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.

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 experimental result shows that the proposed method improves thermal distribution by approximately 28% and 13% compared to TAAR (TAAR -Thermal 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 it increases the area and power consumption compared to previous routing techniques. The layout area is increased by 7% and 11% in comparison with TAAR and PTB3R. The power consumption of the proposed method is 2% higher than that of TAAR and 4% higher than that of PTB3R

[2] 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," 2013 International Symposium on System on Chip (SoC), Tampere, Finland, 2013, pp. 1-4, doi: 10.1109/ISSoC.2013.6675281. keywords: {Routing;Thermal Budget;Proactive Routing;3D NoC;3D IC}

[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.

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 Table: One table maintains local traffic status information, and 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 compared to (Topology-aware Adaptive Routing) [5]. Overall TTQR provides a more uniform temperature distribution across layers. However, TTQR has higher average temperature compared to TAAR.

Another method of addressing thermal issues on NoC is optimizing the design techniques. Dynamic thermal management (DTM) is one of the key techniques which requires accurate thermal information from thermal sensors. Due to the high hardware cost, limited thermal sensors are available which makes the thermal sensor allocation crucial.

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) 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)** 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 model is proposed in [10]. The proposed method improves temperature prediction accuracy by 41.92% to 73.63% compared to the traditional ARMA (Autoregressive Moving Average) model [11]. 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, temperature variance, and maximum 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 runtime or dynamic scenarios

Power and area are also important factors for NoC. A Graph Neural Network (GNN) Framework is proposed [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 = 97.36% and area prediction accuracy = 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.