Sarcouncil Journal of Applied Sciences

ISSN(Online): 2945-3437

Volume- 05| Issue- 11| 2025

Research Article

Received: 10-10-2025 | Accepted: 05-11-2025 | Published: 19-11-2025

Standard Versus Custom HBM4: Design Trade-offs, Performance Gains, and Integration Challenges

Phani Suresh Paladugu

Synopsys, USA

Abstract: High Bandwidth Memory 4 (HBM4) represents a groundbreaking advancement in 3D-stacked DRAM technology, addressing the bandwidth limitations that have long constrained modern accelerators, graphics processing units (GPUs), and high-performance computing (HPC) platforms. The system architects have a critical architectural choice of either adapting the standardized JEDEC-compliant modules or customizing solutions to meet the requirements of the system-on-chip and package co-design solutions. Standard HBM4 configurations offer established reliability, supply chain accessibility, accelerated integration timelines, and predictable cost structures through proven physical layer intellectual property and reference packaging designs. Custom implementations unlock performance differentiation opportunities through optimized die stacking configurations, specialized physical layer tuning, and package-level co-optimization aligned with particular thermal and power delivery constraints, though at substantially elevated development complexity, extended timelines, and increased non-recurring engineering expenditures. The architectural choice fundamentally shapes subsequent design decisions across electrical engineering domains, packaging technology selections, validation methodology requirements, and supply chain strategies. Through-silicon via technology enables vertical interconnect densities exceeding conventional approaches while introducing mechanical stress considerations and keep-out zone constraints. Energy efficiency improvements stem from dramatically shortened interconnect distances, reducing parasitic capacitances. Issues facing integration include the complexity of package substrate co-design, efficient package substrate power delivery network design to meet the demands of transient current, thermal conductivity in the vertically stacked design, signal and power integrity testing with minimal design margins, and the overall test and validation processes. The decision framework is a synthesis of several assessment criteria, such as time-to-market demands, risk tolerance profiles, performance optimization demands, cost-benefit analysis, and organizational capabilities, to facilitate strategic choice between the standard and custom approaches to certain application situations.

Keywords: High Bandwidth Memory, 3D-stacked DRAM, through-silicon vias, heterogeneous integration, advanced packaging.

INTRODUCTION

The insatiable need for more memory bandwidth in modern computing systems has made High Bandwidth Memory (HBM) architectures an inseparable part of accelerators, graphics processing units, and high-performance computing systems. As memory bandwidth increasingly emerges as a critical bottleneck in data-intensive applications, the introduction of HBM4 marks a pivotal milestone in the evolution of 3D-stacked DRAM technology. It features a high-density memory architecture that delivers exceptional aggregate bandwidth within a compact physical footprint, achieved through the strategic use of Through-Silicon Vias (TSVs) and advanced packaging techniques.

The primary motivation behind the adoption of HBM lies in overcoming the bandwidth constraints inherent to traditional memory architectures. Conventional DRAM interfaces face substantial power efficiency challenges, with DDR-based solutions typically consuming around 20pJ (pico joules) per bit at contemporary data rates(Jeddeloh, J., & Keeth, B. 2012). The Hybrid Memory Cube (HMC) architecture, which shares fundamental design concepts with HBM through the use of 3D stacking and through-silicon via (TSV) technology, validated the feasibility of achieving significantly higher bandwidth density while lowering energy consumption to roughly 7pJ per bit (Jeddeloh, J., & Keeth, B. 2012). Utilizing vault structures with 128-bit interfaces operating at 10 Gbps per pin, HMC delivered an impressive aggregate bandwidth of 160 GB/s per package within a compact 31 mm × 31 mm footprint (Jeddeloh, J., & Keeth, B. 2012). The substantial power reduction was primarily attributed to the shortened interconnect paths, as TSV-based vertical connections operate at micrometer-scale distances far shorter than the millimeter to centimeter scale links typical of traditional off-package memory systems. These architectural breakthroughs paved the way for the evolution of High Bandwidth Memory (HBM), demonstrating that 3D-stacked designs could effectively overcome limitations in bandwidth, power, and form factor.

