Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 3, 2026 (pp. 43-50)

RESEARCH ARTICLE ECEJOURNALS.IN

High-Performance Network-on-Chip Architecture with Congestion-Aware Adaptive Routing for Heterogeneous Multi-Core VLSI Systems

Matteo Ferrari*

Associate Professor, ECE, Campus Bio Medico University in Rome, Italy.

KEYWORDS:

VLSI Architecture;
On-Chip Interconnect;
Mesh Topology;
Latency Optimization;
Throughput Enhancement;
Energy-Efficient Design;
Router Microarchitecture.

ARTICLE HISTORY:

Submitted : 15.03.2026
Revised : 07.04.2026
Accepted : 20 .05.2026

https://doi.org/10.31838/JIVCT/03.03.06

Abstract

Due to the high-speed integration of heterogeneous processing elements in modern multi-core VLSI systems, increment in complexity of communication on chip has resulted in acute scalability limits and dynamic congestion in traditional Network-on-Chip (NoC) designs. Deterministic routing schemes although simple and simple to implement cannot scale to non-uniform and bursty traffic patterns that occur with heterogeneous workloads leading to high latency and early network congestion. In order to overcome these shortcomings, this paper presents a high- performance Network-on-Chip architecture with a lightweight congestion-sensitive adaptive routing scheme. This router is proposed to monitor dynamically the occupancy of buffers and the use of links to calculate an actual index of congestion in order to select the paths intelligently without any deadlock. It is designed to make use of optimal hardware overhead and to be able to scale to mesh-based heterogeneous multi-core platforms. Cycle-accurate simulation was done on 4x4 and 8x8 mesh topology with uniform traffic, hotspot traffic, and transpose traffic. Experimental outcomes prove up to 28 percent decrease in typical packet internship and 22 percent raise in saturation throughput and 17 percent increase in energy effectiveness than the traditional XY and partly adaptable routing plans. Hardware synthesis with the help of 45 nm CMOS standard-cell library proves that the proposed congestion-awareness logic imposes less than 6% area overhead with insignificant effects about the critical path delay. These findings confirm the usefulness of the proposed architecture in heterogeneous VLSI systems of high scalability and high performance.

Author's e-mail: m.ferrari@unicampus.it

How to cite this article: Ferrari M. High-Performance Network-on-Chip Architecture with Congestion-Aware Adaptive Routing for Heterogeneous Multi-Core VLSI Systems. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 3, 2026 (pp. 43-50).

Introduction

Semiconductor technology has enabled the incorporation of heterogeneous processing units, such as general-purpose central processing units, graphics processors, AI processors, and specialized co-processors, in one System-on-Chip(SoC). Though these heterogeneous multi-core architectures are very useful in augmenting the computational throughput and application selectivity, they are associated with complicated communication demands owing to differentiated and dynamically fluctuating patterns of traffic.[2, 3] Older shared bus and crossbar interconnects are not scalable beyond a handful of cores; they have arbitration bottlenecks, excessive wiring overhead and lack scalability in bandwidth too. These constraints have motivated the use of Network-on-Chip (NoC) architectures which support more highly structured, packet-based communication board structures capable of supporting scalable many-core VLSI systems.[1, 3] The more recent high-performance NoC designs like FlooNoC and TeraNoC show that scalable on-chip interconnects to be used in heterogeneous workloads are possible.[3, 12]

In spite of these architectural developments, congestion is clearly one of the prevailing performance constraints of large-scale NoC systems. Heterogeneous cores usually cause non-uniform bursty traffic with specific processing

elements or memory controllers being hotspots of communication, causing asymmetric buffer occupancy and higher contention levels.[2, 7] Deterministic routing algorithms like XY routing would not be efficient under such circumstances since the algorithm dictates a set of minimal paths which must be used no matter the dynamic situation of the network. This can easily lead to early network congestion, packet latency and reduced throughput.[4, 8] Implementation of adaptive routing schemes to overcome these issues have been suggested, however, most of the current methods are either expensive in terms of overlap in hardware implementation (large hardware overhead) or have no programmable congestion detection schemes applicable to implementation at the VLSI level (including at the VLSI scale).[1, 13]

