IEEE Open Journal of
Circuits and Systems

Received 7 November 2025; revised 24 December 2025; accepted 31 January 2026.
Date of publication 9 February 2026; date of current version 19 February 2026.

Digital Object Identifier 10.1109/OJCAS.2026.3662252

Thermal-Aware NoC for AI Computing: Tools, Algorithms, and Applications

FAKHRUL ZAMAN ROKHANI[1] (Member, IEEE), MIRCEA R. STAN[2] (Fellow, IEEE), MAURIZIO PALESI[3,4] (Senior Member, IEEE), AND KUN-CHIH CHEN[5] (Senior Member, IEEE)

[1]Universiti Putra Malaysia, Serdang 43400, Malaysia
[2]University of Virginia, Charlottesville, VA 22904 USA
[3]University of Catania, 95124 Catania, Italy
[4]IIT Guwahati, Guwahati 781039, India
[5]National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
This article was recommended by Associate Editor A. James.

CORRESPONDING AUTHOR: K.-C. CHEN (kcchen@nycu.edu.tw)

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC 114-2221-E-A49-170-MY3 and Grant NSTC 113-2640-E-A49-005; and in part by the European Commission through the Chips Joint Undertaking under Grant 101139790 (ECS4DRES).

ABSTRACT The explosive growth of AI workloads elevates on-chip communication and heat dissipation to first-order constraints. This survey paper consolidates thermal-aware Network-on-Chip (NoC) design for AI computing across tools, algorithms, and applications. Concretely, we first assemble a reproducible toolchain that couples cycle-accurate NoC simulators with power/thermal solvers and machine-learning surrogates for fast temperature prediction. We then structure the design space along three design dimensions: sensing strategies, control methodologies, and thermal- and traffic-aware data delivery. Finally, we close the loop among traffic, power, and temperature via an integrated co-simulation workflow, providing practical guidelines for thermal-aware NoC-based AI accelerator designs. Unlike general DNN-accelerator surveys, this survey paper focuses on the thermal–NoC interplay under realistic AI workloads and provides an actionable, closed-loop methodology and tooling for scalable, verifiable evaluation. We conclude with open challenges, scalable yet faithful co-simulation, standardized traces/interfaces, packaging-aware models, and uncertainty-aware surrogates, to guide the path toward thermally resilient, high-throughput AI systems.

INDEX TERMS AI accelerators, network-on-chip (NoC), thermal-aware design, co-simulation.

I. INTRODUCTION

AI HAS evolved from rudimentary to data-driven machine learning (ML) and more recently to deep learning (DL), powered by deep neural networks. The proliferation of the Internet of Things (IoT) and significant advancements in computing performance drive this evolution. DL models, which range from convolutional, recurrent, transformer, and spiking networks, now underpin perception, language, and decision-making across application domains.

Compared to conventional computing workloads, DL exhibits distinctive algorithmic and system characteristics. It is inherently tensor-centric, dominated by highly parallel linear-algebra operations such as iterative General Matrix Multiplications (GEMMs) and convolutions that execute vast numbers of multiply–accumulate operations. Modern DL models involve massive parameter counts and increasingly irregular operator patterns, such as attention mechanisms and sparsity. These characteristics, when combined, require high computational throughput, extensive data movement, and substantial memory bandwidth to maintain performance. This computational intensity, coupled with the increasing complexity of DL models has driven the need for specialized hardware architectures capable of addressing the efficiency and scalability limitations of traditional platforms.

To accelerate deep learning (DL) workloads, specialized accelerators have been developed alongside general-purpose multicore Central Processing Units (CPUs) and Graphics Processing Units (GPUs). These specialized accelerators include Application-Specific Integrated Circuits (ASICs), Field-Programmable Gate Arrays (FPGAs), as well as modified versions of common processors designed specifically for deep neural networks (DNNs). TABLE 1 provides a comparative overview of different hardware architectures accelerating DL workloads. Various surveys have been conducted to summarize existing hardware architectures that accelerate ML and DL workloads, in terms of designs and implementations,

TABLE 1. The comparison of different hardware architectures accelerating DL workloads.

Hardware Architecture	Key Features	Strength	Weakness
Central Processing Unit (CPU)	General purpose core with support for diverse data types and operations, hierarchical caches for memory	Reconfigurable to support diverse DL applications; special ISA and SIMD or vector co-processor capabilities	Poor data specialization; high overhead from complex control circuits.
Graphic Processing Unit (GPU)	Thousands of cores for vector/matrix computations; large shared caches and memory hierarchies	High parallelism; optimized memory for graphics and adaptable to DL; lower overhead than CPUs	Not fully tailored data specialization and memory access for DL applications.
Tensor Processing Unit (TPU)	Custom ASIC with systolic arrays for tensor operations; low-precision data support; dedicated vector/local memory	Excellent data specialization for DL tensors and low-precision; high parallelism via numerous PEs; optimized local memory reducing data movement	Limited flexibility beyond DL tasks; data specialization tied to specific formats; parallelism optimized only for certain DL operations.
Field Programmable Gate Array (FPGA)	Reconfigurable logic for custom data paths; programmable parallelism via logic blocks; customizable memory hierarchies; flexible control logic with some overhead	Reconfigurable to support evolving DL needs; configurable parallelism for specific workloads; ability to optimize local memory layouts	Data specialization requires expertise and reconfiguration time; parallelism lower than fixed ASICs.
Application Specific Integrated Circuit (ASIC)	Fixed custom circuitry for DL operations; specialized data types and units; highly parallel compute arrays; integrated local memories; dedicated minimal-overhead control	Maximal data specialization for target DL tasks; highly parallel with optimized units; highly efficient local memory	Inflexible data specialization if DL evolves; parallelism fixed, not adaptable; local memory designs hard to change; inability to repurpose.

explaining their techniques, and comparing their system-level performance. Sze et al. [1] did an early foundational work to motivate research into efficient DL, especially focusing on Deep Neural Network (DNN), processing, dataflow architectures, hardware architectures, algorithm–hardware co-design, benchmarking metrics, and design trade-offs for improved energy efficiency and throughput. The authors further address energy-efficient embedded machine learning, exploring hardware architectures, algorithms, and technologies to enable adaptive, low-power, near-sensor intelligence [2]. Reuther et al. summarize computational performance and precision of academic and commercial ML devices [3], [4]. Tang et al. investigate neuromorphic devices, linking biological and artificial neural mechanisms, and highlight key challenges guiding future brain-inspired computing research Wang et al. evaluate approximation methods (i.e., quantization and weight reduction) for custom DNN hardware accelerators [5]. Sze et al. comprehensively investigate the efficient processing of DNNs, including a taxonomy of DNN accelerator architectures, key evaluation metrics, and co-design strategies to enhance energy efficiency and performance [6]. Bodiwala et al. review recent trends in efficient DNN hardware, analyzing architectures, algorithm–hardware co-design, and trade-offs to reduce cost and computation without sacrificing accuracy [7]. Chen et al. review recent advances in selected DNN accelerator architectures, covering dataflow optimization, and application-specific designs, and future AI chip design trends [8]. Nabavinejad survey on efficient on-chip interconnection networks for DNN accelerators, focusing on low-power, high-bandwidth communication, reconfigurable architectures, and emerging in-/near-memory processing for enhanced flexibility and performance in edge and IoT applications [9]. Akkad survey on DNN hardware accelerators targeting customizable RISC-V processors, emphasizing architectures for embedded and edge devices. Besides, the authors review DNN acceleration, ISA extensions, quantization, and in-/near-memory computing techniques for efficient low-power AI deployment [10]. Silvano, comprehensively surveys recent DL accelerators for HPC applications, covering GPU/TPU, FPGA/ASIC, NPU, RISC-V, PIM, NVM-based in-memory computing, neuromorphic, photonic, quantum accelerators, with classification of key architectures and emerging technologies [11].

Conversely, the high computing demands of AI lead to increased power density, which causes thermal issues in AI accelerators. Therefore, overheating and thermal inefficiencies are fast becoming critical bottlenecks, threatening the scalability, reliability, and energy efficiency of next-generation AI systems. While hardware accelerators have made significant strides, the integration of thermal-aware design remains a challenge in both research and industry. Thermal-aware NoC design methodologies encompass a variety of algorithms aimed at predicting temperature, managing runtime heat, and optimizing communication to mitigate

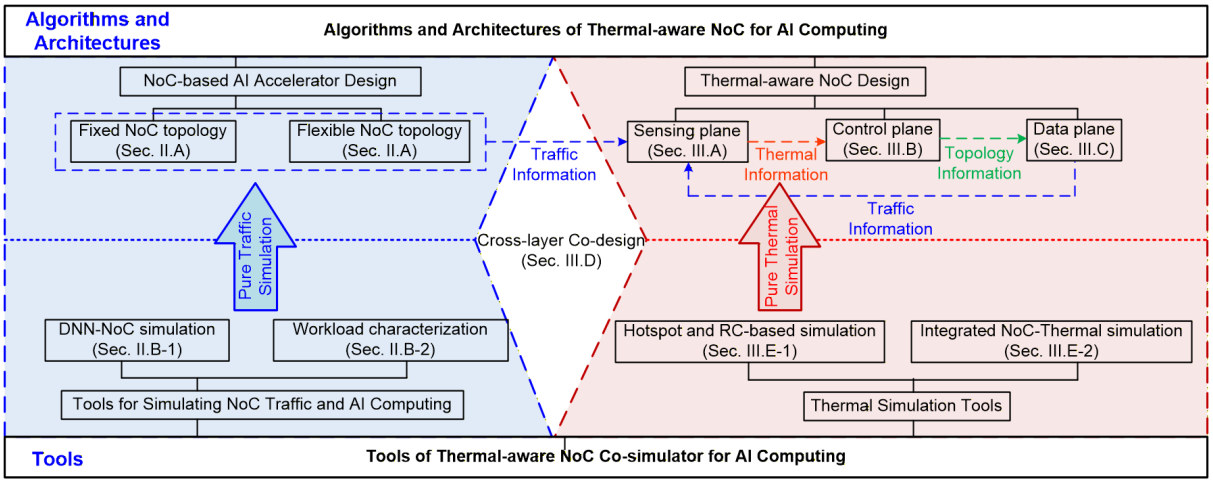


FIGURE 1. The topics discussed in this paper.

thermal hotspots. These challenges are intensified in AI accelerators, where DNN workloads cause uneven power density and localized hotspots. Multicast/broadcast traffic concentrates the load to specific NoC regions, while irregular dataflows from layers such as convolution and attention lead to non-uniform activity across PEs and routers. These challenges are amplified by 3D stacking and chiplet-based designs, where vertical NoCs increase thermal coupling and poor heat dissipation from upper dies or inner chiplets exacerbate hotspot formation. Hence, we investigate thermal-aware NoC design methodologies in the second part.

Temperature prediction techniques are increasingly leveraging machine learning models such as graph convolutional networks (GCNs) [12] to rapidly and accurately estimate thermal maps of complex chiplet and 3D NoC systems. Dynamic thermal management approaches include distributed runtime mechanisms that throttle traffic and reroute packets based on real-time thermal and congestion data, exemplified by schemes like ThermalHerd and traffic- and thermal-aware Q-routing (TTQR) [13] using reinforcement learning to balance load and temperature across layers. Adaptive routing algorithms [14] integrate temperature, traffic state, distance, and path diversity to evenly distribute heat while maintaining high throughput and low latency, often with hardware-efficient implementations to reduce overhead. Thermal-aware memory access [15] and task mapping strategies [16] further reduce peak temperatures by allocating tasks and memory accesses considering both computation and communication energy alongside thermal dissipation characteristics, as demonstrated by cluster-based thermal-aware task allocation algorithms that significantly lower peak chip temperatures and energy consumption in 3D NoCs. A technique exploiting approximate computing to improve DTM efficiency, reducing power, area and cost while ensuring reliable thermal sensing is proposed in [17]. Collectively, these methodologies enable holistic thermal management, crucial for reliable and efficient AI-accelerated NoC systems.

