An Approximate Multiplane Network-on-Chip

Ling Wang*, Yadong Wang† and Xiaohang Wang‡
*†*Harbin Institute of Technology*
*‡*South China University of Technology*
Email: *ling.wang@hit.edu.cn, †ydwang@hit.edu.cn, ‡xiaohangwang@scut.edu.cn

***Abstract*—The increasing communication demands in chip multiprocessors (CMPs) and many error-tolerant applications are driving the approximate design of the network-on-chip (NoC) for power-efficient packet delivery. However, current approximate NoC designs achieve improvements in network performance or dynamic power savings at the cost of additional circuit design and increased area overhead. In this paper, we propose a novel approximate multiplane NoC (AMNoC) that provides low-latency transfer for latency-sensitive packets and minimizes the power consumption of approximable packets through a lossy bufferless subnetwork. The AMNoC also includes a regular buffered subnetwork to guarantee the lossless delivery of nonapproximable packets. Evaluations show that, compared with a single-plane buffered NoC, the AMNoC reduces the average latency by 41.9%. In addition, the AMNoC achieves 48.6% and 53.4% savings in power consumption and area overhead, respectively.**

I. Introduction

Approximation, that is, sacrificing the output accuracy to achieve benefits in performance and a higher energy efficiency, has gained extensive recognition as a solution for satisfying energy-efficient hardware design [1]. Approximate designs rely on the ability of applications to tolerate computations on noisy/erroneous data or imprecision in the computation results. Many machine learning, searching, scientific computing, and multimedia applications are inherently tolerant of approximation [2]. Hence, since some inexactness is acceptable, these applications allow the presence of approximate data in storing, computing or transmitting. These applications, which exhibit some level of error tolerance, motivate the generation of approximate hardware designs to achieve high performance and high energy efficiency.

With increasing on-chip core counts, the network-on-chip (NoC) has emerged as the most competent method for on-chip communications in large-scale parallel systems. An NoC connects varied on-chip components, such as cores, caches and memory controllers, and enables the communications necessary for exchanging data among parallel threads and ensuring data coherence. However, NoCs consume a significant amount of power in modern chip multiprocessors (CMPs) [3]. Thus, energy efficiency has been a primary concern in NoC design [4]. Reducing the NoC power while increasing performance is essential for scaling up the design to larger CMP systems, and approximation techniques that relax the accuracy in exchange for improvements in performance and energy conservation have remarkable potential for research regarding energy-efficient designs.

Recently, some new approximation NoC techniques focusing on providing disproportionate gains in efficiency have been presented by the scientific community [5]–[8]. These designs allow refined energy-quality trade-offs in NoCs by reducing the injection of approximable data [5], [6] or decreasing the supply of energy for approximable data transmission endeavors [7], [8]. Some new microarchitectures have been proposed and integrated into single-plane NoCs to control packet injection or connection circuits in the networks. Nevertheless, although these designs reduce the dynamic NoC power consumption, these accomplishments are obtained at substantial costs, that is, increases in complexity, power leakage and area overhead.

Reducing power consumption while increasing performance requires the efficient use of network resources. Multiplane NoCs have shown their efficiency in total bandwidth usage [4], [9]. Furthermore, multiplane NoCs can be designed with heterogeneous physical subnetworks; as a result, messages are injected into different subnetworks to satisfy different transmission properties. For many approximation-enabled applications, not all messages need to be delivered without loss; that is, approximable data can be transmitted with loss to simplify the NoC architecture and achieve a relatively small area overhead and reduced power consumption. This observation intuitively suggests that instead of having one interconnected plane serving all the on-chip traffic, the NoC may be physically split into two planes: a lossless plane for transmitting nonapproximable traffic and an area- and power-efficient plane that operates with lossy delivery and serves approximable traffic.

