FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 222-227 | e-ISSN: 2347-470X

Communication Latency and Power Consumption Consequence in Multi-Core Architectures and Improvement Methods

Venkata Sridhar .T[1,2], G. Chenchu Krishnaiah[3]
[1]*Department of E&C, VTU, Belgaum, India, venkatasridhar.thatiparthi@gmail.com*
[2]*Department of ETC, IIIT-Bhubaneswar, Bhubaneswar, India, venkatasridhar@iiit-bh.ac.in*
[3]*Department of ECE, ASCET, Gudur, India, gurramchenchukrishnaiah@gmail.com*

***Correspondence:** Venkata Sridhar .T; venkatasridhar.thatiparthi@gmail.com, venkatasridhar@iiit-bh.ac.in

ABSTRACT- The present electronics world has a lot of dependency on processing devices in the current and future developments. Even non-electronic industries have much data to process and are indirectly dependent on processors. The larger the number of processors incorporated into the architecture, will lower the data handling and processing time; thus, efficiency improves. Hence multi-core processors have become a regular part of the design of processing elements in the electronic industry. The large number of processors incorporated into the system architecture results in difficulty in communicating among them without a deadlock or live lock. NoC is a promising solution for communicating among the on-chip processors, provided it is fast enough and consumes less energy. Further, the latency among the multi-core processors should be optimal to stand with the increasing data acquisition and processing in the new/developing operating systems and software. This paper addresses energy efficiency and latency reduction methods/techniques for Multi-core architectures.

Keywords: Latency, Energy efficient, NoC (Network on Chip), NoC Router, Multi-core SoCs (System on Chips).

ARTICLE INFORMATION
Authors: Venkata Sridhar. T and G. Chenchu Krishnaiah;
Received: 25/01/2023; **Accepted**: 20/03/2023; **Published**: 30/03/2023;
E- ISSN: 2347-470X;
Paper Id: IJEER230104;
Citation: 10.37391/IJEER.110130
Webpage-link:
www.ijeer.forexjournal.co.in/archive/volume-11/ijeer-110130.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

The advancement in chip integration technologies has set high upper limits on the number of cores attached to a single processing chip [1]. Further, the user eagerness to have fast processing outputs is an urge in some domestic or security devices; integrating many cores into a single chip is common in present processor design and chip manufacturing. Provided there are certain obstacles to this process, which are 1) communication latency between the cores, 2) the overall power consumption that the connecting devices (Network on Chips) [2] will consume, and the performance is dependent on these [3].

Figure 1 shows the basic Multi-core architecture of a SoC with expanded tile up to connecting router level. Many modern processing chips consist of on-chip integrated elements like processing core, memory, cache, media processing elements, signal processing elements and more. The NoC is the best connecting technology among all of them on-chip. As stated earlier, the significant factors that affect the performance of NoC, which affects the total system (SoC/CMP) performance, are latency and power consumption.

In general, for a NoC router, there are several stages of pipelining [5, 21].

(i) Routing calculations: computes the packet traversal from the source to the destination.
(ii) Allocation of Virtual channel: assigns a buffer to the virtual channel to the early possible router connecting the destination. Arbitration is done when many requests appear for the same virtual channel from different header flits.
(iii) Allocation of switch: after successful virtual channel allocation, the next stage is to allocate the switch to the destination port.
(iv) Switch traversal: after switch allocation, the flit will travel to its destination port via the crossbar.

All the stages are crucial as they define the performance of the NoC router [4]. Even one stage delay can have a significant remark on the whole system. Hence careful design procedures should be followed to adequately control the latency [6, 22] and power consumption in the overall system.