While existing surveys provide valuable overviews of ML and DL accelerators and optimization techniques, none analyze NoC-based AI accelerators from a thermal-aware perspective. Through this work, we aim to provide a holistic perspective on the interplay between thermal management, NoC design, and AI acceleration, paving the way for future innovations in high-performance and thermally resilient AI hardware.

The conceptual framework of this survey is illustrated in FIGURE1. This figure highlights the critical interplay between the two core domains: *NoC-based AI accelerator architecture* and *thermal-aware design strategies*. Specifically, the AI accelerator architecture determines the traffic patterns, which serve as the primary input for thermal analysis. Conversely, thermal constraints necessitate specific algorithmic interventions that feed back into the architectural operation. To address these coupled dynamics simultaneously, we emphasize the role of co-simulation workflows that integrate traffic profiling with thermal modeling. With this manuscript, we provide a holistic roadmap for researchers to navigate the complex design space of thermal-aware AI computing, bridging the gap between architectural demands and thermal reliability. The outline of this survey paper is as follows:

- **NoC-based AI Computing Architecture and Applications:** We explore architectural innovations and practical use cases that leverage thermal-aware NoC designs to enhance AI computing. Three-dimensional (3D) NoC architectures, utilizing through-silicon vias (TSVs) and interposers, offer significant improvements in bandwidth and latency by stacking multiple layers of processing elements. However, this increased integration density intensifies thermal challenges, making algorithmic and architectural solutions—such as intelligent routing and dynamic thermal management—attractive alternatives to purely technological fixes. To address these issues, thermal and traffic simulation

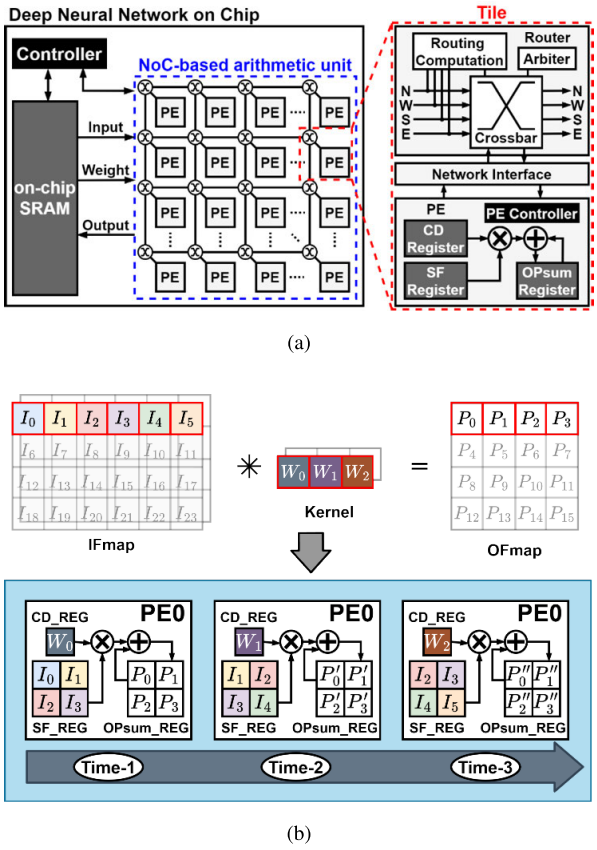


FIGURE 2. (a) The architecture of UFO-NoC uses (b) the reconfigurable PE design to support arbitrary kernel operations [22].

tools (often based on RC-circuit models and platforms like NoCVision) enable designers to capture fine-grained transient temperature and congestion behavior. Furthermore, AI workload simulators that integrate neural network execution models with thermal analysis allow for system-level evaluation of how computation-intensive AI tasks impact communication efficiency. These thermal-aware designs are critical for AI accelerators deployed in deep learning, computer vision, and data analytics. Applications ranging from large-scale inference engines to edge AI platforms benefit from the careful coordination of computation, communication, and thermal management, ensuring sustained performance under demanding workloads.

- **Algorithms and Strategies for Thermal-aware NoC Design:** We focus on key methodologies to mitigate thermal challenges, including thermal monitoring, thermal-aware mapping/routing, and dynamic power management. Simulation frameworks, built upon RC-grid thermal modeling and integrated with traffic analysis, provide the quantitative basis for these algorithms. We discuss strategies such as Reactive Dynamic Thermal Management (RDTM), which throttles overheated nodes to prevent failure, and Proactive Dynamic Thermal Management (PDTM), which utilizes predictive algorithms to adjust system behavior preemptively for higher throughput. Additionally, we examine thermal-aware routing algorithms (e.g., ant colony optimization, beltway routing) that avoid hotspots by balancing traffic load. Finally, task mapping strategies that optimize workload placement according to thermal profiles are discussed as a means to minimize localized heating. These integrated, simulation-driven approaches form the foundation for robust thermal management in AI-accelerated NoC systems.

II. NOC-BASED AI: ARCHITECTURE AND APPLICATIONS

A. ARCHITECTURAL DESIGN of NoC-BASED DNN ACCELERATOR

As accelerators scale to hundreds of thousands of Processing Elements (PEs) and multi-tiered memories, NoCs provide the fabric for multicast/broadcast weights, gather/reduce partial sums, and many-to-many flows while preserving QoS and isolation. NoC-based DNN accelerators leverage on-chip networks to interconnect these PEs, addressing communication bottlenecks and improving efficiency in large-scale AI computing systems. Demonstrated in silicon, Eyeriss [18] achieves energy-efficient spatial dataflow through a reconfigurable on-chip network, while Simba [19] enables scalable multi-chip integration via an NoC-like interconnect fabric. However, modern AI models exhibit extreme diversity in layer shapes (e.g., varying kernel sizes in CNNs) and operator types (e.g., attention mechanisms in Transformers), necessitating architectures that offer both computational flexibility and transport efficiency.

1) RECONFIGURABLE NEURAL PROCESSING UNIT DESIGN

The primary challenge in mapping diverse DNN models to hardware lies in the mismatch between the fixed physical dimensions of the PE array and the variable logical dimensions of the neural network layers (e.g., kernel sizes of 1×1, 3×3, or 7×7). Conventional rigid systolic arrays often suffer from low utilization—referred to as ''dark silicon''—when the kernel size does not perfectly align with the hardware array. To address this, reconfigurable architectures such as MAERI [20] utilized tree-based interconnects to support flexible dataflows. However, tree structures often incur routing congestion at the root nodes.

Leveraging the scalability of NoCs, Chen et al. proposed a reconfigurable design featuring a weight-wise mapping strategy [21]. By decomposing the convolutional filters and mapping them onto NoC-connected PEs, this approach decouples the logical kernel size from the physical PE arrangement. It allows the NoC to route input feature maps (IFMs) to the appropriate PEs regardless of the kernel geometry, significantly improving hardware utilization for irregular layer shapes. Building upon this foundation, the Universal Flexible-Oriented NoC (UFONoC) architecture [22] further extends this flexibility to support arbitrary kernel sizes,

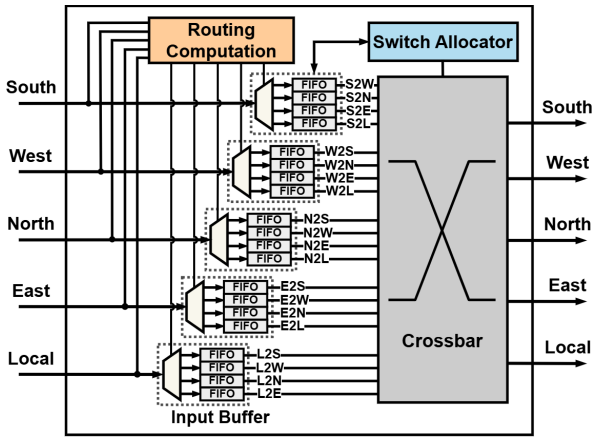


FIGURE 3. The flexible Ultra-NoC router supports hybrid data transmission mechanism [23].

as shown in FIGURE 2. UFONoC employs a kernel-based processing mechanism that dynamically reconfigures the PE clustering based on the target layer's requirements. Crucially, it integrates a hybrid data reuse strategy directly within the reconfigurable units, minimizing redundant fetch operations from global memory. This evolution demonstrates that shifting from rigid arrays to NoC-assisted reconfigurable clusters allows accelerators to maintain high throughput across the wide spectrum of kernel sizes found in modern Computer Vision models.

2) HIGH-PERFORMANCE DATA DELIVERY

Beyond computational flexibility, the efficiency of data delivery, specifically the minimization of data movement and latency, is paramount. DNN workloads generate distinct traffic patterns, including multicast (for weight sharing), unicast (for partial sum accumulation), and scatter-gather (for synchronization). Handling these diverse patterns on a unified fabric requires optimizations at both the protocol and routing levels.

Traditional packet-switched networks often incur high serialization latency. To mitigate this, the UFONoC architecture [22] adopts the industry-standard AXI4-Stream protocol for its NoC data transmission. This protocol streamlining supports high-bandwidth burst transfers and simplifies the interface between the PEs and the memory hierarchy, ensuring that the computation units are not starved of data. Furthermore, the complexity of supporting multiple dataflows (e.g., Input Stationary vs. Weight Stationary) simultaneously necessitates a flexible routing scheme. Chen et al. [23] introduced an Ultra-NoC, which features a Unified Low-Transmission Routing mechanism. Ultra-NoC addresses the limitations of Eyeriss v2 [18], which relies on separate physical networks for different data types. Instead, Ultra-NoC utilizes a consolidated router architecture capable of dynamically switching between unicast, multicast, and broadcast modes, as shown in FIGURE 3. By optimizing the routing path based on the data reuse characteristics of the specific layer, Ultra-NoC significantly reduces transmission hops and energy consumption.

3) SYSTEMATIC MODULAR DESIGN METHODOLOGY

While reconfigurable PEs and optimized routing protocols address specific efficiency bottlenecks, the rapid evolution of DNN architectures calls for a holistic, systematic design methodology that can adapt to entirely new model structures without extensive hardware redesign. Addressing the challenge of design reuse and scalability, Chen et al. [24] proposed a novel Lego-based Neural Network Design Methodology, as shown in FIGURE 4(a). Drawing an analogy to Lego construction, this approach decomposes the complex and heterogeneous operations of various DNN models (e.g., Convolution, Pooling, Activation) into fundamental, standardized computing units termed NeuLego Processing Elements (*NeuLegoPEs*).

Unlike traditional monolithic accelerators where the dataflow is hardwired, the Lego-based methodology utilizes a flexible NoC as the interconnection between these NeuLegoPEs. Designers can create a diverse range of Deep Neural Network (DNN) topologies, ranging from simple linear chains in CNNs to more complex directed acyclic graphs (DAGs) like those used in GoogLeNet, by easily reconfiguring the Network-on-Chip (NoC) routing paths. This process is akin to building various structures using the same set of Lego bricks. Crucially, to support large-scale DNNs that exceed the physical hardware capacity, the authors introduced a Lego Placement Method. This algorithmic strategy optimizes the mapping of virtual neural layers onto physical NeuLegoPEs, maximizing hardware reuse and minimizing communication latency, as shown in FIGURE 4(b). By standardizing the IP interfaces and decoupling the computation logic from the interconnection fabric, this systematic methodology provides a scalable path for generating domain-specific accelerators that remain versatile enough for future algorithm updates.

