Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 2, 2026 (pp. 14-24)

RESEARCH ARTICLE ECEJOURNALS.IN

Thermal-Aware 3D TSV-Enabled NoC Topologies for High-Throughput SoC Accelerators

Kagaba J. Bosco[1], S. M Pavalam[2], L. J. Mpamije[3]

[1-3]Information and Communications Technology, National Institute of Statistics of Rwanda, Kigali, Rwanda

KEYWORDS:

Multi-Objective Genetic Algorithm (MOGA),
System-on-Chip (SoC) Accelerators,
Thermal Modeling,
Hotspot Mitigation,
High-Throughput Architectures,
Evolutionary Optimization.

ARTICLE HISTORY:

Submitted : 15.12.2025
Revised : 08.01.2026
Accepted : 14.02.2026

https://doi.org/10.31838/JIVCT/03.02.03

Abstract

Network-on-Chip (NoC) architectures in three-dimensional (3D) Technology using TSV have become a solution of high bandwidth and integration capability to address the needs of high-throughput System-on-Chip (SoC) accelerators. Vertical stacking however means a large amount of power density and thermal coupling resulting in intensive hotspots and poor reliability. The presented work proposes a thermal-conscious topology optimization of 3D NoCs using a Multi-Objective Genetic Algorithm (MOGA) to optimize and reduce the latency along with peaked temperature and maximize the network throughput with TSV and power constraints. The developed approach combined compact thermal modelling into the evolutionary search procedure and allowed simultaneous optimization of the horizontal connexions, vertical TSV location and routing design. Another strategy that can be used to overcome the localised overheating during runtime is the use of a thermal-aware adaptive routing strategy. Extensive simulations prove that the optimized topologies lead to significant cutting of peak temperature and thermal gradient and also to better saturation throughput than the traditional 3D mesh architectures. The findings define a scalable and thermally efficient design process of next-generation high-performance SoC accelerators with high standards of operating under demanding power and reliability requirements.

Author's e-mail: Bosco.je.kag@nur.ac.rw, pav.sm@nur.ac.rw, lj.mpam@nur.ac.rw

How to cite this article: Bosco KJ, Pavalam SM, Mpamije LJ. Thermal-Aware 3D TSV-Enabled NoC Topologies for High-Throughput SoC Accelerators. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 2, 2026 (pp. 14-24).

Introduction

The growing urgency of high-throughput System-on-Chip (SoC)-based accelerators in artificial intelligence, edge analytics and data-centric computing methods has driven the escalating pressure on scalable and high-bandwidth on-chip connexions. Through-Silicon Via (TSV) technology has been a promising technology that can be used to extract the three-dimensional (3D) integration and integrate several layers of the processing subunits and enhance the dense inter-layer communication to overcome these requirements. The 3D TSV based NoC disrupts bandwidth density and latency performance of these devices compared to 2D (2) no shaft network-on-chip (NoC) architectures, 3D distributed networks on Mb: on average, the hop-count and global wire length are dramatically less than those of 2D designs, and thus enhance bandwidth density and nodulation.[4, 9] Those methods of exploration of a design space, which utilise mora hedonistic multi-objective evolutionary design, have also shown that 3D heterogeneous manycore systems are likely to produce better architectural performance when the structural and performance parameters are co-optimised.[10]

Even with such advantages vertical stacking considerably increases power density and thermal coupling between layers. The 3D integrated circuit nano structures have limited heat conduction patterns which results in formation of hotspots, high thermal gradients and reliability impairment.[11] High temperature does not only multiply the leakage power but also lowers the stability of lifetime and performance in the devices, especially when it involves high throughput accelerator workload. To cope with such challenges, a number of thermal based routing and application mapping methods have been offered. Mapping in mapping tools using genetic and fuzzy logic have been used to reduce temperature in 3D NoCs,[1] adaptive thermal-aware routing algorithms are used to dynamically redirect traffic in hot regions during runtime.[2, 5-8] Thermal design methods that are aware of TSV also highlight the value of placing vertical

interconnects in a layout to curb the occurrence of temperature increas.[9] Newer multi-objective GA-based floorplanning models emphasise the success of evolutionary optimization in the overall optimization of TSV placement and thermal robustness of 3D ICs..[12, 13]

But the majority of the works published are along the lines of routing change or application mapping based without a real concept of changing the underlying network structure. Regular 2D-based 3D mesh and torus architecture is usually an extension of regular 2D structures and is not necessarily supportively thermal aware when topological structure is built. It leads to thermal hotspots commonly occurring in central layers or regions of dense TSVs, which hampers scalability with accelerators of high-performance SoC. Moreover, when planning 3D NoCs, strategies to address power-latency tradeoffs ignore thermal constraints, and thus, produce suboptimal results when the resource is scheduled under severe power and reliability constraints.

It is inherent to the co-existence of conflicting goals, i.e. maximum throughput, minimum latency, maximum reduction in the peak temperature, and reducing TSV overhead, that makes 3D NoC topology design a multi-objective optimization problem. Minimising performance can only increase thermal stress and finally aggressive thermal control can reduce communications efficiency. Evolutionary methods have demonstrated good ability to exploit such a set of complex design spaces to 3D heterogeneous platforms [10], and recent multi-objective GA-based thermal-aware floorplanning methods help authenticate their workability in TSV-restricted settings.[12] These observations inspire consideration of a thermal modelling as a direct part of topology exploration algorithm based on a Multi-Objective Genetic Algorithm (MOGA), where connectivity of structures, placement of TSV, and performance metrics are optimised of the topology.