In this work, we propose an approximate multiplane NoC (AMNoC). The proposed AMNoC includes a lightweight, low-latency, lossy-transmission NoC (the approximate subnetwork, approx-subnet) and a conventional NoC without traffic loss (the lossless subnetwork, lossless-subnet). The approx-subnet is bufferless with a lightweight router architecture that consumes very little area and power. The lossy transmission in the approx-subnet allows flits to be discarded in the event of a conflict and recovers the missing flits after packet transmission; therefore, there is no congestion in the approx-subnet,

This work was supported by funding from the National Key Research and Development Program of China (No. 2017YFC1201201, 2016YFC1202302 and 2017YFSF090117), the National Natural Science Foundation of China (No. 61822108, 61571152 and 61971200), the Natural Science Foundation of Guangdong Province (No. 2018A030313166), Research Grant of Guangdong Province (No. 2017A050501003), Pearl River S&T Nova Program of Guangzhou (No. 201806010038) and the Fundamental Research Funds for the Central Universities (No. 2019MS087).

Corresponding Author: Yadong Wang and Xiaohang Wang.

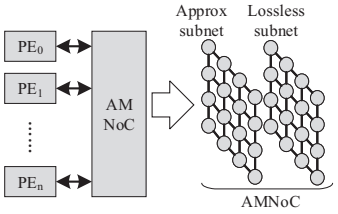


Fig. 1. AMNoC architecture overview.

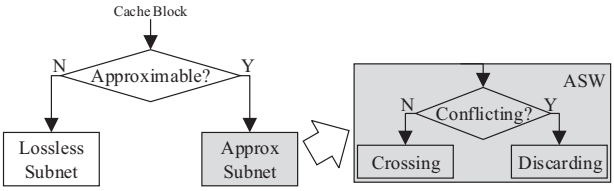


Fig. 2. Approx-subnet operation flowchart.

and packets are transferred with low latency. In addition, the lossless-subnet provides guarantees that nonapproximate flits can arrive at their destinations without loss.

We make the following contributions in this paper:

(1) A novel AMNoC that can achieve low-power and low-latency packet transmission is presented.

(2) We design a low-latency bufferless subnetwork with lossy transmission (approx-subnet), which provides power- and area-efficient delivery for approximable packets and ultra-fast delivery for latency-sensitive packets.

(3) Experiments show that the AMNoC achieves an average latency reduction of 41.9%, an average power reduction of 48.6% and an average area conservation of 53.4% compared to a regular single-plane buffered NoC.

II. Architecture Design For The AMNoC

A. Overview

Multiplane Architecture. In this section, we present the AMNoC design, which consists of a two-plane architecture. Figure 1 shows thea high-level architectural depiction of the AMNoC router. In the AMNoC, which connects the processing elements (PEs), there are two physically separate subnetworks: lossless-subnet and approx-subnet. There is no connection between these two subnetworks, which run independently. The lossless-subnet is a full-functional buffered NoC, while the approx-subnet is a lightweight bufferless NoC that relaxes the transmission accuracy to provide low-latency packet delivery with low power and low area overhead. This two-plane architecture reduces the power consumption of approximable packets while maintaining the lossless transmission of non-approximable packets.

Approximate Design. Bufferless NoCs have been shown to be very effective at conserving power and area under low network loads; in contrast, due to retransmission and misrouting, bufferless NoCs incur a significant performance penalty under high network loads [10], [11]. In the AMNoC, all approximable packets are transmitted by the bufferless approx-subnet, which is designed based on an approximate switch (ASW) design. As a result, instead of being retransmitted or misrouted, the contending flits are discarded in a network conflict. Figure 2 shows the workflow in the AMNoC. There are no buffers at the switch ports in the approx-subnet, so any incoming flits cannot be stopped. In every cycle, flits that arrive at the input ports contend for the output ports. When two or more flits contend for the same output port, only one can be transferred through the output channel, while all other conflicting flits are discarded without being collected. In the approx-subnet, after an approximable packet finishes its transmission, the missing flits will be recovered based on the received flits. Based on the lossy design, the approx-subnet transmits packets without congestion. Through simplified routing and port allocation, the approx-subnet also achieves single-cycle hop delivery, and packets in the approx-subnet can be transmitted with ultra-low latency.

