

HBM4 Hybrid Direct Copper-to-Copper Bonding & Packaging

Engineering Report on 16-High Die Stacking, Sub-Micron Pitch, and Thermal Stress

Published by *Lesty Tech Semiconductor & Packaging Group* | Date: October 20, 2026 | Reference: LESTY-ENG-2026-B6

1. The Interconnect Scaling Wall in High-Bandwidth Memory

As artificial intelligence accelerators transition to trillion-parameter frontier architectures, memory bandwidth is the primary operational bottleneck. High-Bandwidth Memory (HBM3e) achieves 1.2 TB/s per stack using a 1024-bit interface and microbump solder interconnects at a 25-micron pitch. However, scaling to HBM4 mandates a doubled 2048-bit interface across 16-high DRAM die stacks, pushing interconnect pitches down to sub-10 microns. At this geometry, microbump solder balls suffer from severe capillary underfill voiding, solder bridge short-circuits, and intermetallic compound (IMC) brittleness, terminating the viable scaling roadmap for thermal-compression bonding with non-conductive film (TC-NCF).

2. Copper-to-Copper Direct Hybrid Bonding (DBI) Physics

Critical 3D Packaging Reality: Direct Hybrid Bonding eliminates solder bumps entirely. By chemically planarizing alternating dielectric (SiO2 or SiCN) and copper pad surfaces via Chemical Mechanical Polishing (CMP) to sub-0.5nm roughness, room-temperature van der Waals dielectric bonding is initiated. Subsequent thermal annealing at 250°C to 300°C causes the copper pads to expand thermally, forming continuous metallic crystalline grains across the interface at sub-1 micron pitch.

3. Thermal Dissipation and Thermo-Mechanical Shear Stress

While hybrid bonding slashes interconnect height from 15 microns to under 1 micron (reducing vertical electrical parasitic inductance by 85%), stacking 16 active DRAM dies atop a 1,000W AI host ASIC concentrates unprecedented thermal flux. Different coefficients of thermal expansion (CTE) between the silicon dies and interposers generate intense shear stresses during rapid power transients, necessitating advanced microfluidic silicon interposers and optimized structural dummy pads.

ARCHITECTURAL BENCHMARK & PARAMETER TELEMETRY

Packaging Metric	HBM3e Microbump (TC-NCF)	HBM4 Microbump (Advanced MR)	HBM4 Direct Hybrid Bonding
Interface Bus Width	1024-Bit Interface	2048-Bit Interface	2048-Bit Ultra-Wide
Interconnect Pitch	25 - 35 Micron Pitch	15 - 18 Micron Pitch	Sub-1 to 3 Micron Pitch
Interconnect Height	12 - 15 Microns	8 - 10 Microns	< 1 Micron (Near-zero)
Parasitic Inductance	0.12 - 0.18 nH	0.08 - 0.11 nH	< 0.02 nH (-85% drop)
Peak Bandwidth per Stack	1.18 TB/s	1.65 TB/s	> 2.20 TB/s (Record throughput)

