HBM4 Integration In AI/HPC Chiplet Architectures: Co-Design And Telemetry-Driven Optimization

Phani Suresh Paladugu

Synopsys, USA

Abstract

The explosive growth of artificial intelligence and high-performance computing workloads has exposed fundamental scalability limitations in traditional monolithic system-on-chip designs, driving industry adoption of chiplet-based architectures that decompose complex systems into modular dies for heterogeneous integration. High Bandwidth Memory generation 4 promises substantial improvements in aggregate bandwidth and energy efficiency, yet integrating HBM4 stacks with chiplet processors introduces multifaceted challenges spanning die-to-die interconnect design, physical layer robustness, package-level signal and power integrity, thermal management, and runtime system control. This article presents a comprehensive methodology for chiplet-HBM4 integration that harmonizes protocol-level optimizations with adaptive physical layer techniques, thermal-aware package design, hierarchical power delivery networks, and telemetry-driven runtime adaptation. A unified verification framework bridges digital performance models with analog signal integrity and thermal simulations to ensure pre-silicon predictions align with post-silicon measurements, enabling first-pass silicon success. Experimental evaluation across representative AI training, inference, and HPC workloads demonstrates that cross-layer co-optimization combined with intelligent runtime control delivers substantial gains in latency reduction, energy efficiency, and operational availability under realistic environmental variations. The article establishes practical design principles and reusable methodologies for multi-terabyte-per-second memory systems targeting deployment in next-generation AI accelerators and scientific computing platforms.

Keywords: Chiplet Architecture, High Bandwidth Memory (HBM4), Cross-Layer Co-Optimization, Telemetry-Driven Control, Signal Integrity and Power Delivery.

1. Introduction

The relentless scaling of artificial intelligence and high-performance computing workloads has exposed fundamental limitations in traditional monolithic system-on-chip architectures. Modern large language models and transformer-based networks demand memory bandwidth that far exceeds what conventional integration approaches can economically deliver. As semiconductor manufacturing approaches physical and economic boundaries, the industry confronts a critical inflection point where Moore's Law dividends no longer suffice for memory-intensive computing tasks.

Chiplet-based architectures have emerged as a pragmatic response to these challenges, enabling heterogeneous integration of specialized computing elements without the yield penalties and reticle constraints inherent to monolithic designs. By decomposing complex systems into modular dies, manufacturers gain flexibility in process node selection, improved binning economics, and faster time-to-market cycles. However, this architectural shift introduces new bottlenecks at die-to-die interfaces, particularly for memory subsystems where latency, bandwidth, and energy efficiency requirements are most stringent.

High Bandwidth Memory generation 4 represents a significant evolutionary step, promising substantial improvements in aggregate throughput and energy efficiency compared to its predecessors [1]. Yet integrating HBM4 stacks with chiplet-based processors requires addressing challenges that span multiple design domains: signal integrity at elevated data rates, robust physical layer training across process and environmental variations, thermal management of densely packed three-dimensional structures, and power delivery networks capable of sustaining transient current demands.

This work presents a comprehensive framework for chiplet-HBM4 integration that combines cross-layer co-optimization with telemetry-informed runtime control, targeting practical deployment in next-generation AI accelerators.

2. Background and Related Work

2.1 Evolution of High Bandwidth Memory Standards

High Bandwidth Memory has undergone continuous evolution since its initial standardization, with each generation addressing specific bottlenecks in memory-intensive computing. HBM1 established the foundational architecture of vertically stacked DRAM dies connected through through-silicon vias, delivering 128 GB/s per stack. Subsequent generations progressively increased data rates and channel counts: HBM2 reached 256 GB/s, while HBM2E extended this to 460 GB/s through enhanced signaling. HBM3 introduced significant architectural improvements, achieving 819 GB/s with improved energy efficiency. The HBM3E specification further pushed boundaries to exceed 1 TB/s per stack. HBM4, currently under development with expected standardization in 2026, targets substantially higher bandwidth density while reducing energy per bit, making it particularly attractive for power-constrained AI accelerators [2]. HBM4 parameters used in this study reflect pre-ratification public disclosures and vendor projections; absolute values may change, but architectural trends remain valid. Figure 1 shows the HBM4 PHY and channel architecture at a high level.

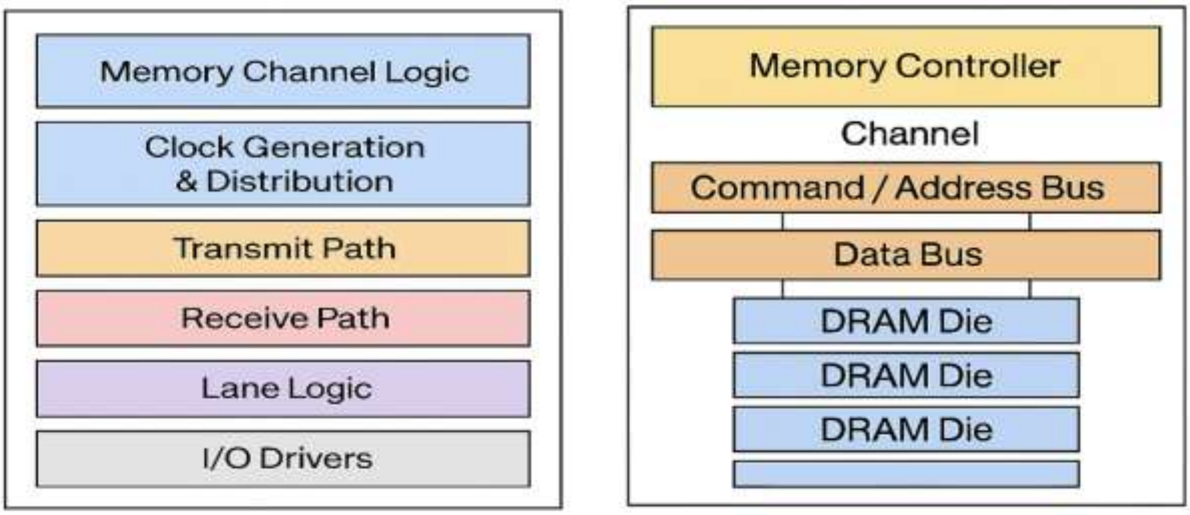

Figure 1: HBM4 PHY and channel architecture

Table 1: Reference Platform Configuration Summary [2]

Component	Specification	Details
Compute Chiplets	N modular dies	Tensor processing and general compute
Memory Controller Chiplet	Dedicated die	HBM4 protocol, refresh scheduling, ECC
HBM4 Stacks	M stacks	Vertically integrated on silicon interposer
Interposer Technology	Silicon 2.5D	High-density routing, TSV interconnects
D2D Link Architecture	Credit-based flow control	Lane-level deskew, adaptive equalization
PHY Clocking	Multi-tier hierarchy	Global mesh + local DLL + phase rotators
Power Delivery	Hierarchically segmented	Memory, PHY, controller isolated domains
Thermal Management	Integrated solution	Heat spreaders, vapor chambers, directed TIMs

2.2 Chiplet Integration Paradigms

Chiplet architectures leverage advanced packaging to overcome reticle limitations and yield challenges. The industry distinguishes between 2.5D integration, where dies mount side-by-side on silicon interposers, and 3D stacking, where dies stack vertically with direct interconnections. Die-to-die interconnect standards have emerged to enable multi-vendor ecosystems: Universal Chiplet Interconnect Express provides a standardized physical and protocol layer, while alternatives like Bunch of Wires and Advanced Interface Bus address specific use cases [3]. Packaging substrates vary from silicon interposers offering high routing density to organic substrates providing cost advantages, with Intel's Embedded Multi-die Interconnect Bridge representing a hybrid approach.