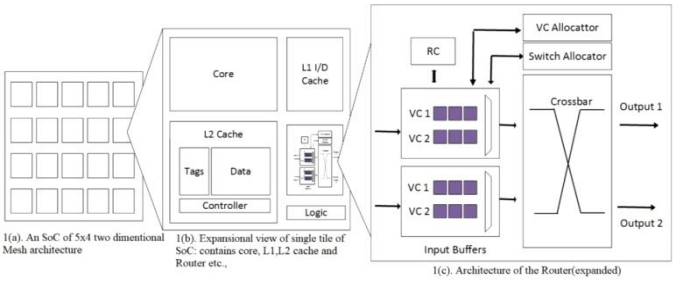

Figure 1: Expanded view of SoC with 5×4 Mesh Architecture

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 222-227 | e-ISSN: 2347-470X

The equation for the latency (L) of a networking packet is defined [7] as in equation (1).

$$L = K.t_r + K.t_w + Dc + P/b \quad (1)$$

K is the hop count, delay in the pipeline of the router is t_r, between router delay is t_w, Dc is the router contention delay, and P/b is the router's stabilization delay.

It is further assumed that a packet that consists of L flits. If L flits are to through H hops are considered in a packet, no-load communication latency, T, is computed on a wormhole Network on chip as in the *equation (2)*.

$$T = T_{lt} \times (H + 1) + T_{router} \times H + L \quad (2)$$

Whereas the latency of the router is T_{router} and the latency of the link transfer is T_{lt}. The general approach to minimize the latency is to reduce the T_{router}, and H. H depends on the network topology diameter and shortest path length average. And both hop count dependent. It is required to do sophisticated research on the router architecture to minimize the T_{router} value.

In multi-core architectures, the higher the shared memory usage is, the higher its power consumption level. High power consumption of chips can cause unstable thermal properties of the system and can cause performance degradation. In general, multiple threads running on different cores don't require the same power all the time. The following are a few suggestible techniques for the architecture parameter developer to achieve multi-core power management.

(i) DVSF (*ii*) Asymmetric cores (*iii*) Variable sized cores (*iv*) Thread motion (*v*) Speculation control and (*vi*) Reconfigurable architectures.

Section 2 surveys the literature on latency and power consumption in multi-core architectures. *Section 3* discusses the suggested remedies to achieve low latency and low power consumption; *Section 4* discusses the results of the proposed models. Finally, *Section 5* gives a conclusion with future scope.

2. LITERATURE SURVEY

The following are a few existing methods *t* reduce the latency and power consumption on multi-core architectures.

2.1 Communication Latency

Mullins et al. [8] discussed a technique of pre-computing arbitration that aims to reduce the delay of critical paths of separable input-first Virtual Channel (VC) allocator. This method discarded the SA stage from the critical path; this way, it is not good enough to handle traffic of non-congested routing, resulting in an ambiguity among the new flits stuck in the unutilized time of the crossbar [24].

Kumar et al. [9] proposed express virtual channels (EVCs). This method skips many routers on the paths that lead to long destinations because of the addition of EVCs, claiming the overall latency reduction. But this method should compromise the large silicon consumption and is further unsuitable for near destinations.

A DPBFP (Dynamic-Priority-Based-Fast-Path) NoC router [10] bypasses SA level through an arbitration request for frequently used paths with flit priority. Arbitration-request will be a one-clock ahead of the next router. But, this method wants more hardware frequency analyzer of the path, etc.

To minimize the routing computations, a look_ahead routing method was introduced [11]. The header flit is attached to the output port by pre-computing one router in advance. The router of NoC conveys to the output port the future flit information at the time of arrival of the header flit in parallel with the computation of the subsequent routing. This method requires more area and hence, consumes power. Vinoda reddy et al. [23] proposed QoS-based clustering for efficient routing to minimize the communication latency, but this method is suitable to single-cluster communication only.

The above-stated literature shows an excellent scope to investigate new methods in this area. Mullins's approach suffers when network congestion occurs, and Kumar's technique uses a large die area, causing more power consumption. DPBFP is more complex in hardware.

2.2 Chip Power Consumption

Weiser et al. [12,13,14] first implemented DVFS on microprocessor power management. In DVFS governing equation of power consumption is as follows.

