

Custom HBM as a System: Co-designing the Base Die, Die-to-Die Power, and Package Thermals for AI Accelerators

Mark Hsin-wei Tseng | Director, Marvell Technology



© 2026 Marvell. All rights reserved.

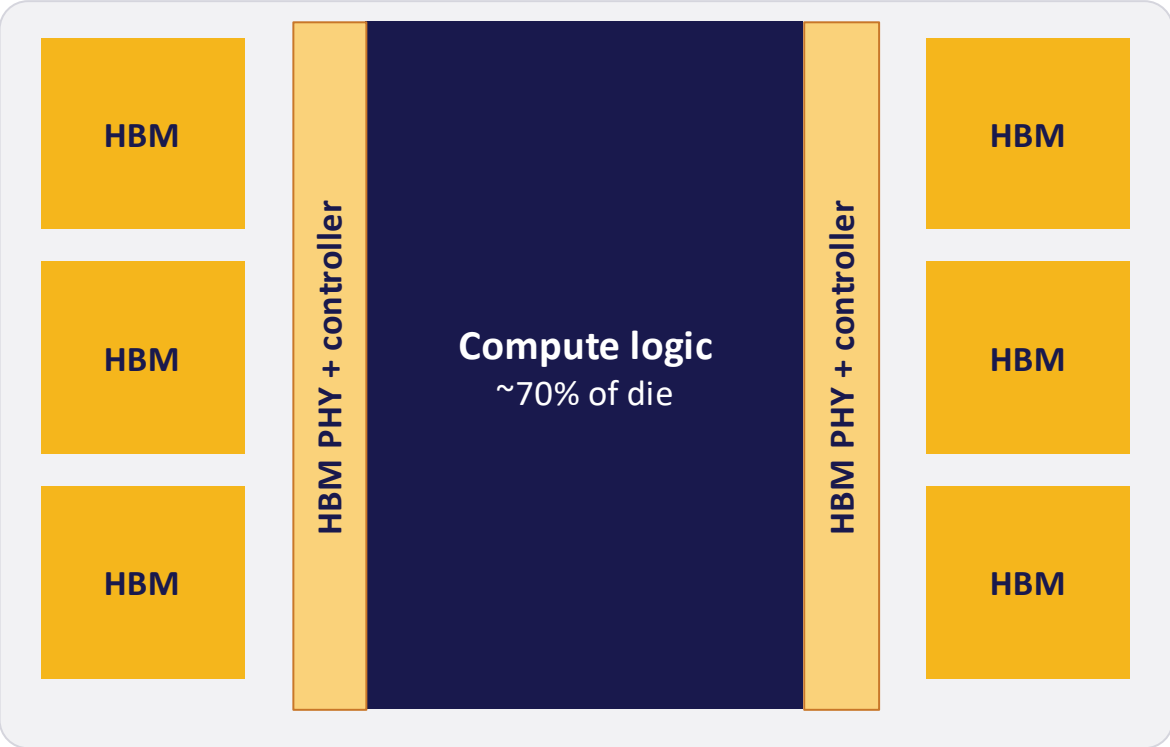
Forward-looking statements

Except for statements of historical fact, this presentation contains forward-looking statements (within the meaning of the federal securities laws) including statements related to future revenue, future earnings, and the success of our product releases that involve risks and uncertainties. Words such as “anticipates,” “expects,” “intends,” “plans,” “projects,” “believes,” “seeks,” “estimates,” “can,” “may,” “will,” “would” and similar expressions identify such forward-looking statements. These statements are not guarantees of results and should not be considered as an indication of future activity or future performance. Actual events or results may differ materially from those described in this presentation due to a number of risks and uncertainties.

Forward-looking statements are only predictions and are subject to risks, uncertainties and assumptions that are difficult to predict, including those described in the “Risk Factors” section of our Annual Reports on Form 10-K, Quarterly Reports on Form 10-Q and other documents filed by us from time to time with the SEC. Forward-looking statements speak only as of the date they are made. You are cautioned not to put undue reliance on forward-looking statements, and no person assumes any obligation to update or revise any such forward-looking statements, whether as a result of new information, future events or otherwise.