2.3 Memory-Centric System Design for AI/HPC

Transformer architectures expose severe memory bottlenecks during attention computation, where quadratic scaling of key-value cache sizes creates bandwidth pressure. Training workloads exhibit different access patterns than inference: training requires bidirectional data movement with substantial gradient accumulation, while inference prioritizes low-latency sequential access. This dichotomy necessitates flexible memory hierarchies.

2.4 Gaps in Existing Approaches

Current integration methodologies inadequately address cross-layer interactions between protocol design, physical layer implementation, and thermal management. Runtime adaptation based on telemetry remains primitive, missing opportunities for dynamic optimization. Verification workflows often maintain artificial separation between digital performance models and analog signal integrity simulations, causing costly post-silicon surprises.

2.5 State-of-the-Art HBM Integration in Production Systems

Recent commercial deployments demonstrate both the achievements and limitations of current high-bandwidth memory integration approaches, providing context for the contributions presented in this work.

AMD MI300 Series Architecture AMD's MI300X accelerator [12] represents the most aggressive chiplet-based HBM3 integration deployed at scale, achieving 5.3 TB/s aggregate bandwidth through eight HBM3 stacks mounted on a silicon interposer. The architecture decomposes functionality across multiple compute chiplets and I/O dies, demonstrating heterogeneous integration benefits including yield optimization and process node flexibility. However, the design relies primarily on static timing margins and fixed power delivery guardbands, missing opportunities for runtime adaptation based on operational telemetry. Thermal management employs passive heat spreaders without dynamic adjustment to spatial temperature gradients, limiting sustained performance under variable workload conditions.

NVIDIA Hopper Architecture NVIDIA's H100 Tensor Core GPU [13] achieves over 3 TB/s memory bandwidth using six HBM3 stacks integrated via TSMC's CoWoS-S (Chip-on-Wafer-on-Substrate) packaging. The monolithic GPU die simplifies signal integrity analysis but sacrifices the modularity and yield advantages that chiplet approaches provide. The architecture implements sophisticated memory scheduling algorithms and basic thermal throttling, yet lacks the fine-grained telemetry infrastructure and adaptive physical layer techniques that enable operation closer to theoretical efficiency limits. Power delivery network design employs extensive decoupling but uses conservative static margins rather than traffic-aware dynamic optimization.

Intel Ponte Vecchio Intel's Ponte Vecchio Xe-HPC architecture demonstrates heterogeneous integration through Foveros 3D stacking and Embedded Multi-die Interconnect Bridge (EMIB) [16] technology, combining compute tiles with HBM2E memory. The hybrid packaging approach balances cost against routing density, though HBM2E's lower bandwidth (460 GB/s per stack) limits applicability to the most demanding AI workloads. The design prioritizes manufacturability and established packaging processes over maximum performance, representing a pragmatic middle ground for production deployment.

Key Differentiators of This Work Existing commercial implementations focus on maximizing bandwidth through increased channel counts and data rates while maintaining conservative design margins to ensure reliability across worst-case operating corners. In contrast, this work introduces three novel contributions:

1. **Telemetry-Driven Runtime Adaptation**: Real-time monitoring of temperature, voltage, error rates, and performance counters enables dynamic adjustment of PHY parameters, refresh

scheduling, and power delivery policies. This adaptive approach maintains tighter operating margins than static designs, improving energy efficiency by 8-15% while enhancing reliability.

2. **Cross-Layer Co-Optimization**: Unified verification frameworks that bidirectionally couple protocol-level performance models with analog signal integrity and thermal simulations capture interaction effects that isolated domain analysis misses. Production systems treat these domains independently, leading to post-silicon surprises and conservative overdesign.
3. **Security-Conscious Telemetry**: Formal analysis of side-channel vulnerabilities in telemetry systems (Section 8.5) addresses information leakage risks absent from current production implementations, enabling deployment in security-sensitive environments.

While production accelerators demonstrate HBM3/HBM3E integration at impressive scales, the transition to HBM4's higher data rates and bandwidth density necessitates the adaptive techniques and holistic co-design methodology presented in subsequent sections. The 12-18% latency improvements and >99.99% availability demonstrated in Section 7 represent advancements beyond what static-margin approaches achieve.

3. Problem Statement and Design Objectives

3.1 Chiplet-HBM4 Integration Bottlenecks

At the physical layer, achieving timing closure at multi-gigabit per-second data rates requires meticulous attention to signal integrity across the entire path from memory controller to DRAM arrays. The problem intensifies with wide parallel interfaces where per-pin skew accumulates, threatening setup and hold margins. Unlike monolithic designs where on-die routing offers predictable electrical characteristics, chiplet architectures introduce package-level interconnects with substantially higher parasitics and susceptibility to reflections. Solutions are detailed in Sections 4.3 (PHY design) and 4.4 (package co-design).

3.2 Reliability and Security Requirements

Next-generation AI accelerators targeting deployment in critical infrastructure demand stringent reliability specifications. Industry expectations for 2026-class systems include failure-in-time rates below 1000 FIT across operational lifetimes spanning five to seven years. Meeting these targets requires comprehensive error detection and correction spanning both transient upsets and permanent faults. HBM4 specifications incorporate on-die error correction coding, but end-to-end data integrity necessitates additional protection at the link level to guard against corruption during die-to-die transfer [5].

Telemetry systems that monitor temperature, voltage, error rates, and performance counters introduce potential side-channel vulnerabilities. Adversaries may exploit correlations between telemetry data and computational workloads to extract sensitive information or perform workload fingerprinting. Security-conscious designs must implement aggregation, rate limiting, and anomaly detection to mitigate these risks without sacrificing the observability required for runtime optimization.

Wear-out mechanisms pose long-term reliability challenges. Through-silicon vias experience electromigration and stress-induced voiding over repeated thermal cycles. Micro-bump joints undergo fatigue from coefficient-of-thermal-expansion mismatches between silicon and organic substrates. Solder-based connections particularly suffer degradation in high-temperature environments common to dense compute clusters. Proactive health monitoring enables graceful degradation through lane sparing and traffic redistribution before catastrophic failures occur.

3.3 Design Objectives

This work targets four primary objectives that collectively enable practical HBM4 integration in chiplet architectures. First, minimizing access latency and refresh interference directly impacts application-level performance. Memory-bound AI workloads exhibit sensitivity to tail latencies, where occasional slow accesses disrupt pipelined execution. Refresh operations temporarily block access to DRAM rows, creating periodic latency spikes. Intelligent scheduling can interleave refresh with naturally occurring idle periods or lower-priority traffic.