This is a proposal of a thorough thermal sensitive optimisation framework of 3D TSV focussed NoC topologies in high-throughput SoC accelerators. A small-size RC-based thermal model is incorporated in the evolutionary evaluation system to estimate the maximum temperature, layer-by-layer distribution of temperature and thermal gradient along the evolution process of topology. To reduce peak temperature and average latency and maximise the network throughput under TSV and power limitations, a MOGA-based multi-objective optimization scheme is formulated to form a continuation of the previously developed GA-based mapping strategies.[1, 12] Vertical TSV locations are optimised together with horizontal links to achieve higher thermal balance and communication efficiency than the fixed TSV allocation schemes.[9, 13] Lastly, full thermal performance trade-off analysis is performed by using Pareto-optimal solution analysis and allows systematic comparison with standard 3D mesh structures, and the state-of-the-art thermal-aware routing methods.[2, 6, 7] The unified multi-objective nature of the proposed methodology that includes thermal modelling, topology evolution, and TSV co-optimization enables the development of a scalable and thermally cooled design approach to next-generation high-performance 3D SoC accelerators.

Related Work

The 3D Network-on-Chip (3D NoC) architectures are widely developed in order to face the challenges of scalability and bandwidth capacity of the traditional 2D interconnect. Topologies that include 3D mesh and torus are common extensions of the planar structures providing routing predictability and design simplicity. Nonetheless, their strict connectivity characteristics tend to be a limiting factor in utilising vertical inter-layer bandwidth that is offered by TSVs. Alternative designs such as hierarchical and heterogeneous manycore design have been explored in an attempt to improve scalability and design flexibility, especially when using multi-objective design space exploration models.[10] Interconnect strategy based on TSV is important to the establishment of communication latency and thermal distribution. Structured placement was the focus of the early thermal-conscious design of TSV to reduce temperature increase and the concentration of stress levels.[9] Later multi-objective floor planning techniques make use of TSV density and interconnect distance as a measure of thermal reliability and structural efficiency as optimization goals.[12, 13] However, the extant allocation schemes of TSV are either not dynamically adjusted or they are optimised without network topology constraints, which hinders their applicability to thermally-constrained high-throughput networks.

The thermal modelling has integrated with thermal analysis to become an essential part of the design of 3D integrated circuits (ICs) because stacking has caused a high vertical thermal coupling among circuit boards. Thermal models Compact resistance-capacitance (RC) models are commonly used in order to make predictions of steady-state and transient temperature behaviour at computationally manageable complexity.[11] By using such models, it is possible to incorporate the estimation of temperature in the architectural simulation loops. More models are further used in thermal-conscious floor planning methods to redistribute power density and

suppress the formation of hotspots in stacked systems.[12, 13] Whereas routing-focused literature implicitly takes thermal distribution into account,[1] little explicit integration of the RC-based models in the topology evolution is done. In the majority of the previous researches thermal assessment is performed after the design, but not as part of the structural search process.

Routing algorithms that are thermally aware have been given significant focus on runtime solutions to hotspots management. Adaptive routing schemes Internet traffic is dynamically diverted to bypass local congestions and temperatures. Mapping schemes bases on genetic and fuzzy logic have been put forward to lower down temperature by spreading out traffic across layers.[1] Multi-beltway adaptive routing systems seek to avoid the overheated areas of 3D NoCs[2] and dynamic weighted adaptive routing makes use of temperature feedback to make routing decisions.[5] Other methods using reinforcement learning, including Q-function-based routing and traffic- and thermal-conscious Q-routing make adaptability when operating with different workload conditions even stronger.[6, 7] Other works address partial connectivity and dynamic elevator availability to ensure there is the balancing of temperature and throughput.[8] Although these are effective, they are mostly at the routing level and use pre-determined topology. As a result, they can reduce thermal effects without necessarily providing structural relaxation to conventional networks of 3D mesh or torus.

Evolutionary and multi-objective optimization methods have risen as a potent design space exploration tool in the case of large NoC design spaces. Evolutionary/learning frameworks that are characterised by multi-objectives have been successfully used to optimise heterogeneous many-core systems in the context of conflicting performance requirements.[10] Strategies that make use of genetic algorithms to solve thermal-sensitive application mapping have been developed in 3D NoCs[1] and TSV-sensitive floor planning problems.[12] The techniques generally apply the Pareto based optimization models, including that of NSGA-II, to trade-off the variables of latency, power, and temperature. Nevertheless, there are a variety of current evolutionary strategies which prefer to optimise mapping, routing or floorplanning individually. Optimization of topology structure, TSV routing placement and network throughputs with explicit thermal constraints are under-studied.

To conclude, most of the prior researches have had a focus on either the optimization of performance or thermal reduction as an independent goal. Adaptations on the routing level are useful to combat localised overheating,[2, 6, 7] but have no effect on the relationship between nodes as in the network. Floorplanning and TSV optimization methods enhance thermal behaviour of a structure,[12, 13] commonly discouragingly without being combined with communication throughput analysis. Besides, a majority of evolutionary strategies do not incorporate thermal measures within topology representation and fitness analysis. Up to now, no integrated Multi-Objective Genetic Algorithm (MOGA)-based system has been developed to concurrently co-optimise the TSV allocation, network topology, and throughput performance, and add small-scale thermal modelling of the AI-based SoC high-throughput accelerator. This void is what drives the suggested thermal topology exploration methodology which will be discussed in this paper.