Building on this foundation, HBM emerged as a JEDEC-standardized technology, offering performance-balanced specifications that gained widespread industry adoption. HBM2 set the

*Corresponding Author: Phani Suresh Paladugu

DOI- https://doi.org/10.5281/zenodo.17652880

baseline with 256 GB/s of bandwidth per stack via 1024-bit interfaces, while HBM2E expanded capabilities to 460 GB/s through higher per-pin speeds and increased die stacking (Semiconductor Engineering). The introduction of HBM3 further advanced performance, achieving bandwidths between 665 GB/s and 819 GB/s with 16-high stacks and 6.4 Gbps data rates (Semiconductor Engineering). The forthcoming HBM4 generation aims for even more ambitious targets approaching 1.5 TB/s per stack enabled by 8 Gbps per-pin rates and improved channel efficiency (Semiconductor Engineering). Across these generations, HBM has consistently leveraged its core architectural strength: positioning memory dies in extremely close proximity to compute logic using advanced packaging methods such as silicon interposers and organic substrates, thereby minimizing signal delay and reducing energy loss (Semiconductor Engineering).

When integrating HBM4, system architects and design engineers face a pivotal architectural decision: whether to adopt standardized, JEDEC-compliant modules that offer proven reliability and robust supply chain support, or to develop custom HBM4 solutions precisely tailored to the unique requirements of a specific system-on-chip (SoC) and co-optimized packaging strategy. This decision has profound implications across multiple fronts, including performance optimization, development risk, manufacturing cost, and time-to-market. This article presents an in-depth analysis of the architectural and engineering trade-offs between standard and custom HBM4 implementations, providing a systematic assessment of design factors, quantitative performance evaluations, and integration challenges.

Table 1: Evolution of High Bandwidth Memory Architectures (Jeddeloh, J., & Keeth, B. 2012; Semiconductor Engineering)

Memory Architecture	Interface Characteristics	Energy Efficiency	Bandwidth Capability	Physical Configuration
Traditional DDR	Off-package planar interconnects	Moderate efficiency with high parasitic loading	Limited by pin count and frequency	Discrete module placement
Hybrid Memory Cube	Vault-based 3D stacking with TSV integration	Significant reduction through shortened paths	Enhanced through vertical integration	Compact cubic footprint
HBM2	JEDEC-standardized wide parallel interface	Improved through proximity placement	Baseline multi-channel architecture	Silicon interposer intégration
HBM2E	Extended stacking with increased rates	Maintained efficiency with higher throughput	Performance scaling through die count	Enhanced vertical configuration
HBM3	Advanced channel organization	Optimized power characteristics	Substantial bandwidth increases	Refined packaging approaches
HBM4	Aggressive per-pin rate targets	Further efficiency refinements	Approaching terabyte-per-second thresholds	Next-generation integration techniques

ARCHITECTURAL FOUNDATIONS AND DESIGN PHILOSOPHY

The fundamental concept of HBM architecture is to vertically stack multiple DRAM dies interconnected through through-silicon vias (TSVs) to a base die, which interfaces with the host processor via a wide, parallel channel architecture operating at relatively low frequencies compared to traditional DDR designs. This architectural paradigm delivers three major advantages: a drastic reduction in interconnect length, enhancing per-bit energy efficiency; exceptionally high aggregate I/O bandwidth enabled by numerous parallel lanes and channels; and a compact physical footprint suitable for complex multi-chip module (MCM) designs.

The three-dimensional integration principle underpinning HBM technology effectively overcomes the inherent limitations of conventional planar memory architectures through vertical stacking facilitated by TSVs. These TSVs serve as the critical enablers of 3D integration, providing