© 2026 Marvell. All rights reserved.

HBM: ~30 % of die area, > 50% of XPU cost



Why HBM crowds the die

- HBM4 doubles the interface to **2048 bits** (32 × 64-bit channels)[3], growing PHY area, power and edge "beachfront."
- More HBM stacks compete with compute logic for the same die edge.
- Memory I/O, not the cores, increasingly limits accelerator scaling.
- At today's memory price, memory BOM cost > 60-80% in many GPGPU/ASICs.

Not to scale. 1. Marvell Industry Analyst Day, Dec 2024, as reported by Futurum, "Marvell Debuts Custom HBM Compute Architecture," Dec 18, 2024. 2. Silicon Analysts, "HBM Memory Pricing and Specifications (2026)," Sep 2026, compiling Epoch AI, Raymond James, TrendForce and SemiAnalysis estimates. 3. JEDEC JESD270-4 HBM4 standard, Apr 2025.

© 2026 Marvell. All rights reserved.

Checklist make high-performance cHBM work

Base die and area

HBM4's 2,048-bit PHY takes ~30% of compute-die area

- ☐ Move PHY and controller into a custom base die
- ☐ Reclaim up to 25% of die area for compute
- ☐ Use the space for up to 30% more compute
- ☐ Speed up customized the base-die interface across memory vendors

D2D / TSV / HCB

UCIe was not built for HBM Next-gen class bandwidth density

- ☐ Cut D2D power
- ☐ Keep one D2D architecture from 2.5D to 3D hybrid bonding
- ☐ Prove 3D hybrid-bonding yield at volume

Package thermals

The HBM stack holds 57% of thermal resistance; HBM Next-gen reach > 2.0 W/mm².

- ☐ Reserve a silicon heat path at the D2D PHY
- ☐ Plan TIM, lid and cold plate with the base die
- ☐ Model heat flux above 2.0 W/mm² for HBM5
- ☐ Validate cooling beyond 100 W/stack and > 2.0 W/mm².

Ecosystem

No single company can build custom HBM alone

- ☐ Align memory makers, foundries, OSATs and XPU designers
- ☐ Share thermal models and test data across the chain
- ☐ Define cHBM with more re-usable modules between different AI applications.

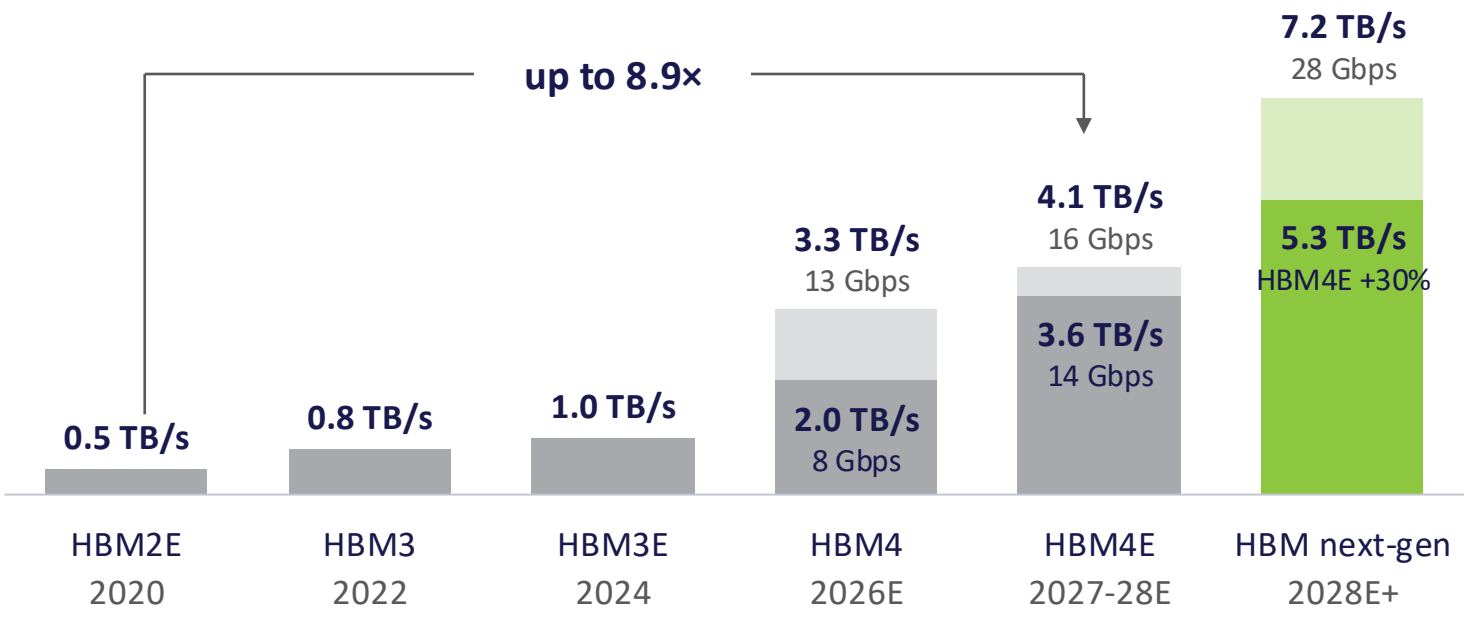
☑ Addressed today ☐ Additional effort needed

