Power, Area and Thermal Prediction in 3D Network-on-Chip using Machine Learning

Abhijith C, Anand M K
Department of Computer Science and Engineering
National Institute of Technology Karnataka (NITK)
Surathkal, India
Email: {abhijithc.242cs003, anandmk.242cs008}@nitk.edu.in

I. Literature Review

The literature survey reviews Machine Learning (ML) and Deep Learning (DL) methods that address Power, Area, and Thermal (PAT) challenges in Network-on-Chip (NoC) systems. While 3D NoCs improve performance by stacking processing elements, they also increase power density, which leads to thermal issues, higher latency, and performance loss. Managing the area of these stacked components is also more complex, adding to the size and cost of the chip. This review gave us useful insights into how ML and DL techniques help solve these PAT challenges.

Chen *et al.* [1] proposed an adaptive machine learning-based Predictive Dynamic Thermal Management (PDTM) technique for thermal management in NoC systems. The approach utilizes an Adaptive Single-Layer Perceptron (ASLP) for predicting temperature, while adaptive reinforcement learning dynamically adjusts the throttling ratio. A revised thermal-traffic co-simulator and MCSL benchmark were used to evaluate the proposed model in simulated traffic patterns within an 8x8 NoC system. The XY routing algorithm was adapted. Results showed that the proposed method reduced the average temperature and maximum temperature error by 0.2%–78.0% and 0.6%–74.1%, respectively. Throughput was improved by 2.4%–43.0%, area overhead was reduced by 3%–9%, and power consumption was enhanced by 11%–21%. Additionally, the model increased hardware efficiency by 28.1%–59.7%.

Shahabinejad *et al.* [2] presented a Q-learning driven thermal-aware algorithm for routing, 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 13%-28% compared to other existing routing algorithms. The proposed method also achieves a 32% improvement in average latency and decreases the number of thermal hotspots by 38%-54%. 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 method increases the area and power consumption compared to previous routing techniques. The overall area is increased by 7%-11% and power consumption by 2%-4%.

Guo *et al.* [3] proposed a method of addressing thermal issues on NoC by optimizing the design techniques. Dynamic thermal management (DTM) 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 algorithm is proposed to initialize thermal sensors, and a Genetic Algorithm (GA) is used for sensor allocation optimization. The method uses an artificial neural network (ANN) to calculate the temperature of nodes without a sensor. 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 methods. The proposed nearest-neighbor-based approach for thermal sensor allocation 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 non-sensor-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.

Liu *et al.* [4] proposed a routing algorithm that incorporates traffic and thermal awareness, which involves a Q-Learning algorithm that uses two Q-tables: one table maintains local traffic status information, while the second table holds thermal information about the network. The proposed method improves latency by an average of 63.6% and throughput by 41.4%. Overall, this method provides a more uniform temperature distribution across layers. However, it has a higher average temperature.

Reza *et al.* [5] proposed a method to address power management in NoCs using reinforcement learning (RL), which configures network resources dynamically based on application needs and system utilization. Specifically, RL is applied to adjust NoC voltage levels, while a neural network (NN) is implemented to identify NoC performance patterns. A concentrated mesh NoC architecture is utilized to reduce the overhead of machine learning techniques. The method was evaluated using the COSMIC and E3S benchmarks, with

performance metrics focusing on throughput and the Energy-Delay Product (EDP). The proposed approach demonstrated a 45% improvement in EDP against the COSMIC benchmark and a 110% improvement against E3S. Throughput was increased by 15% and 10% for COSMIC and E3S, respectively.

DiTomaso *et al.* [6] proposed a method that aims to reduce both static and dynamic power consumption through power-gating techniques. The prediction of link utilization and traffic loads is implemented using a decision tree, and the XY routing algorithm is implemented. The dataset used for training the predictor consists of historical data on link utilization, buffer utilization, and packet type. Power and area estimates were obtained using the DELPHI simulator, which reported a dynamic power consumption of 12.3 mW and a static power consumption of 16.6 mW for the proposed method. The decision tree achieved an accuracy of 46.6% for traffic load prediction and 48.7% for link utilization prediction, which represents improvements of 13.2% and 13.8%, respectively, over a standard predictor.

Li *et al.* [7] proposed a Graph Neural Network (GNN) Framework for predicting the area, power, and performance of NoCs. 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 a power prediction accuracy of 97.36% and an 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.