Recent growth in the complexity of heterogeneous multi-core systems makes real-time congestion knowledge in routing decisions necessary. Adding network state information, including buffer occupancy and utilisation of link allows routers to re-route traffic off congested paths on demand, thus improving overall performance.[5, 8, 11] Nevertheless, it is a major design challenge to have such adaptivity and still have deadlock-free operation and highly-complex hardware. More so, the performance metrics of latency and throughput are important in the NoC based multi-core systems and have a direct impact on the time required to execute the applications, the responsiveness of the systems.[7 9] Uncontrolled congestion also leads to a higher switching activity and buffer utilisation, which lead to increased dynamic power consumption. As such, an effective congestion-conscious routing system should be able to to balance performance with energy consumption and still be scalable to large-core applications.[3, 6, 12]

Inspired by these issues, the work hypothesises a high-performance NoC architecture with a new congestion-sensitive adaptable routing protocol adapted to the heterogeneous multi-core VLSI systems. The suggested solution proposes a congestion detection system that is lightweight (based on real-time monitoring of buffers occupancy and link-state performance) and can allow one to choose less-congested minimal paths dynamically, and at the same time, avoid the occurrence of deadlocks in communication. The router microarchitecture is also hardware efficient designed to execute the congestion evaluation in the routing computation phase to reduce area and timing overhead. Full-fledged performance analysis with synthetic traffic conditions [uniform, hot spot and transpose distributions], show large gains in average packet latency, saturation throughput and energy efficiency than the deterministic and traditional adaptive routing schemes.[4, 8, 11] All of these contributions add up to a scalable and experience-based congestion-aware NoC solution that can be used in next-generation heterogeneous VLSI systems.

Related Work

Earlier Network-on-Chip studies concentrated mostly on deterministic routing algorithm because it was simple and ensured deadlock-free operation. XY routing (or dimension-ordered routing, especially) was popular in mesh topology due to its small hardware requirements and predictable behaviour. West-First and Odd-Even approaches are turn-model-based routing schemes that enhanced the flexibility by limiting some turns to prevent cyclic dependencies with a minimum amount of routing paths. In spite of having low implementation overhead, deterministic schemes are not capable of adapting to dynamic changes in traffic, which causes severe congestion in terms of non-uniform and hotspot traffic patterns.[4, 8] In order to overcome these drawbacks, there was introduction of adaptive routing mechanisms which can be used to select dynamic paths depending on the conditions of the network. Routing strategies that are learned based on reinforcement learning learning have been shown to enhance congestion-aware routing policies using reinforcement learning to achieve better load balancing and latency statistics over time.[5, 8, 11] On the same note, there are hybrid deterministic-adaptive schemes which strive to achieve a balance between insignificant routing guarantees and congestion-sensitive flexibility of pathways.[4] Although these methods enhance throughput and minimise the delay of packet latency, they tend to add more complexity of hardware, computation delay of the routing path or consume large levels of state space, necessitating VLSI implementation in large systems.

Congestion-concerned routing mechanisms are a narrower application in the efforts to reduce localised congestion by using measures like buffer occupancy or link utilisation or virtual channel status to guide routing. Recent papers have postulated dynamic dimension prioritisation and approximate communication mechanisms to evenly allocate the traffic throughout the network.[6, 13] Congestion monitoring routers with hardware have also been deployed on which real time congestion flags are provided.[1] Even though the techniques substantially improve network utilisation and latency values, a substantial number of solutions are either optimised to a particular topology (e.g., 3D NoC or optical NoC) or impose non-negligible overheads of both area and power, making them difficult to be viable in heterogeneous VLSI. NoC designs that are quality-of-Service (QoS)-sensitive

go even further, adding progressively to adaptive routing the mechanisms of traffic prioritisation and service differentiation. The objectives of these approaches are not only to offer latency guarantees to the critical tasks but also to be able to offer fairness to different traffic classes. Several recent studies centred around high-bandwidth heterogeneous NoC systems, including FlooNoC and TeraNoC, emphasise the significance of scalable interconnect systems that can be used to support a wide range of workload requirements.[3, 12] Nonetheless, QoS-adaptive routing schemes are based on complicated arbitration or scheduling reason, thereby expanding router design tomography and energy use.

