Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The Evolution of High-Speed Interfaces and Memory Systems in AI Architectures

Phani Suresh Paladugu
Synopsys

ARTICLE INFO

Received: 17 July 2025

Revised: 05 Aug 2025

Accepted: 18 Aug 2025

ABSTRACT

High-speed interfaces and memory systems form the foundation of modern artificial intelligence architectures, enabling them to meet the rapidly growing computational demands of advanced neural networks. Progress in these domains centers on maximizing data movement efficiency while balancing the trade-offs between bandwidth and power consumption. In SerDes design, key considerations include clocking strategies, signal integrity control, and the physical implementation challenges that directly influence overall system performance. Memory hierarchy optimization requires carefully managing capacity, bandwidth, and power efficiency across multiple technology generations. Emerging solutions—such as processing-in-memory architectures and next-generation non-volatile memories—help reduce data transfer overhead. Together, interface design and memory subsystem advancements create the scalable infrastructure needed to power next-generation AI applications across a wide range of deployment environments.

Keywords: High-speed interfaces, memory hierarchies, SerDes architectures, processing-in-memory, artificial intelligence accelerators

1. Introduction

The rapid advancement of Artificial Intelligence (AI) has placed unprecedented demands on computing infrastructure. As AI workloads grow in both complexity and scale, hardware systems face mounting challenges in data movement, processing efficiency, and power consumption. The performance boundaries of modern AI accelerators are fundamentally shaped by the complex interaction between high-speed interfaces and memory systems. Horowitz's landmark analysis of computing's energy problem underscores why optimizing interfaces and memory is central to AI hardware design. His findings show that performing a 32-bit floating-point operation at 28nm consumes just 0.9 pJ, yet moving that same data a mere 10 mm on-chip requires about 25 pJ—nearly 28 times more energy[1]. The disparity is even greater for off-chip memory: accessing a 32-bit value from DRAM costs roughly 640 pJ, a staggering 711× more than the computation itself[1]. These energy imbalances explain why today's AI accelerators operate at only 5–15% of their theoretical peak efficiency, with most of the loss attributable to data movement across interfaces and memory hierarchies.

Interface design challenges are amplified by the distinctive traffic patterns of transformer-based AI architectures. Studies on processing-in-memory approaches for AI workloads reveal that attention mechanisms produce access patterns with minimal spatial and temporal locality, driving memory bandwidth demands of 35–42 GB/s per TFLOP of compute, substantially higher than the 10–15 GB/s per TFLOP typical of CNNs [2]. For state-of-the-art AI accelerators delivering 64 TFLOPS, this equates to bandwidth requirements exceeding 2.7 TB/s, necessitating advanced SerDes interfaces running at 224 Gbps or 112 Gbps per lane, alongside HBM3E/HBM4 memory subsystems capable of up to 4.8 TB/s aggregate throughput. Such systems commonly employ sophisticated clock and data recovery (CDR) circuits to maintain jitter tolerance below 0.3 UI at these speeds, while adaptive equalization compensates for channel losses exceeding 40 dB at the Nyquist frequency.

Memory hierarchies for AI systems have evolved toward heterogeneous architectures combining multiple technologies. The energy efficiency of these memory accesses varies dramatically across the

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

hierarchy: 5-8 pJ/bit for on-chip SRAM, 8-12 pJ/bit for HBM3E accesses, and 40-60 pJ/bit for off-package memory [1]. Current designs typically feature 8-32 MB of on-chip SRAM (consuming 0.2-0.5 mm² per MB at 7nm), providing 15-25 TB/s of bandwidth to processing elements, complemented by 16-128 GB of HBM3E memory delivering 2.4-4.8 TB/s through 1024-2048 bit interfaces operating at 5.1-6.4 Gbps [2].

Operation Type	Energy Efficiency	Technology Context	Design Impact
Floating-point computation	Minimal energy cost	Advanced process nodes	Computation-optimized designs
On-chip data movement	Moderate energy penalty	Short-distance transfers	Layout optimization critical
Off-chip memory access	High energy overhead	External DRAM interfaces	Memory hierarchy essential
SRAM access	Low energy consumption	On-chip storage	Cache-friendly architectures
HBM access	Moderate energy cost	High-bandwidth memory	Bandwidth-capacity tradeoffs

Table 1: Energy Cost Hierarchy in AI Computing Systems [1,2]

2. System-Level Requirements for High-Speed Interfaces in AI Architectures

The surge in AI model complexity is fueling unprecedented data movement needs, making high-speed interfaces indispensable. At the core of these connections lies SerDes technology, the backbone of modern interconnect systems.