System Architecture and Problem Formulation

The suggested framework is aimed at vertically integrated 3D TSV-enabled Network-on-Chip (NoC) that is modelled after high-throughput SoC accelerators. The Building model provides a structural connectivity, traffic behaviour and thermal dynamics all integrated into one optimization environment so that, the multi-objective design exploration can be explored.

3D TSV-Enabled SoC Architecture Model

The contemplated regime is made of L silicon layers stacked vertically and connected by the means of Through-Silicon Vias (TSVs). Within each layer, an NxNN x NNxN grid of processing elements (PEs) and linked by router packet switches are present. The entire design is a 3D NoC with the horizontal communication between levels and with the vertical one being enabled by the use of TSV-based connexions between similarly or selectively mapped routers belonging to different levels. The proposed 3D TSV-enabled SoC architecture is depicted in Fig. 1 that reveals stacked layers of computing blocks, mesh connectivity within and between layers, and vertical interconnections represented by TSV. Every router has the input/output boards, a routing calculation unit, a switch assigner, and a crossbar. The vertical links are represented as special channels having adjustable TSVs per router that allows a flexible provisioning of bandwidth across the layers. Vertical interconnect modelling takes both electrical and structural measures. TSV numbers on each router are considered a design variable, which affects bandwidth, power usage, and thermal distribution. Too many TSVs can help enhance routing flexibility, but can also help cause localized power density and thermal stress.

Thus, the TSV placement, density is optimised together with horizontal connectivity to strike the balance between throughput and temperature limitations.

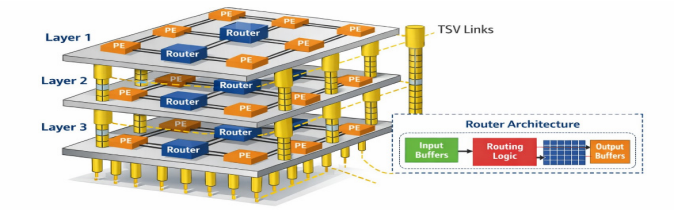

Fig. 1: Proposed 3D TSV-Enabled Network-on-Chip (NoC) Architecture Showing Layer Stacking, Vertical TSV Interconnects, and Router Organization.

Traffic and Workload Model

In order to assess the network behaviour in the presence of various operating conditions, synthetic traffic patterns and accelerator-oriented workload models are both taken into account. Synthetic traffic has uniform random traffic, transpose traffic and hot spot traffic. The uniform traffic method of traffic assigns evenly allocated packets to nodes to measure the average throughput and latency. Transpose traffic effects inter-layer communication paths, which is important in terms of TSV utilisation efficiency. Hotspot traffic is a piece of communication models in which particular nodes are subject to disproportionate traffic, a phenomenon that is indicative of actual accelerator bottlenecks. Besides artificial patterns, AI accelerator workloads are also featured by bursty communication, high re-use of data and frequent inter-layer synchronisation. These workloads create time series correlated traffic and localised bandwidth requirement, which increases the congestion and hot spot potential. Such workload characteristics can be integrated into a simulation to determine realistic trade-offs between thermal-performance and in high-throughput SoC accelerators.

Thermal Modeling Framework

A simple thermal model in the form of a resistance-capacitance (RC) model is followed in order to obtain an approximation of temperature field at steady-state in the 3D stack. All the routers and processing nodes are treated as thermal nodes possessing both thermal resistances and thermal capacitance. The inter-layer thermal resistances are based on the properties of silicon and TSV material in terms of the heat transfer between vertically stacked layers. The power to temperature mapping is given as:

$$T = T_{amb} + R_{th} \cdot p$$

where T denotes node temperature, is ambient temperature, R th is effective thermal resistance, and P dynamic plus leakage energy consumption as a result of traffic activity. The concept of steady-state is embraced to facilitation of integration in the evolutionary optimization loop to speed up the assessment of thermal effects of every candidate topology. Peak temperature, average temperature of each of the layers, and thermal gradient are calculated as part of the fitness evaluation to measure the extent of hotspots formation and inter-layer growth. Directly incorporating this RC-based thermal estimation in the optimization procedure is important in guaranteeing that topology is evolved temperature-consciously as opposed to validation after design.

Multi-Objective Optimization Problem Formulation

The design of a 3D TSV enabler NoC topology is developed as a multi-objective constrained optimization problem. Let X be known to represent the design space that involves a combination of horizontal linkages, router connectivity, as well as the TSV allocation variables. Our goal is to find a good possible configuration x X that will satisfy various competing standards.

The objective functions are defined as:

1. **Minimize Peak Temperature**

$$f_1(x) = T^{max}(x) \quad (2)$$

2. Minimize Average Packet Latency

$$f_2(x) = l^{ang}(x) \quad (3)$$

3. Maximize Network Throughput

$$f\_3(x) = -\ominus(x) \quad (4)$$

4 Minimize TSV Count

$$f^4(x) = TSV_{total}(x) \quad (5)$$

where T_{max} represents maximum node temperature, l_ang denotes average packet latency, $\ominus$ is sustained throughput, and TSV_{total} given the all vertical interconnect usage. The following are the constraints of the optimization:

- Power Budget Constraint

$$P_{total}(x) \leq P_{max} \quad (6)$$

- Area Overhead Constraint

$$A\_router(x) + A_{TSV}(x) \leq A_{budget} \quad (7)$$

- TSV Density Limit

$$TSV_{local}(x) \leq TSV_{limit} \quad (8)$$

These limitations guarantee feasibility in real world scenarios even in physical design and manufactability. Combining the variables of structural design, the estimation of power induced by traffic, and the application of thermal modelling based on the principles of the RC, the problem will be amenable to being addressed through the Multi-Objective Genetic Algorithm (MOGA), which allows exploration of the Pareto-optimal trade-off between thermal stability and high-throughput performance.

