Original Open Access

A Packet Truncation Mechanism in an Approximate Network on Chip

Sina Yousefisadr [1*], Masoumeh Momeni [2]

1. Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
2. School of Electrical Engineering, Electronic Research Center, Iran University of Science and Technology, Tehran, Iran.

Abstract

Network-on-Chips (NoCs) as a standard interconnection impose high latency and excessive power consumption in many-core systems. Emerging data-intensive applications possess a high volume of data movement across the network which deteriorates the network congestion condition. These applications have an intrinsic feature, namely error tolerance, which presents a new communication paradigm. We employ a differential-based approximate method for packet transmission to reduce the packet size on the network. General NoC architectures have a large enough flit channel so that the packet header includes many free bits. As we reduce the packet size by transmitting the difference data on the network, we can accommodate the additional parts of the header to store the difference data that must be transmitted. This approach in data storage and transmission optimizes the packet size, which reduces the network congestion by using the idle space of head flit and employing approximate-based data transmission. We apply this method in 3D NoC due to its low latency architecture. Therefore, we could alleviate 3D NoC thermal challenges. The simulation results show that our approximate-based NoC architecture decreases the latency and dynamic power consumption by 37% and 42% in comparison to traditional 3D NoC, respectively.

Keywords Packet difference transmission, Multi-layer NoCs, Network congestion, Approximation.

Introduction

Three-dimensional Network-on-Chips (3D NoCs) present lower latency and higher performance compared to 2D NoCs. They have been introduced a standard communication platform in many-core systems at the cost of thermal overhead [1]. High volume of data movement imposes network congestion and hotspot creation in data-intensive applications, which increases the temperature in different layers of chip. We must reduce the volume of data transition across the network to improve the system performance in these processing applications. Therefore, low power and low latency interconnection could be used for processing applications.

On the other side, high-speed networks demand solution to control the network traffic [2]. Most of data-intensive application allow some output error [3, 4].

The data-intensive applications impose a heavy traffic on the network. This causes network congestion and thermal issues, which reduces network performance. The heavy volume of data movement demands a solution to address the challenges of NoC. Therefore, low power and low latency interconnection could be used for processing applications. The approximate approach can be applied in the communication as an efficient solution to decrease network traffic volume and obtain energy-saving and performance improvement. The approximate computing trades

* Correspondence:
sinayousefisadr@gmail.com

the accuracy of applications by the energy consumption, since applications possess error- tolerance feature. Approximate communication concepts [5-8] have been introduced to achieve high-throughput and energy-saving in NoC. Approximation can be applied between the cores since the cores allow faulty computation. The additional parts of header flit are also free, which means waste of NoC resources.

In this paper, a communication approach based on the data locality and the error tolerance of the application is presented in 3D NoC that decreases the amount of traffic on NoC. We transfer the differences between the packet data and an adaptive predetermined value in the network. Then, we use the free space of the header flit to store this truncated packet. Therefore, the volume of data transmission and network congestion in today data-intensive applications is decreased. This approach reduces network latency and energy consumption.

Th paper is organized as follows. The recent works related to the NoC challenges would be described in Section II. Section III motivates toward the approximate communication in NoC. The proposed approach applied in 3D NoC is presented in Section IV. Simulation results is discussed in Section V. Eventually, Section VI concludes the paper.

Related works

2D Network-on-Chips consume a large amounts of time and power to transmit data across the network. 3D NoCs possess lower latency compared to 2D NoCs. However, 3D NoCs impose high power consumption and thermal issues. Many studies have been presented to address these challenges that can be classified into network topology, routing mechanism and router architecture [9-11]. A power-aware routing algorithm is presented in [12] to overcome thermal challenges in 3D NOCs.

Approximate communication can improve network performance and power consumption. An approximate method is introduced in [13] which sends only not error resilience packets to obtain energy saving. Many applications in domains like machine learning and image processing are error-resilient and allow data inexactness [3, 4]. Approximate communication approach is used in [5-8] between the nodes to improve energy efficiency and performance while the output of application is in an acceptable range.