Sources: Marvell newsroom (Dec 2024; Aug 2025); JEDEC JESD270-4 and HBM thermal-density reference; IEEE ECTC 2026; SK hynix iHBM release (May 2026); Hot Chips 2026. Details on earlier slides.

© 2026 Marvell. All rights reserved.

HBM bandwidth per stack grows ~9× by 2028

HBM bandwidth per stack



- Per-stack bandwidth grows **~1.7× every 2 years**: 0.5 TB/s (2020) to 3.6–4.1 TB/s (2028E)
- Two ways to get there: **more data I/Os** (1,024 → 2,048 in HBM4) and **faster I/O speed**
- HBM4 moves the base die to a **logic process**, the opening for a custom base die in HBM next-gen

E = projection. Dark segment = low end, light = high end: HBM4 8 Gbps (JEDEC) to 13 Gbps (Hot Chips 2026); HBM4E 14–16 Gbps; HBM next-gen = 16 Gbps HBM4E +30% (industry disclosure, Aug 2026). Industry roadmap cites up to 28 Gbps × 2,048 I/Os (Hot Chips 2026). Source: JEDEC JESD270-4; SK hynix, Aug 2026; Hot Chips 2026; NVIDIA Technical Blog, "NVIDIA NVLink Fusion Brings NVHBM to Next-Generation AI Infrastructure," Aug 2026.

