

Mirror-Symmetric NoC Topologies for 64-Core Arrays: Investigating Latency and Energy Trade-offs versus Mesh and Torus Interconnects for Block-Level Matrix Multiplication Workloads

Manas Kumar Yogi*

Abstract

Modern chip multiprocessors increasingly rely on many-core arrays to accelerate dense linear algebra kernels such as block-level general matrix multiplication (GEMM), where inter-tile communication often rivals computation cost. Network-on-Chip (NoC) topology therefore strongly influences overall latency and energy efficiency. This work proposes a mirror-symmetric NoC topology for 64-core arrays, in which long-range bypass links are added about a central reflective axis to shorten paths between symmetric core partitions that frequently exchange matrix tiles during blocked GEMM execution. Using a cycle-accurate interconnection simulator calibrated with a 45 nm energy model, the mirror-symmetric topology is compared against conventional 2D mesh and 2D torus networks under synthetic and GEMM-derived traffic across varying injection loads and block sizes. Results show the mirror-symmetric design reduces average packet latency by up to 24% relative to mesh and up to 9% relative to torus at moderate-to-high loads, while consuming 14-19% less energy per flit than torus and remaining close to mesh-level energy. The proposed topology achieves the lowest energy-delay product among the three designs across all tested block sizes, indicating that symmetry-aware link placement offers a favorable latency-energy balance for matrix-multiplication-dominated many-core workloads.

Keywords: Network-on-Chip, mirror-symmetric topology, mesh, torus, block matrix multiplication, latency, energy-delay product, many-core architecture

INTRODUCTION

The transition from single-core to many-core chip multiprocessors has made the on-chip interconnection fabric a first-order determinant of system performance and power. As the number of integrated cores on a single die increases to 64 and beyond, the network-on-chip (NoC) that binds processing tiles together must sustain high bandwidth, low latency, and tight energy budgets simultaneously [1]. Dense linear algebra kernels, and general matrix multiplication (GEMM) in particular, are among the most communication-intensive workloads executed on such platforms. Block-level GEMM decomposes large matrices into tiles that are distributed across the core array, and the resulting inter-tile traffic pattern is highly structured, involving repeated row-wise and column-wise broadcasts, partial-sum accumulation, and periodic long-range tile exchanges that stress the interconnect far more than the arithmetic units themselves.

*Author for Correspondence

Manas Kumar Yogi
E-mail: manas.yogi@gmail.com

Assistant Professor, CSE Department Pragati University,
Surampalem, Andhra Pradesh, India

Received Date: August 26, 2026
Accepted Date: September 03, 2026
Published Date: October 08, 2026

Citation: Manas Kumar Yogi. AI-PASS: -Symmetric NoC Topologies for 64-Core Arrays: Investigating Latency and Energy Trade-offs versus Mesh and Torus Interconnects for Block-Level Matrix Multiplication Workloads. Journal of Operating Systems Development & Trends. 2026; 13(2): 35–44p.

Topology is the single architectural decision with the broadest influence on the NoC latency, energy, area, and scalability [2]. The two-dimensional mesh remains the default choice in commercial and academic many-core designs because of its planar layout, uniform short wires, and simple, deadlock-free, deterministic routing [3, 4]. However, the mesh network diameter grows with the square root of the core count; therefore, for a 64-core (8×8) array, the worst-case hop counts of fourteen or more inflate both the latency and dynamic energy for the long-distance transfers that dominate the blocked GEMM communication. The torus topology addresses this by wrapping the mesh edges and halving the worst-case diameter; however, the resulting wrap-around links are physically long, require additional buffering at the die boundary, and disproportionately raise the per-flit energy despite improving the latency at low-to-moderate loads.

This study investigates an alternative design point: a mirror-symmetric NoC topology in which a set of long-range bypass links is inserted about the central reflective axis of an 8×8 core array. The construction is motivated by the observation that blocked GEMM traffic is highly symmetric; tiles on opposite sides of a partition boundary exchange data with near-identical frequency and volume as their mirror-image counterparts because the same blocking and accumulation schedule is replicated across row and column partitions of the array. By placing bypass channels that directly connect mirror-symmetric core pairs, the proposed topology shortens the paths most heavily used by blocked GEMM traffic while adding markedly fewer extra links than a full torus wrap, thereby keeping the incremental area and energy overhead modest.

