Literature Survey Report

Project Topic: ML/DL Based PAT Prediction of 3D NoC

Group Number: 2

Abhijith C
Anand M K

I. Introduction

As multicore systems become more complex, efficient on-chip communication becomes increasingly important. Network-on-chip (NoC) architectures have appeared to be a better alternative to traditional communication systems within the chip. NoCs allow multiple cores and components on a chip to communicate more effectively and at higher speeds.

3D NoC improves performance and reduces latency based on stacking processing elements (PEs). However, the power density of 3D NOC is high due to the large number of PE, which leads to various thermal issues. The thermal issues are one of the reasons for the increase in latency, which leads to performance degradation. Power and thermal issues are closely related, and power consumption will increase along with thermal issues. Efficient area management is another critical design challenge of NoC. Due to the stacking of significant NoC components, the chip's overall size and cost increase. Therefore, the power, area, and thermal (PAT) management of NoC is essential. Machine Learning (ML) and Deep Learning (DL) technologies are widely used to address these issues.

II. Literature Review

This section reviews key Machine Learning (ML) and Deep Learning (DL) methods that address power, area, and thermal (PAT) issues.

The thermal issues of NoC can be addressed in different ways, one of which is by optimizing the routing algorithm. In [1], A Q-learning-based thermal-aware routing algorithm is proposed, which utilizes a Q-table consisting of thermal information to manage routing decisions. Routers select output channels based on Q-values, choosing paths with lower temperatures to optimize thermal distribution. The proposed method improves thermal distribution by approximately 28% and 13% compared to TAAR (Topology Aware Adaptive Routing) [2] and PTB3R (Proactive Thermal Budget-based Beltway Routing algorithm) [3], respectively. Furthermore, the proposed method achieves a 32% improvement in average latency and decreases the number of thermal hotspots by 38% and 54% compared to TAAR [2] and PTB3R [3].Overall, The Q-Thermal method effectively optimizes routing decisions in 3-D NoCs. The approach leads to better thermal balance, reduced hotspots, and improved network performance compared to previous methods. However, the main limitation of the method is that it increases the area and power consumption compared to previous routing techniques. The layout area is increased by 7% and 11% compared to TAAR [2] and PTB3R [3]. The power consumption of the proposed method is 2% higher than that of TAAR [2] and 4% higher than that of PTB3R [3].

Another Q-Learning-based routing algorithm is proposed in TTQR: A Traffic- and Thermal-Aware Q-Routing for 3D Network-on-Chip [4] that uses two Q-tables: one table maintains local traffic status information, while the second table holds global thermal information about the network. The proposed method improves latency by an average of 63.6% and throughput by 41.4% compared to TAAR [2]. Overall, TTQR provides a more uniform temperature distribution across layers. However, TTQR has a higher average temperature compared to TAAR.

Another method of addressing thermal issues on NoC is optimizing the design techniques. Dynamic thermal management (DTM) [5] is an important technique that requires accurate thermal information from thermal sensors. Due to the high hardware cost, limited thermal sensors are available, making thermal sensor allocation an important design challenge. A nearest-neighbor-based initialization algorithm is proposed in [6] to allocate thermal sensors, and a Genetic Algorithm (GA) is used to optimize the initial allocation. The method uses an artificial neural network (ANN) to estimate the temperature of nonsensor-allocated nodes. The proposed method reduces the average temperature error by 17.60%–88.63% and the maximum temperature error by 26.97%–85.92% compared with other state-of-the-art methods [7], [8], [9]. The proposed nearest-neighbor-based thermal sensor allocation method effectively places sensors based on spatial thermal correlation. The use of an artificial neural network (ANN) for temperature reconstruction allows for accurate estimation of temperatures in nonsensor-allocated nodes. However, the method assumes that spatial thermal correlations among cores remain constant across different applications, which may not be valid in all scenarios, potentially impacting temperature reconstruction accuracy.

Proactive Dynamic Thermal Management (PDTM) [10] is another temperature control technique that highly depends on the accuracy of the temperature prediction model. A Long Short-Term Memory (LSTM)-based model for temperature prediction is proposed in [11]. The proposed method improves temperature prediction accuracy by 41.92% to 73.63% compared to the traditional ARMA (Autoregressive Moving Average) model [12]. Additionally, the model can quickly locate new hotspots within 0.075 ms. However, this study is conducted on an 8×8×4 3D NoC system, but it is unclear how well the model scales to larger systems.