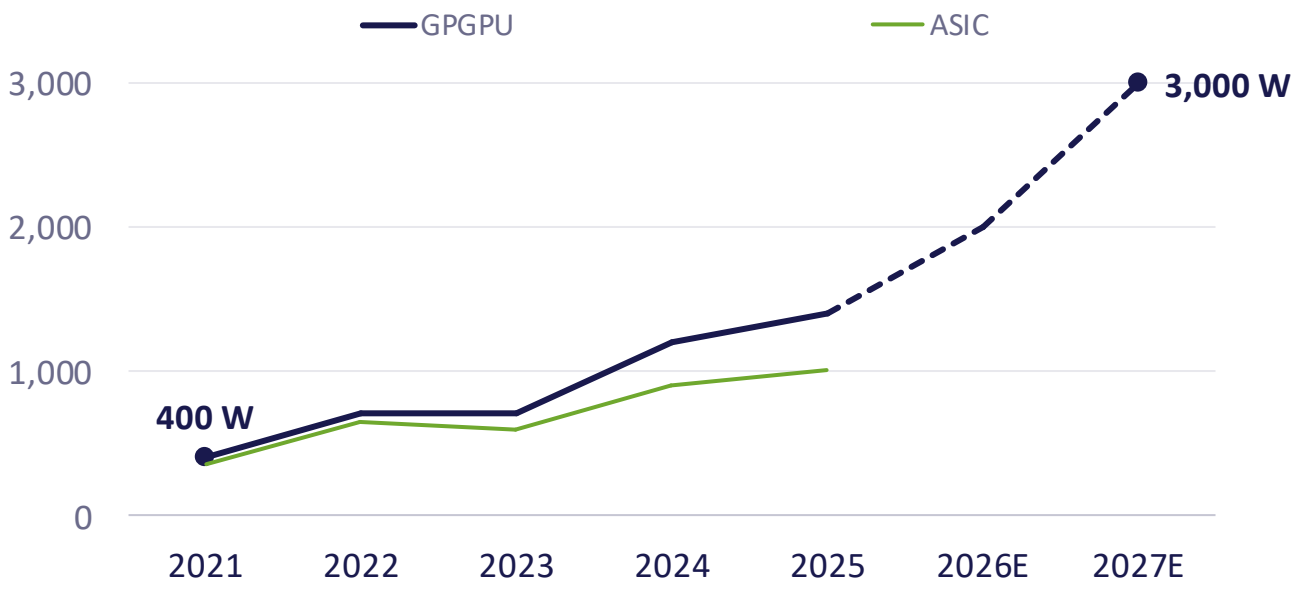
© 2026 Marvell. All rights reserved.

Base-die logic / memory reach heat limit

Base-die logic under the DRAM stack is solvable only through package-level co-design.

- HBM heat flux passes the JEDEC ~1.7 W/mm² reference: HBM5 reaches >2.0 W/mm² at >28 Gbps
- TIM, lid, and integrated cooling must be planned with the compute die, base die, and package — not bolted on

AI chip TDP by type, W



Power density: GPGPU logic die vs. HBM

Year	TDP	Logic die	W/mm²
2021	400 W	~800 mm²	**0.48**
2022	700 W	~800 mm²	**0.86**
2023	700 W	~800 mm²	**0.86**
2024	1,200 W	~1,600 mm²	**0.75**
2025	1,400 W	~1,600 mm²	**0.88**
2026E	2,000 W	~1,600 mm²	**1.25**
—		>28 Gbps I/O (2×)	**>2.0**

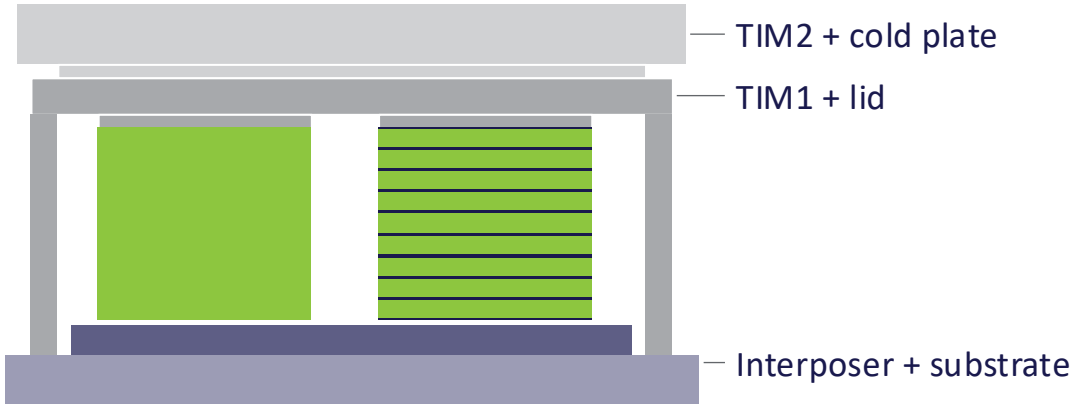
W/mm² = chart TDP ÷ GPGPU logic-die area (excl. HBM and I/O dies); 2024–26 dual-die parts assume ~800 mm² per reticle-size die; 2026–27 are estimates (E).
Source: JEDEC HBM thermal-density reference; SK hynix, "iHBM" Integrated Cooling Elements announcement, May 26, 2026; GPGPU vendor product specifications. / Marvell Internal. / JEDEC 2026 consortium. / S. Han, "HBM Base Die," Hot Chips 2026, Aug 23, 2026.