Reza *et al.* [8] proposed a method to address power issues in NoCs through the application of machine learning (ML) to configure the network. The input dataset includes node and link usage, computation and communication demands, thermal and power consumption, task deadline requirements, and execution time. The target outputs are optimized voltage levels of nodes and link bandwidth settings. Various design and optimization challenges are discussed in the study. The neural network (NN)-based solution is evaluated against traditional non-ML approaches using the COSMIC benchmark, with the XY-routing algorithm employed. Results show a 15% improvement in both throughput and latency, as well as a 6% reduction in energy consumption.

Li *et al.* [9] proposed a Neural Network-based mapping technique that 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.

Chen *et al.* [10] proposed a Predictive Dynamic Thermal Management (PDTM) technique, which proactively manages node temperatures based on predicted thermal information. An artificial neural network (ANN) is employed for temperature prediction. The XY routing algorithm is implemented. The evaluations were performed using a traffic-thermal co-simulator, with performance metrics including average and maximum temperature error, system throughput, and area overhead. The proposed model reduced average temperature error from 37.2% to 62.3%, improved throughput by 9.16%, and reduced area overhead by 18.59% to 22.11%.

Wächter *et al.* [11] proposed a runtime manager incorporating a temperature predictor. The training data for the predictor includes system measurements (temperature, frequency, power consumption) collected while running applications from the PARSEC benchmark suite. Various regression models were trained using this data, and the model with the best performance was integrated into the runtime manager. The models were evaluated based on several metrics, including mean and maximum absolute error, standard deviation, and the Akaike Information Criterion (AIC). After evaluation, the selected model achieved an MAE of 1.13°C, a maximum absolute error of 16.91°C, a standard deviation of 1.31, and an AIC value of 33,222. The proposed runtime management model demonstrated a 10% improvement in energy efficiency and performance; however, it resulted in a doubling of thermal cycling.

Cheng *et al.* [12] proposed a Long Short-Term Memory (LSTM) temperature prediction model for Proactive Dynamic Thermal Management (PDTM), which is a temperature control technique that highly depends on the accuracy of the model. The proposed method improves temperature prediction accuracy by 41.92% to 73.63% compared to the traditional Autoregressive Moving Average (ARMA) model. Furthermore, the model can detect new hotspots in just 0.075 ms. However, the study is carried out on an 8×8×4 3D NoC, and it is unclear how well the model scales to larger systems.

In the review of ML and DL techniques for addressing Power, Area, and Thermal (PAT) challenges in Network-on-Chip (NoC) systems, several promising approaches were found. Methods like Q-learning-based routing algorithm and Predictive Dynamic Thermal Management (PDTM) significantly reduce temperature and improve routing decisions, while methods using ANN help in better design and configuration of NoCs. LSTM and GNN were effectively used for PAT prediction. However, key limitations were identified, including higher area and power consumption, assumptions of constant thermal conditions, and challenges with scaling to larger NoC systems. The summary of the review is given in Table I.

TABLE I: Summary of past studies

Sl.No	Author(s)	Issue Addressed	Approach	Performance Metrics	Results	Observations	Limitations
1	Chen *et al.* [1]	Thermal	Adaptive ML PDTM with ASLP Temp Prediction and RL Control.	Saturation Throughput, Temp Error	Throughput improved by 2.4%–43%, Avg temp errors reduced by 0.2%–78%.	ASLP reduces avg errors, RL improves throughput.	Uses only XY routing, Assumes accurate temp readings.
2	Shahabinejad *et al.* [2]	Thermal	Q-Learning-based routing utilizing thermal Q-values.	Std dev of thermal distribution, Avg latency	Std dev improved by 28% vs. TAAR, Latency improved by 32%.	Optimizes routing for better thermal balance.	Increased area and power consumption.
3	Guo *et al.* [3]	Thermal	Nearest-neighbor algorithm for sensor placement; ANN to estimate temp.	Avg & Max Temp Error	Avg temp error reduced by 17.60%–88.63%.	Effective sensor placement and accurate temp estimation.	Assumes constant thermal correlation among cores.
4	Liu *et al.* [4]	Thermal	Q-Learning routing with local traffic and global thermal info.	Avg latency, Throughput	Latency improved by 63.6%; throughput improved by 41.4%.	More uniform temp distribution and balanced traffic load.	Slightly higher avg temp than TAAR.
5	Reza *et al.* [5]	Power	Dynamic configuration using RL.	Throughput, EDP	Throughput improved by 15% and 10%; EDP improved by 45% and 110%.	Improved throughput and EDP compared to non-ML solutions.	Focus on mesh NoC to reduce overheads.
6	DiTomaso *et al.* [6]	Power	LESSON for reducing static and dynamic power.	Dynamic/static power, Latency	Dynamic power: 12.3 mW; Latency improved by 14%.	Total power saved by 31.7%–85.6%, decision tree accuracy up to 13.8%	Only XY routing used.
7	Li *et al.* [7]	Power	GNN for PPA prediction of NoCs.	Power and Area prediction	Power prediction accuracy: 97.36%; Area prediction: 97.83%.	Fast and accurate PPA prediction.	May struggle with larger systems.
8	Reza *et al.* [8]	Power and Thermal	Heterogeneous NoC configuration using ML (NN).	Throughput, Latency, Energy	Throughput/latency improved by 15%; Energy reduced by 6%.	Improved performance compared to traditional methods.	Only XY routing used.
9	Li *et al.* [9]	Thermal	Neural network mapping technique optimizes temperature distribution.	Avg Temp, Max Temp, Temp Variance	Reduced avg/max temp to 68.3°C and 77.2°C and variance to 5.9°C², improved thermal uniformity.	Uniform temperature distribution	Lacks consideration for dynamic scenarios.
10	Chen *et al.* [10]	Thermal	ML-based PDTM with ANN and RL.	Avg/max temp error, Throughput, Area overhead	Avg error reduced by 37.2%–62.3%, throughput improved by 9.16%, area overhead reduced by 18.59%–22.11%.	Dynamic adaptation to temp behavior.	Only XY routing used.
11	Wächter *et al.* [11]	Thermal	Runtime management (RTM) with temperature prediction.	MAE, Max AE, Std dev, AIC	MAE: 1.13°C, Max AE: 16.91, Std dev: 1.31, AIC: 33222	10% energy and performance improvement, doubled thermal cycling.	High thermal cycling and prediction overhead.
12	Cheng *et al.* [12]	Thermal	LSTM-based temp prediction.	MSE, Prediction Accuracy	MSE: 0.411°C, Accuracy improved by 41.92%-73.63%	Quick hotspot detection (0.075 ms) and better temp prediction vs. ARMA.	Unclear scalability to larger systems.