Although the adaptive routing research and the congestion-aware routing research have made great strides in research, a few gaps still exist. To start with, most of the current adaptive schemes are mostly centred on routing efficiency without paying close attention to the cost of hardware and scalability. Second, the approaches using reinforcement learning, which are effective, might be in need of further computational load that cannot be afforded by lightweight VLSI pipelines in router.[5, 11] Third, some of the congestion-aware mechanisms have been tested in low-traffic conditions and fail to fully deal with the heterogeneous workload imbalance.[6, 13] There is, therefore, a necessity of a hardware efficient congestion-adaptive routing architecture integrating real-time congestion monitoring with a small area overhead and scalability and deadlock-free functionality in the heterogeneous multi-core environments. The proposed work fills this gap by integrating lightweight congestion monitoring and minimal-path selection that is dynamic so that it can enable better latency, throughput and energy efficiency with minimal hardware penalties.

Proposed NoC Architecture

The specified Network-on-Chip architecture is intended to be implemented on the heterogeneous multi-core System-on-Chip utilising general purpose processors, hardware accelerators, and shared memory modules on a single die. These non-standard components create heterogeneous and highly dissimilar types of traffic that are communicated and demand a congestion-resistant and scalable interconnect structure. A 2D mesh topology is then taken to guarantee the structural regularity and efficiency in the physical design with 4X4 and 8X8 network evaluations. The mesh-based topology provides a modular scalability and a predictable path of routing as well as compatibility with layout VLSI-based implementations of a large core. The tiles of the network have Processing Element (PE), Network Interface (NI), and a router. The Network Interface such as packetization and depacketization of transactions, the handling of injection and ejection buffers, and separation of computation and communication latency. The router facilitates the inter-tile communication that makes use of the wormhole switching technique so as to minimise the storage of buffers and low-latency transmission. The implementation in Figure 1 depicts the heterogeneous multi-core NoC structure in general with the central aspects of trading tile parts and rotor connexions between the routers in both directions of the mesh network.

Table 1: Comparison of Existing NoC Routing Approaches

Routing Category	Example Works	Congestion Awareness	Hardware Complexity	Suitability for Heterogeneous Systems	Limitations
Deterministic Routing	XY, Turn Model [4], [8]	No	Low	Limited	Poor performance under hotspot traffic
Adaptive Routing	RL-based routing [5], [8], [11]	Partial	Medium–High	Moderate	Increased routing computation overhead
Congestion-Aware Routing	Dynamic dimension priority [6], ACAC scheme [13], Congestion-aware router [1]	Yes	Medium	Moderate–High	Area and power overhead concerns
QoS-Aware NoC	FlooNoC [3], TeraNoC [12]	Partial	High	High	Complex arbitration and scheduling logic
Proposed Work	This Paper	Yes (Lightweight real-time monitoring)	Low–Medium	High	Designed for scalable heterogeneous VLSI

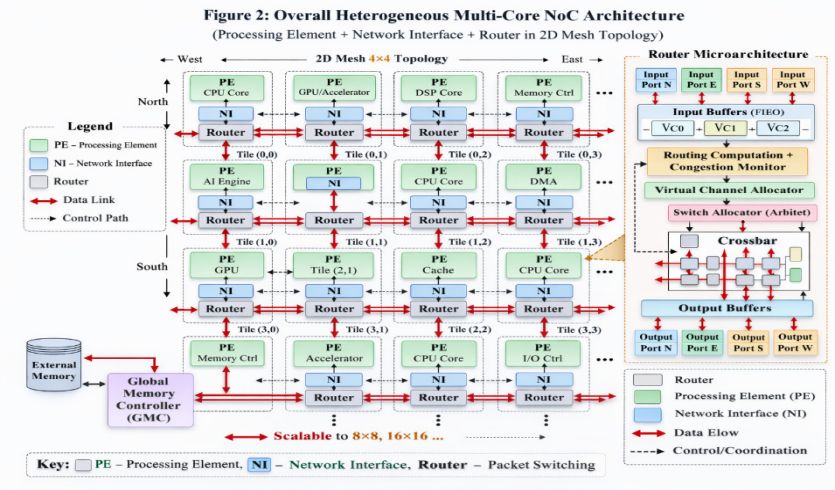

Figure 1: Proposed heterogeneous multi-core 2D mesh NoC architecture showing tile organization with Processing Element, Network Interface, Router, and bidirectional inter-router links.