© 2026 Marvell. All rights reserved.

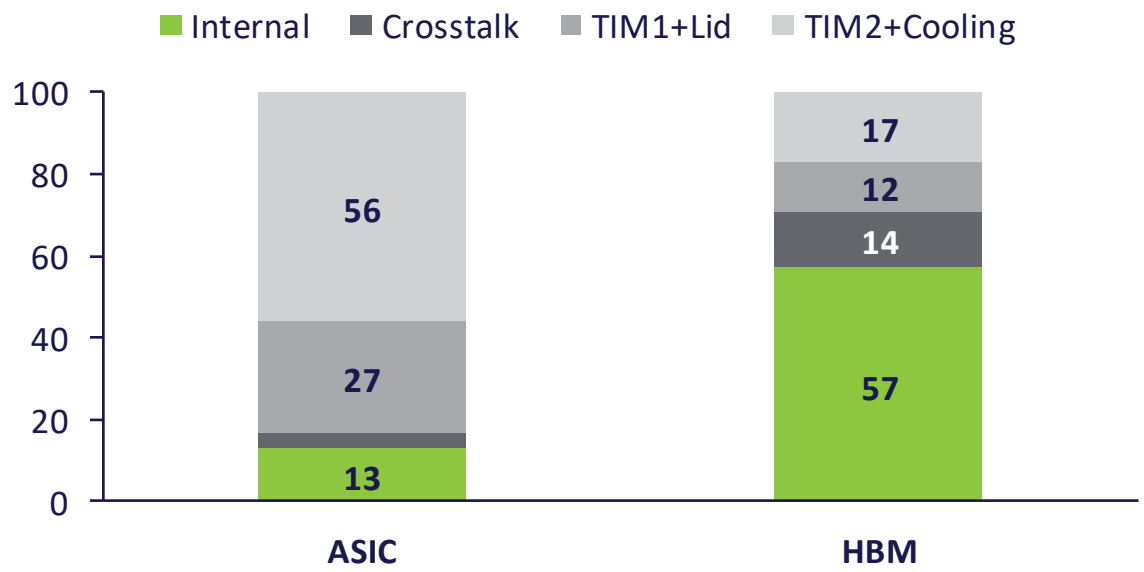
The HBM stack holds 57% of thermal resistance

- In 2.5D packages, HBM internal resistance is **57% of the stackup vs 13% for the ASIC** — the stack, not the logic, is the limiter
- Taller stacks push internal temperature past the acceptable range: **8- → 16-high HBM crosses the thermal budget by 2028**

2.5D package cross-section (heat flows up)



Thermal Resistance Contribution in 2.5D (%)



Source: "Package Thermal Design," 2026 IEEE 76th Electronic Components and Technology Conference (ECTC), Orlando, FL, May 26–29, 2026.

© 2026 Marvell. All rights reserved.

Hybrid bonds cut HBM thermal resistance 23–47%

- **No underfill gap:** Cu–Cu bond removes the ~30 µm µ-bump gap
- **Shorter heat path:** less material between base die and cold plate

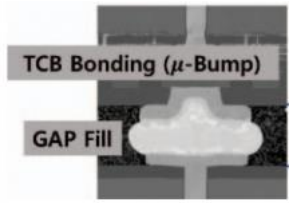
23–47% lower die-to-die thermal resistance than microbump HBM stacks

up to 3× higher vertical thermal conductivity through the bond layer

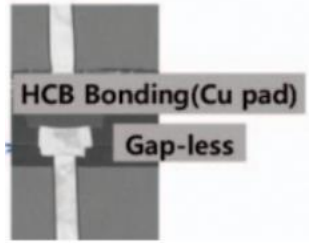
>15% shorter HBM stack height, so less heat builds up in the stack