B. TOOLS for SIMULATING NoC AND AI COMPUTING BEHAVIOR

AI accelerators impose communication patterns and resource demands that are qualitatively different from traditional HPC or general-purpose workloads. Accurate evaluation therefore requires simulators that model not only bit-level NoC behavior but also DNN execution, dataflow/mapping choices, and the resulting multicast/gather traffic that dominates many neural operators. This subsection reviews (1) DNN-aware NoC simulators and execution models, (2) workload-characterization studies that expose the traffic signatures of modern DNNs, and (3) mapping strategies and multicast-heavy communication patterns.

1) DNN-NoC SIMULATORS AND AI EXECUTION MODELING

DNN-aware NoC simulators extend traditional cycle-accurate network models by embedding DNN execution semantics (i.e., operators, dataflows, tiling, and scheduling) so that simulator outputs (i.e., packets, flits, injection timings) reflect real DNN execution rather than synthetic traffic.

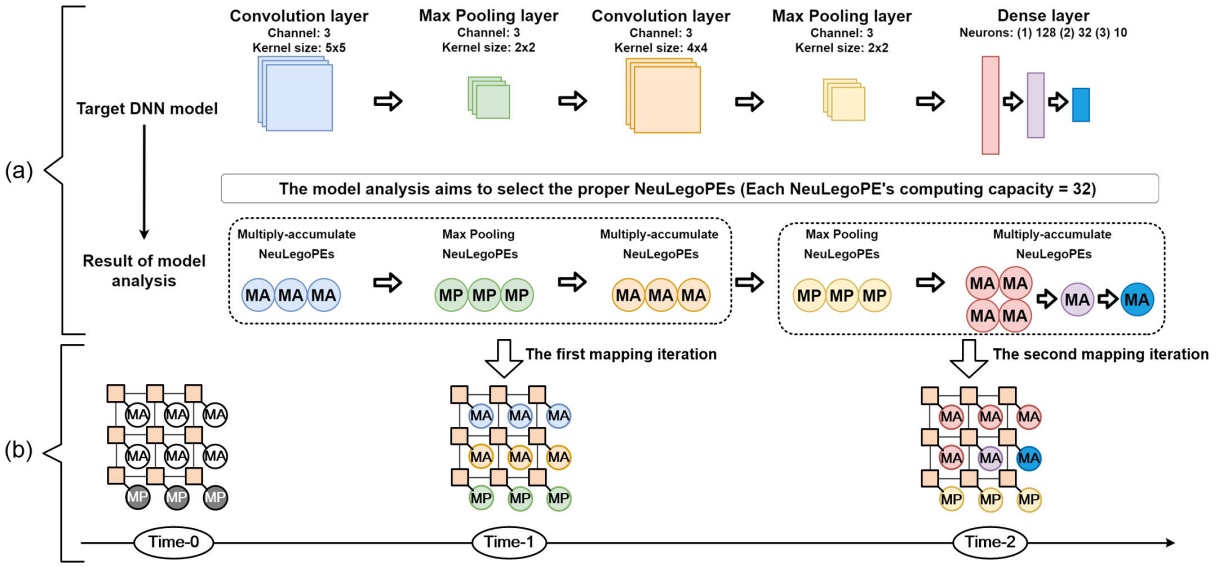


FIGURE 4. The lego-based method (a) analyzes the DNN at design time and (b) use NoC to compute DNNoC at runtime [24].

Representative systems include DNN-Noxim/NN-Noxim [25], which integrate DNN layer semantics into a NoC modeling framework and report metrics such as classification latency, energy consumption, and communication cost under various mappings and NoC configurations. DNN-Noxim produces per-layer traces that are directly usable as inputs to power and thermal backends, making it a practical tool for evaluating how specific layer types (e.g., convolution, GEMM, attention) stress the interconnect.

Complementary toolchains focus on mapping and accelerator evaluation at the operator and tiling level. Timeloop/Accelergy [26] provide a mature stack for exploring mapping choices and estimating compute/communication costs at high fidelity; outputs from Timeloop, including operator execution timelines, data reuse statistics, and bandwidth demands, are frequently used to drive NoC traffic generators or to parameterize more detailed NoC+DNN co-simulators. Timeloop's modeling of memory hierarchy and mapping knobs makes it especially useful when you need to explore alternatives at the DNN-mapping level before committing to cycle-accurate NoC simulations.

For system-level studies that emphasize packaging and multi-chiplet integration, simulators and prototype systems such as Simba [27] demonstrate how chiplet partitioning and inter-chiplet communication strategies affect inference throughput and energy at package scale, and thus motivates the use of chiplet-aware NoC modeling for large DNNs. Such works are invaluable when the evaluation target is package- or system-level performance rather than a single-die microarchitecture.

To bridge these tools into a cohesive research platform, FIGURE 5 illustrates the proposed end-to-end co-simulation workflow. Obviously, the workflow sequentially processes the data from the workload mapping phase to the thermal modeling phase. Use a DNN-aware NoC simulator (e.g., DNN-Noxim) for tight coupling between DNN execution semantics and router-level activity. Use Timeloop/Accelergy upstream to enumerate promising mappings and to produce workload footprints (tile-level bandwidth and reuse statistics) that the NoC simulator will consume. For chiplet- and packet-parallel studies, researchers can rely on scalable NoC engines such as CNSim [28] or 3D-/chiplet-aware extensions of BookSim [29]. These allow the exploration of multi-die and wafer-scale scenarios with manageable runtimes while preserving cycle-level accuracy at the packet granularity. Analytical delay models or trace-driven traffic replay frameworks (e.g., gem5 Garnet extended with DNN traffic traces [30]) can further complement such studies by providing fast approximations for large-scale design-space exploration.

2) WORKLOAD CHARACTERIZATION STUDIES FOR AI ACCELERATORS

Robust simulator input depends on realistic workload models. Recent characterization studies have shown that DNN communication is highly layer-dependent, with markedly different temporal and spatial traffic signatures across convolution, depthwise, fully-connected, and attention layers [31], [32]. For example, convolutional layers often yield high data reuse and bursty multicast/broadcast traffic for weights and partial sums, whereas transformer attention layers produce irregular, often less-reusable activations and different all-to-all or collective communication behaviors. Systematic studies of DNN communication, including measurements across modern transformer and CNN models, provide the trace statistics, such as multicast fraction, per-layer bandwidth, and temporal burstiness, needed to generate realistic

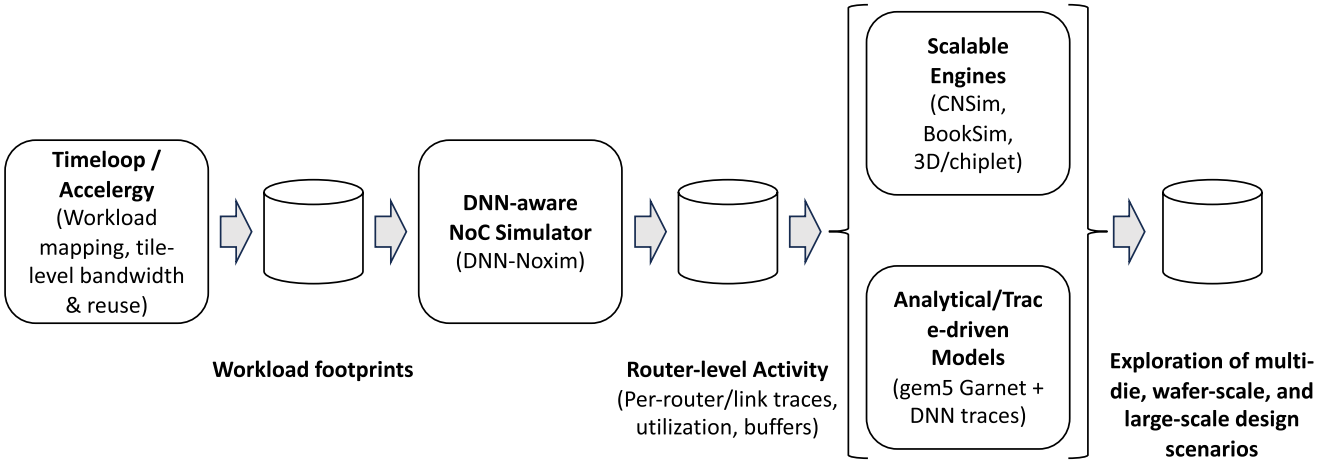


FIGURE 5. Workflow for AI workload–aware NoC simulation.

NoC workloads and to assess how DNNs drive thermal stress in particular regions of a chip or package [33].

Several recent works specifically profile DNN traffic and provide datasets or empirical guidance for simulator inputs. Studies on model-parallel and pipeline-parallel training/serving expose communication costs of tensor and pipeline parallelism (e.g., cross-mesh resharding and collective primitives), which strongly affect interconnect loading for large language models (LLMs) and distributed inference deployments [34]. For example, [35] quantifies link-level and inter-node traffic under ZeRO and data/model/tensor parallelism schemes, and [36] reports traffic overheads and burst distributions for serving transformer models under different parallelization strategies. For accelerator designers, these characterization efforts supply the per-layer and cross-layer traffic fingerprints that should be used instead of synthetic patterns when evaluating thermal and performance trade-offs.

3) MAPPING AND MULTICAST-HEAVY COMMUNICATION PATTERNS

A defining feature of many DNN accelerators is one-to-many (i.e., multicast) traffic for distributing weights and many-to-one (i.e., gather/reduction) patterns for partial sums and gradient aggregation [1], [37], [38]. Multicast, if implemented inefficiently as repeated unicasts, can saturate links and create persistent hotpaths that concentrate both communication energy and thermal dissipation on a small set of routers or tiles. Architectural solutions include reconfigurable multicast trees [39], hardware multicast support [40], hierarchical networks [18], and dataflow-aware mappings that place communicating tiles close together to reduce long-haul traffic [26], [41]. Experimental and simulation studies, including router-level multicast mechanisms and path-based multicast proposals, have shown large improvements in latency and energy when multicast is supported natively or when mappings exploit multicast locality [18].

Mapping strategies, such as output-stationary, weight-stationary, and row-stationary, significantly interact with NoC stress [42], [43]. An output-stationary mapping can concentrate partial-sum accumulation traffic locally, while a weight-stationary mapping may require frequent broadcasting of weights. For transformers and model-parallel deployments, mapping choices also determine whether communication manifests as many small bursts (e.g., attention score exchanges) or as sustained all-reduce traffic (e.g., gradient synchronization) [44]. These mapping-driven differences translate directly into different thermal profiles: concentrated, repeated multicast trees produce localized, sustained dissipation, while widespread, uniform traffic tends to produce milder, more distributed heating.

For practical guidance, it is essential to simulate various mapping strategies for each type of DNN layer, measuring both peak temperatures and hotspot durations. Additionally, evaluate the impact of native multicast support versus repeated unicast on NoC congestion and thermal hotspots by assessing power consumption at each router and inputting this data into tools like HotSpot or similar alternatives. In studies involving multiple chiplets or package-scale designs, consider consolidating frequently communicating layers onto the same chiplet or tile cluster to reduce inter-chiplet traffic and minimize vertical heat coupling. Utilize tools such as Timeloop to explore different mappings and employ a DNN-aware NoC simulator to analyze the resulting traffic patterns and thermal traces.