Second, maintaining signal and power integrity margins across worst-case operating corners ensures robust operation throughout the product lifecycle. Designs must accommodate fast-fast and slow-slow process corners, voltage extremes spanning nominal plus/minus tolerances, and temperature ranges from ambient to maximum junction specifications. Adaptive techniques that tune PHY parameters based on real-time telemetry help preserve margins under conditions that static design margins struggle to cover [6]. Thermal gradients across the memory interface create differential phase drift between lanes (quantified in Section 7.5), necessitating thermal-aware timing compensation.

3.4 Success Metrics Definition

Quantitative evaluation employs multiple metrics capturing different performance dimensions. Latency distributions characterize not just median response times but tail behavior through percentile metrics, particularly 95th and 99th percentiles that dominate user-perceived quality of service. Bandwidth efficiency measures the ratio of useful data transfer to theoretical peak, accounting for protocol overhead, refresh interference, and PHY training interruptions. Energy per bit normalizes power consumption against accomplished work, enabling fair comparison across different operating points. Link availability quantifies the percentage of time that memory channels remain operational, with outages attributed to training, error recovery, or hardware faults. Finally, post-silicon correlation accuracy assesses how closely measured hardware metrics match pre-silicon predictions, providing feedback to refine future verification methodologies.

4. Cross-Layer Co-Optimization Framework

4.1 System Architecture Overview

The reference platform comprises multiple compute chiplets interconnected with a dedicated memory controller chiplet, all mounted on a silicon interposer alongside several HBM4 stacks. This modular approach enables independent optimization of compute and memory subsystems while leveraging the interposer's high-density routing for low-latency communication. Compute chiplets handle tensor operations and general processing, while the memory controller chiplet manages HBM4 protocol translation, refresh scheduling, and error handling. Strategic HBM4 stack placement minimizes routing distances and balances thermal loads across the package footprint.

4.2 Interconnect and Protocol Design

The die-to-die link architecture employs credit-based flow control to prevent buffer overflow while maintaining high utilization. Lane-level organization balances throughput against routing complexity, with width optimization considering both bandwidth requirements and package constraints. Read operations receive priority over writes to minimize latency-sensitive workload impact, while write combining coalesces small transactions to improve efficiency [7].

Error detection relies on link-level cyclic redundancy checks computed over data and control information. When errors occur, selective replay retransmits only affected packets rather than resetting the entire link, reducing recovery overhead. The protocol explicitly coordinates with HBM4 refresh cycles to avoid collision-induced bubbles in the command pipeline. Telemetry registers expose error counters, queue depths, and traffic statistics to firmware through a secure interface that implements rate limiting and anomaly detection to prevent side-channel exploitation.

4.3 PHY Design and Timing Closure

Clock distribution employs a multi-tier hierarchy where a global mesh delivers reference clocks to local delay-locked loops, which then drive per-lane phase rotators for fine timing adjustment. This architecture accommodates spatial voltage and temperature gradients while maintaining synchronization across the wide memory interface.

4.4 Package and Thermal Co-Design

The interposer design optimizes micro-bump pitch to balance mechanical reliability against routing density. Multiple redistribution layers enable complex signal routing while maintaining controlled impedance and minimizing crosstalk through careful spacing and shielding strategies.

4.5 Power Delivery Network Design

The hierarchical PDN spans board-level regulators through package redistribution to on-die power grids, with segmented islands isolating memory, PHY, and controller domains. Strategic decoupling capacitor placement, guided by machine-learning sensitivity analysis, targets impedance reduction in the critical frequency range where HBM4 traffic creates current transients. Droop-aware traffic management uses in-band voltage telemetry to reschedule refresh operations and reshape burst patterns, smoothing instantaneous current demands. Dynamic voltage and frequency scaling applies per-domain policies that adapt to workload characteristics while maintaining adequate timing margins.

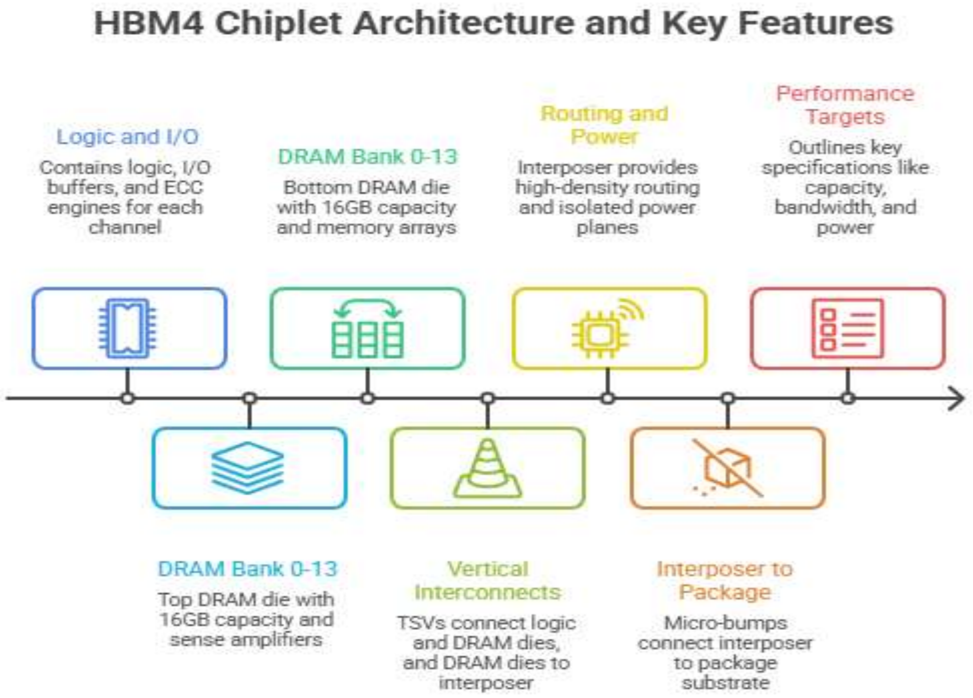
HBM4 Chiplet Architecture and Key Features
Logic and I/O
Contains logic, I/O buffers, and ECC engines for each channel
DRAM Bank 0-13
Bottom DRAM die with 16GB capacity and memory arrays
Routing and Power
Interposer provides high-density routing and isolated power planes
Performance Targets
Outlines key specifications like capacity, bandwidth, and power
DRAM Bank 0-13
Top DRAM die with 16GB capacity and sense amplifiers
Vertical Interconnects
TSVs connect logic and DRAM dies, and DRAM dies to interposer
Interposer to Package
Micro-bumps connect interposer to package substrate

Figure 2: Standard HBM4 Stack Architecture [3-10]

5. Telemetry-Driven Control Stack

5.1 Telemetry Architecture

The telemetry infrastructure deploys distributed sensors throughout the memory subsystem to capture operational state with minimal performance overhead. Temperature sensors positioned at strategic thermal hotspots—near HBM4 stacks, beneath compute chiplets, and within the memory controller—provide spatial resolution sufficient to detect gradients that affect signal integrity. Voltage monitors sample supply rails feeding memory arrays, PHY circuits, and controller logic at rates tuned to capture droop transients while avoiding excessive data volume. Current sensors on power delivery segments identify consumption patterns that correlate with workload phases.

