: A Many core Processor with On-Chip Optical Network" Presented at MA CSAIL
- [18] Sandro Bartolini, Luca Lusnig, Enrico Martinelli (2013). "Olympic: a Hierarchical All-optical Photonic Network for Low-power Chip Multiprocessors." Presented at 16th Euromicro Conference on Digital System Design.
- [19] Paolo Grani (2014). "From Hybrid Electro-Photonic to All-Optical On-chip Interconnections for Future CMPs" Published
- [20] John Kim, James Balfour, William J. Dally (year, month). "Flattened Butterfly Topology for On-Chip Networks." Unpublished.
- [21] Amlan Ganguly, Kevin Chang, Sujay Deb, Partha Pratim Pande, Benjamin Belze, Christof Teuscher (2011, October). "Scalable Hybrid Wireless Network-on-Chip Architectures for Multicore Systems" Published .
- [22] Ruizhe Wu, Yi Wang, Dan Zhao (2010). "A Low-Cost Deadlock-free Design of Minimal-Table Rerouted XY-Routing for Irregular Wireless NoCs" Presented at Fourth ACM/IEEE International Symposium on Networks-on-Chip
- [23] Dominic DiTomaso, Avinash Kodi, Savas Kaya, David Matolak. (2011). "iWISE: Inter-router Wireless Scalable Express Channels for Network-on-Chips (NoCs) Architecture." Presented at 19th Annual IEEE Symposium on High Performance Interconnects
- [24] David W. Matolak, Avinash Kodi, Savas Kaya, Dominic DiTomaso, Soumyasanta Laha, William Rayess (2012). "Wireless Network-on-Chip: Architecture, Wireless Channel, and Device" Published