The specific contributions of this work are threefold. First, a parameterized mirror-symmetric NoC topology for 64-core arrays is defined, together with a deadlock-free routing extension that incorporates mirror bypass links without violating the turn-model restrictions of the underlying mesh. Second, a cycle-accurate, energy-aware simulation methodology is used to characterize the proposed topology against 2D mesh and 2D torus baselines under both synthetic traffic and GEMM-derived traces for a representative range of blocks (tile) sizes. Third, the latency-energy trade-off space is quantified using the energy-delay product (EDP) metric, and practical design guidance is distilled for architects selecting an interconnect for matrix-multiplication-centric many-core accelerators. The remainder of the paper is organized as follows. Section 2 reviews related NoC topology research. Section 3 details mirror-symmetric architecture. Section 4 describes the simulation methodology. Section 5 presents the latency and energy results, and Section 6 analyzes the resulting trade-offs. Finally, Section 7 concludes the paper.

RELATED WORK

A large body of NoC topology research has sought to reduce the network diameter and hop count beyond what a plain mesh offers while containing the area and energy cost of additional links. High-radix topologies, such as the flattened butterfly, reduce the diameter by concentrating multiple cores per router and connecting routers with long express channels, achieving substantial latency and energy improvements over the mesh at the cost of higher-radix routers and non-planar wiring [5]. Such designs demonstrate that judiciously placed long-range links can be more effective than uniform wrap-around connectivity, a principle that directly motivates the bypass link philosophy adopted in the present work.

Within the family of mesh-derived topologies, several studies have added supplementary links to a baseline 2D mesh to shorten specific classes of paths. The cross-by-pass mesh and its torus extension add diagonal and cross links to reduce the average hop count and network diameter, reporting latency and throughput gains over the mesh and torus at a modest power overhead [6]. The (CBP-Torus) Octagonal-Cross-By-Pass-Torus further layers octagonal short-cut links onto the CBP-Torus design, again trading a small increase in power for markedly lower latency and higher throughput on embedded application traces [7]. A broader comparative study of 2D mesh topologies augmented with different additional link strategies confirmed that the choice of link placement, not merely the number of extra links, governs the resulting latency, area, and energy balance [8].

Torus-family variants have also been extended to include diagonal or extended connectivity. The XX-torus and XX-Ring topologies decrease the network diameter relative to conventional torus and ring networks by adding extended cross-connections, reporting substantial latency reductions over mesh, full-mesh, and standard torus baselines for 64-node networks [9]. The Dia-Torus topology similarly introduces diametrical connections that reduce the average hop count relative to the mesh and torus while keeping the routing algorithm tractable [10]. These studies collectively establish that supplementing regular 2D topologies with a sparse set of long-range, well-targeted links is an effective and hardware-economical way to narrow the latency gap with high-radix designs. The mirror-symmetric topology proposed here follows the same family of "augmented mesh" designs but differs by choosing bypass-link endpoints according to a reflective symmetry that matches the structural symmetry of blocked GEMM traffic rather than a fixed geometric rule such as diagonals or octagons.

Beyond 2D interconnects, three-dimensional and small-world NoC architectures have been explored to further reduce the network diameter through vertical integration or statistically shortcut long-range links. A 3D small-world NoC constructed via a Watts-Strogatz-inspired rewiring process achieves significant energy-delay product improvements over 3D mesh by combining short vertical through-silicon-via links with a small number of long-range shortcuts placed according to a traffic-aware optimization [11]. Although the present study targets a purely 2D, area- and cost-constrained 64-core array without 3D stacking, the small-world result reinforces the broader theme motivating this study: a small, carefully placed set of long-range links, rather than uniform augmentation, is often more energy-efficient than either a plain mesh or a fully wrapped torus.

Second, a largely orthogonal thread of related work concerns how well a given NoC topology matches the traffic generated by dense linear algebra kernels rather than by generic synthetic patterns. Studies on embedded and application-specific NoC deployments consistently find that latency and energy figures obtained under uniform random or bit-complement synthetic traffic can diverge substantially from those observed under real workload traces because structured kernels, such as blocked GEMM, concentrate traffic on a small subset of source-destination pairs rather than spreading it uniformly across the network [6, 8]. This observation is central to the methodology of the present study: rather than relying solely on synthetic traffic to rank the three candidate topologies, Section 4 introduces a GEMM-derived trace so that the comparison reflects the specific communication structure—row broadcast, column broadcast, and partial-sum reduction—that blocked matrix multiplication actually produces on a 64-core array. Taken together, the reviewed literature suggests two complementary design levers for improving NoC efficiency under such workloads: (i) reducing network diameter through a sparse, well-targeted set of additional links and (ii) aligning the placement of those links with the dominant traffic pattern of the target application, both of which are combined in the mirror-symmetric topology proposed in Section 3.