$P = CV^2F$, where *F* is the frequency, *C* is the switching capacitance, and *V* is the supply voltage. Thus, *V* and *F* variable parameters can control the total power consumption. This method works fine in the case of variable energy per instruction ratio. Leading developers like Intel; AMD has used it in their 'SpeedStep' and 'PowerNow', respectively. A Set of predefined *V* and *F* values are defined to achieve desired power levels, which are suitable for many cores and larger parallel data loads. But the significant compromise is it affects the system performance as frequency downscales.

Efthymiou et al. [15] proposed non-fixed-size cores. The flexible cores can degrade into smaller ones after a big complex job by continuously disabling the pipeline stages and execution stages. It is like a power gating technique that switches off unused or idle resources. This method is good at some low-core operated scalar loads with no parallelism. But, when higher parallel loads are needed, this method will degrade the system performance somewhat. Intelligent power optimization techniques [25] are greatly appreciated.

I. Engin et al. [16] proposed a convergence technique of fusing simple small cores into a large core on demand dynamically, called Core Fusion, a reconfigurable technique. This proposal is independent of large programming compilers and specific programming. Nagabushanam et al. [25] proposed a power and area optimizing method for VLSI implementations suitable to fewer cores implementation.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 222-227 | e-ISSN: 2347-470X

Core Fusion can work with different workloads and software diversities. When a more extensive parallel data set is processed, it distributes into smaller cores, and if the workload is a larger scalar, then cores fuse into a single powerful large core. In this method, the downfall side is challenging to achieve fusion power and high energy density.

DVFS suffers system performance at lower data rates, and Efthymiou's method performance degrades at higher data rates. Similarly, Engin's method had significant hardware complexity. Hence, an enormous scope is open in power optimization for multi-core architectures.

3. PROPOSED METHODS

After finding the consequences due to communication latency and large power consumptions in the multi-core designs, a few methods are introduced in this section to control them.

3.1 Low Power Techniques

The total consumption power by the chip and its associated connections will cause many undesirable outcomes in the chip manufacturing process. And the same is crucial in multi-core architectures when operating them in real [17]. Controlling or saving the processing element's idle/waiting time while executing different threads is essential. Because all the threads will not consume the same amount of power since different processing threads will have different execution times. A dynamic monitoring and controlling system is required to manage power wastage for both predictive and reactive categories [18].

It is essential to have every multi-core architecture, like SoCs and NoCs, an on-chip built-in control module for efficient power management [19].

This hardware-associated firmware will coordinate with the software and fulfil its job, as shown in *figure 2* and *figure 3*. A proper selection of frequency and voltage is necessary for estimating the power budget, as shown in *figure 4*. And *equation (3)* represents the required frequency for the same. In contrast, F_{req} is the required frequency, $F_{hig.}$ is frequency high, F_{low} is frequency low, $O_{hig.}$is occurrence high, and O_{low} is occurrence low.

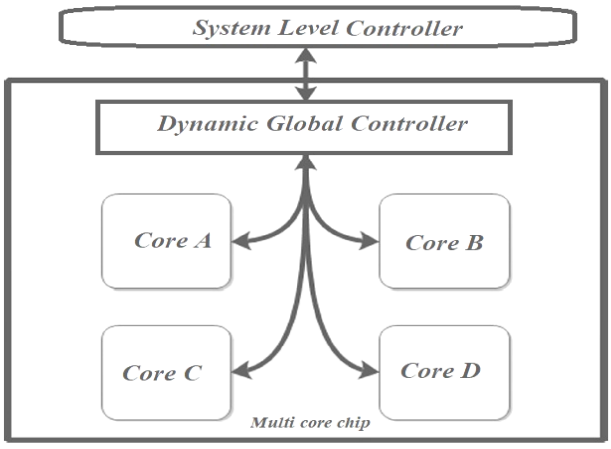

Figure 2: Top-Level Observation of Dynamic Power Managing

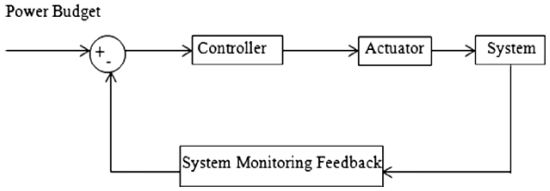

Figure 3: Power Managing System with Closed-Loop Feedback

This paper proposes a congregated dynamic voltage and frequency scaling (CDVFS) with a dynamically configurable global controller to optimize power consumption.