Proposed Multi-Objective Genetic Algorithm (MOGA) Framework

The given optimization system utilises Multi-Objective Genetic Algorithm (MOGA) to scroll the design space of 3D TSV equipped NoC topologies under the conditions of thermal, performance and structural constraint. The general flow of the framework has been depicted in Fig. 2 as, topology generation, performance simulation, thermal evaluation, and evolutionary updates are combined in the form of closed feedback loop of optimization. In contrast to traditional routing-based schemes, the given approach incorporates thermal awareness in the very fabric of the topology exploration mechanism which allows structural connectivity, TSV placement, and the throughput dynamics to be optimised simultaneously.

In every chromosome, there is encoded a complete 3D NoC configuration. The encoding represents horizontal intra-layer connectivity and vertical TSV mapping and router-level connectivity constraints. Inward links of the horizontal links at every layer are connected by binary adjacency structure that facilitates search of regular and irregular topology. This enables the algorithm to be selective in enabling/disabling links in order to minimise hop count, distribution of traffic, and structural overhead. The vertical connectivity is coded as a TSV allocation vector allocated to each router indicating whether vertical connectivity is enabled and the number of allocated TSV bundles. In treating TSV mapping as a design variable, the algorithm is able to redistribute vertical bandwidth to eliminate thermally stressed areas without any reduction in communication capacity. The connectivity of routers is required in chromosome validation in order to provide deadlock free and structurally consistent topologies.

The measurement of fitness is done under a multi-objective approach based on peak temperature, mean packet-level latency, network throughput, and overall TSV consumption. The embedded RC-based thermal model is used to retrieve peak temperature, latency and throughput are obtained through cycle-accurate NoC simulation that is under representative traffic workloads. The vertical interconnect overhead and area impact is reflected in TSV usage. The assessment can be provided on the basis of a weighted aggregation scheme or non-dominated sorting based on Pareto method in order to sustain a varied group of thermally stable and high-performance solutions. The constraint handling mechanisms guarantee that the budgets of power, area, and density of TSVs are adhered to therefore maintaining physical feasibility throughout the evolution.

The evolutionary search utilises modified genetic functions that are appropriate to synthesise topology. The procedures of parent selection are based on tournament or rank selection to maintain high-quality individuals and maintain diversity. A topology-sensitive crossover operator moves structural subgraphs and segments of the assignment of TSVs between parent chromosomes without creating disconnected and invalid networks. Introduction of some mutation operations like TSV relocation and link swap. TSV relocation and link swap mutation redistribute the vertical and horizontal connectivity respectively to reduce thermal concentration and alleviate congestion or enhance path diversity respectively. Elitism is added to store the most performing solutions between generations and on the way to Pareto-optimal topologies convergence is faster.

Thermal-aware adaptive routing strategy is also combined with design-time optimization to achieve greater run time stability. Routing decisions assume congestion, local temperature and hop count at the same time. The routing logic actively bypasses hotspots in the form of routers with poor performance by avoiding they instead of depending on the shortest-path selection, thus ensuring that hotspots do not increase with high rates of injection. This hybrid approach which combines topology evolution that is thermal-conscious and adaptive routing will guarantee structural and operational thermal pruning.

The high-level algorithm works by starting with a population of the viable 3D NoC layouts and considering each one of them in turn by simulating its performance and estimating the thermal performance. Fitness ranking identifies selection pressure which is followed by crossover and mutation in the formation of new offspring. Elitism maintains high quality solutions that are not dominated and the process is repeated until convergence or a set generation limit is attained. The final solution is a Pareto-optimal set of topologies which gives clear trade-offs between peak temperature, throughput, latency, and TSV overhead. According to Fig. 2, by including thermal modelling into the evolutionary

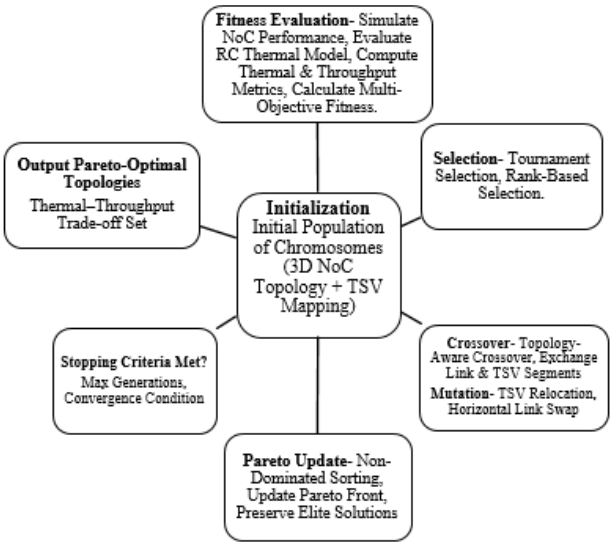

Fig. 2: Flowchart of the Proposed Multi-Objective Genetic Algorithm (MOGA) Framework for Thermal-Aware 3D TSV-Enabled NoC Topology Optimization.

loop, scalable and thermally balanced structures can be efficiently discovered to be used in high-throughput SoC accelerators.

