TECHNICAL DOSSIER // ADVANCED SEMICONDUCTOR PACKAGING & MEMORY SYSTEMS

HBM4 High-Bandwidth Memory Architectural Specification

2048-Bit Physical Interface, Logic Base Die Co-Design, and Hybrid Bonding Physics

Published: October 30, 2026 | Distribution: Open Engineering Community | Document ID: LT-2026-B16-04

EXECUTIVE SUMMARY & MICROARCHITECTURAL SCOPE

High Bandwidth Memory 4 (HBM4) marks the most significant architectural disruption in memory engineering in a decade. By doubling the physical bus width from 1024 bits to 2048 bits and transitioning the base buffer die from legacy memory processes to advanced 4nm/3nm logic foundry nodes, HBM4 delivers over 2.0 to 3.0 TB/s of memory bandwidth per stack while slashing I/O energy consumption below 2.2 picojoules per bit.

Specification	HBM2e	HBM3	HBM3e	HBM4
Physical Bus Width	1024-bit	1024-bit	1024-bit	2048-bit (2x Doubling)
Per-Pin Data Rate	3.6 Gbps	6.4 Gbps	9.6 Gbps	6.4 - 8.0 Gbps
Peak Bandwidth / Stack	460 GB/s	819 GB/s	1.23 TB/s	2.00 - 3.00 TB/s
Max Stack Height	8-High DRAM	12-High DRAM	12-High DRAM	16-High DRAM
Base Die Process	DRAM Buffer Node	DRAM Buffer Node	DRAM Buffer Node	Advanced 4nm / 3nm Logic
Packaging Interconnect	Micro-Bumps (55um)	Micro-Bumps (35um)	Micro-Bumps (25um)	Direct Hybrid Bonding (<1um)

1. The 2048-Bit Wide-Bus Paradigm Shift

Previous HBM generations scaled memory throughput primarily by pushing per-pin signaling frequencies over a static 1024-bit bus. However, operating single-ended I/O lines above 10 Gbps across silicon interposers introduced severe transmission line skin effects, dielectric loss, and crosstalk. HBM4 overcomes this high-frequency wall by doubling the physical pin count to 2048 bits, delivering 2.0+ TB/s at conservative, power-efficient clock frequencies.

2. Advanced Logic Foundry Base Die Co-Design

For the first time in HBM history, the bottom base (buffer) die is fabricated on advanced logic foundry nodes (TSMC N4/N3, Intel 16/3, Samsung 4nm) rather than commodity memory silicon. This allows integrating custom memory controllers, on-die Built-In Self-Test (BIST), advanced power delivery circuits, and even near-memory neural compute accelerators directly beneath the DRAM stack.

3. Direct Copper-to-Copper Hybrid Bonding (DBI)

To accommodate over 10,000 vertical Through-Silicon Vias (TSVs) and 2048 signal pins within standard JEDEC package dimensions, 16-high HBM4 stacks abandon traditional solder micro-bumps. Instead, Direct Bond Interconnect (DBI) creates atomic copper-to-copper bonds surrounded by silicon oxide dielectric at sub-micron pitches, eliminating solder height variance and reducing thermal impedance by 30%.

4. Thermal Dissipation and Power Delivery Architecture

Stacking 16 DRAM layers dissipation upwards of 60W per HBM cube creates severe thermal hot spots. HBM4 resolves this via embedded dummy thermal TSVs, backside power distribution networks, and advanced molded underfill (MUF) materials with ultra-high thermal conductivity (k > 3.0 W/m-K), preventing thermal throttling on multi-kilowatt AI accelerator platforms.

STRATEGIC ENGINEERING TAKEAWAYS & DEPLOYMENT CHECKLIST

- **2048-Bit Physical Interface:** Doubles channel interconnect lines, lowering per-pin clock stress while doubling throughput.
- **Advanced Logic Base Die:** Transitions buffer die to TSMC/Intel 4nm/3nm, enabling custom controllers and BIST.
- **Direct Hybrid Bonding:** Sub-micron copper-copper fusion replaces micro-bumps, unlocking 16-high vertical stacking.
- **Sub-2.2 pJ/bit Efficiency:** Lowers transmission power dissipation by 40% compared to high-frequency HBM3e links.