$$F_{req} = (F_{hig.} * O_{hig.}) + (F_{low} * O_{low}) / (O_{hig} + O_{low}) \quad (3)$$

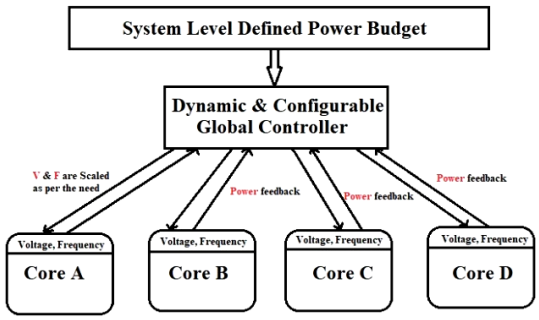

Figure 4: CDVFS High-Level View of Quad Core Processor

3.2 Low Latency Techniques

In this work, we propose a VPLR router with aligned virtual channels (VCs) side by side. The variable priority look ahead router (VPLR) is independent of computing the individual channel requests [23], and they are pre-computed based on its traffic and pre-holding the priorities. In this method, if SA grants any request approval, that particular flits directly pass to the output port.

This method makes traffic monitoring dynamic, and priorities are resolved accordingly. Further, the masked requests are handled not to burden the router. The proposed method is shown in *figure 5*. The proposed method is a composition of Look-ahead routing (LAR) and prioritized output virtual channel (POVC) compositions for input flit request prioritization.

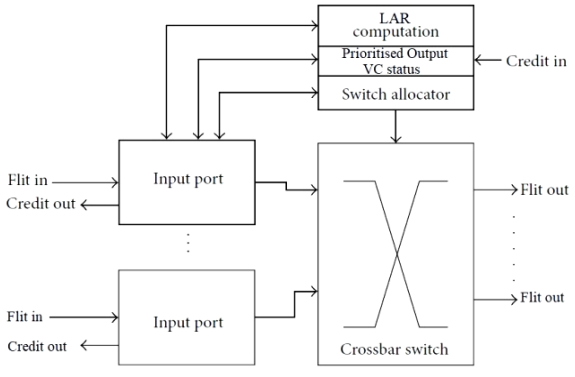

Figure 5: Working Diagram of Proposed VPLR for NoC

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 222-227 | e-ISSN: 2347-470X

4. RESULTS AND DISCUSSION

4.1 Low Power

The input flits are first mapped in the buffer (extra memory) for prioritization. The dynamic power management was implemented on the first-level cache for efficient usage, as shown in *figure 6* below.

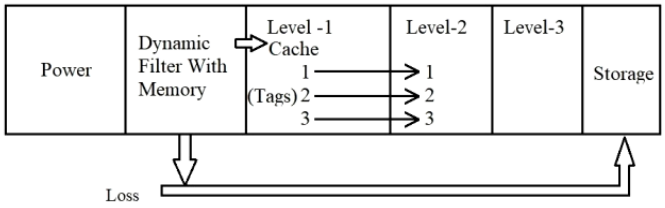

Figure 6: Power Management at First-level Cache

This method processes the power consumption by remembering all the cache levels, and any miss happened is repeated. It does not wait for the tag match and skips it as it is pre-stored. Thus the power efficiency will improve by 47%. Further, the simulation results clearly show that this system can defeat most of the misses by not waiting for tag-matching, and the hit ratio is more than 90%.