II. Research Gaps

- **Availability of Datasets**: Most ML and DL models require a large volume of quality datasets for training. However, the availability of such datasets is limited in the area of PAT prediction for NoC.
- **Integrated Power, Thermal, and Area Prediction Models**: Most existing studies independently focus on power or thermal optimization. Studies on simultaneous analysis of power, area, and thermal are rare.
- **Scalability**: Many studies using ML or DL models have been tested in specific architectures or smaller systems, but their scalability to larger multi-core environments is unclear.
- **Diversity of ML/DL Algorithms:** Only a few types of ML/DL algorithms, such as reinforcement learning and artificial neural networks, have been explored in the context of NoC. The application of other ML/DL algorithms, such as Convolutional neural network (CNN) and other regression algorithms, remains unexplored.

References

[1] K.-C. Chen, Y.-H. Liao, C.-T. Chen and L.-Q. Wang, "Adaptive Machine Learning-Based Proactive Thermal Management for NoC Systems," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 31, no. 8, pp. 1114-1127, Aug. 2023, doi: 10.1109/TVLSI.2023.3282969.

[2] 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,

[3] 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. 2022, doi: 10.1109/JSEN.2022.3218953.

[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] M. F. Reza, "Deep Reinforcement Learning for Self-Configurable NoC," 2020 IEEE 33rd International System-on-Chip Conference (SOCC), Las Vegas, NV, USA, 2020, pp. 185-190, doi: 10.1109/SOCC49529.2020.9524761.

[6] D. DiTomaso, A. Sikder, A. Kodi and A. Louri, "Machine learning enabled power-aware Network-on-Chip design," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, Lausanne, Switzerland, 2017, pp. 1354-1359, doi: 10.23919/DATE.2017.7927203.

[7] 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.

[8] Md Farhadur Reza. 2021. Machine learning for design and optimization challenges in multi/many-core network-on-chip. In Proceedings of the 14th International Workshop on Network on Chip Architectures (NoCArc '21). Association for Computing Machinery, New York, NY, USA, 29–34. https://doi.org/10.1145/3477231.3490427.

[9] 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

[10] K.-C. J. Chen and Y.-H. Liao, "Adaptive Machine Learning-Based Temperature Prediction Scheme for Thermal-Aware NoC System," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seville, Spain, 2020, pp. 1-4, doi: 10.1109/ISCAS45731.2020.9180475

[11] E. W. Wächter, C. de Bellefroid, K. R. Basireddy, A. K. Singh, B. M. Al-Hashimi and G. Merrett, "Predictive Thermal Management for Energy-Efficient Execution of Concurrent Applications on Heterogeneous Multicores," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 6, pp. 1404-1415, June 2019, doi: 10.1109/TVLSI.2019.2896776.

[12] 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.