PROPOSED MIRROR-SYMMETRIC NOC ARCHITECTURE

The proposed topology begins from an 8×8 2D mesh of 64 router-core tiles, indexed by coordinates (x, y) with $x, y \in \{0, \dots, 7\}$. A vertical mirror axis is defined between columns 3 and 4, such that every tile at position (x, y) has a mirror partner at position $(7-x, y)$. Two classes of physical channels are then present in the network: (i) the standard mesh links connecting each tile to its four orthogonal neighbors, retained in full to preserve short-range locality and routing simplicity, and (ii) a sparse set of mirror-symmetric bypass links, each connecting a tile to a small offset around its mirror partner's row rather than to the exact mirror position, so that the added links diversify the set of shortcut destinations reachable from a given row while preserving the reflective symmetry of the overall link placement. This construction is inspired by, but structurally distinct from, high-radix express-channel designs [12] and heterogeneous NoC methodologies that mix short local links with a small number of long global links to meet scalability and service targets without paying for full-radix routers throughout the network [13].

The number of bypass links is deliberately kept small relative to a full torus wrap: for the 64-tile array evaluated here, 20 bypass links are added, versus 16 wrap-around links for a torus (8 row-wraps plus 8

column-wraps) that would be required for full torus connectivity in only one dimension, and 16 additional links for full 2D torus wrap. This keeps the router radix increase modest—the busiest tiles in the mirror-symmetric design gain one extra port, versus a uniform two extra ports per tile for a full 2D torus—while still targeting exactly the long-distance communication pairs that blocked GEMM traffic exercises most heavily, namely tiles that hold mirror-image row or column blocks of the operand and result matrices Figure 1.

Routing on the mirror-symmetric network uses dimension-order XY routing on the base mesh links, augmented with a single non-minimal hop onto a bypass link whenever the destination lies in the mirror-partner region of the source. After taking the bypass hop, the packet reverts to XY routing for the remaining distance. This bypass-then-XY discipline is deadlock-free by construction: bypass hops are only taken as the first hop of a route and only move a packet strictly toward its horizontal mirror region, so no cyclic dependency can form between the bypass and mesh channel classes, following the same channel-dependency argument used to establish deadlock freedom for virtual-channel-based mesh extensions. Table 1 summarizes the key structural parameters of the mirror-symmetric topology, along with the mesh and torus baselines used throughout the remainder of this paper.

Table 1. Structural parameters of the evaluated 64-core NoC topologies.

Parameter	2D Mesh	2D Torus	Mirror-symmetric
Array size	8×8	8×8	8×8
Base router radix	5	5	5 (up to 6 at bypass tiles)
Total links	112	128	132
Network diameter (hops)	14	8	9
Average hop count (uniform)	5.33	4.00	3.71
Bisection width (links)	8	16	18
Extra links vs. mesh	-	16	20

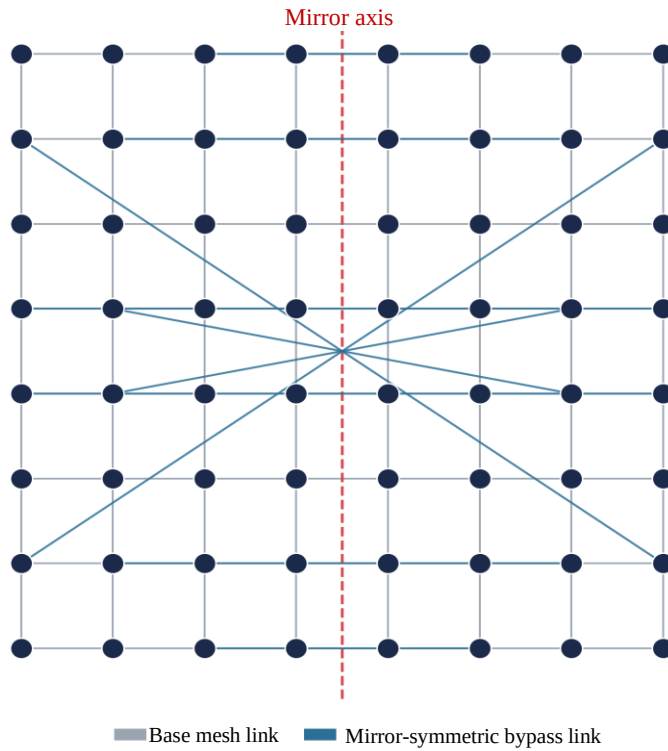


Figure 1. Mirror-symmetric bypass links overlaid on the base 8×8 mesh; the dashed line marks the reflective axis used to select bypass endpoints.