In [5], the traffic volume between the nodes is reduced by presenting three approximate communication methods named compression, synchronization, and prediction. The compression mechanism is used in [6], which approximate data to enhance the compression rate. This method decreases the packet size, which reduces network congestion in NoC. However, it imposes high hardware overhead because of large compression tables. The traffic is regulated in [8] which drops the parts of error-tolerant data based on network congestion. Then, the dropped data is predicted in the network interface (NI) of the destination node. In [14], a reconfigurable router architecture is introduced that switch the supply voltage of routers to a lower voltage swing for approximable data. In [15], an approximate NoC architecture is presented which identifies similar data packets and employs an overlay circuits to transmit them, simultaneously.

In this paper, we introduce an approximate data communication in 3D NoC architecture which remarkably reduces the transmitted packet size on NoC. In the presented approach, we store the difference between the packet data and a predetermined value in the empty space of the packet header flit to reduces the packet size and network congestion.

Challenges and opportunities of NOC

Emerging data-intensive applications include heavy load in NoC, which results in communication bottleneck in many cores on chip. These challenges are described in the following.

A. Communication challenges

By integrating many cores on a chip, communication consumes more time and energy than computation part. Heavy traffic in data-intensive applications has resulted in high latency and power consumption in NoC. This also increases the temperature dissipation.

B. Error tolerant applications

Data-intensive applications in domains including machine learning and big data accepts approximate outputs [3, 4]. Approximate Computing presents a good paradigm for trading the output accuracy by power consumption and time. As applications have error resiliency and the output error is inconsiderable for users, approximate mechanism can be used for improving the communication performance. Therefore, reliability evaluation of output must be discussed in approximate communication paradigm.

C. 3D NoC opportunities

Many cores on a chip and large amount of data movement in emerging applications increases the communication latency. High network diameter in 2D NoCs increases the network latency and degrades the performance of Network-on-Chip. 3D NoCs have lower latency than 2D NoCs do. However, 3D NoCs include thermal problems and high power consumption. Thermal challenges are very important in 3D designs as degrades the performance.

In our proposed paper, we store the differences between the current packet data and the predetermined value in the header flit of the packet to traverse the network for traffic reduction in data-intensive applications.