HBM die-to-die joint, cross-section

Today: µ-bump + underfill (TC-NCF / MR-MUF)



Cu–Cu hybrid bond (3D)



Near-zero gap; heat flows through Cu and oxide

Caveat: 3D logic-on-memory raises power per footprint; the heat path still needs package co-design

Range depends on Cu pad density (≥20% coverage) and bond-interface quality

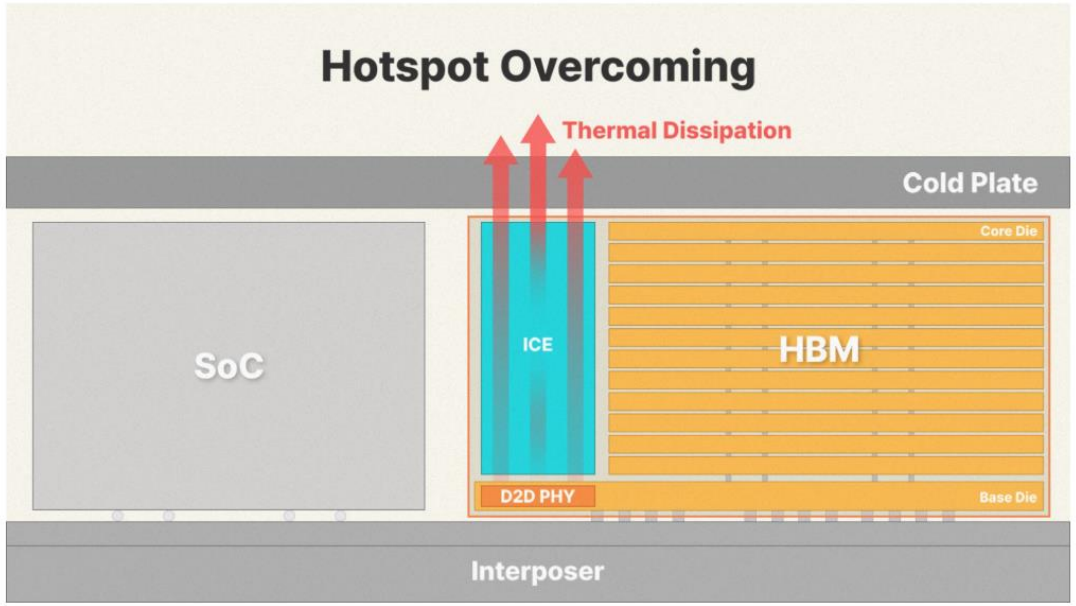
Images: Samsung (ECTC 2024) via The Elec, Jun 2024
Sources: Electronics 14(13):2682 (2025); Onto Innovation (2025)

© 2026 Marvell. All rights reserved.

HBM5 adds a silicon thermal path at the D2D PHY

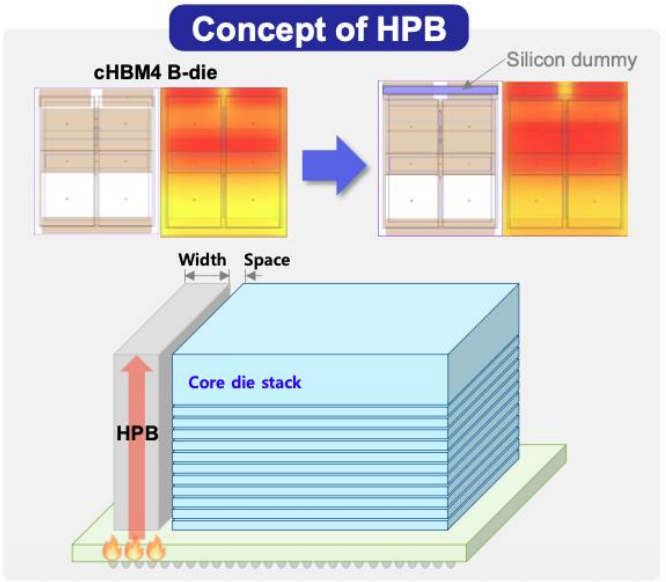
- HBM5 is projected at **~100 W per stack**; the D2D PHY in the base die is its hottest spot
- HBM suppliers add a **dedicated silicon heat path**.