Research by Shao et al. on deep learning accelerator simulation highlights a sharp rise in bandwidth requirements. Using the SimBA framework, they show that transformer models such as BERT generate memory access patterns demanding 24.7–37.2 GB/s per TFLOP of compute, substantially higher than the 9.5–15.8 GB/s per TFLOP typical of CNN models [3]. Their analysis of on-chip network traffic patterns reveals that transformer self-attention mechanisms produce traffic bursts up to 3.2× greater than peak sustainable bandwidth, with spatial locality coefficients of 0.41–0.57 versus 0.76–0.83 for CNNs. This lower locality drives the need for interface designs capable of delivering 1.2–1.8 TB/s aggregate bandwidth for accelerators targeting 32–48 TFLOP/s performance, while keeping router traversal latencies within 1.7–2.3 ns to sustain computational efficiency above 65% of theoretical peak [3].

Designing high-speed interfaces for AI applications requires careful consideration of several system-level factors. Key among these are bandwidth density—measured in Gbps/mm—which directly impacts chip size and overall system cost; power efficiency—typically expressed in pJ/bit—which influences thermal constraints and operating expenses; and performance characteristics that govern real-time processing capability. In a comprehensive survey of low-power AI accelerators, Åleskog et al. analyzed 37 commercial and academic designs, showing that bandwidth density at chip boundaries has risen from 89 Gbps/mm in 2018 to 327 Gbps/mm in today's designs, leveraging advanced packaging [4]. Their findings also indicate that interface power efficiency for cutting-edge data center AI accelerators has improved to 2.1–3.8 pJ/bit at 112 Gbps, while edge AI deployments achieve 1.8–2.4 pJ/bit at lower data rates of 16–32 Gbps [4].

AI workloads exhibit distinctive traffic patterns marked by intensive, often bursty data transfers and variable packet sizes. Shao's simulation studies show that attention mechanisms in transformer models produce packet sizes that vary by 2.7–3.4× within a single forward pass, with temporal correlation coefficients as low as 0.38–0.45 [3]. Such behavior demands specialized flow control schemes capable of dynamically adjusting to fluctuating bandwidth needs while preserving quality of service. In addition,

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

AI systems in edge deployments face stringent power and thermal constraints. Åleskog's extensive survey reports power budgets ranging from 1.8 W for smartphone-class accelerators to 15 W for edge servers, with thermal design power (TDP) limits of 2.5–17 W. These limits, in turn, cap maximum interface power consumption at 0.3–2.1 W [4].

Network Architecture	Bandwidth Demand	Locality Pattern	Design Constraint	Power Classification
Convolutional Neural Networks	Moderate throughput	High spatial locality	Compute-bounded	Standard power envelope
Transformer Models	High throughput	Limited locality	Memory-bounded	Enhanced power delivery
Attention Mechanisms	Burst-intensive	Low correlation	Latency-sensitive	Thermal management
Edge Applications	Low data rates	Variable patterns	Power-constrained	Battery optimization
Data Center Deployment	Maximum bandwidth	Streaming access	Performance-focused	Cooling infrastructure

Table 2: AI Workload Traffic Characteristics and System Requirements [3,4]

3. SerDes Design Considerations for AI Workloads

SerDes design in AI applications necessitates a rigorous focus on several critical factors, foremost among them being the choice of clocking architecture. Designers typically select from forward clocking, embedded clock recovery, or source-synchronous methods according to the application's specific demands.

Recent advances in high-speed SerDes implementation for AI accelerator interfaces demonstrate critical design parameters. State-of-the-art 112 Gbps PAM-4 transceivers achieve bit error rates of less than 10^-12 while maintaining jitter tolerance of 0.3 UI (unit interval) at 28 GHz, utilizing quarter-rate architectures with 4:1 multiplexing to reduce clock distribution challenges [5]. These implementations demonstrate that clock accuracy must be maintained within 1.8-2.2 ps of peak-to-peak jitter to sustain reliable operation, requiring 6-bit phase interpolators with 0.7 ps resolution. The CDR architecture employs second-order loops with an optimized bandwidth of 8.5 MHz, achieving lock times of 156 ns while consuming 67 mW in 7nm FinFET technology. Analysis quantifies the relationship between supply noise and timing jitter, showing that each 10 mV of power supply noise translates to approximately 0.52 ps of deterministic jitter at 112 Gbps, highlighting the importance of power integrity for high-speed interfaces in AI systems where multiple SerDes lanes switch simultaneously [5].