III. THERMAL-AWARE NoC:ALGORITHMS AND STRATEGIES

The design concept of a thermal-aware Network-on-Chip (NoC), as conceptually illustrated in FIGURE 1, operates as a closed-loop feedback system comprising three synergistic components: the *sensing plane*, the *control plane*, and the *data plane*. The *sensing plane* acts as the observer, tasked with capturing real-time spatiotemporal thermal dynamics

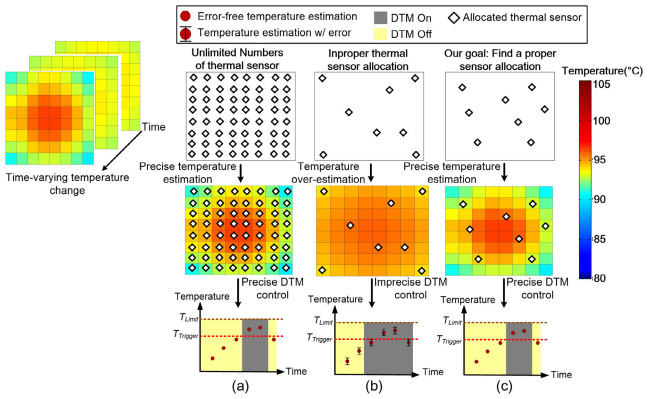


FIGURE 6. Under a time-varying temperature changing scenario, (a) unlimited number of thermal sensors helps to track full-chip temperature and control the DTM precisely; (b) it is hard to find a proper sensor allocation, which leads to improper temperature control; (c) the design goal is to find a proper sensor locations leverage the proper temperature control [46].

and traffic behaviors through physically integrated sensors or virtual telemetry. Based on this sensory data, the *control plane* employs Dynamic Thermal Management (DTM) policies to make high-level decisions, such as frequency scaling or task migration, to regulate the chip's temperature. However, these thermal control actions often create heterogeneous performance zones or unavailable nodes, effectively resulting in an irregular logical topology. Consequently, the *data plane* must dynamically adapt its routing algorithms to navigate this variable topology, ensuring reliable communication and deadlock freedom despite the thermal constraints. The following subsections detail the theoretical foundations and state-of-the-art strategies for each plane.

A. SENSING PLANE: TEMPERATURE AND SENSOR PLACEMENT

The execution of computation-intensive AI workloads on many-core architectures leads to significant power density variations, resulting in severe thermal hotspots and large thermal gradients. To ensure system reliability and prevent thermal runaway, accurate knowledge of the on-chip thermal distribution is a prerequisite for any effective thermal management mechanism. However, deploying a physical thermal sensor for every processing element (PE) is impractical due to substantial area and power overheads. Therefore, the objective is to select a minimal set of optimal sensor locations that can reconstruct the full-chip thermal map with high fidelity, as shown in FIGURE 6. The problem of placing thermal sensors has been proven as an NP-hard problem [45]. Existing literature addressing this challenge can be broadly categorized into 1) statistical-based approaches and 2) information-theoretic approaches.

Statistical approaches primarily exploit the spatial correlation of temperatures across the chip to group PEs with similar thermal behaviors. The core idea is to use historical thermal profiles to cluster PEs and place sensors at representative locations (e.g., cluster centroids). For instance, Chen et al. [46] proposed a correlation-based clustering framework. Based on the correlation between each NoC node along with time, the authors place fewer sensors on some of these correlated nodes and use the sensing temperature information to estimate other location without thermal sensors. While effective for workloads with stable spatial patterns, statistical methods may struggle to capture rapid transient anomalies that deviate from the historical average behavior.

To overcome the limitations of the statistical approach, information-theoretic approaches leverage the sparsity of thermal signals in the frequency or wavelet domain. These methods model the problem of thermal sensor placement as a signal reconstruction problem, often utilizing Compressive Sensing (CS) theory. Chen et al. [47] introduced a low-complexity CS-based framework that selects sensor locations randomly to minimize the mutual coherence of the sensing matrix, thereby maximizing the probability of accurate full-chip thermal reconstruction from sparse measurements. To reduce the computational complexity of full-system temperature distribution reconstruction, they employ the Matrix Inversion Bypass (MIB) property to compute the matrix inversion, which is a popular operation in conventional signal reconstruction methods of the compressive sensing theory (e.g., Orthogonal Matching Pursuit (OMP) and Stagewise Othogonal Matching Pursuit (StOMP)). Although this work reduces the computational complexity significantly, the involved random temperature sampling strategy cannot guarantee the independence of the collected signal (i.e., Restricted Isometry Property (RIP)). Besides, the involved fixed measurement matrix cannot adapt to the time-varying temperature changing situation and may counteract the benefit for full-system temperature reconstruction. To solve this problem, in their latest study, Chen et al. [48] proposed an entropy-based thermal sensor placement strategy combined with adaptive compressive sensing. This work incorporates entropy to quantify the informativeness of each candidate location, which is beneficial for finding locations that fit RIP. In addition, the authors employ LMS-based adaptive filter to adjust the involved measurement matrix dynamically, which can adapt to the time-varying temperature changing situation.

B. CONTROL PLANE: DYNAMIC THERMAL MANAGEMENT

After receiving the information on temperature behavior, the involved Dynamic Thermal Management (DTM) adjusts on-chip activities to ensure thermal safety while maintaining throughput. The DTM strategies generally operate by adjusting the voltage, frequency, or injection rates of the NoC routers based on sensory data. The DTM techniques are broadly classified into two categories based on their timing of intervention: *Reactive DTM (RDTM)* and *Proactive DTM (PDTM)*.

Reactive DTM (RDTM) represents the conventional approach to thermal safety. when a thermal sensor detects that the local temperature has exceeded a predefined

threshold, the DTM controller immediately initiates aggressive cooling measures. Common actuation techniques include global throttling, clock gating, or Dynamic Voltage and Frequency Scaling (DVFS) [49], [50]. For example, Shang et al. demonstrated a history-based frequency scaling method that reacts to thermal violations in localized NoC regions [50]. While RDTM is straightforward to implement and guarantees safety, its major drawback is the *hysteresis effect*. Since thermal violations are only addressed after they occur, the system is often forced to take drastic measures (e.g., stalling the pipeline completely) to dissipate accumulated heat rapidly. Therefore, RDTM typically leads to significant throughput fluctuations and severe performance degradation, making it less ideal for latency-sensitive AI workloads.

To mitigate the performance penalties of reactive schemes, Proactive DTM (PDTM) is used to prevent systems from overheating early. PDTM leverages temperature prediction models to forecast future temperature trends based on current traffic patterns and historical thermal data. By anticipating a potential hotspot before it breaches the safety limit, the controller can apply gentler, fine-grained adjustments (e.g., slightly reducing the injection rate or pre-routing traffic) to smooth out the thermal profile with minimal performance impact. Early PDTM works focused on physical models; for instance, Chen et al. proposed an RC-based temperature prediction scheme [51]. By modeling the thermal correlations between adjacent nodes as an RC circuit, this method effectively predicts future temperatures in 3D NoCs, allowing for timely flow control that avoids the heavy penalty of full throttling. More recently, with the increasing complexity of workload behaviors, data-driven approaches have gained prominence. Chen et al. in [52] introduced an adaptive Machine Learning-based thermal management framework. Unlike static predictors, this approach utilizes online learning to adaptively update its prediction model in response to varying workload characteristics, thereby achieving superior thermal compliance and throughput compared to traditional look-ahead policies.

C. DATA PLANE: TRAFFIC- AND THERMAL-AWARE ROUTING

As discussed in the previous section, the Control Plane employs DTM mechanisms (e.g., throttling or power gating) to regulate temperature. These thermal control actions dynamically alter the availability or performance of network nodes, effectively transforming the physical regular mesh into a logically *Non-Stationary Irregular Mesh (NSI-mesh)* [53]. In an NSI-mesh, the network topology and node status change rapidly over time due to the on-off cycling of thermal management. This dynamic irregularity poses significant challenges for traditional routing algorithms, which typically assume a static topology. A standard routing algorithm unaware of these thermal-induced changes may route packets into throttled regions, leading to severe congestion, increased latency, or even deadlocks. To address these challenges, routing strategies in thermal-aware NoCs have evolved through three distinct paradigms: 1) Traffic-aware routing, 2) Thermal-aware routing, and 3) Joint Traffic- and Thermal-aware routing.

Traditional traffic-aware adaptive routing algorithms focus solely on network congestion status (e.g., buffer occupancy) to make routing decisions. The primary goal of algorithms in this category is to guarantee packet delivery and minimize latency by navigating around throttled (disabled) nodes, relying primarily on traffic and topology status rather than direct temperature values. Chen et al. proposed a Topology-Aware Adaptive Routing (TAAR) in [53]. TAAR dynamically adjusts its routing mode based on real-time topology availability and regional traffic information. It seeks to increase path diversity to bypass throttled regions. However, TAAR predominantly favors non-congested minimal routing regions to maintain performance. Consequently, when multiple flows are forced to detour around a thermal hotspot, they often converge onto the few remaining available minimal paths. To address this problem, Chen et al. subsequently proposed a Traffic- and Thermal-aware Adaptive Beltway Routing (TTABR) [54]. TTABR introduces the concept of "beltway paths" (non-minimal paths) to significantly expand lateral path diversity. By assessing current traffic conditions, the algorithm dynamically selects between minimal routing paths and non-minimal beltway paths, effectively spreading traffic loads away from the central congested zones. However, TTABR lacks a direct mechanism to sense the temperature of the beltway nodes. As a result, it cannot prevent packets from being routed through potential thermal hotspots that are currently low in traffic but high in temperature, posing a risk to thermal safety.

To actively mitigate thermal hotspots, the thermal-aware routing prioritizes thermal metrics over traffic metrics. A representative approach is Dynamic Thermal-Balance Routing (DTBR), proposed by Liu et al. [55]. DTBR aims to homogenize the thermal profile across the chip by adaptively steering packets toward regions with lower temperatures based on neighbor monitoring. While DTBR successfully flattens thermal gradients, it suffers from a critical synchronization mismatch (i.e., the thermal time constant of silicon is significantly larger (slower) than the packet transmission rate). Consequently, the routing path remains fixed during a thermal sensing period, causing packets to continuously converge on a cool region until it becomes heavily congested. This cool-spot congestion phenomenon occurs because the algorithm lacks real-time awareness of rapid traffic fluctuations. On the other hand, to address the need for a more predictive metric, Kuo et al. [14] introduced a Proactive Thermal-Budget-Based Beltway Routing (PTB^{3}R). This method employs a novel metric called Mean Time to Throttle (MTTT), which quantifies the remaining active time of a node before it hits the alarm threshold. MTTT is derived from the margin between the current temperature and the critical limit, divided by the temperature consumption rate. PTB^{3}R prioritizes paths with

TABLE 2. Qualitative comparison of adaptive routing algorithms in thermal-aware NoCs.