SK hynix iHBM



Silicon ICE in the D2D PHY: **~30% lower thermal resistance**

HBM5 + Heat Path Block (HPB)



Silicon block **beside the core-die stack** pulls PHY heat to the cold plate

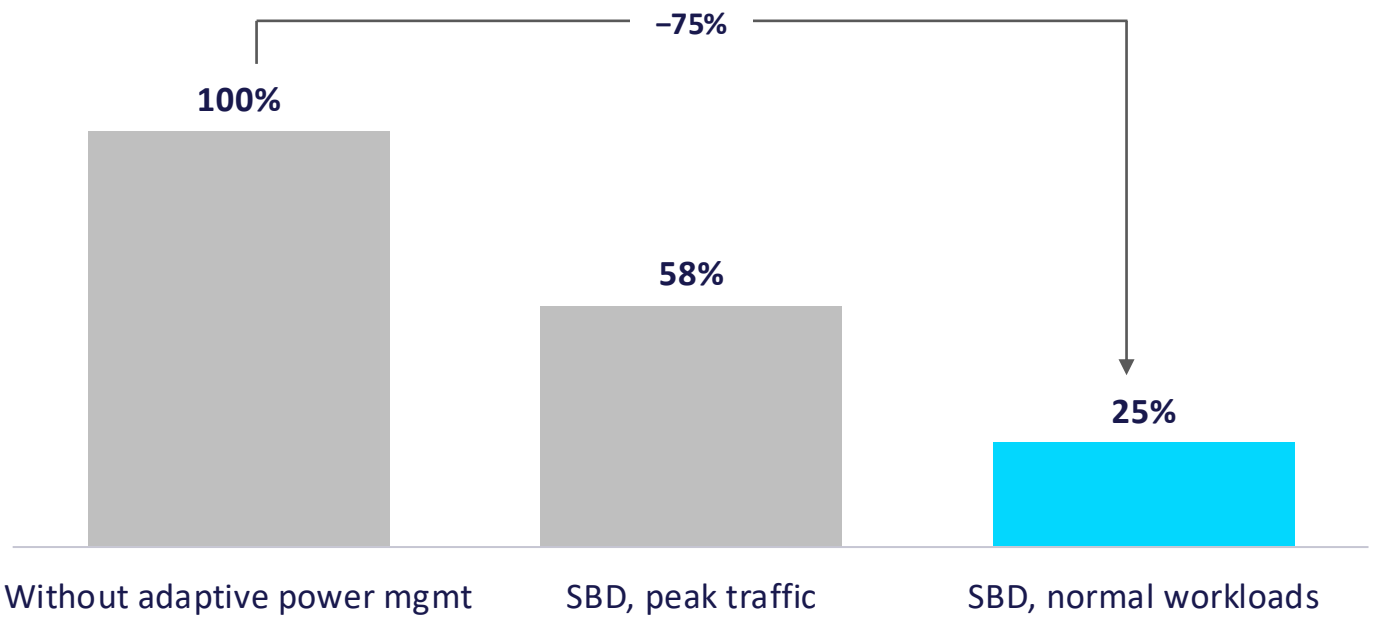
Images: SK hynix (left), industry mockup (right). Sources: SK hynix iHBM press release (May 26, 2026); HBM5 HPB mockup, Hot Chips 2026.

© 2026 Marvell. All rights reserved.

SBD cuts D2D power by up to 75%

For HBM, energy per bit sets the budget, not bandwidth density: every pJ/bit saved goes back to compute.

D2D power, % of an interface without adaptive power management



Density is already enough

HBM5 needs ~6.4 Tbps/mm (7.2 TB/s over 9 mm beachfront)

–42% even at peak traffic

Power falls at full HBM bandwidth, not just when idle

Power tracks traffic

Savings are largest under bursty AI-inference traffic

75% power reduction applies to normal workloads (42% at peak traffic). Source: Marvell, “Marvell Unveils Industry’s First 64 Gbps/wire Bi-Directional Die-to-Die Interface IP in 2nm,” press release, Aug 2025.