Signal integrity challenges are paramount in high-speed SerDes design, especially as data rates reach the 112 Gbps PAM-4 signaling regime. Recent studies offer detailed insights into the equalization demands for high-speed SerDes functioning within typical AI system environments [6]. Modern 112 Gbps PAM-4 receiver implementations demonstrate that channels with insertion loss of 35.7 dB at Nyquist frequency require a combination of CTLE providing 12.4 dB of peaking and 9-tap DFE consuming 124 mW to achieve vertical eye-opening of 28.6 mV (14.3% of nominal eye height). Analysis across multiple channel configurations reveals that crosstalk degrades receiver sensitivity by 3.4-4.2 dB in dense routing environments typical of AI accelerator packages, necessitating careful layout with guard traces maintaining minimum separation of 2.8× the trace width [6]. For power integrity, measurements show that PDN impedance must remain below 0.18 Ω from 10 MHz to 5.6 GHz to maintain supply ripple within 12 mV for a 0.8V supply voltage, requiring 6.8-8.4 µF of on-die decoupling capacitance per SerDes macro to prevent supply-induced jitter from exceeding 0.03 UI.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

These concerns extend to physical implementation, where floor planning and placement constraints significantly impact performance. Modern SerDes implementations demonstrate that placement must consider both signal and power integrity, with optimal configurations placing transmit and receive circuits within 325 μm of I/O pads to minimize on-die routing parasitics [5]. Measurements show that each additional 100 μm of on-die routing introduces 0.23 dB of insertion loss and 3.4 ps of propagation delay variation across process corners. Analysis further reveals that differential pairs must maintain a consistent impedance of 85-100 Ω with less than 3% mismatch between positive and negative traces to prevent mode conversion that degrades signal integrity [6]. The increasing integration density in modern AI chips further complicates these challenges, with implementations utilizing silicon interposer technology providing 9.2× higher routing density than organic substrates while reducing channel loss by 0.28 dB/mm.

Design Parameter	Performance Target	Implementation Challenge	Signal Integrity Factor	Power Consideration
Bit Error Rate	Ultra-low error tolerance	Noise immunity	Jitter management	Supply stability
Clock Recovery	Phase accuracy	Loop bandwidth	Timing precision	CDR power consumption
Equalization	Channel compensation	Adaptive algorithms	Eye diagram optimization	Equalizer overhead
Physical Layout	Impedance matching	Routing constraints	Crosstalk mitigation	PDN design
Packaging Integration	High density	Thermal management	Loss minimization	Decoupling requirements

Table 3: SerDes Design Parameters for AI Applications [5,6]

4. Memory Hierarchy Optimization for AI Systems

The memory subsystem represents the most important component in modern AI architecture, where memory bandwidth bottlenecks often serve as major performance limiters. Designing an effective memory hierarchy for AI workloads requires a deep understanding of specific access patterns generated by various neural network topologies.

Kwon and colleagues' research on MAERI demonstrates the criticality of memory hierarchy optimization in AI accelerators. Their detailed analysis shows that dataflow patterns in CNNs can achieve reuse factors of 28-196× for weights and 4-18× for activations when properly mapped to hardware, while less regular networks like transformers exhibit reuse factors of only 3.7-12.3× [7]. Their reconfigurable architecture implements a three-level memory hierarchy with 3.2KB register files distributed across 168 processing elements, supported by 256KB of shared L1 scratchpad providing 205.8 GB/s bandwidth at 1.2ns access latency and 2MB L2 buffer delivering 78.3 GB/s at 3.5ns latency. This hierarchy enables compute utilization of 68.9-73.4% across diverse neural networks, compared to just 41.2-46.8% for designs with less optimized memory systems. MAERI's implementation in 28nm technology achieves energy efficiency of 7.2-11.3 TOPS/W, with detailed power breakdown revealing that data movement consumes 62.7% of total energy, 35.4% in on-chip networks, and 27.3% in memory accesses [7].