A neural network-based mapping technique proposed in [13] optimizes temperature distribution by mapping NN layers to appropriate nodes of NoC based on their computational loads. The layer with the highest load is mapped onto dies closest to the heat sink, which optimizes temperature distribution. The model is tested with different neural networks and reduces average temperature. The temperature distribution across the NoC is more uniform, leading to improved thermal management. However, the proposed approach primarily focuses on offline inference scenarios, which lack consideration for dynamic scenarios.

Power and area are also important factors for NoC. A Graph Neural Network (GNN) Framework is proposed in [14] for predicting the power, performance, and area (PPA) of Network-on-Chips. The method models NoCs as attributed graphs and uses GNNs to learn patterns that affect PPA, such as traffic patterns and congestion. The proposed method provides power prediction accuracy of 97.36% and area prediction accuracy of 97.83%. However, the proposed method demonstrates effective performance only up to a certain number of cores, and the model may struggle in larger systems.

TABLE I
*

Summary of Literature Survey

S.No	Title	Author	Year	Issue Addressed	Approach	Performance Metrics	Results	Observations
1	Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network On-Chips	N. Shahabinejad, H. Beitollahi	2020	Thermal Distribution	Q-Learning Algorithm	Thermal Balance, Latency	28% improvement in thermal distribution	Effective optimization for routing decisions
2	TTQR: A Traffic- and Thermal-Aware Q-Routing for 3D Network-on-Chip	Liu et al.	2022	Latency and Throughput	Two Q-tables for local traffic	Latency, Throughput	63.6% improvement in latency	Uniform temperature distribution across layers
3	A Nearest-Neighbor-Based Thermal Sensor Allocation and Temperature Reconstruction Method	M. Guo et al.	2022	Thermal Sensor Allocation	Nearest-Neighbor + ANN	Average and Maximum Temperature Error	17.60%–88.63% reduction in average error	Effective spatial thermal correlation
4	LSTM-based Temperature Prediction and Hotspot Tracking for Thermal-aware 3D NoC System	T. Cheng et al.	2021	Temperature Prediction	Long Short-Term Memory	Prediction Accuracy	41.92% to 73.63% improvement in accuracy	Quickly locates hotspots
5	TTNNM: Thermal- and Traffic-Aware Neural Network Mapping on 3D-NoC-based Accelerator	Xinyi Li et al.	2024	Temperature Distribution	Neural Network Mapping	Average Temperature	More uniform temperature distribution	Effective for offline scenarios
6	NoCeption: A Fast PPA Prediction Framework for Network-on-Chips Using Graph Neural Network	F. Li et al.	2022	Power, Performance, and Area	Graph Neural Network Framework	Power and Area Prediction Accuracy	97.36% power accuracy, 97.83% area accuracy	Performance limited to a certain number of cores
7	Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network On-Chips	N. Shahabinejad, H. Beitollahi	2020	Thermal Distribution	Q-Learning Algorithm	Thermal Balance, Latency	28% improvement in thermal distribution	Effective optimization for routing decisions
8	Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network On-Chips	N. Shahabinejad, H. Beitollahi	2020	Thermal Distribution	Q-Learning Algorithm	Thermal Balance, Latency	28% improvement in thermal distribution	Effective optimization for routing decisions
9	Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network On-Chips	N. Shahabinejad, H. Beitollahi	2020	Thermal Distribution	Q-Learning Algorithm	Thermal Balance, Latency	28% improvement in thermal distribution	Effective optimization for routing decisions
10	Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network On-Chips	N. Shahabinejad, H. Beitollahi	2020	Thermal Distribution	Q-Learning Algorithm	Thermal Balance, Latency	28% improvement in thermal distribution	Effective optimization for routing decisions
11	Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network	N. Shahabinejad, H. Beitollahi	2020	Thermal Distribution	Q-Learning Algorithm	Thermal Balance, Latency	28% improvement in thermal distribution	Effective optimization for routing decisions

III. Research Gap

Summarize the key findings from your literature survey. Discuss any gaps in the research and suggest areas for future research.