Publisher: SARC Publisher

vertical electrical connections through the silicon substrate with typical diameters of 5 to 10 micrometers and aspect ratios between 5:1 and 10:1 (Athikulwongse, K. *et al.,* 2010). However, the physical realization of TSVs imposes notable design constraints particularly the need to maintain keep-out zones around each via to mitigate mechanical stress-induced failures and electrical interference with nearby circuitry (Athikulwongse, K. *et al.,* 2010). Design guidelines typically require exclusion zones extending 10 to 15 micrometers radially from the TSV center, restricting standard cell placement and routing in these regions (Athikulwongse, K. *et al.,* 2010). Consequently, TSV arrays can occupy approximately 5 to 8% of the total die area, depending on via pitch and density specifications (Athikulwongse, K. *et al.,* 2010). Furthermore, mechanical stress effects extend beyond the immediate exclusion zones, with finite element simulations indicating stress propagation up to 50 to 100 micrometers from TSV locations during thermal cycling (Athikulwongse, K. *et al.,* 2010). Despite these limitations, TSV technology enables ultra-dense vertical interconnects exceeding 10,000 connections per square millimeter when distributed across the die surface representing several orders of magnitude improvement over conventional wire-bonded or flip-chip interconnection methods.

The remarkable energy efficiency achieved by 3D-stacked memory architectures arises primarily from the substantial reduction in interconnect lengths and the associated parasitic capacitances. In conventional off-chip memory systems, signal paths extend across package substrates, printed circuit boards, and connectors, accumulating capacitive loads of 15 to 25 picofarads per signal line and necessitating impedance matching networks that further elevate power consumption (Black, B. *et al.,* 2006). By contrast, three-dimensional die stacking shortens these interconnects to micrometer-scale distances rather than millimeters or centimeters thereby reducing wire capacitance by factors of 10 to 100x, depending on implementation geometry (Black, B. *et al.,* 2006). Quantitative studies on 3D processor memory integration have demonstrated energy reductions of 10 to 100x for memory access operations compared to traditional off-chip DRAM, with power savings as high as 94% observed for processor to L2 cache communication when implemented via TSV-based interconnects instead of conventional 2D routing (Black, B. *et al.,* 2006). These gains scale with both access frequency and data transfer volume, making 3D integration particularly advantageous for bandwidth-intensive workloads where memory systems dominate power consumption (Black, B. *et al.,* 2006). Additionally, the reduced capacitance enables faster signal transitions and improved timing margins, allowing higher data rates without the need for complex equalization or signal-conditioning circuits thereby simplifying design and further lowering energy costs.

HBM4 extends these advantages with per-pin data rates reaching up to 8 Gbps, enhanced channel architectures supporting up to sixteen independent 64-bit channels per stack, and advanced packaging technologies such as hybrid bonding that further minimize parasitic effects and increase interconnect density. Standardized JEDEC-compliant designs ensure interoperability and ecosystem compatibility, while custom implementations allow fine-tuning of architectural parameters to align with specific workload characteristics and thermal design constraints.

Table 2: Three-Dimensional Integration Technology Characteristics (Athikulwongse, K. *et al.,* 2010; Black, B. *et al.,* 2006)

Integration Aspect	TSV Implementation	Design Constraints	Performance Impact	Energy Considerations
Vertical Interconnection	Micron-scale via structures penetrating silicon	Keep-out zones around via locations	High-density connection capability	Minimal parasitic capacitance
Mechanical Effects	Stress field propagation from thermal cycling	Exclusion regions for circuit placement	Die area utilization impact	Reduced switching energy
Electrical Properties	Low-resistance vertical pathways	Aspect ratio limitations	Superior to wire-bonding approaches	Dramatic capacitance reduction
Integration	Orders of magnitude	Pitch and spacing	Enhanced	Power

Publisher: SARC Publisher

Density	improvement	requirements	bandwidth potential	consumption benefits
Thermal Behavior	Vertical heat conduction paths	Stress management requirements	Signal integrity improvements	Energy efficiency gains

COMPARATIVE ANALYSIS OF DESIGN TRADE-OFFS

Selecting between standard and custom HBM4 implementations requires careful evaluation across both engineering and business dimensions. Each approach offers distinct advantages and challenges that must be weighed against project requirements and organizational priorities.