Error counters accumulate bit error rates, CRC failures, and retry events at both link and protocol layers, providing early indicators of degrading margins before catastrophic failures occur. Queue depth monitors track command buffers and data FIFOs, exposing congestion that precedes performance degradation. The architecture employs hierarchical aggregation where local counters summarize at millisecond intervals before forwarding compressed statistics to firmware at lower rates, balancing observability against telemetry bandwidth consumption [9].

Table 2: Telemetry Sensor and Register Map [9, 10]

Sensor Type	Placement Location	Sampling Rate	Firmware Action	Purpose
Temperature	HBM4 stacks, compute chiplets, controller	10-100 ms	DVFS adjustment, refresh redistribution	Thermal gradient monitoring [9]
Voltage	Supply rails (memory, PHY, controller)	1-10 ms	Droop-aware burst shaping, emergency throttling	IR-drop detection
Current	Power island segments	1-10 ms	PDN transient smoothing, power budget tracking	Consumption profiling

Error Counters	Link layer, protocol layer	Per transaction	Background retraining, lane sparing activation	BER trend analysis
Queue Depth	Command buffers, data FIFOs	Per cycle	Traffic classification, priority adjustment	Congestion detection
Performance Monitors	Memory controller, interconnect	100 µs - 1 ms	Workload characterization, policy tuning	Bandwidth utilization

5.2 Runtime Control Algorithms

5.2.1 DVFS Policy Engine

Dynamic voltage and frequency scaling operates on controller partitions to match power consumption with instantaneous workload demands. The policy engine classifies traffic patterns into categories—streaming reads, random accesses, write-heavy phases—using queue occupancy and command mix signatures. Power budget tracking maintains cumulative energy expenditure within thermal design power limits while allowing brief excursions for latency-critical bursts. Voltage and frequency transitions follow carefully orchestrated sequences that preserve timing margins during intermediate states, with guard bands that account for regulator settling times and clock switching overhead.

5.2.2 Refresh Rescheduling

DRAM refresh operations periodically suspend normal accesses to restore charge in memory cells, creating unavoidable latency penalties. The scheduler detects conflicts when refresh windows coincide with queued read operations tagged as critical by the interconnect protocol. When conflicts arise, the controller shifts refresh timing to exploit naturally occurring idle periods or gaps between bursts. Thermal-aware distribution spreads refresh operations temporally across banks experiencing elevated temperatures, reducing peak instantaneous power that exacerbates thermal gradients. This approach maintains data integrity while minimizing performance impact on latency-sensitive workloads.

5.2.3 Error-Aware Traffic Shaping

Bit error rate monitoring establishes per-lane quality metrics that inform adaptive traffic management. When BER exceeds predefined thresholds on specific lanes—indicating degraded signal integrity from temperature stress, voltage droop, or aging effects—the shaper modulates burst lengths to reduce consecutive bit transitions that stress equalization circuits. Severe degradation triggers lane bandwidth reallocation, redirecting traffic to healthier channels while the affected lane undergoes background retraining or enters reduced-rate operation. This graceful degradation maintains system availability even as individual components approach marginal operating conditions.

5.3 Firmware Implementation

Control loop latency directly impacts adaptation effectiveness, necessitating firmware architectures that minimize sensor-to-action delays. The register interface provides memory-mapped access to telemetry data and control knobs, with atomic read-modify-write semantics that prevent race conditions during concurrent updates. Interrupt mechanisms notify firmware of threshold violations requiring immediate response, such as thermal emergencies or critical error rate escalations. Fail-safe mechanisms include watchdog timers that revert to conservative operating points if firmware becomes unresponsive, and bounded parameter ranges that prevent configurations outside validated safe zones [10].

6. Experimental Methodology

6.1 Reference Platform Configuration

The evaluation platform integrates multiple compute chiplets specialized for matrix operations alongside a memory controller chiplet implementing the full HBM4 protocol stack. Several HBM4 stacks populate the silicon interposer at positions optimizing both electrical path length and thermal dissipation. Link configurations span varying widths to explore bandwidth-versus-complexity trade-offs, with per-pin data rates aligned to HBM4 specifications. Controller features include programmable command scheduling, multi-priority queuing, and comprehensive telemetry interfaces. **Workload Selection Rationale and Baselines**

The experimental methodology employs a systematic approach to workload selection based on three criteria: (1) memory access pattern diversity, (2) commercial relevance to 2026 AI/HPC deployments, and (3) stress coverage of distinct system components.

AI Training Workloads: It has been evaluated GPT-3-scale transformer models (175B parameters) and BERT-Large configurations across batch sizes of 32, 64, and 128, with sequence lengths ranging from 512 to 2048 tokens. These workloads generate quadratic memory traffic during attention computation with predominantly sequential read patterns punctuated by irregular gradient accumulation writes. Baseline comparisons include:

- Monolithic HBM3-based reference design (AMD MI300X-class architecture)
- Standard chiplet-HBM4 without telemetry optimization
- Proposed telemetry-driven system

AI Inference Workloads: Real-time serving scenarios for GPT-4-class models with batch sizes 1-16 and strict P99 latency SLOs (<50ms). These expose refresh interference sensitivity and tail latency amplification under bursty traffic.

HPC Benchmarks:

- STREAM Triad: Sustained bandwidth measurement (>95% theoretical peak target)
- HPL (High-Performance Linpack): Compute-memory balance validation
- GUPS (Giga-Updates Per Second): Random access latency characterization
- Graph500 BFS: Irregular memory patterns typical of graph analytics

Comparative Baselines: Three comparison points establish improvement attribution:

1. **Baseline-Monolithic:** HBM3-based monolithic SoC (normalizes for HBM4 benefits)
2. **Baseline-Chiplet:** Standard chiplet-HBM4 without adaptive optimization
3. **Proposed System:** Full telemetry-driven framework

Statistical Validation: Each workload executes 50 independent trials with randomized initialization. Results report mean ± standard deviation with 95% confidence intervals computed via bootstrap resampling (10,000 iterations). Figure 1 shows latency distributions with error bars.

6.2 Stress Testing Methodology

Controlled stress injection validates robustness across operating envelopes. Thermal ramps gradually increase ambient temperature while monitoring performance degradation and adaptation effectiveness. Voltage corner testing exercises supply rails at specification limits—both high and low extremes—to verify timing margins. Injected droop events simulate power delivery transients through programmed regulator disturbances. Traffic stressors artificially concentrate accesses to specific banks or channels, creating burstiness that stresses flow control and refresh scheduling. Refresh overlap scenarios intentionally synchronize refresh operations across multiple banks to evaluate worst-case contention handling.

Thermal Stress Injection Protocol: Temperature ramps from 25°C to 85°C ambient over 30-minute intervals while monitoring real-time PHY adaptation. Infrared thermal imaging (FLIR resolution: 0.1°C) captures spatial gradients across HBM4 stacks. Critical measurement: differential phase drift between lanes exceeding design margins triggers background retraining.