Multiplane Packet Transmission. As an approximate multiplane design, approximable packets are transmitted through the approx-subnet, while nonapproximable packets are injected into the lossless-subnet. However, some critical latency-sensitive messages are nonapproximable, and their transmission latencies greatly influence the overall application performance [4], [9]. In particular, the control packets are all single-flit and nonapproximable. In the AMNoC, we also inject these critical messages into the approx-subnet to provide an opportunity for the ultra-fast delivery of these messages; however, these critical messages are also injected into the lossless-subnet to guarantee the lossless delivery of nonapproximable data. Hence, any critical messages dropped by the approx-subnet will still arrive at its destination through the lossless-subnet. Therefore, the dropped critical messages in the approx-subnet do not need to be recovered. Critical messages include all control packets and the first flit in each data packet since the first flit of a data packet contains the initially requested word. When the first flit arrives at the destination node, the processor can continue executing with the received word before the rest of the packet has arrived. In addition, the first flit of an approximable packet is also injected into the lossless-subnet; this prevents all the flits in an approximable packet from being discarded in the approx-subnet, which may cause the data request or synchronization to fail. At least one flit exists in each approximable packet, which can reach the destination node. In the AMNoC, the approx-subnet carries approximable data packets, control packets and the first flits of nonapproximable data packets, while the loss-subnet transmits control packets, nonapproximable data packets and the first flits of approximable data packets, as shown in Figure 3. Although some messages are transmitted twice, requiring more energy, the overall power overhead is decreased. The reason for this decrease is that packets are transmitted with a much lower power in the bufferless approx-subnet than in regular buffered transmission, and the approximate transmission also reduces energy consumption. The AMNoC therefore provides

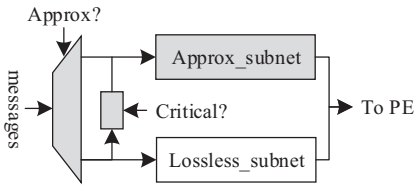


Fig. 3. Multiplane packet transmission in the proposed AMNoC.

low-latency transfer for critical messages and minimizes the power consumption of approximable messages.

B. Approx-subnet Design

Bufferless Router. The AMNoC is designed as a 2D mesh topology, which is commonplace in modern systems. Each router consists of five input and output ports: one for each of the four cardinal neighbor connections and one for the local connection. In each neighbor connection, only a latch is used to store an incoming flit from the neighbor direction at any cycle. There are no buffers in any of the four neighbor connections, so incoming flits cannot be stopped. All arriving flits contend for output ports and are forced to be routed out in the next cycle*. The injection and ejection port is equipped with some buffers since the packets in local routers can be injected into the approx-subnet only when their ideal output slot is free. Hence, packets in injection ports can be buffered instead of being discarded in allocation conflicts. In addition, ejection buffers are necessary for receiving the arrived packets. In the approx-subnet, the local injection is also controlled by a timer. Multiflit packet injection could be interrupted by flits from other input ports. This would incur additional complexity in receiving the packets to determine the completion of multiflit packet transfer. The timer works to limit the injection time of each multiflit packet. When the first flit of a multiflit packet is injected, the timer is turned on. Each of the remaining flits must be injected in one cycle; otherwise, the remaining flits will be discarded. All multiflit packets in the approx-subnet are approximable, and the discarded flits can be recovered in the destination node. In addition, the bufferless router supports the ASW design, and the ASW quickly determines whether the incoming flits are to be transferred or discarded.

ASW. In the bufferless router, each flit contains some header bits for independent routing and port arbitration. In the approx-subnet, flits are transmitted based on the ASW design, which employs XY routing and simplified port allocation. XY routing computes the productive output port of each flit based only on the destination address, which is stored in the header bits as X and Y values, and provides each packet with a unique transmission path; thus, the flits in a multiflit packet are routed independently but without disorder at the destination node.