Category	Algorithm	Decision Metric	Routing Strategy	Path Diversity	Pros (+) / Cons (-)
Traffic- and Topology-Aware	TAAR [54]	Traffic + Topology	Adaptive minimal routing	Low	(+) Guarantees delivery in NSI-mesh (-) Congestion in minimal routing regions
	TTABR [55]	Traffic + Topology	Adaptive beltway routing	High	(+) Relaxes minimal path constraint (-) Blind to thermal hotspots
Thermal-Centric	DTBR [56]	Neighbor temperature	Gradient-based	Medium	(+) Flattens thermal distribution (-) Cool-spot congestion (slow thermal response)
	PTB3R [14]	Thermal budget	Beltway routing	High	(+) Proactive use of thermal headroom (-) Historical power $\neq$ Instant traffic status
Joint Traffic- and Thermal-Aware	GTDAR [57]	Game utility (Delay + Temperature)	Game-Theoretic	High	(+) Pareto-optimal trade-off (Delay vs. Temperature) (-) High computational complexity

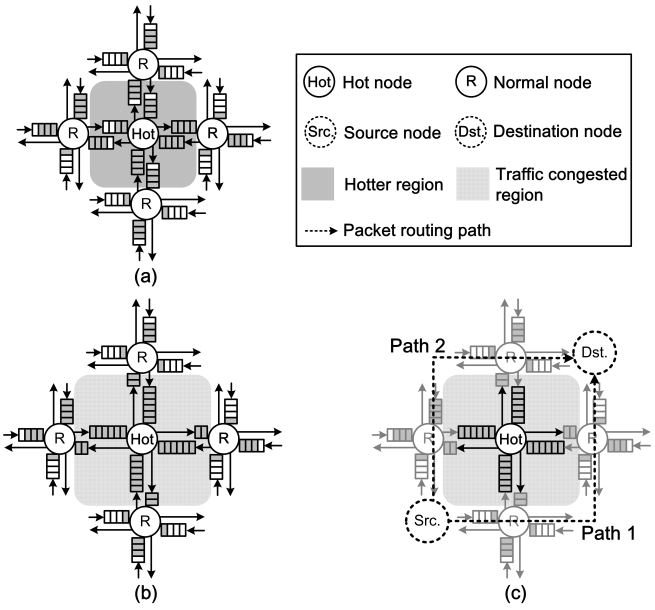


FIGURE 7. The control flow of GTDAR. (a) When a hot node occurs, (b) the GTDAR first adjust the buffer length around the hot node to synchronize the information on thermal and traffic, which (c) distributes the packet delivery to increase the routing path diversity [56].

larger MTTT values, effectively routing traffic through nodes that have a higher thermal buffer. However, a fundamental limitation persists, which is the temperature consumption rate utilized in MTTT is essentially a reflection of historical power usage rather than instantaneous traffic status. As a result, PTB^{3}R may inadvertently guide packets into currently congested regions that historically had low activity (and thus high MTTT), thereby worsening network contention. These examples highlight that relying solely on thermal sensors or historical thermal trends is insufficient for maintaining high throughput in bursty AI workloads.

The state-of-the-art solution lies in the co-optimization of both metrics. These algorithms utilize a composite cost function, typically a weighted sum of thermal stress and buffer occupancy, to select the optimal path. Chen et al. [53] proposed a Topology-Aware Adaptive Routing (TAAR) scheme specifically designed for NSI-mesh environments. TAAR broadcasts the throttling status of nodes to maintain a consistent view of the irregular topology, allowing the routing logic to bypass throttled regions proactively while balancing traffic load. Moving beyond heuristic approaches, recent research has adopted advanced mathematical models to solve this multi-objective optimization problem. For instance, Chen et al. introduced a Game-based Thermal-Delay-aware Routing (GTDAR) strategy [56], as shown in FIGURE 7. By modeling the routing decision as a game between traffic agents competing for network resources, GTDAR effectively balances the trade-off between thermal reliability

(avoiding hotspots) and communication performance (minimizing delay), demonstrating superior adaptability in complex 3D NoC systems compared to single-objective approaches.

To provide a clear overview of the evolution and trade-offs of these routing mechanisms, TABLE 2 presents a qualitative comparison. As observed, purely traffic-aware schemes (e.g., TAAR, TTABR) excel at handling congestion but risk violating thermal constraints. Conversely, purely thermal-centric approaches (e.g., DTBR, PTB3R) effectively manage heat distribution but often induce performance penalties due to ''cool-spot congestion'' or synchronization mismatches. Joint optimization strategies like GTDAR represent the state-of-the-art, effectively balancing these conflicting objectives, albeit at the cost of higher design complexity.

IV. THERMAL SIMULATION TOOLS

Thermal simulation is indispensable for evaluating the reliability and efficiency of NoC-based systems. While traffic simulators provide insights into communication bottlenecks, only thermal models can reveal whether such bottlenecks translate into hotspots or thermal emergencies under realistic workloads. A range of tools has been developed to address this challenge, spanning from compact RC-based solvers to fully integrated NoC+thermal frameworks and emerging machine learning surrogates.

A. HotSpot AND COMPACT RC-BASED SOLVERS

One of the most widely adopted thermal tools in computer architecture research is HotSpot [57]. HotSpot models the chip as a network of thermal resistances and capacitances (i.e., RC network) corresponding to functional blocks and heat flow paths. Given power traces as input, it computes both steady-state and transient temperature distributions, producing fine-grained thermal maps. Its strengths include accuracy, ease of integration, and support for multi-layer 3D ICs through extensions in recent versions. However, HotSpot requires external power data generated by performance simulators or power models such as McPAT [58] or ORION3.0 [59], and it does not natively capture the bidirectional coupling between traffic activity and temperature. As such, HotSpot is often used as a post-processing step: NoC simulators generate router-level power traces, which are then fed into HotSpot to obtain temperature profiles, as shown in FIGURE 8. This modular approach is flexible but can be computationally expensive for large design-space studies, motivating tighter integration of thermal solvers into NoC frameworks.

In addition to HotSpot, newer compact solvers such as TIMiTIC [60] support fast design-space exploration of 3D ICs with integrated microchannels for cooling, enabling quicker assessment of cooling designs with good thermal fidelity. Likewise, ATSim3D [61] improves accuracy in 3D-IC thermal simulation by handling nonlinear thermal conductivity and leakage power, using a hybrid global-local model to reduce runtime while maintaining $<3\%$ error relative to full-fidelity solvers.

B. INTEGRATED NoC+THERMAL SIMULATORS

To overcome the limitations of decoupled flows, researchers have developed simulators that combine cycle-accurate NoC models with power and thermal engines. A representative example is PAT-Noxim [62], an extension of the open-source Noxim simulator [63], which incorporates router-level power estimation and couples it with HotSpot-like thermal analysis. This integration allows designers to evaluate latency, throughput, energy, and temperature within the same framework, making it possible to study thermal-aware routing and mapping algorithms directly. Similarly, extensions to BookSim2.0 [29] integrate power and thermal modeling with cycle-accurate traffic simulation, enabling evaluation of 3D NoC architectures under combined performance, power, and thermal constraints [64]. These integrated tools highlight how localized traffic surges can translate into hotspots in specific routers or layers, providing insight that standalone thermal or traffic simulators cannot deliver.

Integrated simulators are also evolving to include runtime thermal management policies. For example, Chen et al. [52] propose an adaptive perceptron model plus reinforcement learning to predict future temperature and pick throttling ratios in a live NoC, thereby closing the loop between traffic, temperature, and control policies.

The main limitation of these integrated frameworks is their scalability. As NoCs expand to thousands of nodes or chiplet-based architectures, cycle-accurate co-simulation of power, thermal, and traffic can become prohibitively slow. This challenge has sparked interest in surrogate models that approximate thermal behavior without invoking full RC solvers at every iteration.

C. CASE STUDY: END-TO-END THERMAL PROFILING OF CNN VS. TRANSFORMER

To demonstrate the utility of the proposed toolchain and to elucidate the complex relationship between network traffic and on-chip temperature, we conducted a comparative case study. We mapped two distinct workloads. which are ResNet-50 CNN and BERT Transformer, onto an 8×8 Mesh NoC-based accelerator. The ResNet represents spatial convolution workloads; the BERT represents temporal attention-based workloads. The power traces generated by DNN-Noxim were fed into HotSpot (grid model) to produce transient temperature profiles. The simulation workflow integrates DNN-Noxim for traffic generation and HotSpot for thermal modeling.

FIGURE 9(a) and FIGURE 9(b) illustrate the normalized traffic density for ResNet-50 and BERT, respectively. The traffic distribution of ResNet-50 is relatively dispersed. Due to the sliding-window nature of convolution operations and efficient spatial mapping, the communication load is distributed across the processing elements (PEs),

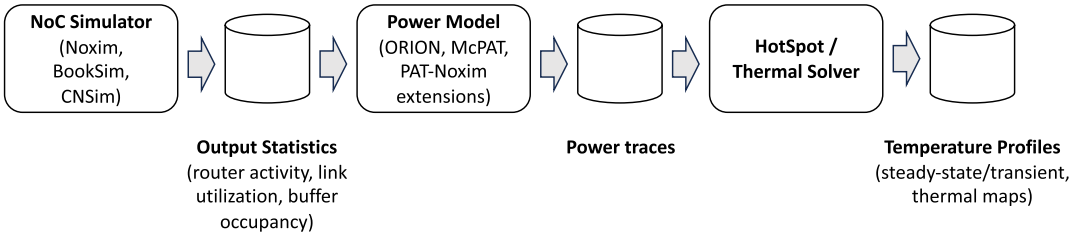


FIGURE 8. Example of workflow for NoC-thermal co-simulation.

avoiding severe congestion in any single region. In contrast, the Transformer workload exhibits a highly concentrated traffic pattern. The attention mechanism involves heavy matrix multiplications that require extensive all-to-all communication, leading to significant traffic congestion in the central routers of the mesh. FIGURE 9(c) and FIGURE 9(d) present the corresponding steady-state thermal maps for ResNet-50 and BERT, respectively. According to ResNet-50, consistent with its traffic profile, the thermal map shows a uniform temperature distribution with no critical hotspots, indicating effective load balancing. The thermal profile of BERT reveals a prominent hotspot in the center of the NoC platform, with peak temperatures significantly higher than the peripheral regions. This aligns with the high power density observed in the computation-heavy attention layers.

A critical observation from comparing the traffic maps with the thermal maps is the non-linear correlation between traffic volume and temperature. In other words, the locations of peak traffic do not always perfectly align with the locations of peak temperature. This discrepancy stems from the fundamental physical difference between network traffic and thermal dynamics. Traffic is often a transient phenomenon. A router may experience an extremely high burst of packets for a short duration (e.g., during a specific synchronization phase). However, if this burst is short-lived, it may not generate enough sustained heat to raise the local temperature significantly due to the thermal time constant of the silicon. On the other hand, temperature is an accumulative manifestation of energy. It represents the integration of power consumption over time. A region will only become a thermal hotspot if it experiences sustained high activity (e.g., traffic or computation) that exceeds the heat dissipation capability of the package.

With these analysis above, we can find that focusing solely on instantaneous traffic counters for thermal management is insufficient, as it may lead to false alarms from throttling due to short traffic bursts that are thermally safe. Conversely, relying solely on temperature sensors introduces a detection lag due to thermal inertia. Therefore, effective thermal-aware NoC design requires a joint traffic-thermal co-analysis. Advanced routing algorithms or management techniques must consider both the instantaneous traffic pressure to prevent congestion and the historical power accumulation to prevent hotspots.

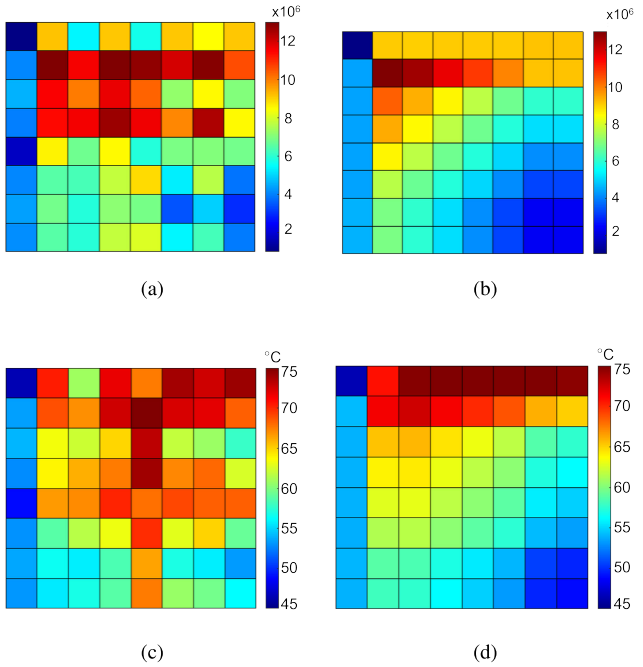


FIGURE 9. Comparison of (a) ResNet-50 traffic distribution, (b) BERT traffic distribution, (c) ResNet-50 temperature distributions, and (d) BERT temperature distribution for CNN and Transformer workloads on an 8 x 8 Mesh NoC.