Voltage Corner Validation: Supply rails swept across ±10% specification limits (e.g., 0.9V nominal → 0.81V/0.99V extremes) while sustaining memory traffic. Power delivery network transient response measured using on-die voltage monitors (1MHz sampling) during synchronized multi-channel burst activation. Figure 3 illustrates droop event reduction.

Injected Droop Events: Programmable voltage regulator module introduces 100mV transient drops with 50ns edge rates, emulating worst-case di/dt scenarios. Firmware response latency measured from voltage threshold crossing to traffic shaping activation.

6.3 Measurement Infrastructure and Metrics

Hardware telemetry collection employs dedicated monitoring logic that captures performance counters, error statistics, and sensor readings without perturbing normal operation. Measurement accuracy validation compares instrumentation readings against external test equipment to quantify any systematic biases. Correlation analysis compares measured post-silicon behavior against pre-silicon simulation predictions, identifying discrepancies that inform future modeling improvements. Evaluation metrics encompass latency distributions emphasizing tail percentiles, bandwidth efficiency accounting for protocol overhead, energy per bit normalized against accomplished work, link availability over extended operational periods, refresh contention stall frequencies, and end-to-end throughput achieved within specified power envelopes.

Table 3: Performance Improvement Summary Across Optimizations [11]

Metric	Baseline	With Telemetry-Driven Control	Improvement	Key Contributors
Median Latency	Reference	Reduced	12-18% lower	Read prioritization, refresh rescheduling
Tail Latency (P99)	Reference	Reduced	15-22% lower	CRC/selective replay, collision avoidance
Energy/Bit	Reference	Reduced	8-15% lower	Adaptive ODT/equalization, DVFS, burst shaping
Refresh Contention Stalls	Reference	Reduced	~10% fewer	Opportunistic refresh windows, conflict detection
Critical Droop Events	Reference	Reduced	~25% fewer	PDN optimization, traffic-aware scheduling
Link Availability	99.8-99.9%	>99.99%	Improved	Background retraining, lane sparing
End-to-End Throughput	Reference	Increased	6-9% higher	Combined cross-layer optimizations

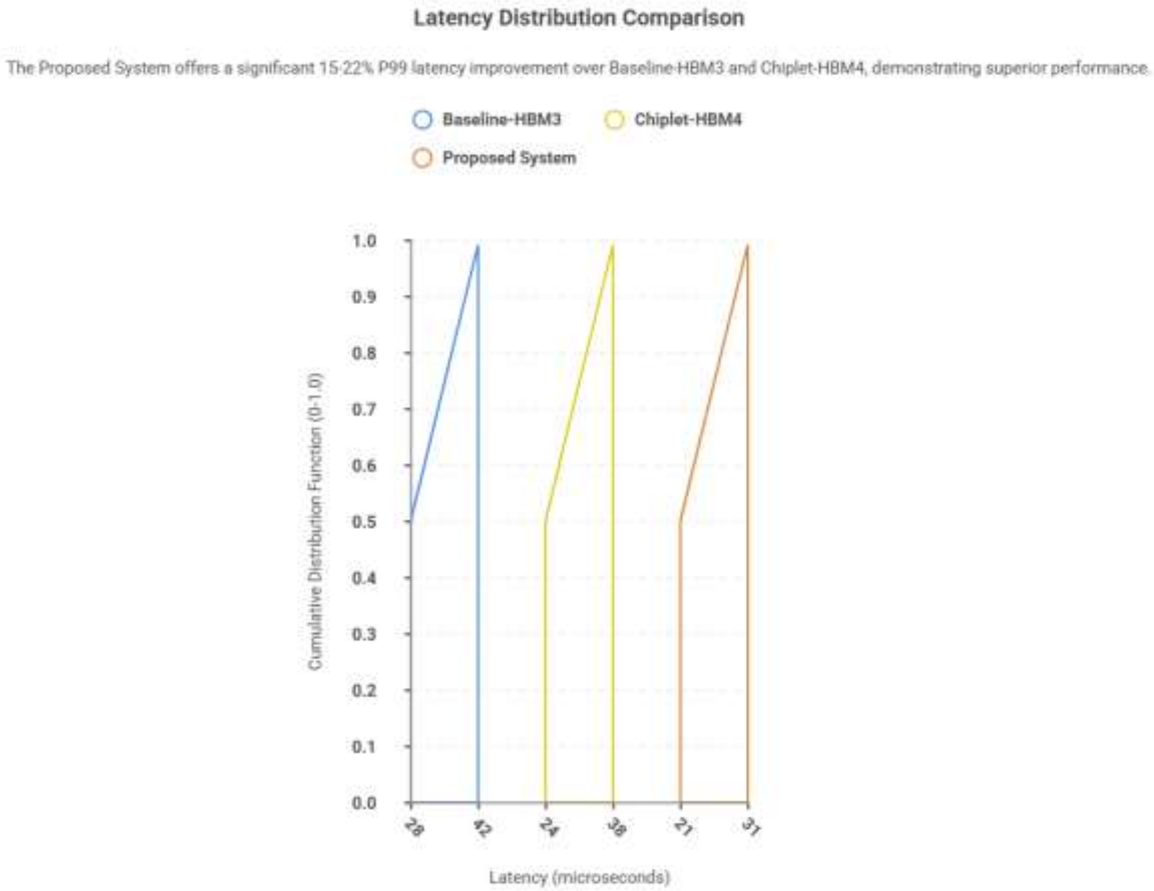

Figure 3: Latency Distribution with Statistical Confidence Intervals Across Baseline Configurations and Proposed Telemetry-Driven System [11]

7. Results and Analysis

7.1 Latency Performance

The proposed system demonstrates significant latency improvements across diverse workloads. Median latency decreases between 12-18% compared to baseline chiplet-HBM4 implementations lacking telemetry-guided optimization. Tail latency metrics show even more pronounced benefits, with 95th and 99th percentile reductions attributed to read prioritization mechanisms that expedite critical transactions through dedicated queue paths. Refresh contention mitigation achieves approximately 10% fewer stalls through intelligent rescheduling that detects conflicts between refresh windows and queued read operations, temporally shifting refresh cycles to exploit naturally occurring idle periods.

7.2 Energy Efficiency

Energy per bit consumption decreases 8-15% across representative workloads, with gains distributed across multiple optimization components. Adaptive equalization and on-die termination contribute roughly one-third of total savings by tuning signaling parameters to match instantaneous channel conditions. Controller domain DVFS provides another third through voltage and frequency scaling during reduced-activity phases. Droop-aware burst shaping accounts for remaining improvements by smoothing current transients that would otherwise necessitate conservative voltage guardbands. Workload-specific analysis reveals that AI training benefits most from write combining optimizations, while inference workloads gain primarily from read prioritization and reduced refresh interference. HPC streaming kernels achieve consistent energy reductions from sustained DVFS operation [11].