4.2 Low Latency

For the evolution of our work, we compared the proposed VPLR with the existing router MCONNECT [20]. The result analysis discussed in this section for low latency is the work done on the Field Programmable Gate Array boards CYCLONE-V-SX-C6 and STRATIX-V-EB using four VCs and four flits per one VC. *Figure 7* shows the logic cost utilization and comparison with respect to VCs of two, four and eight, with five ports, thirty-two payloads and a buffer width of four.

This shows that even at higher data rates and LC utilization, the ratio is the same. Thus the effective power consumption is minimal.

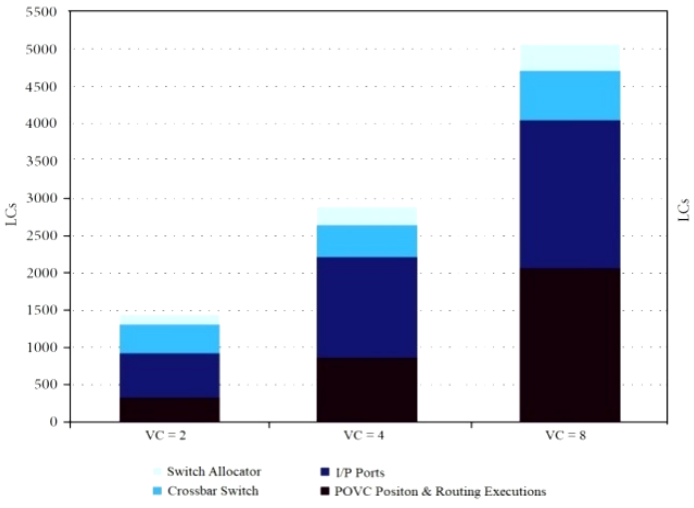

Figure 7: Logic Budget Estimation for 2, 4, and 8 VCs

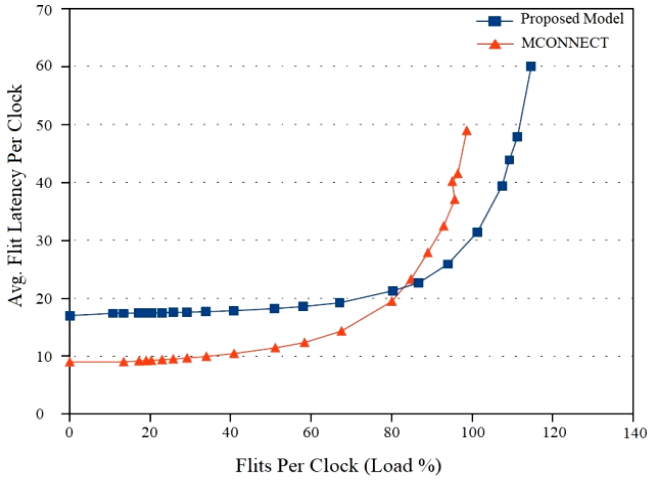

Figure 8(a): Proposed Model and MCONNECT, Load Performance under URT

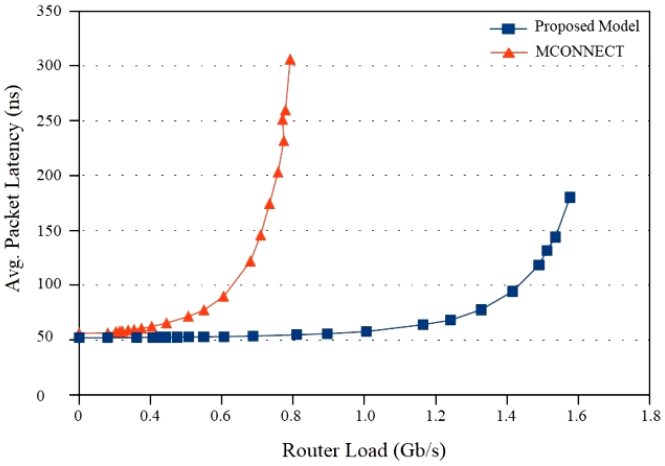

Figure 8(b): Load Delay Execution of Proposed Model (120MHz) w.r.t MCONNECT (85MHz)