V. EMERGING TRENDS AND OPEN CHALLENGES

A. THERMAL CHALLENGES IN EMERGING CHIPLET-BASED INTEGRATION

Chiplet-based architectures are rapidly emerging as a scalable alternative to monolithic SoCs, enabling higher integration density, heterogeneous technology integration, and improved yield for large AI accelerators [19], [65]. Instead of consolidating all compute and memory resources on a single die, chiplet systems distribute processing, memory, and I/O across multiple smaller dies interconnected using 2.5D interposers or 3D stacking. This architectural shift is increasingly important for AI and edge workloads where multimodal sensing, near-sensor inference, and heterogeneous accelerators must coexist under strict performance, thermal, and energy constraints. In these environments, the conventional NoC naturally extends to a multi-die Network-on-Package (NoP), adding new dimensions of communication, mapping, and thermal behavior that cannot be captured by single-die models.

A first emerging consideration is workload partitioning across chiplets [66], especially for multimodal sensing pipelines where vision, audio, tactile, and inertial data may be processed on separate dies. Unlike traditional on-chip partitioning, chiplet-level partitioning must jointly consider cross-die link bandwidth, latency, and power density, as well as the thermal implications of concentrating workloads in vertically or laterally adjacent dies. These challenges highlight the need for mapping and scheduling strategies that explicitly model chiplet locality and die-to-die constraints.

Chiplet-based systems also introduce nonuniform and asymmetric vertical thermal coupling due to variation in die thickness, interposer materials, and distance to the heat sink [67]. Upper-layer chiplets tend to exhibit poorer thermal dissipation paths, and hotspots in one die can propagate into neighboring dies, creating strongly coupled thermal behaviors. Consequently, DTM, routing policies, and predictive thermal control must account for package-level heat flow rather than treating each die as thermally independent.

The communication fabric further complicates system behavior. Die-to-die interfaces typically offer lower bandwidth, higher latency, and higher energy per bit compared to on-die NoCs, creating communication bottlenecks for real-time or monitoring-intensive workloads. These bottlenecks become more severe when thermal throttling affects specific chiplets or interconnect channels. This interplay between traffic, thermal constraints, and package topology motivates new routing, flow-control, and QoS mechanisms tailored to multi-chiplet environments [68], [69].

Finally, chiplet-based accelerators are increasingly deployed in safety-critical domains such as autonomous driving, medical imaging, and industrial robotics. These applications require reliable chiplet-to-chiplet NoC/NoP connectivity with fault detection, redundancy, thermal anomaly awareness, and guaranteed timing even under partial failures or thermal stress. Ensuring predictable and certifiable behavior in such heterogeneous, strongly coupled systems remains an open research challenge.

Taken together, these trends motivate the need to revisit thermal-aware NoC research from a broader, chiplet- and package-level perspective. The remainder of this section builds on this motivation by outlining (i) emerging trends that shape chiplet-aware NoC and thermal co-design, (ii) open challenges that limit current solutions, and (iii) a research agenda that integrates workload, thermal, and communication considerations across dies, packages, and system layers.

B. EMERGING TRENDS

One visible trend is the growing role of machine learning surrogates for physics-based solvers. Compact RC models such as HotSpot remain widely used, but recent graph neural network and operator-learning approaches demonstrate that accurate temperature fields can be predicted orders of magnitude faster, with acceptable error bounds [61], [70], [71], [72]. These surrogates are increasingly adopted as intermediate fidelity layers in co-simulation workflows, enabling design-space exploration and even online thermal control loops that would be infeasible with purely physics-based solvers.

A second trend concerns scalability in the era of chiplets and multi-die integration. Packet-parallel NoC simulators and hierarchical models are now used to handle the thousands of routers and heterogeneous links present in large chiplet fabrics [28], [73], [74]. These tools are often paired with realistic AI workloads, moving beyond synthetic traffic patterns toward traces extracted from DNN execution.

Another clear direction is the shift toward tighter integration and closed-loop co-simulation. Instead of the traditional trace-driven workflow, modern frameworks increasingly allow thermal estimates to feed directly back into routing algorithms, DVFS schemes, or throttling policies during the same run. For instance, there are platforms with traffic-thermal mutual coupling in 3D NoCs [75], and control-studies like ControlPULP which embed thermal feedback into power management loops in many-core systems [76]. Hybrid strategies, where fast surrogates provide online feedback while high-fidelity solvers are called occasionally for correction, are becoming a pragmatic compromise between accuracy and runtime.

The workload landscape itself is also changing. Whereas CNN benchmarks once dominated, transformers and LLMs now drive system design, bringing communication patterns that include bursty attention exchanges and collective synchronization. Characterization studies for these models are beginning to appear, and simulators are being adapted to capture their unique traffic signatures [77], [78].

Finally, tool interoperability is emerging as a community priority. Researchers are moving away from ad hoc integration scripts toward modular ecosystems in which NoC simulators, power/thermal engines, and workload generators interoperate through standardized interfaces. For example, ATLAHS [79] provides a multi-backend simulation toolchain that accepts real application traces and supports multiple simulation engines; CoMeT [80] offers APIs for configuring workloads, thermal policies, and DVFS/migration policies in a plug-and-play fashion; and Ratatoskr [81] integrates multiple abstraction levels under a common interface for PPA (power, performance, area) studies. This modularity promises more reproducible studies and easier sharing of workloads, models, and results across research groups.

C. OPEN CHALLENGES

Despite this progress, several obstacles limit current practice. A persistent difficulty is the balance between accuracy and scalability. Full-cycle, tightly coupled co-simulation across thousands of routers remains prohibitively slow, while surrogates, though fast, can fail when applied outside their training distribution. Developing robust hybrid approaches with built-in uncertainty estimation is still an open research frontier.

Another challenge lies in temporal and spatial coupling. NoC simulators produce activity at cycle-level granularity, while thermal solvers advance on millisecond scales. If activity traces are aggregated too coarsely, peak hotspots are smoothed out; if kept too fine, runtimes explode. Designing principled coarsening strategies that preserve thermal relevance remains a key issue.

The lack of standardized datasets and interfaces is a limitation. Unlike CPU/GPU benchmarking, there is no widely accepted corpus of AI traffic traces combined with floorplans, package parameters, and reference thermal stacks. Because of this, results across groups are difficult to compare, and tool evaluation often depends on private or synthetic traces.

Packaging effects further complicate matters. Thermal coupling across stacked dies, interposer impedance, and cooling solutions critically affect temperature distributions but are frequently simplified. Without validated parameters, simulation studies risk drifting away from real silicon behavior. The absence of systematic validation platforms exacerbates this issue, since most academic groups cannot access prototype chiplets for measurement.

Finally, the role of thermal simulation in reliability and security is underexplored. Temperature gradients not only affect performance but can leak information or accelerate wear-out. Integrating anomaly detection, covert-channel analysis, and security-aware thermal modeling into NoC simulation flows is a nascent but important area [82].

D. DIRECTIONS FOR THE RESEARCH AGENDA

Several directions stand out as particularly impactful. First, the community would benefit from benchmarking infrastructures: standardized DNN and LLM traces, floorplans, and package descriptions distributed in open repositories. These datasets should become the basis for evaluating thermal- and workload-aware NoC simulators, in the same way that SPEC or MLPerf serve other communities. For example, ONNXim [83] uses the ONNX model format to enable cross-framework workload evaluation; SIAM [73] provides chiplet + NoC + memory interconnect benchmarks including mesh/tree layouts; and more recent efforts like LLMServingSim [84] offer a co-simulation infrastructure with flexible back-ends, which could form parts of such standardized platforms.

Second, investment in robust surrogate methodologies is needed. Hybrid pipelines where surrogates flag low-confidence cases and trigger targeted physics-based simulations represent a promising balance between speed and fidelity (e.g. DeepOHeat-v1 [70], ARO [85]). Active learning loops, in which surrogates improve progressively as new workloads are explored, would further increase reliability, which can be seen in ARO's multi-fidelity and active learning strategies, and surrogate-assisted placement in [86].

Third, the development of APIs and interchange formats for traces, power maps, and floorplans would accelerate interoperability. With standard interfaces, researchers could swap one simulator or solver for another without rewriting integration code, lowering the barrier to cross-validation and reproducibility. The Unified Power Model (UPM) standard from Si2 provides a semantics for exchanging power models among IP, system, and EDA tool providers [87], while tools like NISTT enable standardized non-intrusive trace capture in SystemC-TLM flows [88], and Netrace supplies a common format and library for NoC message trace/replay [89]. These efforts show that parts of the stack are already moving toward modular, standardized components.

Another important step is to build cross-layer evaluation suites that measure not only thermal and performance metrics in isolation but also the impact of runtime policies. For example, design studies should quantify how much throughput is sacrificed when hotspot duration must be capped, or how effective thermal-aware routing is under transformer workloads. Prior work such as [90] has demonstrated performance penalties when scheduling tasks under thermal constraints. Similarly, adaptive routing strategies like the Q-function-based traffic-and-thermal-aware routing in 3D NoCs show throughput or latency trade-offs when avoiding hotspots dynamically [91], and works on hotspot reduction via packet-reordering show that performance gains can come at the cost of added routing complexity or latency under certain traffic patterns [92].

Finally, open measurement and emulation platforms would be invaluable. FPGA-based emulators or test boards with controllable heaters could provide reference traces, enabling researchers to validate their simulators without needing access to full-scale industrial prototypes. By combining such measurement with open datasets and interoperable tools, the community can move toward a reproducible and credible foundation for future NoC and AI accelerator research.

VI. CONCLUSIONS AND FUTURE WORK

This survey has provided a holistic review of thermal-aware NoC design for AI computing, integrating architectural innovations, algorithmic strategies, and simulation methodologies. We analyzed how flexible NoC architectures enable adaptive dataflow, routing, and mapping across diverse AI workloads such as CNNs, GNNs, and transformers. Furthermore, we investigate cross-layer thermal management from different research angles, including spanning sensing, control, and routing planes, to ensure sustained performance and reliability. The unified toolchains combining DNN-aware NoC simulators, power/thermal solvers, and ML-based surrogates have demonstrated strong potential for accurate and scalable co-design exploration. Future work will emphasize hybrid physics- and learning-based thermal models, packaging- and reliability-aware evaluation, and standardized datasets for reproducible benchmarking. Moreover, extending thermal-aware NoC principles to chiplet- and edge-AI architectures will be key to achieving energy-efficient, secure, and thermally resilient next-generation AI accelerators.

REFERENCES

[1] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[2] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2018, pp. 1–8.

[3] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and benchmarking of machine learning accelerators," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2019, pp. 1–9.

[4] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey of machine learning accelerators," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2020, pp. 1–12.

[5] E. Wang et al., "Deep neural network approximation for custom hardware: Where we've been, where we're going," *ACM Comput. Surv.*, vol. 52, no. 2, pp. 1–39, May 2019.

[6] V. Sze, Y. Chen, T.-J. Yang, and J. Emer, *Efficient Processing of Deep Neural Networks* (Synthesis Lectures on Computer Architecture). San Rafael, CA, USA: Morgan & Claypool, 2020.

[7] S. Bodiwala and N. Nanavati, "Efficient hardware implementations of deep neural networks: A survey," in *Proc. 4th Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2020, pp. 31–36.

[8] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A survey of accelerator architectures for deep neural networks," *Engineering*, vol. 6, no. 3, pp. 264–274, Mar. 2020.