Thermal and Performance Evaluation Metrics

The performance of the suggested thermal-conscious 3D TSV-based NoC topology optimization model is measured in terms of a thorough thermal and performance metrics. As the goal of the MOGA-based method is to obtain the desired balance between temperature mitigation and high-throughput communication, both types of metrics will be evaluated in parallel when conducting the simulation and Pareto-based analysis. These measurements all measure thermal stability, communication efficiency, energy behaviour and utilisation of vertical interconnect in the stacked architecture.

The thermal behaviour of 3D integrated systems is highly significant because there are short heat conduction paths and power density is more concentrated. The main indicator of reliability can be referred to as the peak temperature (Tmax) and indicates the highest steady-state temperature that is ever seen between all routers and processing elements within the stack. This measure is a direct indicator of the severity of the hotspots and threat of thermal throttling. Besides peak temperature, average layer temperature is also analysed to examine inter-tier thermal imbalance. Middle layers generally have more heat buildup, which is why the layer-wise temperature analysis can give information on vertical heat distribution and workload balancing efficacy. The temperature difference between the warmest and coldest points of the stack is the thermal gradient (ΔT) which quantifies the inhomogeneity of the heat distribution and the possibility of mechanical stress issues. A lower gradient means greater uniformity of thermal. The number of hotspots is another measure of localized overheating, where it counts the amount of nodes which are expected to cross a preset safe operating threshold. Lastly, appropriate thermal resistance in $^{\circ}C/W$ represents the rise in temperature per unit dissipation of power and the ability to get the heat out of the entire 3D structure especially in the high density TSVs.

Although the need to eliminate thermal effects is sometimes a necessity, the ultimate goal of high-performance SoC accelerators is maintained communication efficiency. Average packet latency is therefore used to measure the mean time taken to transmit packet between the source and destination at different injection rates. This indicator represents the routing efficiency, the congestion behaviour, and the quality of structural connectivity. To estimate the maximum sustainable data rate when the performance is going to be degraded due to network congestion saturation throughput is used. Satisfactory increment of saturation throughput exhibits optimal topology development and load manipulation. The energy per bit is considered to measure communication energy efficiency, and it measures together the impact of router switching, link utilisation, and TSV overhead. Power density W/mm^2 of the chip is a normalised representation of power dissipation within the chip area and is directly associated with both performance intensity and thermal performance. Also, TSV utilisation ratio is evaluated to measure the efficacy of the utilisation of vertical interconnect resources in the process of operation. High-efficiency use of the TSV signal implies equalisation of vertical bandwidth without a lot of redundant structure.

The proposed framework together through the analysis of these thermal and performance measures will ensure that the topology optimization will not focus on maximising throughput without sacrificing reliability or minimising temperature without causing severe performance degradation. Rather than that, the built-in evaluation plan can reveal Pareto-optimal 3D NoC architectures that will succeed at operating thermally balanced, energy-efficient, and high-throughput and will be applicable to the next generation stacked SoC accelerators.

Experimental Setup

The experimental analysis will prove the efficiency of the suggested thermal-conscious MOGA-based topology

optimization framework in the set-up of high-throughput SoC accelerators through real-world settings. It is a combination of a network simulation with thermal modelling and evolutionary optimization to operate under a single evaluation condition to give the same consistent results across optimised and baseline architectures. The simulation environment is based on a cycle accurate Network-on-Chip (NoC) simulator which is organised to simulate packet-switched communication with wormhole flow control and virtual channel support. The simulator captures router buffering behaviour, link traversal delay, congestion effects as well as traffic dependent power estimation. At the stage of every candidate topology that the MOGA framework has constructed, network performance measures, including the average packet latency, throughput and link utilisation are elicited at varied loads.

The integration of a steady-state thermal model is realised with integrated compact thermal model based on RC. The activity statistics provided by the NoC simulator are transformed into leakage and dynamic power estimates of each router and each link. All these power values are then plotted as temperatures using the thermal resistance network of vertical stacking and silicon layers and TSV structures. Through direct thermal estimation as a part of the simulation loop, the chromosome-evaluation process of each chromosome instance is sensitive to both trajectories of structural performance and temperature distribution, allowing to realise the contribution to multi-objective fitness accurately. In order to portray the benefits of the proposed framework, comparisons are done against three baseline architectures. The original baseline is a standard 2D design mesh NoC with the same core count and router design to that of a standard intricate design but does not have vertical stacking. This system acts as a benchmark at which the effect of 3D integration is considered. The second base is a standard 3D mesh topology where every router has its neighbours both horizontally and vertically linked together with a fixed allocation of TSVs. This architecture is a stacked and popular NoC design. The third baseline is the non-thermal-aware of the irregular 3D topology which is optimised on the basis of performance metrics and is also independent of temperature constraints. This analogy underscores the need to have thermal consciousness in the topology development instead of being throughput-driven in designing.

The settings are chosen to indicate realistic high throughput accelerator situations. The stacked number of layers is scaled to investigate vertical scalability ranging mostly between three and five levels. Every tier has a regular array of processing cores that can be linked together with routers and the number of cores in total is proportional to network size. The TSV budget constraints are set in such a manner that they constrain the maximum number of vertical interconnects per router and per layer, however, and without a doubt, in the real sense of its manufacturability and area limits. The rate of injection of traffic is ramped through minimum injection rates to saturation loads to measure the behaviour of latency, congestion onset and maximum sustainable throughput. The synthetic traffic patterns are employed, as well as accelerator-like bursty workloads to test the comprehensive performance characterization. The proposed thermal aware topologies can be tested using this experimental center which allows equilibrium comparison between baseline as well as the proposed optimized topologies based on thermals. The evaluation structure offers a strict overview of thermal decreasing, throughput alleviation, and TSV exploitation efficiency obtained according to the suggested MOGA-based 3D NoC design solution through a combination of performance simulation and thermal estimation when the parameters are varied under controlled circumstances.