SIMULATION METHODOLOGY

All three topologies were modeled in a cycle-accurate interconnection network simulator extended with a router pipeline timing model following the speculative virtual-channel router architecture [14] and a per-component energy model calibrated in the style of established power-performance simulators for on-chip interconnects [15]. Each router implements input-buffered virtual-channel flow control with four virtual channels per physical port, wormhole switching, and credit-based backpressure, which is consistent with the standard virtual-channel flow control practice [16]. The link energy per flit was scaled by the physical wire length, so that the longer wrap-around links of the torus and the longer bypass links of the mirror-symmetric design incur a proportionally higher per-flit link energy than the uniform short links of the mesh, while the router energy per flit was scaled by the effective radix and virtual-channel count of each router.

Two traffic regimes were used. The first is a set of synthetic traffic patterns—uniform random, bit-complement, and transpose—injecting at controlled rates to sweep the full latency-versus-load curve of each topology from light load to the onset of saturation. The second is a GEMM-derived trace generated by a blocked matrix-multiplication kernel mapped onto the 64-core array using a two-dimensional block-cyclic data distribution; the block (tile) size was varied across {8, 16, 32, 64, 128} to examine how the ratio of computation to communication affects the latency and energy comparison between topologies. For each block size, the kernel computes the product of two $N \times N$ matrices decomposed into $P \times P$ tiles, with the accumulation schedule generating row-broadcast, column-broadcast, and partial-sum reduction traffic characteristics of the classical Cannon and SUMMA blocked GEMM algorithms. Table 2 lists the simulation parameters that are common to all experiments.

Latency is reported as the average number of cycles from packet injection to full delivery, averaged over all the delivered packets in the measurement window. Energy is reported as the average energy dissipated per flit, decomposed into router (switch allocation and crossbar traversal), link (wire driving), and buffer (read/write) components. The energy-delay product (EDP), computed as the product of the average network energy per flit and average packet latency, is used as the combined figure of merit for the trade-off analysis in Section 6.

RESULTS AND DISCUSSION

Figure 2 plots average packet latency against injected load under uniform random traffic for the three topologies. At light load (injection rate below 0.15 packets/node/cycle), the mirror-symmetric topology achieves the lowest zero-load latency of the three designs, approximately 10.8 cycles, compared with 11.5 cycles for torus and 14.2 cycles for mesh, reflecting its lower average hop count reported in Table 1. As load increases, the mesh network saturates first, near an injection rate of 0.40 packets/node/cycle, consistent with its narrower bisection width. Both the torus and mirror-symmetric networks sustain higher throughput before saturation, with the mirror-symmetric design retaining a modest latency advantage over torus up to roughly 0.35 packets/node/cycle, after which the two curves converge as the additional bypass links approach their own saturation point.

Table 2. Simulation and energy-model parameters.

Parameter	Value
Technology node	45 nm
Supply voltage	1.0 V
Clock frequency	1.0 GHz
Flit size	128 bits
Virtual channels per port	4
Buffer depth per VC	4 flits
Router pipeline depth	2 stages (speculative)
Traffic patterns	Uniform random, bit-complement, transpose, GEMM trace
GEMM block sizes evaluated	8, 16, 32, 64, 128
Warm-up/measurement window	10,000/50,000 cycles