Memory optimization begins with characterizing workload requirements across multiple dimensions: capacity needs, bandwidth demands, access latency sensitivity, and energy efficiency constraints. Chi et al. demonstrate through their PRIME architecture that memory-bound neural network operations have

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

fundamentally different characteristics compared to compute-bound operations. Their measurements across seven benchmark networks reveal that fully-connected layers require 21.3-38.7× more memory bandwidth per operation compared to convolutional layers while exhibiting 83-97% lower arithmetic intensity [8]. PRIME addresses these challenges through a ReRAM-based processing-in-memory architecture that achieves 644.2 GOPS/W for FC layers—a 23.4× improvement over GPU implementations. Their memory hierarchy analysis shows that access patterns in FC layers exhibit row locality of only 0.17-0.25 and column locality of 0.32-0.41, making traditional cache hierarchies ineffective. By integrating computation directly within memory arrays, PRIME reduces data movement energy by 95.4% while increasing effective memory bandwidth by 8.5× compared to conventional architectures. Implementation in 32nm technology yields computational density of 78.4 GOPS/mm² with memory capacity density maintained at 2.17 Mb/mm² [8].

This multidimensional optimization problem has driven the development of specialized memory hierarchies incorporating multiple technologies. Kwon's analysis quantifies tradeoffs between SRAM, embedded DRAM, and emerging non-volatile memories, demonstrating that optimal hierarchies for large language models allocate 12-15% of chip area to memory subsystems with capacity ratios of approximately 1:8:64 across the three levels [7]. Recent progress in compute-in-memory architectures further complicates this landscape, with implementations achieving 9.7× higher energy efficiency compared to GPU platforms for memory-bound operations through fine parallelization across 2048 memory banks operated at 1.2 GHz [8]. The memory hierarchy design process must carefully balance these techniques, considering not only raw performance metrics but also system-level constraints related to power budgets, thermal boundaries, and cost targets.

Optimization Strategy	Architecture Type	Data Reuse Efficiency	Implementation Benefit	Energy Advantage
Hierarchical Caching	Traditional memory	High weight reuse	Standard interfaces	Moderate savings
Reconfigurable Dataflow	Flexible mapping	Variable reuse patterns	Adaptive throughput	Network optimization
Processing-in-Memory	Integrated compute	Eliminates movement	Novel architectures	Maximum efficiency
Non-volatile Storage	Weight stationarity	Persistent storage	Density advantages	Standby elimination
Compression Techniques	Parameter encoding	Effective capacity	Bandwidth multiplication	Access reduction

Table 4: Memory Architecture Optimization Strategies [7,8]

5. Advanced Memory Solutions for AI Workloads

The diversity of AI applications has driven the development of specialized memory solutions to address their unique data processing demands. Training tasks, due to their large datasets and complex gradient computations, typically prioritize memory capacity and bandwidth over latency. For these applications, HBM3E/HBM4 technology has emerged as the leading solution, offering capacity up to 64 GB per package with bandwidth exceeding 2 TB/s per stack in the latest iterations. Inference workloads, conversely, often operate under more stringent power and cost constraints, adopting LPDDR and GDDR technologies that offer more favorable performance-per-watt and performance-per-dollar metrics.

Beyond traditional memory technologies, several emerging solutions hold promise for meeting the unique demands of AI systems. Processing-in-Memory (PIM) architectures integrate computation directly within memory blocks, minimizing data movement and thereby reducing the energy costs that

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

dominate many AI workloads. The rise of non-volatile memory technologies—such as Magnetoresistive RAM (MRAM), Resistive RAM (ReRAM), and Phase-Change Memory (PCM)—offers promising paths toward achieving high memory density while lowering standby power consumption. These technologies also support innovative architectural strategies, such as weight-stationary designs, which help minimize costly data movement. Additionally, specialized memory compression techniques corresponding to statistical properties of neural network parameters have demonstrated significant effective capacity improvements with minimal computational overhead, further expanding the effective memory capacity available to AI systems.