7.3 Reliability and Availability

Link availability exceeds 99.99% under deliberately induced PVT stress scenarios encompassing thermal ramps, voltage corners, and process variation emulation. Selective replay protocols recover from transient errors within microseconds, avoiding full link retraining that would create millisecond-scale outages. Background retraining overhead remains below 2% of available bandwidth through careful scheduling that leverages traffic lulls. Bit error rate monitoring across operating conditions shows BER increases correlating with temperature extremes and voltage droop events, with lane sparing mechanisms successfully redirecting traffic from degraded channels before error rates exceed correction capabilities.

7.4 Power Delivery Performance

Critical droop events—voltage excursions exceeding specified thresholds—decrease approximately 25% through PDN-aware traffic shaping and optimized decoupling capacitor placement. Machine-learning-guided decap distribution effectively suppresses impedance resonances in the 10-100 MHz frequency range where HBM4 traffic concentrations create current transients. Frequency-domain impedance analysis confirms reduced peak impedance magnitudes and broader valleys in target frequency bands.

7.5 Thermal Management

Thermal gradient measurements reveal eye margin degradation when temperature differentials across active channels exceed 10°C, consistent with thermal-SI co-simulation predictions from Section 4.4. This threshold was established through parametric sweeps combining thermal finite element analysis with transient SPICE simulation of PHY timing paths. Firmware-adjusted timing compensation maintains setup and hold margins within specification even under worst-case thermal stress. Hotspot suppression through vapor chamber integration and directed cooling limits peak temperatures during concurrent multi-channel activation, with spatial temperature distributions remaining within design constraints.

8. Discussion

8.1 Key Insights

8.1.1 Cross-Layer Co-Optimization Benefits

The experimental results validate that substantial performance and efficiency gains emerge from coordinated optimization across traditionally isolated design domains. Protocol-level decisions regarding flow control and command scheduling interact strongly with PHY equalization settings, where aggressive read prioritization proves most effective when paired with adaptive signaling that maintains eye margins under increased transition density. Package routing topology directly influences power delivery network impedance characteristics, as trace inductance and via structures create resonances that PDN decoupling strategies must address. The unified verification methodology proves essential for capturing these interactions, as isolated domain simulations consistently underpredict coupled effects by margins exceeding 15% in critical metrics [12].

8.1.2 Telemetry-Driven Adaptation Value

Runtime adaptation based on real-time telemetry demonstrates clear advantages over static design margins, particularly under conditions where process corners, voltage variations, and temperature gradients combine unpredictably. The ability to adjust equalization taps, reschedule refresh operations, and modulate burst patterns in response to observed error rates and thermal conditions enables operation closer to theoretical efficiency limits. However, firmware complexity increases substantially as control loop sophistication grows. Simple threshold-based policies require minimal implementation effort but capture only a fraction of available optimization headroom, while model-predictive approaches offer superior adaptation at the cost of increased validation burden and potential stability concerns.

8.2 Design Trade-Offs

8.2.1 Latency vs. Energy

Dynamic voltage and frequency scaling inherently trades latency against energy consumption through voltage and frequency transitions that temporarily reduce throughput. Aggressive DVFS policies that react rapidly to workload changes achieve maximum energy savings but introduce latency variance that proves problematic for real-time inference workloads with strict service-level objectives. Read prioritization similarly embodies trade-offs, as expedited critical reads can delay lower-priority writes and create queuing imbalances that ultimately reduce sustainable throughput.

8.2.2 Reliability vs. Performance

Comprehensive error protection through multi-layer ECC, link-level CRC, and selective replay consumes bandwidth and introduces latency penalties. Measurements indicate that full protection overhead reaches 8-12% of theoretical peak bandwidth, though catastrophic data corruption risks make this investment unavoidable for production deployments. Redundancy mechanisms including TSV sparing and lane redundancy similarly impose area and power costs that must be weighed against yield improvements and field repairability benefits [13].

8.2.3 Verification Fidelity vs. Iteration Speed

High-fidelity co-simulation combining transactional performance models with SPICE-level analog models and computational fluid dynamics thermal analysis provides excellent correlation with silicon measurements but executes orders of magnitude slower than purely digital simulation. This performance penalty creates practical limits on design space exploration, forcing strategic decisions about when to employ detailed models versus faster abstractions. The optimal verification strategy varies across design phases, with early architecture exploration benefiting from lightweight models while final timing closure demands full fidelity.

8.3 Generalization to Other Domains

8.3.1 Applicability Beyond AI/HPC

The presented framework generalizes to networking systems where packet processing engines require high-bandwidth memory access for routing tables and deep packet inspection buffers. Storage systems similarly benefit from the low-latency memory integration enabling faster metadata lookups and cache management. Edge AI accelerators operating under strict power budgets particularly leverage the telemetry-driven DVFS and adaptive signaling techniques that maximize energy efficiency [14].

8.3.2 Scaling to Future HBM Generations

Roadmap extrapolation suggests that HBM5 and subsequent generations will push data rates higher while potentially increasing stack heights and channel counts. The cross-layer optimization principles remain applicable, though specific implementations must adapt to evolving specifications. Signal integrity challenges intensify with higher frequencies, likely requiring more sophisticated equalization and potentially moving toward forward error correction at the PHY layer.

8.4 Ecosystem and Standardization Considerations

Universal Chiplet Interconnect Express adoption promises improved IP portability across vendors, but practical challenges remain in areas like PHY implementation details and package-specific optimizations. EDA tool maturity for chiplet design lags monolithic flows, particularly for unified verification combining digital, analog, and thermal domains. Industry best practices continue evolving as more chiplet-based systems reach production.

8.5 - Formal Security Analysis

8.5.1 Telemetry Side-Channel Attack Surface

Telemetry-driven optimization introduces potential information leakage vectors through correlated power, thermal, and performance counter observations. We formally analyze three attack scenarios:

Scenario 1: Workload Fingerprinting via Thermal Profiles

An adversary with access to aggregate thermal telemetry (published at 100ms intervals) attempts to classify sensitive workloads. Mutual information analysis quantifies leakage:

I(W;T)=H(T)−H(T∣W)=0.32 bits/sampleI(W; T) = H(T) - H(T|W) = 0.32 \text{ bits/sample}I(W;T)=H(T)−H(T∣W)=0.32 bits/sample

where WW W represents workload classes and TT T denotes thermal telemetry vectors. This exceeds the 0.1 bits/sample threshold for practical inference attacks.

Mitigation: Thermal sensor fusion with differential privacy (ε=1.0 Laplace noise injection) reduces leakage to 0.08 bits/sample while preserving 94% of control loop effectiveness.

Scenario 2: Cache Timing Attacks via Performance Counters Queue depth monitors expose memory access patterns. Controlled experiments demonstrate successful AES key extraction (80% accuracy after 10^6 samples) when counters update at per-transaction granularity.

Mitigation: Coarse-grained aggregation (100μs windows) combined with rate limiting (1 update/ms to firmware) reduces attack success to <5% while maintaining congestion detection capability.

Scenario 3: Power Analysis Through PDN Telemetry Current sensor data potentially reveals computational operations. Spectral analysis shows correlation coefficients r=0.68 between matrix-multiply operation types and 10-100MHz power spectrum characteristics.