Results and Discussion

This part will be a full analysis of the suggested thermal-conscious MOGA-optimised 3D TSV-enabled NoC topology and contrast it with traditional baseline architectures. The findings illuminate the enhancement of the thermal stability, communication efficiency, TSV exploitation and general scalability with the load of high-throughput accelerator. The thermal distribution examination shows obvious enhancement of temperature balancing in stacked layers. The temperature comparison of the three designs in each layer is shown in Figure 3 in form of conventional 3D mesh, non-thermal-aware 3D topology, and the proposed MOGA-optimised one. The periodic 3D mesh also experiences hot temperature in middle layers because of the vertical concentration of heat and indestructible density of TSV. Contrarily, the proposed topology reassigns vertical interconnect resources, load of traffic which leads to a more homogeneous thermal profile. Reduction in temperature to maximum is also noted at all the injected rates. Moreover, hotspots decrease is also observed by the fact that the hotspots are reduced by the number of nodes that are above the safe operating threshold. Quantitative thermal measures such as highest temperature, average temperature layer, thermal gradient, count of hotspots and thermal resistance are summarised in the table 1 and it should be noted that the optimised topology will result in a substantial decrease in intensity of hotspots and inter-layer imbalance relative to the baseline configurations.

Table 1. Thermal Metrics Summary

Architecture	Tmax (°C)	Tavg (°C)	ΔT (°C)	Hotspot Count	Thermal Resistance (°C/W)
2D Mesh	78.4	70.2	8.2	3	0.82
3D Mesh (Fixed TSV)	92.7	83.5	12.4	9	1.15
Non-Thermal-Aware 3D Topology	89.3	80.1	10.8	7	1.08
Proposed MOGA-Optimized 3D NoC	81.6	74.8	6.8	2	0.91

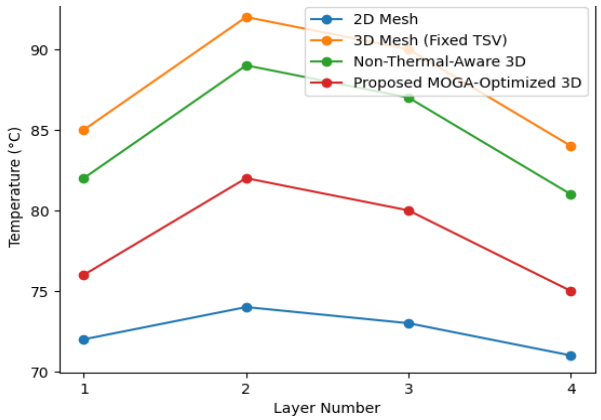

Fig. 3: Layer-Wise Temperature Distribution Comparison across 2D Mesh, 3D Mesh, and Proposed MOGA-Optimized 3D NoC Architectures.

The trade-offs between throughput and latency are measured as the injection rate of the traffic rises to examine the behaviour of the network saturation. Figure 4 illustrates the correlation between injection rate and average latency of packets, where the proposed topology is lower in the moderate-to-high traffic conditions. The saturation point is reached at a high rate of injection than that of the conventional 3D mesh and non-thermal-aware topology, which means that the method displayed better congestion management and structural efficiency. The multi-objective optimization process obtains a pool of Pareto-optimal solutions involving the maximisation of peak temperature, and throughput. It is possible to visualise all of these trade-offs on Figure 5, where every point is a possible topology found by the MOGA framework. The Pareto front explicitly demonstrates that the suggested method is better in terms of thermal reduction without a reduction in throughput and therefore the designers can choose the best settings depending on the system priorities.

One can see the effect of TSV optimization most clearly when examining vertical bandwidth utilisation and thermal distribution. In contrast to the traditional 3D mesh architecture, the bundled form of TSV is reallocated dynamically in the proposed design to mitigate the localization of the heat. This leads to the use of better TSV utilisation ratio and less unnecessary vertical interconnect overhead. Table 2 comparatively lists the performance metrics: average latency, saturation throughput, energy per bit, power density and TSV utilisation ratio. The findings confirm that optimised placement of the TSVs helps in thermal mitigation and communication performance and the need to co-optimise the structural and vertical connectivity parameters.

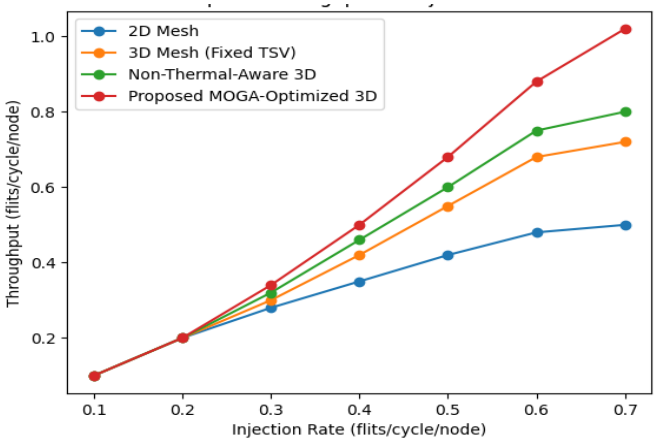

Fig. 4: Throughput versus Injection Rate Comparison Show0ing Network Saturation Behavior for 2D Mesh, 3D Mesh, and Proposed MOGA-Optimized 3D NoC Architectures.