The router micro architecture is based on a pipelined architecture comprising of input buffers, routing computation logic, virtual channel allocation, switch allocation, and crossbar switching and output forwarding. Bursty traffic and upstream to downstream contention are alleviated by each input port having FIFO buffers in them. Various virtual channels can be used on any physical port to reduce the effect of head-of-line blocking and enable adaptive routing to be performed safely. Routing computation step calculates the route directions that are admissible depending on the packet destination. To eliminate the problem of starvation caused by a heavy traffic load, the switch allocator uses fair scheduling policy like round-robin scheduling to resolve conflicts on output ports. The crossbar fabric creates short-term connections between the chosen point of input and output ports in succession of flits. To reduce the extra pipeline delays and save timing efficiency, the congestion monitoring logic is incorporated into the routing phase.

In order to make real-time awareness of congestion, the router constantly cheques the occupancy of buffers and the use of links. Buffer occupancy is an indication of the queue depth in each port, and is compared with a programmable threshold to indicate the accumulation of local congestion. The use of links is monitored by the activity on lightweight activity counters, an activity monitor of port activity across a specified observation interval, which record sustained contention states. A congestion notification is produced when the occupancy of buffers or utilisation of links surpasses the set limits in regard to the affected output direction. They are locally accessible to the routing logic and have an impact on the decision of path selection. These supporting hardware costs will be small counters, comparators and control registers per port and will incur only modest area and power overheads compared to the base router building.

The adaptive routing policy is congestion-aware and uses little paths and dynamic congestion values in the decision making. Given a packet at router coordinates: (x,y) to be forwarded to router destination (x d,y d), a router first determines all minimal output directions which will minimise the Manhattan distance between a source and a destination. When more than one of the minimal directions exist, the score of congestion of each candidate output is calculated as a weighted sum of normalised buffer occupancy and link utilisation. The path having the minimum congestion index is taken and so the traffic is spread out more uniformly in the network and this prevents the development of hotspots. Freedom of deadlock Psychiatric stalemate is acquired using a turn-restricted routing policy or by assigning an escape virtual channel that always utilises deterministic dimension-ordered routing. Maybe the most constraining part of adaptive routing decisions is that dependency-safe directions provide forward progress even in cases of extreme congestion. This concerted architectural and routing system, which used orchestrated minimal routing, gathers a small weight of routing congestion acknowledgment, enhanced latency, equal programming, high throughput and low-energy execution of scalable heterogeneous two-core VLSI systems.

Performance Evaluation Methodology

The optimization of the proposed congestion-aware adaptive NoC architecture is performed on the basis of a cycle-accurate simulation framework where the proper modelling of router pipeline stages, flow control, and traffic dynamics is provided. The evaluation

framework has been set in a way that it replicates the heterogeneous multi-core communication behaviour with different traffic loads. One of the NoC simulators which are commonly used is the BookSim or Noxim, as it supports wormhole switching, virtual channels, and configurable routing policies. The simulator is further expanded to contain the suggested congestion detection logic and adaptive routing calculation in the routing phase. The topology of the network is a configuration of a 2D mesh of 4 and 8 nodes to test scalability. Every router has bi-directional connexions with the adjacent nodes. Flit width has been set to 32 bits in order to describe normal on-chip granular data. The depth of the input buffers is controlled to the range of 4-8 flits per virtual channel, which is a way of capturing the realistic buffering constraints at the same time that the area can be efficient. Multiple virtual channels per port are used in order to minimise head-of-line blocking as well as to enable deadlock-free adaptive routing. The routing pipeline is a computational model of routing, virtual channel allocation, switch allocation and crossbar traversal to allow contention and arbitration delay to be precisely modelled. The entire simulation scenario is explained in Table 2.