[9] S. M. Nabavinejad, M. Baharloo, K.-C. Chen, M. Palesi, T. Kogel, and M. Ebrahimi, "An overview of efficient interconnection networks for deep neural network accelerators," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 10, no. 3, pp. 268–282, Sep. 2020.

[10] G. Akkad, A. Mansour, and E. Inaty, "Embedded deep learning accelerators: A survey on recent advances," *IEEE Trans. Artif. Intell.*, vol. 5, no. 5, pp. 1954–1972, May 2024.

[11] C. Silvano et al., "A survey on deep learning hardware accelerators for heterogeneous HPC platforms," *ACM Comput. Surv.*, vol. 57, no. 11, pp. 1–39, Nov. 2025.

[12] L. Chen, W. Jin, and S. X.-D. Tan, "Fast thermal analysis for chiplet design based on graph convolution networks," in *Proc. 27th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Salt Lake City, UT, USA, Jan. 2022, pp. 485–492.

[13] H. Liu, X. Chen, Y. Zhao, C. Li, and J. Lu, "TTQR: A Traffic- and thermal-aware Q-routing for 3D network-on-chip," *Sensors*, vol. 22, no. 22, p. 8721, Nov. 2022.

[14] 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," in *Proc. Int. Symp. Syst. Chip (SoC)*, Oct. 2013, pp. 1–4.

[15] S. Pandey and P. R. Panda, "NeuroTAP: Thermal and memory access pattern-aware data mapping on 3D DRAM for maximizing DNN performance," *ACM Trans. Embedded Comput. Syst.*, vol. 23, no. 6, pp. 1–30, Sep. 2024.

[16] M. Maatar, K. N. Dang, and A. B. Abdallah, "Thermal-aware task-mapping method for 3D-NoC-based neuromorphic systems," in *Proc. 6th Int. Conf. Electron. Technol. (ICET)*, May 2023, pp. 1067–1076.

[17] S. Rahimipour, W. N. Flayyih, N. A. Kamsani, S. J. Hashim, M. R. Stan, and F. Z. B. Rokhani, "Low-power, highly reliable dynamic thermal management by exploiting approximate computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 10, pp. 2210–2222, Oct. 2020.

[18] Y.-H. Chen, T.-J. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 292–308, Jun. 2019.

[19] Y. S. Shao et al., "Simba: scaling deep-learning inference with chiplet-based architecture," *Commun. ACM*, vol. 64, no. 6, pp. 107–116, 2021.

[20] H. Kwon, A. Samajdar, and T. Krishna, "MAERI: Enabling flexible dataflow mapping over DNN accelerators via reconfigurable interconnects," in *Proc. 23rd Int. Conf. Architectural Support Program. Lang. Operating Syst. (ASPLOS)*, 2018, pp. 461–475.

[21] K.-C. Chen and Y.-S. Liao, "A reconfigurable deep neural network on chip design with flexible convolutional operations," in *Proc. 15th IEEE/ACM Int. Workshop Netw. Chip Architectures (NoCArc)*, Oct. 2022, pp. 1–5.

[22] K.-C. Chen, Y.-S. Liao, W.-R. Syu, H.-C. Chen, and P.-C. Shen, "An arbitrary kernel size applicable universal flexible-oriented network on chip (UFONoC) neural network accelerator with AXI4-stream data transmission protocol," in *Proc. IEEE Int. Conf. Omni-Layer Intell. Syst. (COINS)*, Aug. 2025, pp. 1–6.

[23] K.-C.-J. Chen, H.-H. Peng, and P.-C. Shen, "Ultra-NoC: Unified low-transmission routing assisted NoC for high-flexible DNN accelerator," in *Proc. IEEE 37th Int. System-on-Chip Conf. (SOCC)*, Sep. 2024, pp. 1–5.

[24] K.-C. Chen, C.-K. Tsai, Y.-S. Liao, H.-B. Xu, and M. Ebrahimi, "A lego-based neural network design methodology with flexible NoC," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 11, no. 4, pp. 711–724, Dec. 2021.

[25] K.-C. J. Chen, M. Ebrahimi, T.-Y. Wang, Y.-C. Yang, and Y.-H. Liao, "A NoC-based simulator for design and evaluation of deep neural networks," *Microprocessors Microsyst.*, vol. 77, Sep. 2020, Art. no. 103145.

[26] A. Parashar et al., "Timeloop: A systematic approach to DNN accelerator evaluation," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Mar. 2019, pp. 304–315.

[27] Y. S. Shao et al., "Simba: Scaling deep-learning inference with multi-chip-module-based architecture," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 14–27.

[28] Y. Feng, Y. Wei, D. Xiang, and K. Ma, "Evaluating chiplet-based large-scale interconnection networks via cycle-accurate packet-parallel simulation," in *Proc. USENIX Conf. Usenix Annu. Tech. Conf.*, 2024, pp. 731–747.

[29] N. Jiang et al., "A detailed and flexible cycle-accurate network-on-chip simulator," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Apr. 2013, pp. 86–96.

[30] N. Agarwal, T. Krishna, L.-S. Peh, and N. K. Jha, "GARNET: A detailed on-chip network model inside a full-system simulator," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw.*, Apr. 2009, pp. 33–42.

[31] H. Chu, X. Zheng, L. Liu, and H. Ma, "NnPerf: Demystifying DNN runtime inference latency on mobile platforms," in *Proc. 21st ACM Conf. Embedded Networked Sensor Syst.*, Nov. 2023, pp. 125–137.

[32] M. Musavi, E. Irabor, A. Das, E. Alarcón, and S. Abadal, "Communication characterization of AI workloads for large-scale multi-chiplet accelerators," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2025, pp. 1–5.

[33] Y. Zhuang et al., "On optimizing the communication of model parallelism," in *Proc. Mach. Learn. Syst.*, 2022, pp. 526–540.

[34] H. Wang et al., "Towards domain-specific network transport for distributed dnn training," in *Proc. 21st USENIX Symp. Networked Syst. Design Implement.*, 2024, pp. 1421–1443.

[35] B. Hanindhito, B. Patel, and L. K. John, "Bandwidth characterization of DeepSpeed on distributed large language model training," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, May 2024, pp. 241–256.

[36] L. Xu, K. K. Suresh, Q. Anthony, N. Alnaasan, and D. K. Panda, "Characterizing communication patterns in distributed large language model inference," in *Proc. IEEE Symp. High-Performance Interconnects (HOTI)*, Aug. 2025, pp. 1–11.

[37] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2017, pp. 1–12.

[38] A. Parashar et al., "SCNN: An accelerator for compressed-sparse convolutional neural networks," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2017, pp. 27–40.

[39] Y. Ouyang, J. Wang, C. Sun, Q. Wang, and H. Liang, "URMP: Using reconfigurable multicast path for NoC-based deep neural network accelerators," *J. Supercomput.*, vol. 79, no. 13, pp. 14827–14847, Apr. 2023.

[40] M. Ebrahimi, M. Daneshtalab, P. Liljeberg, J. Plosila, J. Flich, and H. Tenhunen, "Path-based partitioning methods for 3D networks-on-chip with minimal adaptive routing," *IEEE Trans. Comput.*, vol. 63, no. 3, pp. 718–733, Mar. 2014.

[41] A. Das, E. Russo, and M. Palesi, "Multi-objective hardware-mapping co-optimisation for multi-DNN workloads on chiplet-based accelerators," *IEEE Trans. Comput.*, vol. 73, no. 8, pp. 1883–1898, Aug. 2024.

[42] H. Kwon, P. Chatarasi, M. Pellauer, A. Parashar, V. Sarkar, and T. Krishna, "Understanding reuse, performance, and hardware cost of DNN dataflow: A data-centric approach," in *Proc. 52nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Oct. 2019, pp. 754–768.

[43] G. Krishnan, S. K. Mandal, C. Chakrabarti, J.-S. Seo, U. Y. Ogras, and Y. Cao, "Impact of on-chip interconnect on in-memory acceleration of deep neural networks," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 18, no. 2, pp. 1–22, Apr. 2022.

[44] Q. Anthony et al., "Demystifying the communication characteristics for distributed transformer models," in *Proc. IEEE Symp. High-Perform. Interconnects (HOTI)*, Aug. 2024, pp. 57–65.

[45] G. Liu, M. Fan, and G. Quan, "Neighbor-aware dynamic thermal management for multi-core platform," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2012, pp. 187–192.

[46] 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.

[47] K.-C. Chen, H.-W. Tang, C.-H. Wu, and C.-H. Chen, "Thermal sensor placement for multicore systems based on low-complex compressive sensing theory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 11, pp. 5100–5111, Nov. 2022.

[48] K.-C. Chen, C.-H. Chen, L.-Q. Wang, and C.-C. Wang, "Entropy-based thermal sensor placement and temperature reconstruction based on adaptive compressive sensing theory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, early access, Oct. 28, 2025, doi: 10.1109/TCAD.2025.3626515.

[49] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *Proc. 27th Int. Symp. Comput. Archit.*, Mar. 2000, pp. 83–94.

[50] L. Shang, L. Peh, A. Kumar, and N. K. Jha, "Thermal modeling, characterization and management of on-chip networks," in *Proc. 37th Int. Symp. Microarchitecture (MICRO)*, 2004, pp. 67–78.

[51] K.-C. Chen, E.-J. Chang, H.-T. Li, and A.-Y. Wu, "RC-based temperature prediction scheme for proactive dynamic thermal management in throttle-based 3D NoCs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 206–218, Jan. 2015.

[52] K.-C. Chen, Y.-H. Liao, C.-T. Chen, and L.-Q. Wang, "Adaptive machine learning-based proactive thermal management for NoC systems," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 31, no. 8, pp. 1114–1127, Aug. 2023.

[53] K.-C. Chen, S.-Y. Lin, H.-S. Hung, and A. 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.

[54] K.-C. Chen, C.-C. Kuo, H.-S. Hung, and A. A. Wu, "Traffic- and thermal-aware adaptive beltway routing for three dimensional network-on-chip systems," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 1660–1663.

[55] F. Liu, H. Gu, and Y. Yang, "DTBR: A dynamic thermal-balance routing algorithm for network-on-chip," *Comput. Electr. Eng.*, vol. 38, no. 2, pp. 270–281, Mar. 2012.

[56] K.-C. Chen, "Game-based thermal-delay-aware adaptive routing (GTDAR) for temperature-aware 3D network-on-chip systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 9, pp. 2018–2032, Sep. 2018.

[57] K. Skadron, M. R. Stan, K. Sankaranarayanan, W. Huang, S. Velusamy, and D. Tarjan, "Temperature-aware microarchitecture: Modeling and implementation," *ACM Trans. Archit. Code Optim.*, vol. 1, no. 1, pp. 94–125, Mar. 2004.

[58] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2009, pp. 469–480.

[59] A. B. Kahng, B. Lin, and S. Nath, "ORION3.0: A comprehensive NoC router estimation tool," *IEEE Embedded Syst. Lett.*, vol. 7, no. 2, pp. 41–45, Jun. 2015.

[60] P. Zajac, "TIMiTIC: A C++ based compact thermal simulator for 3D ICs with microchannel cooling," in *Proc. 25th Int. Workshop Thermal Investigations ICs Syst. (THERMINIC)*, Sep. 2019, pp. 1–6.

[61] Q. Wang, T. Zhu, Y. Lin, R. Wang, and R. Huang, "ATSim3D: Towards accurate thermal simulator for heterogeneous 3D-IC systems considering nonlinear leakage and conductivity," in *Proc. 2nd Int. Symp. Electron. Design Autom. (ISEDA)*, May 2024, pp. 618–623.