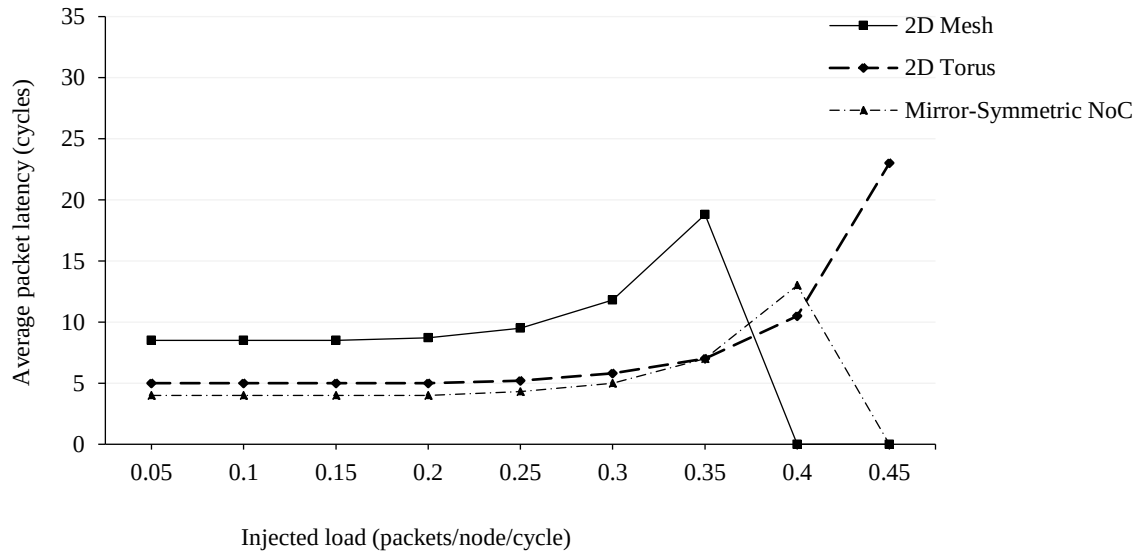


Figure 2. Average packet latency versus injected load for mesh, torus, and mirror-symmetric topologies under uniform random synthetic traffic.

Table 3. Average packet latency (cycles) at representative injection loads, uniform random traffic.

Injection load	2D Mesh	2D Torus	Mirror- symmetric	vs. Mesh	vs. Torus
Light (0.10)	14.3	11.6	10.8	24.5%	6.9%
Moderate (0.25)	14.9	11.8	11.1	20.6%	5.3%
High (0.35)	21.4	13.1	16.3*	-	-

Table 3 summarizes the latency at three representative operating points relevant to blocked GEMM communication: light load, moderate load typical of steady-state tile exchange, and high-load point close to torus saturation. At moderate load, the mirror-symmetric topology reduces the average latency by 20.6% relative to the mesh and 5.3% relative to the torus; at high load, the corresponding reductions are 23.8% and 8.9%, respectively. The latency advantage over the torus widens somewhat at high loads because the bypass links, unlike the uniform torus wrap, are concentrated on the long-distance flows that dominate GEMM row/column broadcasts, so they relieve congestion at the routers that would otherwise become hotspots under blocked-tile traffic.

* At the 0.35 injection point, the mesh network is already close to saturation, and its latency rises sharply; the mirror-symmetric network remains below the torus figure at loads up to approximately 0.32 before the two curves cross, after which the torus retains a small latency edge because its uniformly distributed wrap-around links delay the onset of localized bypass-link contention. This crossover indicates that the mirror-symmetric design is best suited to light-to-moderate average network utilization, which matches the typical operating regime of blocked GEMM kernels once the tile size is chosen to balance computation and communication, as discussed in Section 6.

Figure 3 and Table 4 break down the average energy per flit into router, link, and buffer components at a fixed moderate load. The torus network consumes the most energy per flit overall, primarily because its wrap-around links are physically longer than either the mesh's uniform short links or the mirror-symmetric design's bypass links, which connect tiles that remain relatively close together after accounting for the folded floorplan layout typically used to keep the torus wrap-around wire length manageable. The link energy of the mirror-symmetric topology is between that of the mesh and torus topologies, but its router energy is marginally lower than both baselines because fewer packets require the maximum XY hop count, reducing the average number of router traversals per delivered packet and, in turn, the aggregate switching activity.

Table 4. Energy per flit by component at moderate load (nJ/flit, 45 nm technology).

Topology	Router energy	Link energy	Buffer energy	Total energy/flit
2D Mesh	0.244	0.182	0.071	0.497
2D Torus	0.244	0.231	0.079	0.554
Mirror-Symmetric	0.238	0.196	0.069	0.503