In the ASW design, arbitration is based only on a default priority and the approximable notation. The default priority follows a rule that a flit traveling straight has a higher priority than a flit that is turning. The injection port has the lowest priority since the flits can be buffered first in the injection port. Therefore, the default priority is north $>$ south $>$ west $>$ east $>$ injection port. The approximable notation is also stored in the header bits as '1' for approximable and '0' for nonapproximable. Moreover, approximable flits have a higher priority than nonapproximable flits since discarding approximable flits may incur errors in the result, while nonapproximable flits dropped in the approx-subnet are still transmitted by the lossless-subnet.

Figure 4 shows the ASW design. The required signals are denoted by $V_{direction}$, $A_{direction}$, $X_{direction}$, and $Y_{direction}$, which correspond to the valid bit ('1' for an incoming flit and '0' for none), the approximable notation, and the destination X and Y values, respectively. The east and west output structure, north and south output structure and ejection port structure are shown in Figure 4(a), (b), and (c), respectively. Flits are transmitted first along the west (if $X_{direction} > X_r$) or east (if $X_{direction} < X_r$) direction and then along the north (if $Y_{direction} < Y_r$) or south (if $Y_{direction} > Y_r$) direction. Approximable flits always have a higher priority in 2-to-1 multiplexers; the default priority is further hardcoded by three 2-to-1 multiplexers in Figure 4(b) and (c). Moreover, Figure 4 shows that any flit traverses at most three 2-to-1 multiplexers from the input port to its productive output port, which reduces the critical path delay. Additionally, the routing, allocation and traversal in the router are combined, and thus, flits pass through each hop in only a single step. Therefore, the approx-subnet achieves single-cycle hop delivery without congestion.

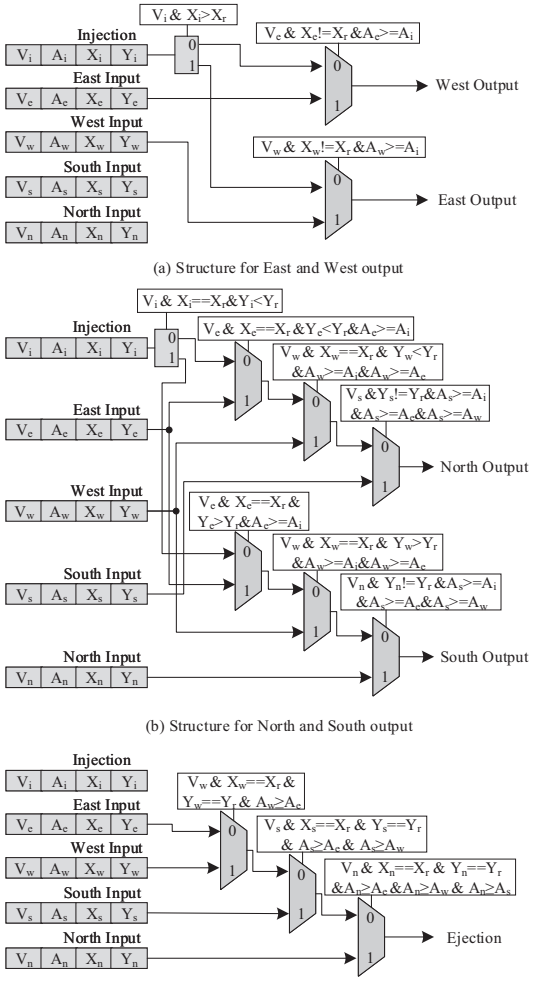
The ASW design greatly simplifies route computation and port arbitration. Packets are transmitted through a single-cycle latency pipeline, and all conflicting flits are dropped instead of creating congestion. The approx-subnet is thus naturally free of deadlocks.