In early-stage decision-making, time-to-market considerations often dominate. Standard HBM4 modules accelerate integration through validated design methodologies, established supplier qualification processes, and minimal intellectual property negotiations. Design teams benefit from mature development ecosystems, including extensive documentation, reference designs, and technical support. Integration timelines for advanced packaging depend on the complexity of heterogeneous requirements: standard configurations typically move from specification to production in 18 to 24 months, whereas custom solutions can extend to 24 to 36 months due to additional design iterations and validation cycles. Custom designs inherently require longer development periods, as they involve specialized validation, extra design iterations, and coordination across vendors for non-standard specifications. Complexity grows further when multiple die types with differing process technologies, thermal coefficients, and electrical characteristics must coexist within a single package (Lau, J. H. 2022).

Risk profiles differ significantly between the two approaches. Standard solutions carry lower first-silicon risk, benefiting from well-characterized yield metrics, established manufacturing processes across multiple qualified suppliers, and extensive field experience that informs reliability expectations. Custom HBM4 development introduces higher risk factors, including yield uncertainty from non-standard TSV layouts or modified die geometries, potential single-source supply chain vulnerabilities, and more complex validation of untested configurations. Advanced 2.5D and 3D packaging techniques face warpage challenges, with package bow measurements reaching 200 to 400 micrometers during reflow when integrating large dies with mismatched coefficients of thermal expansion (Lau, J. H. 2022). Such warpage can generate mechanical stresses exceeding 100 MPa at critical interfaces, risking delamination, cracking, or solder joint failures. Thermal management also becomes more demanding with stacked dies, where vertical heat flux densities can surpass 100 W/cm² in high-performance computing setups, requiring sophisticated thermal interface materials and heat-spreading solutions to maintain junction temperatures within safe limits (Lau, J. H. 2022).

Electrical and physical design considerations highlight additional performance trade-offs. Standard PHY implementations ensure robust operation across a range of SoC platforms through conservative design margins and vendor-validated parameters. In contrast, custom PHYs allow aggressive per-pin data rate optimization and tailored equalization techniques aligned with specific package characteristics. Embedded Multi-die Interconnect Bridge technology exemplifies high-density localized interconnects, achieving 55-micrometer routing and bump pitches, delivering bandwidth densities exceeding 2 TB/s per millimeter of bridge width (Mahajan, R. *et al.*, 2019). This architecture employs fine-pitch copper pillar bumps (40 micrometer diameter on 55 micrometer pitches), achieving approximately 8× higher connection density than conventional organic substrate routing while preserving signal integrity at multi-gigahertz frequencies (Mahajan, R. *et al.*, 2019). The bridge itself is only 400 micrometers thick and integrates into standard organic substrates via localized silicon embedding, enabling heterogeneous interconnect solutions that balance high-density die-to-die communication with cost-efficient global routing (Mahajan, R. *et al.*, 2019). Electrical characterization shows insertion losses below 1 dB per millimeter at 10 GHz, supporting high-speed serial protocols and wide parallel memory interfaces without complex equalization circuits (Mahajan, R. *et al.*, 2019).

Cost considerations also differ fundamentally. Standard HBM4 benefits from economies of scale, whereas custom solutions incur substantial upfront expenses for design, validation, and supplier qualification. Feature flexibility varies accordingly: custom implementations allow specialized capabilities such as asymmetric

Publisher: SARC Publisher

channel allocation and application-specific error correction, optimized for targeted workloads.

Table 3: Standard Versus Custom HBM4 Implementation Comparison (Lau, J. H. 2022; Mahajan, R. *et al.*, 2019)

Design Dimension	Standard Approach	Custom Approach	Risk Factors	Performance Differentiation
Development Timeline	Accelerated through proven methodologies	Extended through iterative optimization	First-silicon success variability	Incremental improvements possible
Supply Chain	Multiple qualified suppliers available	Limited suppliers, have to collaborate early in the development cycle	Yield uncertainty considerations	Specialized feature enablement
Package Integration	Reference interposer designs	Non-standard routing topologies	Warpage and stress challenges	Optimized electrical paths
Thermal Management	Established interface materials	Tailored thermal solutions	Heat flux density constraints	Sustained bandwidth enhancement
Physical Layer	Conservative vendor-validated margins	Aggressive tuning for specific conditions	Signal integrity complexity	Higher data rate potential
Interconnect Technology	Standard interposer routing	Advanced bridge architectures	Manufacturing process control	Superior bandwidth density

PERFORMANCE CHARACTERISTICS AND INTEGRATION CHALLENGES

Performance analysis reveals that custom HBM4 implementations can achieve meaningful bandwidth and efficiency improvements over standard configurations, though realizing these gains requires navigating substantial integration complexity. Effective bandwidth scales with per-pin data rate, channel count, and utilization efficiency, with custom solutions potentially delivering improvements through aggressive PHY tuning, optimized channel allocation, and reduced command overhead from specialized protocols.