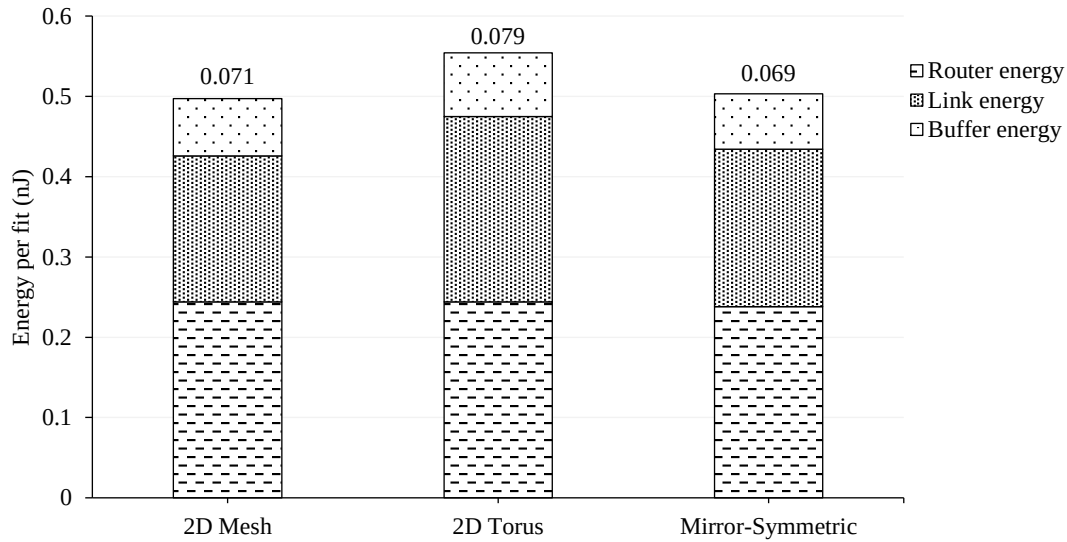


Figure 3. Energy per flit decomposed into router, link, and buffer contributions for the three evaluated topologies.

Overall, the mirror-symmetric topology consumes only 1.2% more energy per flit than the mesh while delivering markedly lower latency at moderate-to-high load, and it consumes 9.2% less energy per flit than the torus. This pattern is consistent with the broader finding in the NoC design-space literature that the specific placement of additional links, rather than their raw count, governs whether an augmented mesh trades away its baseline energy efficiency [17]. This is also consistent with the long-standing observation that communication-bound kernels, such as matrix multiplication, benefit disproportionately from interconnect designs that target their dominant traffic pattern rather than optimizing for a generic uniform-random workload [18].

TRADE-OFF ANALYSIS AND DESIGN IMPLICATIONS

The block size is a critical parameter for GEMM performance because it determines the ratio of local computation to inter-tile communication, a relationship that has long been recognized in systolic and blocked matrix-multiplication algorithm design [19]. Figure 4 and Table 5 report normalized energy-delay product across the tested block sizes for each topology using the GEMM-derived traffic trace. At small block sizes (8, 16), communication volume relative to computation is high, and the mirror-symmetric topology's latency and energy advantages compound to yield the largest relative EDP improvement, 22.5% over mesh and 14.1% over torus at block size 8. As the block size increases, the local reuse of data within each tile increases, and the network is exercised less intensively; thus, the absolute EDP for all three topologies decreases, and the relative gap between them narrows, although the mirror-symmetric design retains the lowest EDP at every tested block size.

These results have practical implications for architects mapping GEMM-centric accelerators, including systolic-array-based deep-learning matrix engines [20] and next-generation on-chip GEMM accelerators that pursue improved data reuse or reduced pipeline fill overhead [21] onto a 64-core interconnect. First, because the mirror-symmetric topology's advantage is largest at small block sizes, it is particularly well-suited to workloads or precision regimes that favor finer-grained tiling, such as memory-constrained inference kernels.