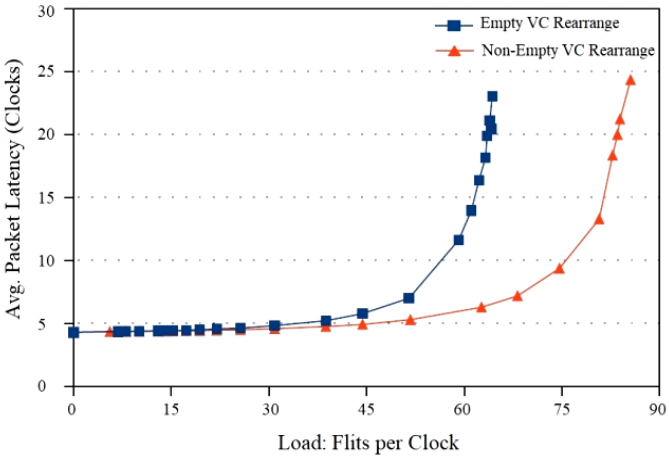

Figure 8(c): Router Performance for Empty VC w.r.t Non-Empty VC Re-arrangement under URT

Figure 8(a) represents the load delay execution between the proposed model and the MCONNECT under Uniform Random Traffic (URT) with respect to clock cycles. *Figure 8(b)* shows

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 222-227 | e-ISSN: 2347-470X

the load delay execution for the proposed model and MCONNECT at 120MHz and 85MHz, respectively, under URT. *Figure 8(c)* shows the performance of rearrangement of non-empty VCs (2 VCs under URT). This shows average packet latency with no load and full load conditions. The proposed model achieved 15% less latency at full load under uniform random traffic. And more than 45% less latency on full load at variable traffic.

The *table-1* shows the utilization of Hardware (HW) resources utilisation of the proposed router implemented on CYCLONE-V-SX-C6 compared with MCONNECT for both 1-Clk and 2-Clks. And, the *table-2* shows the utilization of HW resources of the proposed router implemented on STRATIX-V-EB in compared with MCONNECT for both 1-Clk and 2-Clks.

The proposed method of a dynamic filter with memory at first level cache for power optimization reduced the buffer area by 54% and power by 47.46%.

Table 1. Utilization of HW-resources of Proposed Model with MCONNECT on CYCLONE-V-SX-C6

Board → (4 Flits per VC : 4 VCs)	CYCLONE-V-SX-C6		
	Proposed	MCONNECT (1-clk)	MCONNECT (2-clks)
Max. Frequency	120MHz	66MHz	85MHz
Memory blocks	6(1.3%)	NA	NA
Logic cells	2710(2.3%)	5710(5%)	5980(5.3%)

Table 2. Utilization of HW-resources of Proposed Model with MCONNECT on STRATIX-V-EB

Board → (4 Flits per VC : 4 VCs)	STRATIX-V-EB		
	Proposed	MCONNECT (1-clk)	MCONNECT (2-clks)
Max. Frequency	180MHz	95MHz	150MHz
Memory blocks	6(0.65%)	NA	NA
Logic cells	2610(1.4%)	5671(3.1%)	5521(3%)

Table 3. Power expenditure of Proposed Model with existing model

Router Size	Dynamic power(µW)		Static power(µW)		Total power(µW)	
	Existing method	Proposed method	Existing method	Proposed method	Existing method	Proposed method
2×2	44.75	23.47	0.110	0.057	44.86	23.52
4×4	202.27	106.19	1.871	0.969	204.14	107.16
8×8	818.47	430.51	7.521	3.936	825.99	434.46

The proposed model was implemented on Xilinx Vivado ML2021.1. The power expenditure of proposed and existing methods is compared at 2×2, 4×4 and 8×8 folded torus topology. It is found that the proposed method saved a total power up to 47%.