Package and substrate co-design emerges as a primary integration challenge for custom HBM4 implementations. Non-reference routing topologies on interposers or embedded bridge solutions introduce signal routing congestion, via density constraints, and mechanical stress considerations that may induce package warpage. Engineering costs escalate substantially compared to reference designs, requiring advanced electronic design automation tools and specialized packaging expertise. TSV interposer technology represents the most cost-effective integration platform for heterogeneous 3D integration, enabling fine-pitch interconnects with line width and spacing capabilities down to 2 micrometers and supporting routing densities exceeding 10,000 connections per square millimeter (Lau, J. H. 2011). The interposer substrate typically measures between 100-200 micrometers in thickness after thinning processes and incorporates multiple redistribution layers, commonly 2-5 metal levels, fabricated using wafer-level processing with features far exceeding organic substrate capabilities (Lau, J. H. 2011). TSV structures within interposers exhibit diameters ranging from 5-10 micrometers with aspect ratios between 5:1 and 10:1, enabling vertical electrical connections through the silicon substrate with parasitic capacitance below 50 femtofarads per via and resistance typically under 100 milliohms (Lau, J. H. 2011). The manufacturing cost structure for TSV interposers proves favorable compared to alternative 3D integration approaches, with wafer-level processing enabling economies of scale and cost per interposer ranging from $15-40, depending on size, complexity, and production volume (Lau, J. H. 2011). However, thermomechanical reliability challenges arise from coefficient of thermal expansion mismatches between silicon interposers at 2.6 ppm/°C and organic package substrates at 15-17 ppm/°C, generating shear stresses at interfaces during thermal excursions from -40°C to 125°C that can reach 40-70 MPa (Lau, J. H. 2011). These stress concentrations necessitate careful underfill material selection and package design optimization to prevent solder joint fatigue and delamination failures over operational lifetimes spanning billions of thermal cycles (Lau, J. H. 2011).

Publisher: SARC Publisher

Power delivery network design confronts substantial challenges from the large transient current demands characteristic of HBM stacks. Custom implementations must carefully position voltage regulator modules and decoupling capacitors to manage voltage droops within acceptable margins. Advanced power delivery networks for high-performance computing systems must maintain target impedance specifications below 1 milliohm across frequency ranges from DC to 1 GHz to support instantaneous current transients exceeding 500 amperes with voltage regulation tolerances under ±5% (Murari, K. *et al.*, 2025). The power distribution impedance characteristics directly impact signal integrity and timing margins, with inductive impedance components at high frequencies creating resonant peaks that require strategic placement of decoupling capacitors ranging from large bulk capacitors providing low-frequency stability to small on-die capacitors addressing multi-hundred megahertz transients (Murari, K. *et al.*, 2025). Comprehensive simulation of transient IR drop phenomena becomes mandatory when supporting multiple power domains with distinct voltage requirements per die, particularly in 3D-stacked configurations where power delivery paths traverse vertical interconnects with cumulative resistance and inductance (Murari, K. *et al.*, 2025). The complexity multiplies when accommodating dynamic voltage and frequency scaling schemes requiring voltage transitions of 200-300 millivolts within microsecond timescales while maintaining load regulation (Murari, K. *et al.*, 2025).

Thermal management represents a critical constraint, with heat extraction paths encountering significant impedance through multiple die layers. Signal and power integrity verification intensifies for custom PHY implementations, requiring on-chip equalization and adaptive training algorithms. Test methodologies expand significantly, while reliability phenomena, including TSV fatigue, necessitate extensive qualification campaigns. Firmware integration introduces complexity as memory controller software must accommodate vendor-specific timing and training protocols diverging from standard implementations.