References

[1] N. Shahabinejad and H. Beitollahi, "Q-Thermal: A Q-Learning-Based Thermal-Aware Routing Algorithm for 3-D Network On-Chips," in IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 10, no. 9, pp. 1482-1490, Sept. 2020, doi: 10.1109/TCPMT.2020.3018176. keywords: Three-dimensional displays;Routing;Heat sinks;Thermal management;Manufacturing;Two dimensional displays;Thermal sensors;3-D network-on-chip (3-D NoC);packet routing;Q-learning;Q-routing;thermal management,

[2] K.-C. Chen, S.-Y. Lin, H.-S. Hung and A.-Y.-A. Wu, "Topology-aware adaptive routing for nonstationary irregular mesh in throttled 3D NoC systems", IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 10, pp. 2109-2120, Oct. 2013.

[3] C.-C. Kuo, K.-C. Chen, E.-J. Chang and A.-Y. Wu, "Proactive thermal-budget-based beltway routing algorithm for thermal-aware 3D NoC systems", Proc. Int. Symp. Syst. Chip (SoC), pp. 1-4, Oct. 2013.

[4] Liu, H.; Chen, X.; Zhao, Y.; Li, C.; Lu, J. TTQR: A Traffic- and Thermal-Aware Q-Routing for 3D Network-on-Chip. Sensors 2022, 22, 8721. https://doi.org/10.3390/s22228721

[5] A. K. Coskun, J. L. Ayala, D. Atienza, T. S. Rosing, and Y. Leblebici, "Dynamic thermal management in 3D multicore architectures," in Proc. Design, Autom. Test Eur. Conf. Exhib., Apr. 2009, pp. 1410–1415.

[6] M. Guo, T. Cheng, X. Li, L. Li and Y. Fu, "A Nearest-Neighbor-Based Thermal Sensor Allocation and Temperature Reconstruction Method for 3-D NoC-Based Multicore Systems," in IEEE Sensors Journal, vol. 22, no. 24, pp. 24186-24196, 15 Dec.15, 2022, doi: 10.1109/JSEN.2022.3218953.

[7] K.-C. Chen, H.-W. Tang, Y.-H. Liao, and Y.-C. Yang, "Temperature tracking and management with number-limited thermal sensors for thermal-aware NoC systems," IEEE Sensors J., vol. 20, no. 21, pp. 13018–13028, Nov. 2020.

[8] J. Ranieri, A. Vincenzi, A. Chebira, D. Atienza, and M. Vetterli, "Near-optimal thermal monitoring framework for many-core systemson- chip," IEEE Trans. Comput., vol. 64, no. 11, pp. 3197–3209,Nov. 2015.

[9] M. Guo, T. Cheng, L. Li, and Y. Fu, "Optimized method for thermal tracking in 3D NoC systems by using ANN," in Proc. 18th Int. SoC Design Conf. (ISOCC), Oct. 2021, pp. 111–112.

[10] T. Wegner, M. Gag, and D. Timmermann, "Impact of proactive temperature management on performance of networks-on-chip," in Proc. Int. Symp. Syst. Chip (SoC), Oct. 2011, pp. 116–121.

[11] T. Cheng, H. Du, L. Li and Y. Fu, "LSTM-based Temperature Prediction and Hotspot Tracking for Thermal-aware 3D NoC System," 2021 18th International SoC Design Conference (ISOCC), Jeju Island, Korea, Republic of, 2021, pp. 286-287, doi: 10.1109/ISOCC53507.2021.9613862.

[12] A. K. Coskun, T. S. Rosing, and K. C. Gross, "Utilizing predictors for efficient thermal management in multiprocessor socs," IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 28, no. 10, pp.1503–1516, 2009.

[13] Xinyi Li, Wenjie Fan, Heng Zhang, Jinlun Ji, Tong Cheng, Shiping Li, Li Li, and Yuxiangfu Fu. 2024. TTNNM: Thermal- and Traffic-Aware Neural Network Mapping on 3D-NoC-based Accelerator. In Proceedings of the Great Lakes Symposium on VLSI 2024 (GLSVLSI '24). Association for Computing Machinery, New York, NY, USA, 364–369. https://doi.org/10.1145/3649476.3658703

[14] F. Li, Y. Wang, C. Liu, H. Li, and X. Li, "NoCeption: A Fast PPA Prediction Framework for Network-on-Chips Using Graph Neural Network," *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Antwerp, Belgium, 2022, pp. 1035-1040, doi: 10.23919/DATE54114.2022.9774525.