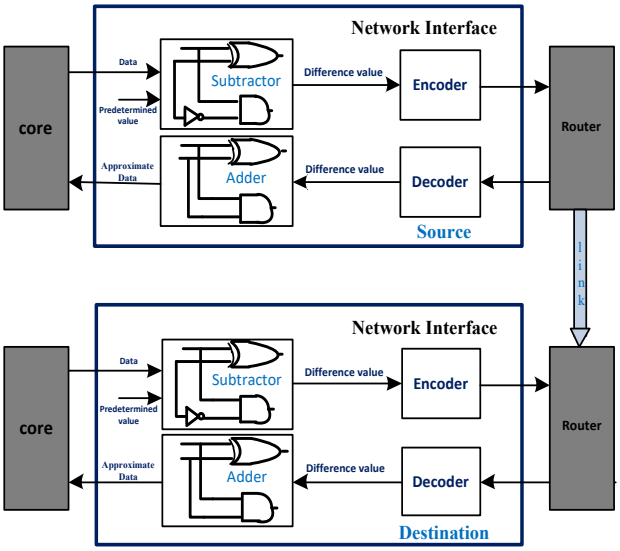

Fig 1. Implementation of the proposed communication method in NoC

Trancated packet across the network due to error tolerance of applications

The data-intensive applications imposes a heavy data transmission in the network. This volume of traffic results in communication bottleneck and NoC performance degradation. Thus, a solution is required to overcome these challenges of NoCs. These applications accept some specified output error since the users cannot distinguish this output variation. We explore this feature and transmit the truncated packet across the network. Our communication approach is applied in 3D NoC architecture. It reduces the transmitted packet size on NoC as the compression technique. In this communication approach, we store the differences between the packet data and a predetermined value in the empty space of the header flit. Therefore, the packet length is considerably reduced in the network. This method can significantly control the congestion in NoC and improve the network performance. Therefore, the volume of traffic and congestion in data-intensive applications is significantly reduced which results in the great performance achievement and energy saving in such time-consuming and power-hungry interconnections. The long length packet which would be transmitted across the network, is subtracted from the predetermined value and the differences between them is transmitted to the destination node. This approach is based on the data locality and the error tolerance nature of the applications. Then, we place this data difference in the free space of the header flit to optimize the resource utilization. It should be mentioned that the predetermined value would be updated after a determined cycles. As the applications can tolerate some output error, the difference value is not a precise value.

As shown in Fig. 1, to implement the proposed communication method in NoC, we change the Network Interface of NoC. The difference data that is stored in the header flit, would be transmitted across the network. This packet with short length would be passed the network. It would be added with the predetermined value to be transmit to the processing core at the destination. This approximate value embedded in free space of the header flit is acceptable since the applications allow faulty output which are not noticeable for end users.

Table 1. Evaluation Parameters

On chip network Parameters	4 × 4× 3 3D-mesh 3 Virtual Channel Wormhole Switching XYZ Routing
System Parameters	48 Cores at 2GHz L1 Instruction Cache and Data Cache with 32KB size 4-way associative L2 Cache with 16 MB size

Performance evaluation

A. Simulation results

The proposed packet truncation mechanism is simulated in 3D NoC architecture using full system simulator, gem5 [16].

We use Garnet [17] on-chip network model in our simulation. The simulation configuration and parameters are summarized in Table I. The DSENT simulator [18] is added to the Gem5 to measure the latency and power consumption. PARSEC benchmark suite [19] is employed for performance evaluation of NoC.

PARSEC benchmarks have some approximate region which allows output error. We apply the proposed truncated packet transmission to the variable of these regions. There are quality metrics related to each of the benchmarks to calculate the value of the error. In blackscholes as an example, the error is determined by the percentage of prices with the error greater than 1%. The other benchmarks quality metric are explained in Table II.

B. Results Analysis

We evaluate two metrics of the energy efficiency and network latency in proposed NoC architecture. We store difference data flits into header flit and transmit the truncated packet across the network based on the data locality

Table 2. PARSEC benchmark suite [19]

Name	Domain	Quality evaluation metric
Blackscholes	financial analysis	percentage of prices with the error greater than 1%
Swaption	pricing of a portfolio of swaptions	error between the approximated prices with its exact ones
X264	video encoding	peak signal-to-noise ratio
Caneal	machine learning	difference between the routing cost for the approximated and precise
Ferret	image search	the percentage of the image not searched out in exact output
Vips	image processing	error of RGB image pixel values

and the error tolerance of the application. Therefore, traffic volume and network congestion is decreased. Also, we optimize the resource utilization of the NoC by using the idle parts of the header flit. Both of these techniques lead to packet size reduction, which is a real bottleneck in data-intensive applications. As shown in Fig. 2, the proposed method reduces network latency by 37% on average compared to baseline 3D NoC for different benchmarks of PARSEC suit. In other words, our proposed mechanism performance is in similar way of compression mechanism, while it does not impose high hardware overhead. Thus, network congestion is significantly reduced for data-intensive application in many-core systems, which means low latency achievement and high energy-saving. Fig. 3 shows the energy efficiency of our presented communication mechanism, which is improved on average by 42% compared to baseline 3D NoC. As the results indicate, the presented communication technique can significantly improve both the energy efficiency and the latency in the 3D NoC. Our solution decreases network congestion, and avoids creating hotspot node. Therefore, it reduces thermal challenges upper layer of 3D NoC.