5. ACKNOWLEDGMENTS

I thank my working place, IIIT-Bhubaneswar, which provided me with the flexibility and facilities to conduct my research as an Assistant Professor of the ETC department. Further, I thank my research bringing university VTU, Belgaum, for permitting me as a researcher. I specially acknowledge my supervisor Dr G. Chenchu Krishnaiah, and the doctoral committee for giving me timely suggestions.

Conflicts of Interest: We declare that we do not have any funding for this research in full or partial.

6. CONCLUSION

The proposed power efficiency and latency reduction models are working well and are suitable to fit the multicore architectures. The discussion of chapter-4, a and b sections shows that the proposed model has improved power optimization and latency reduction. This research has certain limitations; as such, it dynamically occupies a bit more die area depending on the traffic density in a few cases. This needs to be further investigated as a future expansion. Additionally, the utilization of resources for both CYCLONE-V-SX-C6 and STRATIX-V-EB implementations shows that the proposed model has achieved a maximum of 47.46% optimized to the existing model. Proposed low-power methods have a limitation of complex prioritization if all the router ports have a full load and additional buffers cause an increase in latency. Hence more scope is there for further investigations for more power and latency optimization.

REFERENCES

[1] Gaha, H.I., Balti, M. 2022. Novel Bi-UWB on-Chip Antenna for Wireless NoC, Micromachines, *MDPI AG, Vol. 13 No. 2*, p. 231.

[2] Joardar,B. K., Kim R. G., Doppa, J. R., Pande, P. P. D. Marculescu and R. Marculescu, 2019. Learning-Based Application-Agnostic 3D NoC Design for Heterogeneous Manycore Systems, *IEEE Trans., Computers, , Vol. 68, no. 6*, pp. 852-866.

[3] Zou, K., Wang, Y., Cheng, L., Qu, S., Li, H., Li X., 2022. CAP: Communication-Aware Automated Parallelization for Deep Learning Inference on CMP Architectures, *IEEE Trans., Computers, Vol. 71, no. 7*, pp. 1626-1639.

[4] Psarras, A., Moisidis, S., Nicopoulos, C., Dimitrakopoulos, G., 2017. Networks-on-Chip with Double-Data-Rate Links, *IEEE Trans. Circuits Syst I Regul Pap, Vol. 64, no. 12*, pp. 3103-3114.

[5] Kumar, S., Jantsch, A., Soininen J.P., Forsell, M., Millberg, M., Oberg, J, Kari, T., Hemani, A., 2002. Network on chip architecture and design methodology, in Proceedings of the IEEE Computer Society Annual Symposium on VLSI, Pittsburgh, Pa, USA, pp. 105– 112.

[6] Chen P., Liu, W., Hui, C. Li, S., Li, M., Yang, L., Nan, G., 2021. Reduced Worst-Case Communication Latency Using Single-Cycle Multihop Traversal Network-on-Chip, IEEE Trans., Computer-Aided Design of Integrated Circuits and Systems, *Vol. 40, no. 7*, pp. 1381-1394.

[7] Martin, S., Florian, B., Jens, S., Kaspaki, E., 2012. A Statically Scheduled Time-Division-Multiplexed Network-on-Chip for Real-Time Systems, in Proceedings of *IEEE/ACM Sixth International Symposium on Networks-on-Chip*, Lyngby, Denmark, pp. 152-160, doi: 10.1109/NOCS.2012.25.

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 11, Issue 1 | Pages 222-227 | e-ISSN: 2347-470X

[8] Mullins,R., West, A., Moore, S., 2004. Low-latency virtual channel routers for on-chip networks, in Proceedings of the 31st Annual International Symposium on Computer Architecture, ACM, vol. 32, pp. 188–197.

[9] Kumar, A., Peh, L. S., Kundu, P., Jha, N. K., Express virtual channels: towards the ideal interconnection fabric, in Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA '07), ACM, June 2007., pp. 150–161.

[10] Park, D., Das, R., Nicopoulos, C., Kim, J., Vijayakrishnan, N., Iyer, R., Das, C.R., 2007. Design of a dynamic priority-based fast path architecture for on-chip interconnects, in Proceedings of the 15th Annual IEEE Symposium on High- Performance Interconnects, pp. 15–20.