Characteristic	HBM (Training)	GDDR6 (Cloud Inference)	LPDDR5 (Edge Inference)	Emerging Technologies
Primary Advantage	Extreme bandwidth	Cost-performance balance	Power efficiency	Novel capabilities
Key Limitation	Cost and integration complexity	Power consumption	Limited bandwidth	Maturity and reliability
Form Factor	Stacked die with interposer	Traditional BGA package	Mobile-optimized package	Various (embedded/discrete)
Ideal Application	Large model training	Batch inference	Real-time edge processing	Specialized workloads
Access Pattern Support	High-throughput streaming	Burst-oriented access	Low-duty cycle operation	In-situ computation
Scaling Direction	Vertical integration	Higher signaling rates	Power optimization	Density and integration
Thermal Consideration	Active cooling required	Managed thermal solution	Passive cooling sufficient	Low-temperature sensitivity
Cost Structure	Premium technology	Moderate cost	Consumer volume pricing	Early adoption premium
Future Evolution	Increased stack height	Improved signaling efficiency	Further power reduction	Computational capabilities
Implementation Challenge	System Integration	Signal integrity	Power delivery	Novel programming models
Emerging Variant	HBM3E/HBM4/ HBM-PIM	GDDR7	LPDDR5X/6	Hybrid memory architectures
Architectural Impact	Enables model scaling	Bandwidth/capacity balance	Enables edge deployment	Blurs the compute/memory boundary

Table 5: Advanced Memory Technologies for AI Workloads

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conclusion

Building high-performance AI systems fundamentally relies on high-speed interfaces and advanced memory architectures capable of handling the massive data movement demands of modern neural networks. This article explores key design principles across both areas, emphasizing the intricate interplay between system-level requirements, SerDes design choices, and memory hierarchy optimization. As AI workloads continue to grow in both complexity and scale, foundational technologies will play a critical role in pushing the practical performance limits of next-generation AI accelerators. Future research will likely focus on several key areas: advanced SerDes architectures surpassing 112Gbps per lane with improved energy efficiency; novel memory hierarchies that strike a balance between capacity, bandwidth, and power constraints; and emerging interconnect technologies aimed at narrowing the widening gap between compute capability and memory access speeds. Additionally, co-design approaches that simultaneously optimize algorithms, architecture, and circuit implementation will become increasingly essential, especially as the benefits of traditional technology scaling diminish. By mastering the core principles outlined in this article and advancing these research directions, designers can enable hardware architectures capable of meeting the demanding computational requirements of next-generation AI applications.

References

[1] Mark Horowitz, "1.1 Computing's energy problem (and what we can do about it)," IEEE, 2014. https://ieeexplore.ieee.org/document/6757323
[2] Masoud Daneshtalab; Mehdi Modarressi, "Hardware Architectures for Deep Learning," The Institution of Engineering and Technology, 2020. https://digital-library.theiet.org/doi/book/10.1049/pbcs055e
[3] Yakun Sophia Shao, et al., "Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture," ACM Digital Library, 2019, https://dl.acm.org/doi/10.1145/3352460.3358302
[4] Christoffer Åleskog et al., "Recent Developments in Low-Power AI Accelerators: A Survey," 2022. https://www.mdpi.com/1999-4893/15/11/419
[5] Jihwan Kim et al., "A 112-Gb/s PAM-4 56-Gb/s NRZ Reconfigurable Transmitter With Three-Tap FFE in 10-nm FinFET," IEEE. 2018. https://ieeexplore.ieee.org/document/8500752
[6] Luke Wang et al., "A 64Gb/s PAM-4 transceiver utilizing an adaptive threshold ADC in 16nm FinFET," IEEE, 2018, pp. 96-98. https://ieeexplore.ieee.org/document/8310208
[7] Hyoukjun Kwon, et al., "MAERI: Enabling Flexible Dataflow Mapping over DNN Accelerators via Reconfigurable Interconnects," ACM Digital Library, 2018. https://dl.acm.org/doi/10.1145/3173162.3173176
[8] Ping Chi et al., "PRIME: A Novel Processing-in-memory Architecture for Neural Network Computation in ReRAM-Based Main Memory," IEEE, 2016, https://ieeexplore.ieee.org/document/7551380
[9] Partstack, "HBM3E Advancements for AI and HPC Applications," 2024. https://partstack.com/blog/hbm3e-advancements-ai-hpc/
[10] HPCwire, "JEDEC Releases HBM4 Standard to Advance AI and HPC Memory," 2025. https://www.hpcwire.com/off-the-wire/jedec-releases-hbm4-standard-to-advance-ai-and-hpc-memory/