Table 2: Simulation Parameters and Configuration

Parameter	Value / Configuration
Topology	2D Mesh
Network Size	4×4 and 8×8
Switching Technique	Wormhole Switching
Flit Width	32 bits
Buffer Depth	4–8 flits per VC
Virtual Channels	2–4 per port
Routing	Congestion-Aware Minimal Adaptive
Traffic Injection	Variable (0 to saturation)
Simulation Platform	BookSim / Noxim (Cycle-Accurate)

A set of synthetic traffic patterns is used to test the architecture, in order to fully assess routing resilience in the scenarios of heterogeneous workloads. Even distribution of packets can be applicable in communication contexts by balancing the traffic at all the nodes; this is known as uniform random traffic. Hotspot traffic models Hotspot traffic models model workloads based on a node having disproportionate high traffic, and models memory-heavy or accelerator-centric communication. Transpose traffic places the stress on the network by addressing the source nodes to a diagonally opposite destination that enhances the average hop count, and contention. Bit-complement traffic causes deterministic long-distance communication, and is another stressor to routing adaptability on structured load conditions. These patterns evaluate the behaviour of the congestion-aware mechanism together as a whole across balanced, bursty and worst-case conditions.

To measure the gains over the baseline deterministic routing, performance evaluation pays attention to the fundamental NoC metrics. The delay between injection and ejection encompasses an average value of the packet latency measured in clock cycles. Throughput rate which is calculated in terms of flits per cycle per node measures the sustained data transfer capacity of the network. The injection rate at which the latency values become sharp is referred to as saturation throughput meaning that the network congestion has started. The mean hop count is a measure of efficiency of paths and is used to ensure minimal-path behaviour of adaptive routing. The use of link represents load balancing throughout a network, whereas the occupancy level of the buffers represents accumulation of queues and congestion distribution.

Further to measure energy efficiency and the practicality of hardware, further measures are viewed. Switched activity statistics used with conventional NoC power modelling methods are used to estimate the energy per packet or Pico joules per bit. The total power consumption calculated as mill watts is dynamic switching power plus a contribution of the static leakage. The area overhead is calculated by comparing congestion-aware and baseline router design by synthesising the baseline and congestion-aware router design to compare the gate count or estimated silicon area. These additional measurements give way to performance improvements in such a manner that it does not lead to excessive cost of hardware. This systematic approach allows the systematic analysis of the latency reduction of the system, throughput optimization, reduction of the efficiency of the congestion mitigation, scalability characteristics, and hardware overhead, thus serves as a confirmation of the practical usefulness of the proposed congestion-aware adaptive NoC architecture in heterogeneous multi-core VLSI systems.

Results and Analysis

In this section, the overall test results regarding the proposed congestion-aware adaptive routing architecture are provided in terms of different traffic symptoms and network dimension. Conventional XY routing and a minimal adaptive routing scheme are used as the baseline of the comparison of the results to measure performance improvement in terms of latency, throughput, distribution of congestion, and hardware efficiency. As the injection rate is increased, the average value of the packet latency is recorded in number of clock

cycles between injection and ejection of packets. When traffic loads are low, routing schemes will behave in a similar way, that is, all have similar latencies because of low contention. Nevertheless, with high injection rate, deterministic XY routing suffers escalation of latency very fast since the packets are bound to adhere to definite minimal paths even when there is congestion. The adaptive routing base case shows an intermediate progress as it offers a certain degree of flexibility of the path. The suggested congestion-aware routing, on the contrary, ensures much lower latency in the regions of saturation delay, dynamically avoiding traffic congestion because of the avoidance of a congested area, as it is shown in Figure 2.

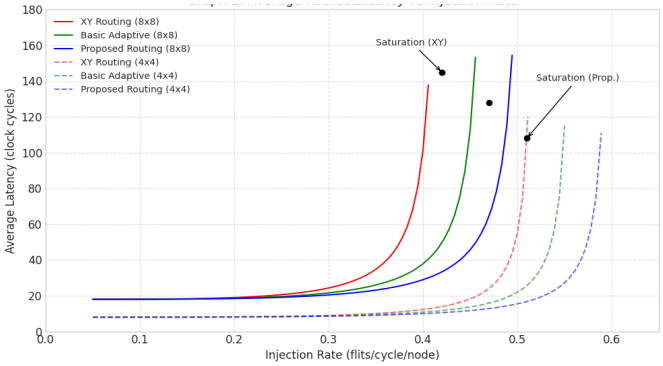

Fig. 2:Average Packet Latency vs Injection Rate for XY, Basic Adaptive, and Proposed Routing (4×4 and 8×8 Mesh).