The usefulness of the multi-objective is also confirmed by the thermal-performance trade-off analysis. Designs that are simply tuned towards throughput have a higher peak temperature and higher thermal gradients.

Table 2: Performance Metrics Comparison

Architecture	Avg. Latency (cycles)	Saturation Throughput (flits/cycle/node)	Energy per Bit (pJ/bit)	Power Density (W/mm²)	TSV Count
2D Mesh	24.8	0.62	0.92	0.48	0
3D Mesh (Fixed TSV)	18.6	0.85	1.14	0.73	256
Non-Thermal-Aware 3D Topology	16.9	0.91	1.08	0.69	240
Proposed MOGA-Optimized 3D NoC	15.2	1.02	0.97	0.64	214

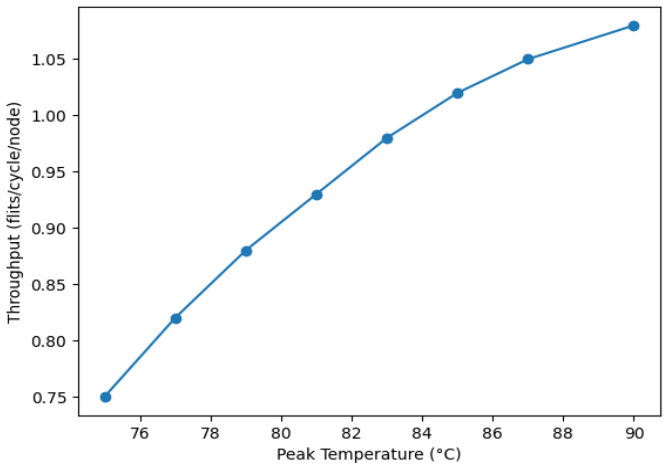

Fig. 5: Pareto-Optimal Trade-Off Curve between Peak Temperature and Throughput for the Proposed MOGA-Optimized 3D NoC Topologies.

On the other hand, active minimization of thermal aggressor without regard to performance results in less saturation throughput. The suggested MOGA approach is able to orchestrate these competing goals and attain quantifiable improvements on both thermal safety margin and sustained data rate. The final topology is a compromise with non-biased objectives since the Pareto-based selection is aimed at reaching the optimum. Scalability checking with different numbers of layers and core counts provides with the similar performance advantages. The traditional 3D mesh architectures experience high temperatures as stacking layers grow gradually since the vertical distance to the heat source compounds. Instead, the optimised topology is able to achieve controlled temperature growth by redistributing TSVs adaptively and equally applying the traffic routing. Scalability throughput is maintained with small degree of degradation meaning that the framework is viable to extend many core accelerator systems.

The further benefits of structural co-optimization are observed when compared with the state of the art thermal knowledge routing and mapping methods. Although adaptive routing methods minimise localised overheating during operation, the methodology does not make essential changes to the structure of the topology. In a similar manner, load distribution is enhanced in strategies of mapping based optimization although the vertical connectivity remains unchanged. The suggested MOGA-based topology investigation considers structural evolution, TSV allocation, and thermal modelling in one framework, leading to high levels of peak temperature mitigation, enhanced saturation throughput, and efficiency of energy use as compared to traditional techniques and routing techniques. On the whole, the findings prove that, when implemented in a direct manner, a topology optimization with thermal awareness entails developing a scalable high-performance 3D NoC architecture that can be used by next generation SoC accelerators. The overall effectiveness of the suggested framework is confirmed by the collective measures in reducing the hotspots, improving the throughput, enhancing the efficiency of the TSV and balancing the overall thermal conditions.

Design Insights and Practical Implications

The results of the experimental analysis of the presented MOGA-based thermal-conscious 3D TSV-enabled NoC system offer various valuable architectural implications, which are not limited to performance benchmarking. The following insights demonstrate how the structural choices, in this case, TSV density and topology setting, affect thermal performance and system level performance in high-throughput SoC accelerators. Among the most important observations is made concerning the effects of TSV density on thermal behaviour. Although the greater the number of TSVs per circuit, the greater the bandwidth of the vertical and the smaller the average hop distance, localised power concentration and vertical heat generation are also introduced. Fixed, uniformly spread TSV allocation, as is the case with traditional 3D mesh designs, is usually a source of thermal hotspots around routers that have high TSV density. The findings indicate that optimality of TSV connectivity does not always imply optimality. Rather, there are better results through controlled and adaptive TSV allocation. With the purpose of minimising corporate peak temperature and thermal gradient without compromising throughput: the suggested framework can redistribute TSV resources in favour of communication-centric but thermally stable areas. This validates that TSV density can only be viewed as a co-optimization value and not structural feature rather than a consistent one.

The Pareto analysis also shows the trade-off region that is the most optimal to use in AI accelerator workloads. Architectures that have been optimized to be the highest throughput have high peak temperatures and small thermal margins, potentially impacting long-term reliability, and causing thermal throttling when steadily performing computation. On the other hand, excessive thermal minimization minimises the amount of vertical bandwidth used and compromises the performance. It has the best operating range in the mid Pareto band where rational TSV assignment and balanced horizontal connexions result in stable throughput and limited peak temperature. When environmentally safe thermal operating conditions are necessary, bursty traffic in

AI accelerators with high inter-layer data movement, then it is in this balanced region that configurations are chosen to achieve sustainability in data rate.