C. Packet Recovery

In previous approximate designs, some annotation frameworks have been proposed that label sections of the approximable data [12], [13]. We manually annotate benchmarks in a fashion similar to these methods. In the AMNoC, a multiflit data packet is annotated as approximable only when the packet stores words of the same type (integer or floating point) and only if those words are all approximable.At the destination node, packet recovery is implemented to approximate the missing data in an approximable data packet. Two stages are required to recover a data packet: (1) determine the missing flits of a data packet and (2) approximate the values in the missing flits.

In the approx-subnet, flits of an approximable data packet are injected (or dropped) within a single-cycle interval. Due to the ASW design, these flits are transmitted in succession. Therefore, at the destination node, flits from the same approximable packet can be easily gathered together. The location of a flit in the packet is also stored in its header bits and transmitted along with the flit. After an approximable packet

*If an incoming flit is successfully allocated to a output port, it will be transferred to its neighbor router or ejection buffer; otherwise, it will be discarded.



(a) Structure for East and West output

(b) Structure for North and South output

(c) Structure for Ejection

Fig. 4. ASW design.

is transmitted†, the missing flits can be determined based on the locations of the received flits.

Previous studies have proposed many value approximation designs, such as last value, stride, FCM [14], VTAGE [15], etc. Reducing the complexity and enhancing the accuracy are the two main challenges of value approximation. In this context, the above mentioned methods are either insufficiently accurate [14] or excessively complex, thereby incurring a high power consumption and a high overhead [15]. We choose linear interpolation due to its low complexity and reasonable accuracy. Linear interpolation requires only one addition to perform value approximation. The data in an approximable packet are generally fetched from successive memory blocks; hence, the data in a packet (*e.g.*, the adjacent pixels in an image) are considerably similar. The data of dropped flits are most relevant to the preceding and following flits. Therefore, the linear interpolation error is small. If the dropped flit is the first flit or last flit, its approximation is to copy the received flit that is the closest to it.

†The transmission of an approximable packet is considered complete when the last flit of the packet reaches the destination, when the maximum waiting time (<packet size) of the first flit received at the destination node is reached, or when the first flit transmitted by the lossless-subnet has been received.

TABLE I
NoC CONFIGURATIONS

	baseline	lossless-subnet	approx-subnet
Channel Width	16 bytes	8 bytes	10 bytes
Virtual Channels	4 × 4	4 × 4	none
Data Packet	5 flits	9 flits	8 flits
Control Packet	1 flit	1 flit	1 flit
Pipeline Stages	3	3	1
Injection/Ejection		16 flits	
Routing Algorithm		XY	
Network Size		8 × 8	

TABLE II
FULL-SYSTEM SIMULATION SYSTEM CONFIGURATIONS

Number of cores	64 (x86 ISA)
L1 D cache (private)	128 KB, 64-byte lines, 4 cycles
L1 I cache (private)	128 KB, 64-byte lines, 4 cycles
L2 cache (shared)	512 KB slice/node, 64-byte lines, 8 cycles
Caching protocol	MESI
DRAM controllers	4 (on nodes 0, 16, 32, and 48)
DRAM size	2 GB, 45 cycles

III. METHODOLOGY

NoC Configurations. We evaluate the effectiveness of the AMNoC in comparison with a baseline single-plane buffered network.

The configurations are listed below:

- **Baseline**: In this configuration, the NoC is a single lossless buffered NoC with 16-byte channels, as in Table I.
- **AMNoC**: This NoC is composed of a lossy bufferless plane (approx-subnet) and a regular buffered plane (lossless-subnet). The approx-subnet has 10-byte channels (8 bytes for data and 2 bytes for header bits), while the lossless-subnet operates with 8-byte channels. The total channel width of the AMNoC is 18 bytes, but it uses 16 bytes for data transfer, which is the same as the baseline NoC. The approximate multiflit data packets and critical single-flit packets are injected into the approx-subnet, while the nonapproximate multiflit data packets and single-flit control packets are carried by the lossless-subnet. Data packets in the buffered NoC must be packaged with a head flit for routing, while a head flit is not required in the bufferless NoC. Thus, the data packet size in the approx-subnet is 1 flit smaller than that in the lossless-subnet.