[11] Galles, M., Spider: a high-speed network nterconnect, *IEEE Micro*, 1997. *Vol. 17, no. 1*, pp. 34–39.

[12] Weiser, M., Brent, W., Alan, D., Scott, S., Scheduling for reduced CPU energy. *Mobile computing. US: Springer*, 1996, P 449-71.

[13] Chao, S.J., Hyun,Y.S., Wook, J.J., 2015. A powersaving DVFS algorithm based on operational intensity for embedded systems. *IEICE Electr Exp*.

[14] Zheng, L., Ren, S., Quan, G., 2015. Energy minimization for reliability-guaranteed real-time applications using DVFS and checkpointing techniques. *J Syst Architec.*

[15] Aristides, E., Jim, D. G., 2002. Adaptive pipeline depth control for processor power-management. IEEE international conference on computer design: VLSI in computers and processors.

[16] I.Engin et al. 2017. Dynamic power management technique in Multi-core Architectures: A survey study, *Ain Shams Engineering Journal, Vol. 8, Issue-3*., pp 445-456.

[17] Li, H., Tian, Z., Xu, J., Rafel, K.V.M., Wang, Z., Zhifei, W., et al, 2020. Chip-Specific Power Delivery and Consumption Co-Management for Process-Variation-Aware Manycore Systems Using Reinforcement Learning, *IEEE Trans., on Very Large Scale Integration (VLSI) Systems*, *Vol. 28, no. 5*, pp. 1150-1163.

[18] Bircher, W.L., John, L., 2012. Predictive power management for multi-core processors. In: Computer architecture. Springer, Berlin, Heidelberg.

[19] You, Y., Chang, Y., Wu, W., Guo, B., Luo, H., Liu, X., Liu, B., Jhao, K., He, S., Li, L., Guo, D., 2022. New paradigm of FPGA-based computational intelligence from surveying the implementation of DNN accelerators. *DAES, Vol. 26, Issue 1*, pp 1–27., doi.org/10.1007/s10617-021-09256-8

[20] Shaheen, M., Fahmy, H., Mostafa, H., 2018. Modified CONNECT: New Bufferless Router for NoC-Based FPGAs, IEEE 61st International Midwest Symposium on Circuits and Systems (*MWSCAS*), pp. 424-427.

[21] Venkata Sridhar, T., Krishnaiah, G.C., 2022. Integrated Sensory Throughput and Traffic-Aware Arbiter for High Productive Multicore Architectures, *Journal of Sensors, Vol. 2022*, pp. 1–14, doi: 10.1155/2022/2911777.

[22] Sridhar, T.V., Krishnaiah, G.C., 2022. Multicore Architectures and Their Applications in Image Processing, in *High-Performance Medical Image Processing*, Sanjay Saxena, Sudip Paul, 1st Edn., Apple Academic Press, New York, pp. 63–93, doi: 10.1201/9781003190011-4.

[23] G. Vinoda Reddy, Kavitha Thandapani, N. C. Sendhilkumar, C. Senthilkumar, S. V. Hemanth, S. Manthandi Periannasamy and D. Hemanand, 2022. Optimizing QoS-Based Clustering Using a Multi-Hop with Single Cluster Communication for Efficient Packet Routing. IJEER, Vol. 10(2), pp. 69-73. DOI: 10.37391/IJEER.100203.

[24] Jenila and R. Aroul Canessane, 2022. Cross Layer Based Dynamic Traffic Scheduling Algorithm for Wireless Multimedia Sensor Network. IJEER Vol. 10(2), pp.399-404. DOI: 10.37391/IJEER.100256.

[25] M Nagabushanam, Skandan S, Rushita M, Sushmitha S Kumar and Swathi K, 2022. Optimization of Power and Area Using VLSI Implementation of MAC Unit Based on Additive Multiply Module. IJEER, Vol. 10(4), pp. 1099-1106. DOI: 10.37391/IJEER.100455.