Table 4: Package-Level Integration Challenges (Lau, J. H. 2011; Murari, K. *et al.*, 2025)

Integration Element	Interposer Technology	Power Distribution	Critical Parameters	Design Complexity
Substrate Architecture	Fine-pitch routing capability	Multi-domain voltage delivery	Line width and spacing limits	Advanced fabrication processes
Vertical Connections	TSV-based through-silicon paths	Decoupling capacitor networks	Parasitic resistance and capacitance	Via density optimization
Thermal Expansion	Silicon-organic CTE mismatch	Transient current management	Stress concentration locations	Reliability engineering
Manufacturing Cost	Wafer-level processing economies	Component placement strategy	Unit cost variability factors	Design-for-manufacturability
Electrical Performance	Low-loss signal propagation	Impedance control requirements	Frequency dependent characteristics	Electromagnetic modeling
Reliability Concerns	Thermal cycling fatigue	Voltage regulation tolerance	Operational lifetime targets	Qualification testing

DESIGN METHODOLOGY AND DECISION FRAMEWORK

Establishing rigorous design and verification methodologies is critical for the successful integration of custom HBM4 solutions. Architectural modeling serves as the foundation, employing system-level simulation frameworks to estimate bandwidth and latency requirements under realistic workload conditions. These simulations help pinpoint performance bottlenecks and inform resource allocation decisions throughout the memory hierarchy.A structured decision framework integrates these insights, guiding whether standard or custom HBM4 configurations best align with project objectives. Standard implementations are typically optimal when time-to-market pressures are high and production volumes fit within existing ecosystem offerings. In contrast, custom HBM4 solutions become justified when workloads require specialized architectures, production scales support non-recurring engineering (NRE) investments in the range of $25–60 million, and the organization possesses the necessary signal integrity and

Publisher: SARC Publisher

packaging expertise.Cost-benefit sensitivity analyses further strengthen the decision process by quantifying engineering costs, per-unit price differentials, anticipated performance gains, and overall business value realization. Practical recommendations include early prototypes, whether standard or custom, to obtain silicon prototypes (or FPGA emulation of controller behavior) early memory interactions are often the cause of late integration issues. Invest in SI/PI tooling and test infrastructure, high-fidelity modeling shortens iterations. Plan for firmware adaptability, make training parameters updatable in the field to tune margins post-packaging. Use staged customization, start with standard stacks and gradually introduce custom tuning (PHY, training algorithms) before committing to custom die or package geometry. Negotiate with suppliers, custom work is more successful if suppliers are committed partners; involve packaging and DRAM vendors early.

Table 5: Decision Framework for standard vs custom HBM4

Metric	Standard HBM4	Custom HBM4
Time-to-Market	Fast (reference IP and validated stacks)	Slower (requires design co-optimization)
Design Risk	Low (mature vendors, proven yields)	Higher (yield and validation risks)
Performance Potential	Moderate (JEDEC-constrained)	High (PHY and stack tuning possible)
Power Efficiency	Good, limited tuning options	Potentially excellent (optimized for SoC)
Thermal Optimization	Reference cooling, limited flexibility	Fully optimized for local thermal zones
Cost (NRE + BOM)	Low NRE, predictable unit cost	High NRE, variable per-unit cost
Supply Chain Flexibility	High (multi-vendor sourcing possible)	Low (often single-source supply)
Test/Validation Complexity	Moderate (standard training firmware)	High (requires advanced test flows)
Custom Feature Support	Limited (fixed JEDEC features)	High (custom ECC, channel tuning)
Integration Difficulty	Low (reference interposer designs)	High (complex co-design required)