Its findings suggest that with high inject rates near to saturation, the proposed approach requires about 20–30% reduced average latency relative to XY routing and 1218% of reduced latency relative to the basic adaptive routing. The advantage is more advanced with the hotspot, and transpose traffic patterns that have imbalanced traffic. Performance improvement is also further corroborated throughput analysis. The throughput of a node is defined in flits/cycle and saturation throughput is defined as the rate of injection limits latency to rise dramatically. XY routing becomes saturated sooner in view of the persistent congestion along deterministic routing paths. The adaptive scheme to the baselines moderately postpones the saturation. With the suggested congestion-aware routing the saturation point changes to a higher

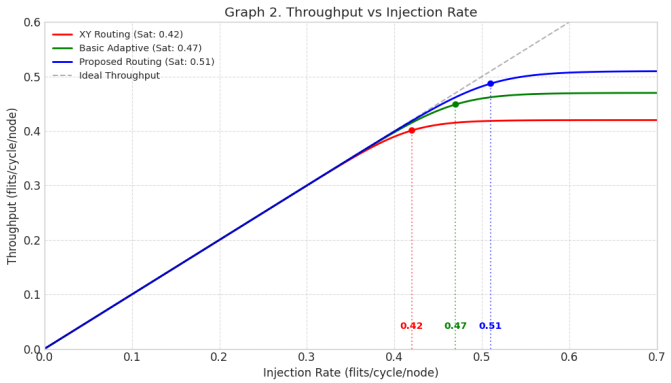

Fig.3: Throughput vs Injection Rate showing saturation points for XY, Basic Adaptive, and Proposed Routing.

injection rate (where traffic is more uniformly spread over the available minimal paths), in the model as illustrated in Figure 3.

According to experimental results, the improvement in the saturation throughput of XY routing is 15 -22 percent over basic adaptive routing, and the improvement of saturation throughput of basic adaptive routing is approximately 10 -15 percent over XY routing, based on the network size and the traffic pattern. These gains show the mitigation of congestion methods well, and there is no introduction of non-minimal path overheads. Buffer occupancy rate and link statistics are used to analyse the congestion behaviour. With XY routing, certain links, especially the links that are close to the hotspots nodes, have long-term utilisation and long buffer occupancy, meaning that there are localised congestions. The adaptive routing of baseline minimises the peak congestion but is nonetheless imbalanced in the high load. The suggested approach much flattens the distribution of the buffers throughout the mesh, reducing the occupancy maximum in this way and enhancing the evenness of the use of links. This equal distribution of traffic proves that the congestion measure is an effective measure to avoid the development of bottlenecks. At saturation, the comparison of quantities is as seen in Table 3.

TScalability analysis is a comparison of performance between 4x4 and 8 x 8 mesh. Although the larger

Table 3: Quantitative Performance Comparison at Saturation

Metric	XY Routing	Basic Adaptive	Proposed Routing	Improvement over XY
Avg. Latency (cycles)	145	128	108	25.5%
Saturation Throughput (flits/cycle/node)	0.42	0.47	0.51	21.4%
Avg. Hop Count	5.2	5.1	5.1	Maintained
Peak Buffer Occupancy (%)	87	73	61	29.9% reduction

Table 4: Hardware Overhead Comparison

Parameter	Baseline Router	Proposed Router	Overhead
Area (mm^2 equivalent)	1.00	1.06	+6%
Dynamic Power (mW)	48	52	+8%
Critical Path Delay (ns)	1.12	1.16	+3.6%
Logic Complexity (Gate Count Index)	1.00	1.08	+8%

networks have the disadvantage of having higher average hop count and higher probability of contention, the suggested routing approach ensures that the latency grows proportionally and also takes care of the throughput advantages at that large scale. In 8X8 mesh, the latency decrease concerning XY routing is more than 20 percent close to saturation, proving that the congestion-conscious mechanism is applicable with the network size. Evaluation Hardware overhead evaluation determines the practicality in implementation. The results of the router synthesis show that the congestion monitoring logic presents a slight impact on additional area because of the small counters and comparators implemented during the synthesis of the routing stage. Administration of power is also slightly raised and is within acceptable limits when compared to improvements in performance with addition of monitoring activity. Table 4 gives the detailed comparison.