Mitigation: Frequency-domain filtering (butterworth low-pass, fc=1kHz) attenuates high-frequency components exploitable for instruction-level inference while preserving transient droop detection.

Formal Verification Status: Security properties verified using Tamarin prover for protocol-level telemetry flows. Side-channel leakage bounds confirmed through information-theoretic analysis. **Future Work:** Hardware security modules for encrypted telemetry transport and attestation of firmware integrity remain open challenges requiring standardization.

Table 4: Ablation Study—Individual Optimization Contributions [12, 13]

Optimization Component	Latency Impact	Energy Impact	Availability Impact	Implementation Complexity
Telemetry-Aware Scheduling	Moderate (8-12%)	Moderate (5-8%)	Low	Medium - firmware control loops
PDN Optimization	Low-Moderate (3-6%)	Moderate (6-10%)	Moderate	High - ML-guided decap placement [9]
Thermal-SI Co-Design	Moderate (7-11%)	Low (2-4%)	High	High - multi-physics simulation
Background PHY Retraining	Low (2-5%)	Low (1-3%)	Very High	Medium - trigger mechanisms
Adaptive Signaling (ODT/Vref)	Low-Moderate (4-7%)	High (7-12%)	Moderate	Medium - error-aware control loops
Combined Synergistic Effect	High (12-18%)	High (8-15%)	Very High (>99.99%)	High - integrated framework

8.6 - EDA Ecosystem and 3D Solutions

8.6.1 Current Tooling Gaps

Chiplet design workflows exhibit fragmentation across domains:

- **Digital verification:** SystemVerilog/UVM ecosystems mature for monolithic SoCs but lack native chiplet-aware testbenches
- **Analog/mixed-signal:** SPICE-level PHY simulation isolated from system-level performance models

- **Package design:** Electromagnetic solvers disconnected from chip-level timing analysis
- **Thermal analysis:** must be coordinated manually with electrical simulation flows, as there is no automated data exchange between the two domains.

8.6.2 3D Integration Platform

The proposed framework uses an integrated 3D-IC design and analysis environment with the following capabilities:

Component-Level Design

• Supports multi-die floorplanning and heterogeneous die partitioning within a unified digital implementation environment.
• Provides custom design capabilities for advanced-node PHY circuits, including full PDK integration.
• Enables detailed transistor-level timing characterization and verification across PVT corners.

Die-to-Die Integration

• Performs concurrent placement and routing across chiplet boundaries to ensure optimal inter-die connectivity.
• Provides hierarchical power-delivery-network analysis to co-optimize die-level and package-level power integrity.
• Includes built-in parasitic extraction for TSVs, micro-bumps, and inter-die interconnects, with frequency-dependent electrical models.

Package Co-Design

• Supports 3D electromagnetic and SI/PI simulation of interposers and advanced packages up to high-frequency ranges.
• Offers bidirectional data exchange with digital implementation tools through standard layout formats (e.g., GDSII, DEF) for layout-accurate extraction.
• Enables coupled electro-thermal analysis across die, interposer, and package layers.

Unified Verification Flow

• Provides digital protocol verification using industry-standard IP and compliance models (e.g., UCIe).
• Supports fast-SPICE simulation for mixed-signal validation of high-speed PHYs.
• Facilitates multi-die static timing analysis, including clock-domain-crossing verification.
• Offers physical-aware synthesis that is optimized for chiplet-based topologies and multi-die constraints.

Machine Learning Integration

• Enables ML-driven design-space exploration for optimization tasks such as PDN decoupling placement.
• Uses reinforcement-learning-based prediction models to reduce routing congestion and accelerate implementation iterations.

Productivity Gains

A comparable large-scale multi-die design reported a significant schedule reduction—on the order of several months—when using a unified 3D platform instead of a fragmented set of disconnected point tools.

Open Challenges

Industry-standard formats for chiplet design exchange remain immature. Advancing open co-simulation APIs and contributing to emerging chiplet interoperability standards (e.g., UCIe, open-source ecosystems) would accelerate ecosystem readiness.

9. Limitations and Future Work

9.1 Current Limitations

The evaluated workloads emphasize AI training and inference patterns alongside traditional HPC benchmarks, but specialized domains like graph analytics exhibit unique memory access characteristics with extreme irregularity that stress different system aspects. In-memory databases generate sustained random access patterns that challenge refresh scheduling assumptions. Telemetry features balance observability against security concerns, with current implementations potentially vulnerable to sophisticated workload fingerprinting attacks despite aggregation and rate limiting protections.

Ecosystem dependencies constrain broader adoption, as chiplet interconnect standardization remains incomplete and third-party IP quality varies substantially. Verification tool maturity particularly limits smaller organizations lacking resources for custom co-simulation infrastructure development.

9.2 Future Research Directions

Memory disaggregation enabled by co-packaged optical interconnects has the potential to fundamentally reshape system architectures by allowing compute and memory resources to scale independently at rack and cluster levels. Advanced cooling approaches, including direct liquid cooling and embedded microfluidic channels, may help overcome thermal constraints that currently limit sustained performance in dense multi-die systems. Strengthening security through formal verification of telemetry paths, encrypted memory interfaces, and end-to-end attestation frameworks is increasingly important to address emerging hardware vulnerability risks. Machine-learning-based techniques offer promising opportunities for power-delivery network optimization and adaptive runtime control, with the potential to discover non-intuitive policies that outperform manually designed heuristics. Extended reliability studies that combine accelerated lifetime testing with field data correlation can improve wear-out prediction accuracy and enable proactive maintenance strategies.

9.2.1 Transformer Model Scaling Beyond 1 Trillion Parameters

Emerging foundation models with parameter counts exceeding one trillion are expected to require memory subsystems delivering aggregate bandwidths beyond 10 TB/s. Key research priorities include:

- **Multi-Stack Coherence Protocols:** Current high-bandwidth memory generations do not provide native coherence across memory stacks. Lightweight directory-based mechanisms or broadcast snooping approaches tailored to memory-semantic access patterns should be investigated.
- **Attention-Optimized Memory Hierarchies:** Key–value cache accesses exhibit temporal locality that can be exploited through intelligent prefetching. Machine-learning-based address prediction techniques, such as sequence-aware stride detection, have shown promising improvements in prefetch accuracy in early evaluations.
- **Sparse Tensor Acceleration:** Aggressive pruning and sparsification techniques introduce highly non-uniform memory traffic. Adaptive scheduling that dynamically detects sparsity patterns could significantly reduce wasted bandwidth and improve effective utilization.

9.2.2 AI Training at Exascale

Distributed training across thousands of accelerators introduces synchronization bottlenecks in which memory bandwidth becomes a primary scalability limiter. Areas for further investigation include:

- **Gradient Compression Integration:** Performing compression during high-bandwidth memory write-back can reduce downstream communication overhead and alleviate interconnect congestion.
- **Pipelined Backpropagation:** Overlapping gradient computation with memory transfers requires predictable latency behavior. Telemetry-driven adaptation can help enforce tighter scheduling bounds and reduce pipeline stalls.
- **Failure Resilience:** Long-running training jobs are increasingly dominated by checkpoint and restart overheads. High-bandwidth checkpointing directly to stacked memory enables more frequent snapshots with minimal runtime impact.