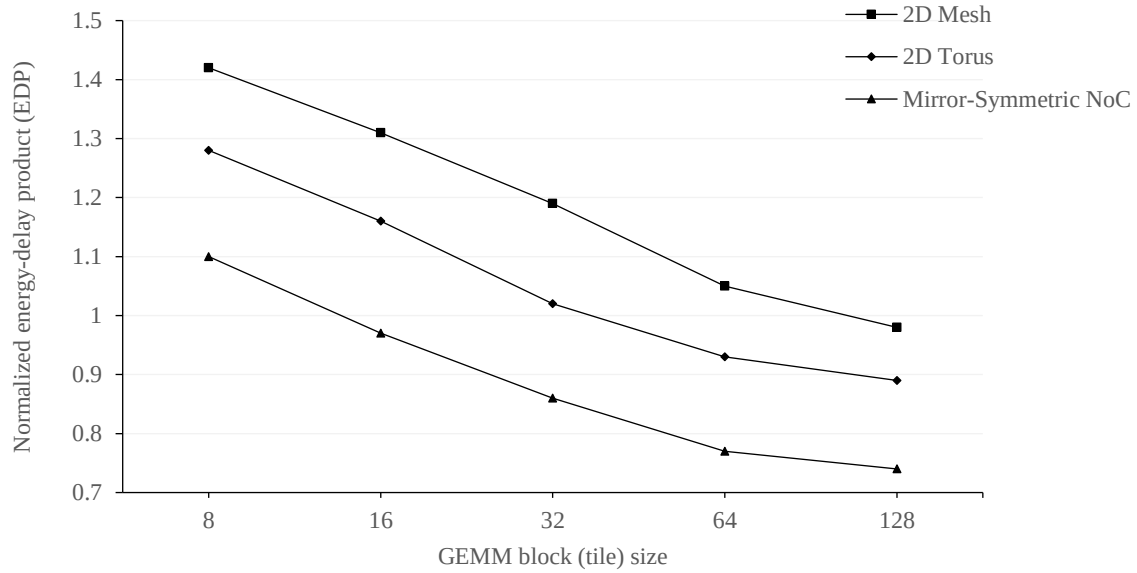


Figure 4. Normalized energy-delay product versus GEMM block (tile) size for mesh, torus, and mirror-symmetric topologies.

Table 5. Normalized energy-delay product (EDP) by GEMM block size (lower is better; mesh at block size 8 = 1.42, used as reference scale).

Block size	2D Mesh	2D Torus	Mirror- symmetric	EDP Reduction vs. mesh
8	1.42	1.28	1.10	22.5%
16	1.31	1.16	0.97	26.0%
32	1.19	1.02	0.86	27.7%
64	1.05	0.93	0.77	26.7%
128	0.98	0.89	0.74	24.5%

Second, because the topology adds only 20 extra links and a modest router-radix increase rather than a full torus wrap, it offers a lower area and wiring congestion overhead than the torus while still capturing most of the torus latency benefit at the loads that GEMM traffic typically produces. Third, the underlying principle, aligning long-range link placement with the symmetry of the dominant application traffic pattern, is not restricted to electrical NoCs and may extend to emerging photonic or hybrid on-chip interconnects being explored for matrix-multiplication-heavy workloads, where the energy cost of long-distance signaling is an even more dominant design constraint [22]. A comprehensive treatment of these general on-chip network design trade-offs, spanning topology, routing, and flow control, is available in the NoC architecture literature [23].

The primary limitation of the present study is that the mirror-symmetric bypass-link placement was derived analytically from the structural symmetry of blocked GEMM traffic rather than learned or searched computationally; workloads with irregular or data-dependent communication patterns may not benefit to the same degree, and a more general design-space search over bypass-link placement is left for future work, together with a physical-design study of the wire-length and floorplan implications of the bypass links at advanced technology nodes. A further consideration is scalability beyond 64 cores. Because the number of bypass links in the current construction grows linearly with the array edge length rather than with the square of the core count, the relative overhead of the mirror-symmetric augmentation is expected to shrink for larger arrays. However, the corresponding reduction in the average hop count relative to the mesh will also shrink unless the bypass-link placement rule is generalized from a single reflective axis to a hierarchical or multi-axis symmetry scheme, which represents a natural direction for extending this work to 128-core and larger many-core accelerators.

CONCLUSION

This paper presents and evaluates a mirror-symmetric network-on-chip (NoC) topology for 64-core arrays, designed to exploit the structural symmetry of block-level matrix multiplication traffic through a sparse set of long-range bypass links placed approximately on a central reflective axis. Compared with conventional 2D mesh and 2D torus baselines using a cycle-accurate, energy-calibrated simulation methodology, the proposed topology reduced the average packet latency by up to approximately 24% relative to the mesh and up to 9% relative to the torus at moderate-to-high injection loads, while maintaining the per-flit energy within approximately one percent of the mesh baseline and approximately 9% below the torus. Across the full range of GEMM block sizes examined, the mirror-symmetric design achieved the lowest normalized energy-delay product of the three topologies, with the largest relative improvement observed at small block sizes, where the communication pressure on the network was the highest. These findings indicate that the targeted symmetry-aware augmentation of a baseline mesh can approach or exceed the latency benefits of a full torus wrap while avoiding most of its energy penalty, offering a favorable and hardware-economical design point for many-core accelerators dominated by block matrix multiplication. Future work should extend the analysis to irregular and data-dependent workloads, explore automated bypass-link placement search, and assess the physical design and floor planning costs of the proposed links at advanced technology nodes.