CONCLUSION

HBM4 stands as a transformative architectural solution to the growing challenge of memory bandwidth limitations that increasingly constrain performance in data-intensive computing systems. The strategic decision between adopting standard or custom HBM4 implementations carries significant ramifications across technical, economic, and competitive dimensions extending well beyond raw performance considerations. Standard HBM4 configurations offer predictable performance, lower development and supply chain risks, and faster integration through established ecosystems of validated IP, trusted supplier networks, and comprehensive reference designs. These advantages make standardized solutions particularly attractive for organizations seeking rapid market entry, moderate production volumes, and conservative risk profiles. Conversely, custom HBM4 implementations enable meaningful differentiation in performance and efficiency through tailored die stacking architectures, optimized channel configurations, aggressive PHY tuning, and co-optimized packaging strategies. Quantitative analyses indicate that such custom designs can deliver superior bandwidth and energy efficiency through improved thermal management and power delivery networks. However, these benefits require significant non-recurring engineering (NRE) investments, complex supply chain coordination, and extensive verification efforts to ensure functional integrity across non-standard configurations.

Integration challenges in custom solutions span multiple engineering domains, including package co-design, power delivery optimization, thermal regulation of stacked dies, signal and power integrity validation, test infrastructure development, and software stack adaptation. Successfully addressing these demands requires specialized expertise, advanced simulation capabilities, and strong collaboration with packaging and memory vendors. Organizations without these competencies or those unwilling to assume elevated risk should carefully weigh potential performance advantages against the associated design and validation complexities. A pragmatic strategy involves adopting a phased customization model beginning with standard HBM4 stacks while progressively introducing tailored enhancements at the physical layer, firmware, and training algorithm levels. This

Publisher: SARC Publisher

incremental approach mitigates risk, builds internal expertise, and validates performance improvements through empirical data before committing to fully custom die or package designs. Ultimately, a structured decision framework should guide the evaluation of standard versus custom HBM4 options, considering workload demands, business models, engineering maturity, and competitive positioning. For many design teams, hybrid approaches combining standardized memory stacks with advanced PHY and firmware optimizations offer substantial performance benefits while maintaining manageable risks and preserving supply chain flexibility.

As HBM4 technology continues to mature and its adoption broadens across AI, HPC, and advanced graphics domains, the ecosystem is expected to evolve toward greater standardization of high-end features that were once exclusive to custom designs. Nonetheless, the fundamental trade-off between broad interoperability and workload-specific optimization will remain, ensuring that both standard and custom HBM4 solutions continue to hold distinct strategic value across diverse segments of the high-performance computing landscape.

REFERENCES

1. Jeddeloh, J., & Keeth, B. "Hybrid memory cube new DRAM architecture increases density and performance." *2012 symposium on VLSI technology (VLSIT)*. IEEE, (2012).
2. Semiconductor Engineering, "High Bandwidth Memory,"
3. Athikulwongse, K., Chakraborty, A., Yang, J. S., Pan, D. Z., & Lim, S. K. "Stress-driven 3D-IC placement with TSV keep-out zone and regularity study." *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, (2010).
4. Black, B., Annavaram, M., Brekelbaum, N., DeVale, J., Jiang, L., Loh, G. H., ... & Webb, C. "Die stacking (3D) microarchitecture." *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*. IEEE, (2006).
5. Lau, J. H. "Recent advances and trends in advanced packaging." *IEEE Transactions on Components, Packaging and Manufacturing Technology* 12.2 (2022): 228-252.
6. Mahajan, R., Sankman, R., Aygun, K., Qian, Z., Dhall, A., Rosch, J., ... & Salama, I. "Embedded Multi-die Interconnect Bridge (EMIB) A Localized, High Density, High Bandwidth Packaging Interconnect." *Advances in Embedded and Fan-Out Wafer-Level Packaging Technologies* (2019): 487-499.
7. Lau, J. H. "The most cost-effective integrator (TSV interposer) for 3D IC integration system-in-package (SiP)." *International Electronic Packaging Technical Conference and Exhibition*. Vol. 44618. (2011).
8. Murari, K., Bhushan, R., Parida, S. K., Singh, S. N., & Soman, S. A. (Eds.). "Recent Advances in Power Systems." *Springer*, (2025).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Paladugu, P. S. "Standard Versus Custom HBM4: Design Trade-offs, Performance Gains, and Integration Challenges." *Sarcouncil Journal of Applied Sciences* 5.11 (2025): pp 63-70.

Publisher: SARC Publisher