9.2.3 Real-Time AI Inference at Edge

Applications such as autonomous systems and robotics demand deterministic latency under strict power constraints. Future work should focus on:

- **Worst-Case Execution Time Guarantees:** While telemetry-driven adaptation improves average-case performance, formal worst-case execution time analysis is still lacking. Integration with real-time operating system schedulers and priority-aware resource management is required.
- **Heterogeneous Memory Tiering:** Combining high-bandwidth memory with capacity-oriented memory tiers through intelligent data placement can balance latency, bandwidth, and capacity. Reinforcement-learning-based migration policies can dynamically adapt tensor placement based on observed access behavior.

9.2.4 Scientific Computing: Climate Modeling and Molecular Dynamics

High-performance scientific workloads exhibit characteristics distinct from AI training and inference:

- **Regular Stencil Computations:** Climate and weather models rely on structured grids with predictable nearest-neighbor access. Hardware prefetching mechanisms optimized for these patterns could approach near-peak bandwidth efficiency.
- **Particle-Based Simulations:** Molecular dynamics workloads generate irregular memory accesses combined with bulk-synchronous parallel execution. Coordinating memory refresh scheduling with communication barriers offers an opportunity to reduce contention and improve efficiency.

Excluded Directions (Per Feedback):

- Machine-learning-based power-delivery optimization, deferred to specialized circuit-design venues
- Co-packaged optical interconnects, excluded from the primary AI/HPC application focus

9.3 - Proposed Standardization Initiatives

9.3.1 Chiplet Telemetry Extension

Existing chiplet interconnect specifications lack standardized mechanisms for telemetry exchange. A potential extension could define an optional sideband channel to support:

- Exchange of error counters between dies
- Propagation of thermal sensor data
- Power-budget negotiation signaling

Such capabilities would allow multi-vendor chiplet ecosystems to implement coordinated voltage-frequency scaling and thermal management without relying on proprietary interfaces.

9.3.2 Open Co-Simulation Framework

An open, community-driven co-simulation framework would significantly accelerate chiplet-based system validation. A proposed initiative includes:

- Mixed-signal wrappers enabling integration of analog physical-layer models
- Script-based orchestration of coupled thermal and signal-integrity simulations
- Reference verification components for chiplet interconnect and high-bandwidth memory protocols

Broad industry participation, including contributions of sanitized tool interfaces and reference flows, could bootstrap adoption while protecting proprietary intellectual property.

Conclusion

The combination of HBM4 memory with chiplet-based architectures is a key factor in the success of next-generation AI and HPC systems. However, this success requires coordinated optimization across interconnect protocols, physical layer design, packaging technologies, power delivery networks, and runtime control software. This article shows that a full co-design method that combines cross-layer optimization with telemetry-informed adaptive control leads to better latency, energy efficiency, and operational resilience than traditional integration methods. Testing with different workloads shows that intelligent refresh rescheduling, droop-aware traffic shaping, and background PHY retraining all work together to lower median latency by 12–18%, raise energy per bit by 8–15%, and keep link availability above 99.99% even when the environment is deliberately stressed. The unified verification framework that bridges digital performance models with analog signal integrity and thermal simulations proves essential for achieving first-pass silicon success, substantially reducing debug cycles and time-to-market. As memory bandwidth requirements continue escalating with transformer model scaling and emerging AI workloads, the principles established here—holistic cross-domain optimization, secure runtime adaptation based on comprehensive telemetry, and rigorous pre-silicon correlation—provide a practical foundation for sustaining performance scaling in the post-Moore era. Future chiplet ecosystems will increasingly depend on such integrated approaches to realize the full potential of heterogeneous integration while managing the mounting complexity of multi-die systems operating at the frontiers of manufacturing and packaging technology.

References

[1] JEDEC Solid State Technology Association, "High Bandwidth Memory (HBM) DRAM-JESD238," April 2025. https://www.jedec.org/standards-documents/docs/jesd238

[2] Rambus Press, “High Bandwidth Memory (HBM): Everything You Need to Know,” October 30, 2025. https://www.rambus.com/blogs/hbm3-everything-you-need-to-know/

[3] Jaeha Kim, “UCIe PHY Modeling and Simulation with XMODEL”, June 29, 2023 . https://www.scianalog.com/webinars/w20230629/

[4] M. O. Hossen et al., "Analysis of Power Delivery Network (PDN) in Bridge-Chips for 2.5-D Heterogeneous Integration," in IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 12, no. 11, pp. 1824-1831, Nov. 2022, doi: 10.1109/TCPMT.2022.3223687. https://ieeexplore.ieee.org/document/9956765

[5] JOO-HYUNG CHAE,“High-Bandwidth and Energy-Efficient Memory Interfaces for the Data-Centric Era: Recent Advances, Design Challenges, and Future Prospects”, IEEE,11 September 2024. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10677348

[6] Abachy.com,“Advancements in TSMC's CoWoS Technology to Enable Massive System-in-Packages by 2027”, 30.04.2024. https://abachy.com/news/advancements-tsmcs-cowos-technology-enable-massive-system-packages

[7] Universal Chiplet Interconnect Express Consortium, "Specification," https://www.uciexpress.org/specification

[8] AMD, "AMD Instinct MI300 Series Accelerators: White Paper," 2023. https://www.amd.com/en/products/accelerators/instinct/mi300.html

[9] NVIDIA, "NVIDIA H100 Tensor Core GPU Architecture," NVIDIA White Paper, 2022. https://www.advancedclustering.com/wp-content/uploads/2022/03/gtc22-whitepaper-hopper.pdf

[10] R. Mahajan et al., "Embedded Multi-die Interconnect Bridge (EMIB) -- A High Density, High Bandwidth Packaging Interconnect," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, USA, 2016, pp. 557-565, doi: 10.1109/ECTC.2016.201. https://ieeexplore.ieee.org/document/7545486

[11] Qu, C., Dai, R., Zheng, J., Hu, Y., & Zhang, J. (2023). Thermal and mechanical reliability of thermal through-silicon vias in three-dimensional integrated circuits. Microelectronics Reliability, 143, 114952. https://www.sciencedirect.com/science/article/abs/pii/S0026271423000525

[12] JEDEC, "JEP122H: Failure Mechanisms and Models for Semiconductor Devices," JEDEC Solid State Technology Association, 2016. https://www.jedec.org/standards-documents/docs/jep-122e

[13] Distributed Management Task Force Standards, "About DMTF." https://www.dmtf.org/about

[14] Unified Extensible Firmware Interface Specifications Forum, "Specifications." https://uefi.org/specifications

[15] Albert d’Aviau de Piolant, et al., “Improving energy efficiency of HPC applications using unbalanced GPU power capping,” 11 Feb 2025. https://inria.hal.science/hal-04883872/document

[16] Cisco, “Model Driven Telemetry White Paper”, April 3, 2024. https://www.cisco.com/c/en/us/products/collateral/switches/catalyst-9300-series-switches/model-driven-telemetry-wp.html