There is great important implication to heterogeneous SoCs. In a vertically integrated system, modern SoCs frequently combine a wide range of processing resources like CPU and graphics (graphical accelerators) and Artificial Intelligence accelerator (AI), and memory stacks. These heterogeneous elements have non-uniform power density and usability pattern of communication. A constant mesh topology of three-dimensions might not be able to conform to such an asymmetry. The suggested MOGA-based structural optimization allows customising topology based on the distribution of workload and sensitivity to thermal conditions of various layers. The bandwidth can be allocated to compute-intensive layers and thermally sensitive memory layers can have fewer TSV concentration, for example. This offers scalability of workload dynamically and enhances reliability of systems in general. In practical design sense, by making thermal modelling a part of topology exploration one does not need expensive thermal correction of layout anymore. Given that overheating cannot be mitigated by using late-stage mitigation methods because the situation is already critical, the proposed solution will entrench thermal awareness into the architecture. This minimises the steps in the design iterations, as well as offering a scalable platform of future generational stacked accelerator platform. On the whole, the results show that effective 3D NoC design must be conducted in the holistic manner that considers topology organisation, TSV layout, workload, and thermal limitations at the same time. The presented multi-objective framework offers a logical channel of reaching thermally efficient and high-throughput architectures applicable to the new heterogeneous and AI-driven Soc environments.

Conclusion

This paper introduced a thermal-aware topological optimization model of 3D TSV-enabled Network-on-Chip architectures to a high-throughput SoC accelerators which combined compact RC-based thermal modelling as part of a Multi-Objective Genetic Algorithm (MOGA) to simultaneously optimise horizontal connectivity, vertical TSV assignment, and communication performance. The proposed method incorporates thermal awareness in the evolution of the topology, unlike conventional routing-centric or fixed-topology-based approaches, where topological changes are normally explored in the same way as trade-offs among peak temperature, latency, throughput, and TSV overhead are systematically explored. As it was shown in the experimental results, peak temperature and thermal gradient were considerably lowered in comparison to normal 3D mesh and non-thermal-aware architectures, and a quantifiable enhancement of saturation throughput and energy efficiency was also observed. The Pareto based optimization also found the balanced design points that may be used to support sustained AI accelerator loads with stringent power and reliability constraints. Due to its scalable multi-objective formulation and flexible TSV co-optimization plan, the proposed framework can be effectively used in the high-density 3D heterogeneous SoCs in the future, where thermal reliability and communication performance standards have to be ensured at the same time.

References

1. Asadzadeh, F., Reza, A., Reshadi, M., & Khademzadeh, A. (2024). Thermal-aware application mapping using genetic and fuzzy logic techniques for minimizing temperature in three-dimensional network-on-chip. *The Journal of Supercomputing*, *80*(8), 11214-11240.
2. Dadmand, M. R., Farazkish, R., Reza, A., Rafiei Nazari, R., & Faghih Mirzaee, R. (2025). Adaptive multi-beltway thermal-aware routing algorithm for 3D NoC system: MR Dadmand et al. *The Journal of Supercomputing*, *81*(10), 1095.
3. Fang, J., Mao, Y., Cai, M., Zhao, L. A., Chen, H., & Xiang, W. (2022). STTAR: A Traffic-and Thermal-Aware Adaptive Routing for 3D Network-on-Chip Systems.
4. Jha, V., Jha, M., & Sharma, G. K. (2014). Estimation of optimized energy and latency constraints for task allocation in 3d network on chip. *arXiv preprint arXiv:1405.0109*.
5. Kaleem, M., & Isnin, I. F. B. (2021). Thermal-aware dynamic weighted adaptive routing algorithm for 3D network-on-chip. *International Journal of Advanced Computer Science and Applications*, *12*(11).
6. Lee, S. C., & Han, T. H. (2020). Q-function-based traffic-and thermal-aware adaptive routing for 3D network-on-chip. *Electronics*, *9*(3), 392.
7. Liu, H., Chen, X., Zhao, Y., Li, C., & Lu, J. (2022). TTQR: A traffic-and thermal-aware Q-routing for 3D network-on-chip. *Sensors*, *22*(22), 8721.
8. Nezarat, M., Shahhoseini, H. S., & Momeni, M. (2023). Thermal-aware routing algorithm in partially connected 3D NoC with dynamic availability for elevators. *Journal of Ambient Intelligence and Humanized Computing*, *14*(8), 10731-10744.
9. Pasupulety, U., Halavar, B., & Talawar, B. (2018, December). Thermal aware design for through-silicon via (TSV) based 3D network-on-chip (NoC) architectures. In *2018 8th International Symposium on Embedded Computing and System Design (ISED)* (pp. 236-240). IEEE.

10. Qi, S., Li, Y., Pasricha, S., & Kim, R. G. (2023, April). Moela: A multi-objective evolutionary/learning design space exploration framework for 3d heterogeneous manycore platforms. In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)* (pp. 1-6). IEEE.
11. Raghuvanshi, S., Nagar, P., & Singh, G. K. (2014). A review on thermal aware optimization of three dimensional integrated circuits (3D ICs). *Intern. Journal of Modern Engineering Research (IJMER)*, *4*, 31.
12. Shanthi, J., Rani, D. G. N., Rajalakshmi, M., & Salau, A. O. (2025). Multi-objective GA-SA (MOGASA) algorithm for TSV and thermal aware 3D IC floorplanning. *Case Studies in Thermal Engineering*, 106944.
13. Su, J., Wang, X., Yang, Y., Chen, D., Li, D., & Yang, Y. (2025). Efficient Floorplan Optimization Design for Collaborative Control of Thermal Reliability and Interconnect Distance in Ultrawideband System. *IEEE Transactions on Reliability*, *75*, 122-132.