Full-System Simulation. For a full-system evaluation, we use a modified event-driven many-core simulator [16]. Table II lists the full-system simulation parameters.

Benchmarks. The benchmarks used for evaluating the performance are selected from the Princeton Application Repository for Shared-Memory Computers (PARSEC) [17]. We configure each application with 64 threads and run the benchmarks for 100 million instructions with a small input size.

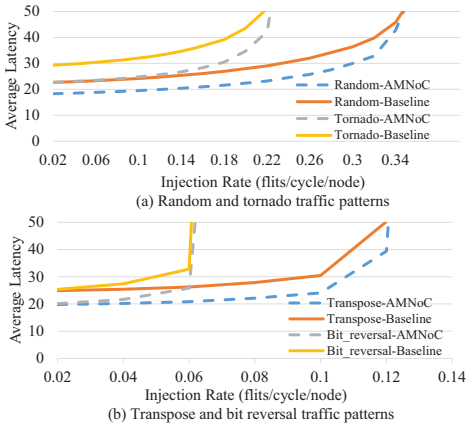


Fig. 5. Average latency curves under synthetic traffic patterns.

IV. Experimental Results

A. Evaluation with Synthetic Traces

Average Latency. We use synthetic traces to evaluate the communication latency with varying traffic loads. We generate traces such that there are 30K data messages in the NoCs. The synthetic traces are then studied with different traffic pattens and different injection rates. In the baseline NoC, each data message contains 5 flits. In the AMNoC, a nonapproximable message is injected into the lossless-subnet as a 9-flit packet, while an approximable message is injected into the approx-subnet as an 8-flit packet, and its first flit is also injected into the lossless-subnet. We randomly select 50% of the data messages to be approximable. Figure 5 shows the average latency for the baseline NoC and AMNoC under different synthetic traffic patterns. The AMNoC shows a significant decrease in average latency compared to the baseline NoC; this is because the approximable messages in the approx-subnet are transmitted with low latency. The AMNoC and baseline NoC saturate around the same injection rate under different traffic patterns; the reason for this phenomenon is that the lossless-subnet and the baseline network have a similar architecture, and the selection of 50% approximable packets generates the same injection rate in both the lossless-subnet and the baseline network.

Dropped Flit Ratio. Some flits are dropped in the AMNoC, and the drop rate will influence the output quality of applications. Here, we evaluate the dropped flit ratio, which is the percentage of dropped flits among all transmitted flits, as shown in Figure 6. The dropped flit ratio is plotted until network saturation is reached. As illustrated, all the dropped flit ratios are less than 14%. This means that most data can be transmitted without approximation.

B. Evaluation with Full-system Analysis

Latency and Speedup. In this section, we evaluate the average packet latency, system runtime and dropped flit ratio on a variety of PARSEC benchmarks. Figure 7 shows the normalized average latency and speedup for the benchmarks

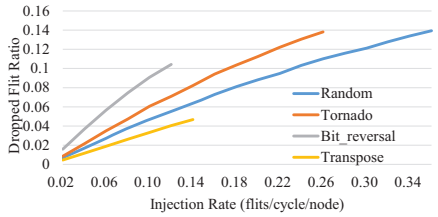


Fig. 6. Dropped flit ratios in the AMNoC.