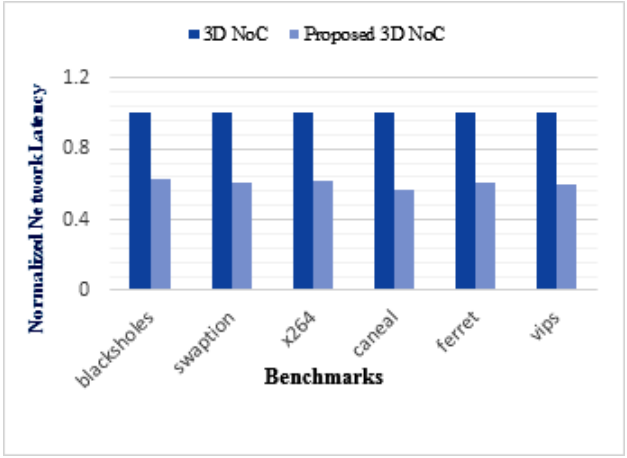

Fig 2. Normalized network latency across different benchmarks.

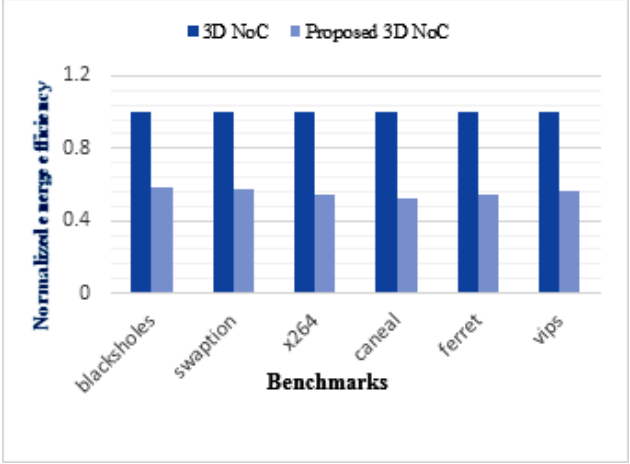

Fig 3. Normalized network energy efficiency improvement across different benchmarks.

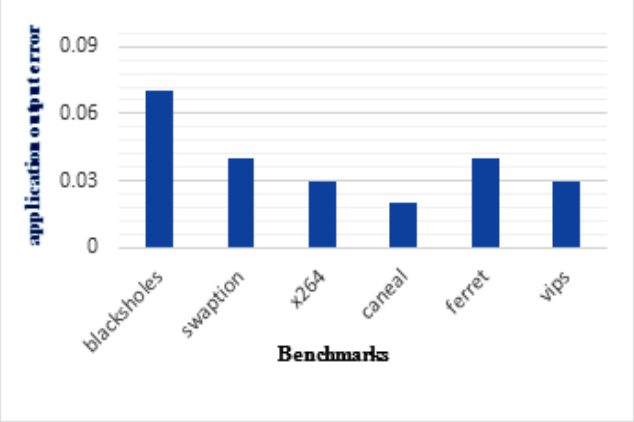

Fig 4. Application output error

Conclusion and future works

Data-intensive applications impose a large amount of data movement in many-core systems, which results in network congestion and hotspot creation in NoC. This means high latency and excessive power consumption. On the other side, these applications have interesting error tolerance feature due to inconsiderable output variation. By employing this feature, we store difference data flits into header flit and transmit the packet differences across the network based on the approximate approach. This new communication approach in which the packet size is significantly reduced is applied in 3D NoC due to lower latency architecture. Our solution significantly decreases the congestion in NoC for today's data-intensive application. Therefore, it reduces network latency and energy consumption, and thermal issues in 3D NoC. Our approximate data communication mechanism decreases

the latency and dynamic power consumption by 37% and 42% compared to baseline 3D NoC, respectively.
As future work, mapping algorithms can be applied in 3D NoC in a way that the approximate part of application would be mapped in upper layer with low power resources. It would balance the thermal distribution for 3D NoC architecture. Therefore, workloads can be mapped in a way that the number of voltage convertors in each layer are reduced to achieve more energy efficiency.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations
Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: May. 2023 Accepted: Aug. 2023
Published online: Sep. 2023