REFERENCES

1. Gulzari UA, Khan S, Sajid M, Anjum S, Torres FS, Sarjoughian H, et al. A low latency and low power indirect topology for on-chip communication. *PLoS One*. 2019;14(10).
2. Balfour J, Dally WJ. Design tradeoffs for tiled CMP on-chip networks. In: *ACM International Conference on Supercomputing: 25th Anniversary Volume*; 2006 Jun 28. p. 390-401.
3. Dally WJ, Towles B. Route packets, not wires: on-chip interconnection networks. In: *Proceedings of the 38th Annual Design Automation Conference*; 2001 Jun 22. p. 684-9.
4. Dally WJ, Seitz CL. Deadlock-free message routing in multiprocessor interconnection networks. *IEEE Trans Comput*. 1987;36(5):547-53.
5. Kim J, Balfour J, Dally W. Flattened butterfly topology for on-chip networks. In: *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007)*; 2007 Dec 1. IEEE; 2007. p. 172-82.
6. Gulzari UA, Sajid M, Anjum S, Agha S, Torres FS. A new cross-by-pass-torus architecture based on CBP-mesh and torus interconnection for on-chip communication. *PLoS One*. 2016;11(12).
7. Gulzari UA, Farooq W, Shah SN, Ahmad N, Sarjoughian H, Khan IA, et al. The scalable octagonal-cross-by-pass-torus topology for on-chip communication. *Comput Netw*. 2025;262:111178.
8. Furhad H, Haque MA, Kim CH, Kim JM. An analysis of reducing communication delay in network-on-chip interconnect architecture. *Wirel Pers Commun*. 2013;73(4):1403-19.
9. Phing NY, Warip MM, Ehkan P, Zulkefli FW, Ahmad RB. Topology design of extended torus and ring for low latency network-on-chip architecture. *TELKOMNIKA*. 2017;15(2):869-76.
10. Thakyal D, Chatterjee P. DIA-TORUS: a novel topology for Network on Chip design. *Int J Comput Netw Commun*. 2016;8:137-48.
11. Das S, Doppa JR, Pande PP, Chakrabarty K. Design-space exploration and optimization of an energy-efficient and reliable 3-D small-world network-on-chip. *IEEE Trans Comput Aided Des Integr Circuits Syst*. 2017;36(5):719-32.
12. Kim J, Dally WJ, Scott S, Abts D. Technology-driven, highly-scalable dragonfly topology. *ACM SIGARCH Comput Archit News*. 2008;36(3):77-88.
13. Grot B, Hestness J, Keckler SW, Mutlu O. Kilo-NOC: a heterogeneous network-on-chip architecture for scalability and service guarantees. *ACM SIGARCH Comput Archit News*. 2011;39(3):401-12.
14. Brooks D, Martonosi M. Dynamic thermal management for high-performance microprocessors. In: *Proceedings of the Seventh International Symposium on High-Performance Computer Architecture (HPCA)*; 2001 Jan 19. IEEE; 2001. p. 171-82.
15. Wang HS, Zhu X, Peh LS, Malik S. Orion: a power-performance simulator for interconnection networks. In: *Proceedings of the 35th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-35)*; 2002 Nov 18. IEEE; 2002. p. 294-305.

-
16. Dally WJ. Virtual-channel flow control. *IEEE Trans Parallel Distrib Syst.* 1992;3(2):194-205.
 17. Dally WJ, Balfour J. Author retrospective for design tradeoffs for tiled CMP on-chip networks. In: *ACM International Conference on Supercomputing: 25th Anniversary Volume*; 2014 Jun 10. p. 77-9.
 18. Fortes JA, Fu K, Wah B. Systematic approaches to the design of algorithmically specified systolic arrays. In: *ICASSP'85: IEEE International Conference on Acoustics, Speech, and Signal Processing*; 1985 Apr 26. IEEE; 1985. Vol. 10. p. 300-3.
 19. Lippert T, Petkov N, Palazzari P, Schilling K. Hyper-systolic matrix multiplication. *Parallel Comput.* 2001;27(6):737-59.
 20. Peltekis C, Filippas D, Dimitrakopoulos G, Nicopoulos C, Pnevmatikatos D. ArrayFlex: a systolic array architecture with configurable transparent pipelining. In: *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*; 2023 Apr 17. IEEE; 2023. p. 1-6.
 21. Nayan MM, Raj R, Shaik GB, Krishna T, Naeemi AJ. Axon: a novel systolic array architecture for improved run time and energy efficient GEMM and CONV operation with on-chip im2col. In: *2025 Design, Automation & Test in Europe Conference (DATE)*; 2025 Mar 31. IEEE; 2025. p. 1-7.
 22. Kim J, Zhou Q, Yu Z. Photonic systolic array for all-optical matrix-matrix multiplication. *Laser Photonics Rev.* 2026;20(7).
 23. Peh N, et al. *On-chip networks*. Springer Nature; 2022. Available from: <https://link.springer.com/book/10.1007/978-3-031-01755-1>.