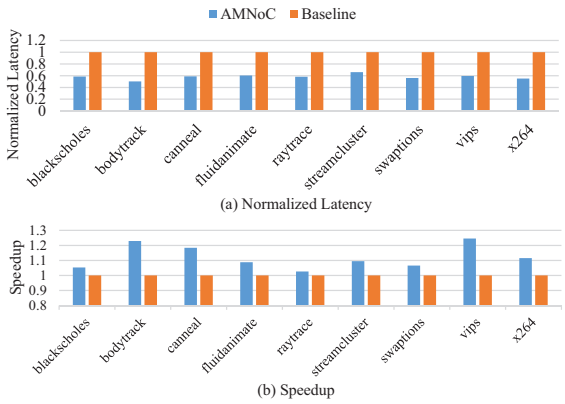


Fig. 7. Network latency and runtime for the full-system analysis.

compared against the baseline NoC with 50% approximable data packets. For all the benchmarks, the AMNoC reduces the latency by an average of 41.9% with respect to the baseline NoC and achieves an average 1.12× speedup.

Sensitivity. Next, varying approximable packet ratios are studied to investigate their impacts on the AMNoC performance. Figure 8 shows the normalized packet latency as the percentage of approximable flits is varied by 25%, 50% and 75%. Evidently, in most applications the average latency decreases as the percentage of approximable packets increases; this relationship is observed because a high approximable packet ratio increases the percentage of packets transmitting in the approx-subnet with low latency. The more approximable packets there are, the greater the improvement in the average latency. For 'streamcluster' 'vips' 'x264', the average latency decreases when approximable flit ratio increases from 25% to 50% but increases when approximable flit ratio increases from 50% to 75%. This is due to that 75% approximable flits brings more workload to the approx-subnet and cause more waiting time before packet injection. When the increased waiting time exceeds the benefits of transmitting time, the packet latency becomes even greater. However, it is clear that the average latency has improved significantly in all experiments, which shows the effectiveness of the AMNoC design.

Approximated Data Ratio. The data that is recovered in destination node is called approximated date. Figure 9 shows the percentage of approximated data under different approximable packet ratios. When the approximable packet ratio is 25%, the percentage of approximated data can be less than 0.1%. For a 50% approximable packet ratio, The

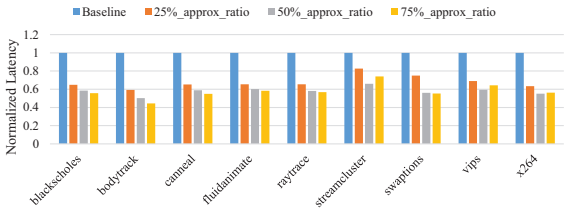


Fig. 8. Approximable packet ratio sensitivity analysis.

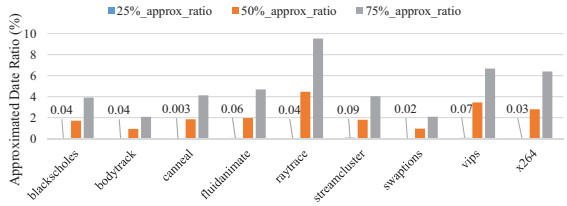


Fig. 9. The percentage of recovered data in benchmarks.

proportion of approximated data in all applications is less than 5%. Even when applications are annotated with 75% approximable data packets, their approximated data ratios can still be less than 10%. These results show that only a very small amount of data needs to be recovered and most data can be transmitted losslessly in AMNoC. This is due to the low dropped flit ratio in the AMNoC.

C. Power and Area

We model the power consumption and area overhead using the Design Space Exploration for Network Tool (DSENT) [18] with a 45 nm complementary metal oxide semiconductor (CMOS). The results are shown in Table III. We use the average injection rate obtained in our full-system simulation across all benchmarks to simulate the dynamic power consumption. The AMNoC achieves 48.6% power savings and reduces the area overhead by 53.4%.

V. Conclusion

In this paper, we develop a lightweight energy-efficient multiplane architecture for an approximate NoC, denoted AMNoC. The AMNoC is equipped with a lossy bufferless NoC and a lossless buffered NoC. The lossy bufferless NoC is designed for simplicity and allows flits to be dropped in the presence of contention. In contrast, the lossy design provides single-cycle hops and no-congestion delivery for approximable packets and latency-sensitive packets, while the buffered NoC guarantees the lossless transmission of nonapproximable packets. Experiments show that the AMNoC reduces the average latency by 41.9% and achieves 48.6% and 53.4% savings in power consumption and area overhead, respectively, compared to a regular buffered NoC while maintaining more than 90% data being transmitted losslessly.