DOI: 10.22034/ASAS.2023.412533.1034

REFERENCES
[1] K. Manna, and J Mathew, Design and Test Strategies for 2D/3D Integration for NoC-based Multicore Architectures. Springer Nature, 2020.
[2] T.S.Das, P.Ghosal, and N.Chatterjee, "VCS: a method of in-order packet delivery for adaptive NoC routing," Nano Communication Networks, 28, p.100333, 2021.
[3] S.Mittal, "A survey of techniques for approximate computing," ACM Computing Surveys, 2016.
[4] Q. Xu, T. Mytkowicz, and N. Sung Kim, "Approximate computing," A survey: IEEE Design & Test, Vol 33, No. 1, pp. 8-22, 2016.
[5] Filipe Betzel et al., "Approximate Communication: Techniques for Reducing Communication Bottlenecks in Large-Scale Parallel Systems", ACM Computing Survey, 2018.
[6] Rahul Boyapati, et al., "APPROX-NOC: A Data Approximation Framework for Network-On-Chip Architectures", ISCA Conference, pp. 666–677, 2017.
[7] Y. Chen, M. F. Reza, and A. Louri, "DEC-NOC: An Approximate Framework Based on Dynamic Error Control with Applications to Energy-Efficient NoCs", ICCD Conference, pp. 480–487, 2018.
[8] L. Wang, X. Wang, and Y. Wang, "ABDTR: Approximation-Based Dynamic Traffic Regulation for Networks-on-Chip Systems", ICCD Conference, pp. 153– 160, 2017.
[9] A.S.Kumar, and B.Naresh Kumar Reddy, "An Efficient Real-Time Embedded Application Mapping for NoC Based Multiprocessor System on Chip," Wireless Personal Communications, 128(4), pp.2937-2952, 2023.
[10] T.Pullaiah, K.Manjunathachari, and B.L.Malleswari, "B-NIS: Performance analysis of an efficient data compression technique for on-chip communication network," Integration, 89, pp.83-93, 2023.
[11] M.Trik, et.al "A new adaptive selection strategy for reducing latency in networks on chip," Integration, 89, pp.9-24, 2023. M. Elahi, et.al, "LTD-Router: Low Latency Traffic Distributed FPGA Based Router Architecture Using Dedicated Paths," ICEE Conference, May 2018.
[12] M. Nezarat, H.S. Shahhoseini, M. Momeni, "Thermal-Aware Routing Algorithm in Partially Connected 3D NoC with Dynamic Availability for Elevators", Journal of Ambient Intelligence and Humanized Computing, pp. 10731-10744, 2023.
[13] Y. Chen, M. F. Reza, and A. Louri, "DEC-NOC: An Approximate Framework Based on Dynamic Error Control with Applications to Energy-Efficient NoCs", ICCD Conference, pp. 480–487, 2018.
[14] A. B. Ahmed, et al., "AxNoC: Low-power Approximate Network-on- Chips using Critical-Path Isolation", NOCS'18, 2018.
[15] V. Y. Raparti and S. Pasricha. Dapper, "Data aware approximate noc for gpgpu architectures", IEEE/ACM Int'l Symp on NOCS, pp. 1–8, 2018.
[16] N. Binkert, et al., "The Gem5 Simulator", SIGARCH Comput. Archit. News, pp. 1–7, 2011.
[17] N. Agarwal, et al., "GARNET: A detailed on-chip network model inside a full-system simulator", ISPASS. 33–42, 2009.
[18] C. Sun, et al., "DSENT - A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling", NOCS, pp. 201–210, 2012.
[19] C. Bienia, et al., "The parsec benchmark suite: Characterization and architectural implications", PACT, 2008.