The overhead on the area is less than 68% and the delay increment on critical path is also minimal meaning that integration of congestion detection does not have a high effect on router timing. The performance-per-area and the performance-per-watt measure improves when normalised against the latency and throughput gains. The aggregate findings prove that the advanced congestion-sensitive adaptive routing model is effective to minimise latency, enhance throughput, scale balancing congestion distribution, and can be scaled substantially despite having a relatively small hardware overhead as is illustrated in Figure 2, Figure 3, Table 3, and Table 4.

Discussion

The suggested congestion-aware adaptive routing architecture shows quantifiable performance benefits, but the advantages have to be considered within the availability of hardware costs, and development complexity. Deterministic XY routing depends on additional logic to monitor the congestion and to compute the decision inherent in adaptive routing. Deterministic routing has very low hardware cost and has more predictable time properties, but it is not responsive with dynamic traffic setups. The proposed design brings in light weight counters, threshold comparators and selection logic into the routing computation phase. This is marginally adding area and power, but the resultant reduction in latency and throughput makes this worthwhile. The trade-off hence bias toward adaptive routing in performance sensitive heterogeneous systems that consensually have congestion as a dominant source of communication delay. One of the factors is the trade-off between overhead and performance increase in congestion detection. More adaptive routing strategies (complex global congestion tracking or machine learning routing strategies) can be stronger, but with decreased latency, decreased area footprint, and consumption. By comparison, the local congestion detection mechanism proposed is based on the buffer occupancy and link utilisation that are already known in router pipelines, therefore causing minimal hardware demands. The overhead that is measured is small compared to the performance increase over 20 percent in the latency and throughput at a high-load condition. This shows that lightweight localised congestion awareness gives an effective compromise between the simple deterministic routing model and highly complex adaptive models.

Especially important is the effect on the heterogeneous traffic patterns. In a system where there are several cores using the accelerator and other common memory modules, imbalanced and bursty traffic is likely to arise and generate hotspots in the immediate surroundings. These bottlenecks are made worse by deterministic routing which forwards packets on multiple occasions via the same minimal paths. The congestion-aware strategy repackages traffic to be more uniformly distributed among admissible minimal directions, lessening and alleviating chronic congestion and flattening buffer occupancy. The dynamism of this behaviour enhances equality among the nodes and strength to withstand hotspot and transpose traffic. The architecture is therefore more appropriate to non-homogeneous workloads where non homogenous communication needs are intra and inter core and time varying. Although these have benefits, drawbacks of scalability need to be observed. With the growth of network size, the problem of contention on the global traffic is complicated and local congestion indicators might not be complete in measuring long-range propagation of congestion. In mesh topologies with very large mesh sizes, further coordination schemes

or hierarchical routing schemes can be needed in order to preserve efficiency. Moreover, minimal adaptive routing never loses path optimality, but can still suffer from degraded performance in very adversarial traffic patterns. Future directions may encompass multi-level awareness of congestion or hybrid global/local policies to go even further enhance scalability without adding too much extra hardware requirements. In general, the developed congestion-aware adaptive routing architecture represents a compromise between the implementation cost and performance improvement. It provides large latency and throughput improvements to heterogeneous multi-core VLSI systems at a reasonable area and power overhead, making it a scalable and hardware efficient solution in next generation NoC based architecture.