TABLE III
Router power and area comparison

	Total Power	Total Area
AMNoC	19.11 mW	0.053 mm^2
Basline	37.15 mW	0.11 mm^2

References

[1] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 62, 2016.

[2] A. Raha, S. Venkataramani, V. Raghunathan, and A. Raghunathan, "Quality configurable reduce-and-rank for energy efficient approximate computing," in *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*. EDA Consortium, 2015, pp. 665–670.

[3] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar, "A 5-ghz mesh interconnect for a teraflops processor," *IEEE Micro*, vol. 27, no. 5, pp. 51–61, 2007.

[4] Z. Li, J. San Miguel, and N. E. Jerger, "The runahead network-on-chip," in *High Performance Computer Architecture (HPCA), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 333–344.

[5] L. Wang, X. Wang, and Y. Wang, "Abdtr: Approximation-based dynamic traffic regulation for networks-on-chip systems," in *Computer Design (ICCD), 2017 IEEE International Conference on*. IEEE, 2017, pp. 153–160.

[6] R. Boyapati, J. Huang, P. Majumder, K. H. Yum, and E. J. Kim, "Approx-noc: A data approximation framework for network-on-chip architectures," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ser. ISCA '17. New York, NY, USA: ACM, 2017, pp. 666–677. [Online]. Available: http://doi.acm.org/10.1145/3079856.3080241

[7] V. K. Rajanna and M. Alioto, "Low-swing links with dynamic energy-quality trade-off for error-resilient applications," in *2019 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2019, pp. 1–4.

[8] G. Ascia, V. Catania, S. Monteleone, M. Palesi, D. Patti, and J. Jose, "Improving energy consumption of noc based architectures through approximate communication," in *2018 7th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 2018, pp. 1–4.

[9] A. K. Abousamra, R. G. Melhem, and A. K. Jones, "Deja vu switching for multiplane nocs," in *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on*. IEEE, 2012, pp. 11–18.

[10] M. Hayenga, N. E. Jerger, and M. Lipasti, "Scarab: A single cycle adaptive routing and bufferless network," in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, 2009, pp. 244–254.

[11] C. Fallin, C. Craik, and O. Mutlu, "Chipper: A low-complexity bufferless deflection router," in *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*. IEEE, 2011, pp. 144–155.

[12] J. S. Miguel, M. Badr, and N. E. Jerger, "Load value approximation," in *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2014, pp. 127–139.

[13] A. Sampson, W. Dietl, E. Fortuna, D. Gnanapragasam, L. Ceze, and D. Grossman, "Enerj: Approximate data types for safe and general low-power computation," *SIGPLAN Not.*, vol. 46, no. 6, pp. 164–174, 2011.

[14] B. Goeman, H. Vandierendonck, and K. De Bosschere, "Differential fcm: Increasing value prediction accuracy by improving table usage efficiency," in *High-Performance Computer Architecture, 2001. HPCA. The Seventh International Symposium on*. IEEE, 2001, pp. 207–216.

[15] A. Perais and A. Seznec, "Practical data value speculation for future high-end processors," in *High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium on*. IEEE, 2014, pp. 428–439.

[16] X. Wang, M. Yang, Y. Jiang, P. Liu, M. Daneshtalab, M. Palesi, and T. Mak, "On self-tuning networks-on-chip for dynamic network-flow dominance adaptation," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 13, no. 2s, p. 73, 2014.

[17] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The parsec benchmark suite: Characterization and architectural implications," in *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '08. New York, NY, USA: ACM, 2008, pp. 72–81.

[18] C. Sun, C.-H. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *2012 IEEE/ACM Sixth International Symposium on Networks-on-Chip*. IEEE, 2012, pp. 201–210.