[62] A. Norollah, D. Derafshi, H. Beitollahi, and A. Patooghy, "PAT-noxim: A precise power & thermal cycle-accurate NoC simulator," in *Proc. 31st IEEE Int. Syst.-Chip Conf. (SOCC)*, Sep. 2018, pp. 163–168.

[63] V. Catania, A. Mineo, S. Monteleone, M. Palesi, and D. Patti, "Energy efficient transceiver in wireless network on chip architectures," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2016, pp. 1321–1326.

[64] B. Halavar, U. Pasupulety, and B. Talawar, "Extending BookSim2.0 and HotSpot6.0 for power, performance and thermal evaluation of 3D NoC architectures," *Simul. Model. Pract. Theory*, vol. 96, Nov. 2019, Art. no. 101929.

[65] B. Vinnakota, "The open domain-specific architecture: Next steps to production," in *Proc. 8th Annu. ACM Int. Conf. Nanosc. Comput. Commun.*, Sep. 2021, pp. 1–5.

[66] M. Odema, L. Chen, H. Kwon, and M. A. Al Faruque, "SCAR: Scheduling multi-model AI workloads on heterogeneous multi-chiplet module accelerators," in *Proc. 57th IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Nov. 2024, pp. 565–579.

[67] H. Sharma, J. Doppa, U. Ogras, and P. Pande, "Designing high-performance and thermally feasible multi-chiplet architectures enabled by non-bendable glass interposer," *ACM Trans. Embedded Comput. Syst.*, vol. 24, no. 5s, pp. 1–28, Sep. 2025.

[68] Y. Ma, L. Delshadtehrani, C. Demirkiran, J. L. Abellan, and A. Joshi, "TAP-2.5D: A thermally-aware chiplet placement methodology for 2.5D systems," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Feb. 2021, pp. 1246–1251.

[69] M. Palesi, E. Russo, A. Das, and J. Jose, "Wireless enabled inter-chiplet communication in DNN hardware accelerators," in *Proc. IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)*, May 2023, pp. 477–483.

[70] X. Yu et al., "DeepOHeat-v1: Efficient operator learning for fast and trustworthy thermal simulation and optimization in 3D-IC design," 2025, *arXiv:2504.03955*.

[71] D. Miao, G. Duan, D. Chen, Y. Zhu, and X. Zheng, "Real-time temperature prediction for large-scale multi-core chips based on graph convolutional neural networks," *Electronics*, vol. 14, no. 6, p. 1223, Mar. 2025.

[72] M. Hasanzadeh, K. Khalil, C. Sturton, and A. Patooghy, "HeatSense: Intelligent thermal anomaly detection for securing NoC-enabled MPSoCs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, early access, Jan. 22, 2026, doi: 10.1109/TCAD.2026.3657198.

[73] G. Krishnan et al., "SIAM: Chiplet-based scalable in-memory acceleration with mesh for deep neural networks," *ACM Trans. Embedded Comput. Syst.*, vol. 20, no. 5s, pp. 1–24, Sep. 2021.

[74] H. Zhi et al., "A methodology for simulating multi-chiplet systems using open-source simulators," in *Proc. Eight Annu. ACM Int. Conf. Nanosc. Comput. Commun.*, Sep. 2021, pp. 1–6.

[75] A.-Y.-A. Wu, K.-C.-J. Chen, and C.-H. Chao, "Thermal/traffic mutual-coupling co-simulation platform for 3D network-on-chip (NoC) designs," in *Proc. 10th Int. Workshop Netw. Chip Architectures*, Oct. 2017, pp. 1–2.

[76] A. Ottaviano et al., "ControlPULP: A RISC-V on-chip parallel power controller for many-core HPC processors with FPGA-based hardware-in-the-loop power and thermal emulation," *Int. J. Parallel Program.*, vol. 52, nos. 1–2, pp. 93–123, Feb. 2024.

[77] S. Cheng et al., "Thorough characterization and analysis of large transformer model training at-scale," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 52, no. 1, pp. 39–40, Jun. 2024.

[78] A. Sun et al., "BurstAttention: An efficient distributed attention framework for extremely long sequences," 2024, *arXiv:2403.09347*.

[79] S. Shen, T. Bonato, Z. Hu, P. Jordan, T. Chen, and T. Hoefler, "ATLAHS: An application-centric network simulator toolchain for AI, HPC, and distributed storage," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Nov. 2025, pp. 349–367.

[80] L. Siddhu et al., "CoMeT: An integrated interval thermal simulation toolchain for 2D, 2.5D, and 3D processor-memory systems," *ACM Trans. Archit. Code Optim.*, vol. 19, no. 3, pp. 1–25, Aug. 2022.

[81] J. M. Joseph, L. Bamberg, I. Hajjar, B. R. Perjikolaei, A. García-Ortiz, and T. Pionteck, "Ratatoskr: An open-source framework for in-depth power, performance, and area analysis and optimization in 3D NoCs," *ACM Trans. Model. Comput. Simul.*, vol. 32, no. 1, pp. 1–21, Sep. 2021.

[82] H. Huang, X. Wang, Y. Jiang, A. K. Singh, M. Yang, and L. Huang, "Detection of and countermeasure against thermal covert channel in many-core systems," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 41, no. 2, pp. 252–265, Feb. 2022.

[83] H. Ham et al., "ONNXim: A fast, cycle-level multi-core NPU simulator," *IEEE Comput. Archit. Lett.*, vol. 23, no. 2, pp. 219–222, Jul. 2024.

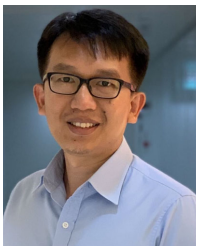
[84] J. Cho, M. Kim, H. Choi, G. Heo, and J. Park, "LLMServingSim: A HW/SW co-simulation infrastructure for LLM inference serving at scale," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Sep. 2024, pp. 15–29.

[85] M. Wang et al., "ARO: Autoregressive operator learning for transferable and multi-fidelity 3D-IC thermal analysis with active learning," in *Proc. 43rd IEEE/ACM Int. Conf. Comput.-Aided Design*, 2025, pp. 1–9.

[86] Q. Zhang, X. Liang, N. Xu, and Y. Chen, "Fast thermal-aware chiplet placement assisted by surrogate," 2025, *arXiv:2504.03808*.

[87] Silicon Integration Initiative (Si2). (2023). *Unified Power Model (UPM) V2.1*. Accessed: Sep. 27, 2025. [Online]. Available: https://si2.org/si2-openstandards/

[88] N. Bosbach, J. M. Joseph, R. Leupers, and L. Jünger, "NISTT: A non-intrusive SystemC-TLM 2.0 tracing tool," in *Proc. IFIP/IEEE 30th Int. Conf. Very Large Scale Integr. (VLSI-SoC)*, Oct. 2022, pp. 1–6.
[89] J. Hestness, B. Grot, and S. W. Keckler, "Netrace: dependency-driven trace-based network-on-chip simulation," in *Proc. 3rd Int. Workshop Netw. Chip Architectures*, Dec. 2010, pp. 31–36.
[90] Y. Lee, K. G. Shin, and H. S. Chwa, "Thermal-aware scheduling for integrated CPUs-GPU platforms," *ACM Trans. Embedded Comput. Syst.*, vol. 18, no. 5s, pp. 1–25, Oct. 2019.
[91] S. C. Lee and T. H. Han, "Q-function-based traffic- and thermal-aware adaptive routing for 3D network-on-chip," *Electronics*, vol. 9, no. 3, p. 392, Feb. 2020.
[92] Z. Li et al., "Hotspots reduction for GALS NoC using a low-latency multistage packet reordering approach," *Micromachines*, vol. 14, no. 2, p. 444, Feb. 2023.



FAKHRUL ZAMAN ROKHANI (Member, IEEE) was with STMicroelectronics Catania, Intel Penang Design Center, and Huawei Technologies as a Visiting Professor and a Visiting Scholar with the ASIC & Systems State Key Laboratory, Fudan University, China, and a Visiting Professor at several universities. He is currently an Associate Professor with the Faculty of Engineering, Universiti Putra Malaysia. His current research interests include low-power/energy-efficient system-on-chip (SoC) design and automation, IoT system integration, and sensors for food quality applications. He has (co)-authored more than 130 peer-reviewed articles, one country policy paper, nine books/book chapters, three magazine articles, and 25 intellectual properties. In 2023, he co-founded a startup company focusing on low-power IPs. He serves as the Vice President for Education and Communications for the CAS Society (2023–2026) and chairs the CASS Institute for Global Education. He is a Technical Committee Member of VLSI Systems and Applications and Circuits and Systems Education and Outreach. On the publication front, he is/has been contributing as a Senior Associate Editor of IEEE Transactions on Circuits and Systems—I: Regular Papers and CASS Newsletter, a Guest Editor of IEEE Transactions on Circuits and Systems—II: Express Briefs, and the technical program/publication chair/track chairs/embedded workshop chair of several IEEE CASS conferences. At the IEEE level, he serves on the steering committee for IEEE Future Direction—Global Semiconductors.

MIRCEA R. STAN (Fellow, IEEE) received the Diploma degree from Politehnica University, Bucharest, Romania, in 1984, and the M.S. and Ph.D. degrees from UMass Amherst in 1994 and 1996, respectively. He is currently teaching and doing research in the areas of AI hardware, processing in memory, cyber-physical systems, computational RFID, spintronics, and nanoelectronics. Since 1996, he has been with the ECE Department, UVa, where he is also the Virginia Microelectronics Consortium (VMEC) Endowed Chair. He received the 2024 A. Richard Newton Technical Impact Award in Electronic Design Automation and the 2018 Influential ISCA Paper Award and was a co-author on best paper awards at ISQED24, ASILOMAR19, LASCAS19, SELSE17, ISQED08, GLSVLSI06, ISCA03, and SHAMAN02. He is the Editor-in-Chief of IEEE Transactions on Very Large Scale Integration (VLSI) Systems; and a member of ACM, Eta Kappa Nu, Phi Kappa Phi, and Sigma Xi.

MAURIZIO PALESI (Senior Member, IEEE) is currently an Associate Professor with the Department of Electrical, Electronics, and Computer Engineering, University of Catania, Italy. His research focuses on manycore architectures, on-chip/in-package interconnection networks, domain-specific hardware accelerators, and quantum computing. He serves as the Steering Committee Chair for the International Workshop on Network on Chip Architectures (NoCArc) and has been the general chair of multiple workshops and conferences. Additionally, he is an Associate Editor of IEEE Transactions on Computers and *ACM Transactions on Design Automation of Electronic Systems* and has guest-edited over 20 special issues in leading academic journals. He is a member of ACM.

KUN-CHIH (JIMMY) CHEN (Senior Member, IEEE) is currently a Full Professor and an Electric Junior Chair Professor with the Institute of Electronics, National Yang Ming Chiao Tung University (NYCU). His research interests include multiprocessor SoC (MPSoC) design, neural network learning algorithm design, reliable system design, VLSI/CAD design, and smart manufacturing. He heads many services in IEEE Circuits and Systems Society, such as the IEEE Journal on Emerging and Selected Topics in Circuits and Systems Guest Editor and the General Chair of NoCArc 2020. He has received several prestigious national and international awards, including the Ta-You Wu Memorial Award of NSTC (i.e., Early Career Award in Taiwan), Chinese Institute of Electrical Engineering (CIEE) Outstanding Youth Electrical Engineer Award, Taiwan IC Design Society Outstanding Young Scholar Award, and IEEE Tainan Section Best Young Professional Member Award. Under his leadership, his research team became the first in Taiwan to win the IEEE ISCAS Best Student Paper Award and IEEE TVLSI Best Paper Award.