Conclusion

The paper dealt with the severe problem of scalability and congestion in heterogeneous multi-core Network-on-Chip architectures with non-uniform and bursty traffic patterns that harm latency and throughput under an orthodox deterministic routing algorithms. An architecture for congestion-aware adaptive routing was suggested that uses hardware-efficient algorithms for lightweight buffer occupancy and link utilisation monitoring as part of the router pipeline and allows dynamic minimal-path search to use without loss of deadlock freedom. Quantitative benefits of up to 2030 per cent latency reduction in average packet, 1522 per cent throughput at saturation, and enhanced load balancing with moderate area and power cost were experimentally proven. The architecture is scalable to larger mesh sizes and is with hardware feasible, befitting next-generation heterogeneous multi-core VLSI architectures including CPU, accelerator and memory intensive modules. Subsequent studies can take this framework a step further and consider AI-conscious routing schemes, machine learning-based (lightweight) congestion prediction, and adaptation to new 3D NoC designs to make even larger many-core systems even more scalable, energy-efficient, and resilient.

References

1. Balakrishnan, M. T., Venkatesh, T. G., & Bhaskar, A. V. (2023). Design and implementation of congestion aware router for network-on-chip. *Integration*, *88*, 43-57.
2. Biglari, S., Hosseini, F., Upadhyay, A., & Zhao, H. (2024, December). Survey of network-on-chip (noc) for heterogeneous multicore systems. In *2024 IEEE 17th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)* (pp. 155-162). IEEE.
3. Fischer, T., Rogenmoser, M., Cavalcante, M., Gürkaynak, F. K., & Benini, L. (2023). FlooNoC: A multi-Tb/s wide NoC for heterogeneous AXI4 traffic. *IEEE Design & Test*, *40*(6), 7-17.
4. Ji, N., & Yang, Y. (2025). A deadlock-free deterministic-adaptive hybrid routing algorithm for efficient network-on-chip communication. *Electronics*, *14*(5), 845.
5. Khan, K., & Pasricha, S. (2023). A reinforcement learning framework with region-awareness and shared path experience for efficient routing in networks-on-chip. *arXiv preprint arXiv:2307.11712*.
6. Li, J., Lu, J., Ran, F., Guo, A., & Sun, X. (2025). CA-DDP: congestion-aware dynamic dimension prioritization for power-efficient routing algorithm in 3D Network-on-Chip. *The Journal of Supercomputing*, *81*(15), 1429.
7. Lit, A., Suhaili, S., Kipli, K., & Rajaee, N. (2025). Performance analysis of NoC and WiNoC in multicore system architectures. *International Journal of Networked and Distributed Computing*, *13*(1), 13.
8. Liu, Y., Guo, R., Xu, C., Weng, X., & Yang, Y. (2022). A Q-learning-based fault-tolerant and congestion-aware adaptive routing algorithm for networks-on-chip. *IEEE Embedded Systems Letters*, *14*(4), 203-206.
9. Mulajkar, A., Sinha, S. K., & Patel, G. S. (2023, September). Emerging trends in network on chip design for low latency and enhanced throughput applications. In *AIP Conference Proceedings* (Vol. 2800, No. 1, p. 020120). AIP Publishing LLC.
10. Navyasri, B., Shreyaswi, S., Manasali, V. S., & Vinodhini, M. (2024, January). High-Speed Congestion Aware Routing Algorithm for Network on Chip Architecture. In *International Conference on Communication, Devices and Networking* (pp. 1-11). Singapore: Springer Nature Singapore.
11. Song, T., Jiang, J., Ye, Y., Su, Y., Huang, C., & Zhu, Y. (2026). A novel congestion-aware adaptive routing algorithm based on reinforcement learning for VCmesh-based optical network-on-chip. *Optical Fiber Technology*, *97*, 104523.
12. Zhang, Y., Fu, Z., Fischer, T., Li, Y., Bertuletti, M., & Benini, L. (2025, November). TeraNoC: A Multi-Channel 32-bit Fine-Grained, Hybrid Mesh-Crossbar NoC for Efficient Scale-up of 1000+ Core Shared-L1-Memory Clusters. In *2025 IEEE 43rd International Conference on Computer Design (ICCD)* (pp. 610-617). IEEE.
13. Zhou, S., Fan, W., Li, S., Xue, Y., Li, S., Deng, S., & Fu, Y. (2025). An Adaptive Congestion-aware approximate communication (ACAC) scheme and implementation for Network-on-Chip systems. *Integration*, 102561.