© 2026 Marvell. All rights reserved.

SBD D2D enables a custom, low-power base die

What simultaneous-bidirectional (SBD) D2D brings to a customized base die

High performance

64 Gbps/wire

Simultaneous-bidirectional signaling on every D2D wire
With bandwidth guard with 15% or higher.

High capacity

15% PHY area

Smaller D2D PHY frees base-die area for custom logic and memory

Lower power

Up to 75% less

D2D power in normal workloads; up to 42% less at peak traffic

High connectivity

> 7 TB/s

UCIe density for main-die-to-main-die and main-die-to-I/O-die links

Source: Marvell — 64 Gbps/wire Bi-Directional D2D Interface IP in 2nm (Aug 2025); Hot Chips 2025

© 2026 Marvell. All rights reserved.

Checklist make high-performance cHBM work

Base die and area

HBM4's 2,048-bit PHY takes ~30% of compute-die area

- ☐ Move PHY and controller into a custom base die
- ☐ Reclaim up to 25% of die area for compute
- ☐ Use the space for up to 30% more compute
- ☐ Speed up customized the base-die interface across memory vendors

D2D / TSV / HCB

UCIe was not built for HBM Next-gen class bandwidth density

- ☐ Cut D2D power
- ☐ Keep one D2D architecture from 2.5D to 3D hybrid bonding
- ☐ Prove 3D hybrid-bonding yield at volume

Package thermals

The HBM stack holds 57% of thermal resistance; HBM Next-gen reach > 2.0 W/mm².

- ☐ Reserve a silicon heat path at the D2D PHY
- ☐ Plan TIM, lid and cold plate with the base die
- ☐ Model heat flux above 2.0 W/mm² for HBM5
- ☐ Validate cooling beyond 100 W/stack and > 2.0 W/mm².

Ecosystem

No single company can build custom HBM alone

- ☐ Align memory makers, foundries, OSATs and XPU designers
- ☐ Share thermal models and test data across the chain
- ☐ Define cHBM with more re-usable modules between different AI applications.

☑ Addressed today ☐ Additional effort needed

Sources: Marvell newsroom (Dec 2024; Aug 2025); JEDEC JESD270-4 and HBM thermal-density reference; IEEE ECTC 2026; SK hynix iHBM release (May 2026); Hot Chips 2026. Details on earlier slides.

© 2026 Marvell. All rights reserved.

Checklist make high-performance cHBM work

Base die and area

HBM4's 2,048-bit PHY takes ~30% of compute-die area

- ☑ Move PHY and controller into a custom base die
- ☑ Reclaim up to 25% of die area for compute
- ☑ Use the space for up to 30% more compute
- ☐ Speed up customized the base-die interface across memory vendors

D2D / TSV / HCB

UCIe was not built for HBM Next-gen class bandwidth density

- ☑ Cut D2D power
- ☑ Keep one D2D architecture from 2.5D to 3D hybrid bonding
- ☐ Prove 3D hybrid-bonding yield at volume

Package thermals

The HBM stack holds 57% of thermal resistance; HBM Next-gen reach > 2.0 W/mm².

- ☑ Reserve a silicon heat path at the D2D PHY
- ☑ Plan TIM, lid and cold plate with the base die
- ☑ Model heat flux above 2.0 W/mm² for HBM5
- ☐ Validate cooling beyond 100 W/stack and > 2.0 W/mm².

Ecosystem

No single company can build custom HBM alone

- ☑ Align memory makers, foundries, OSATs and XPU designers
- ☑ Share thermal models and test data across the chain
- ☐ Define cHBM with more re-usable modules between different AI applications.

☑ Addressed today ☐ Additional effort needed

Sources: Marvell newsroom (Dec 2024; Aug 2025); JEDEC JESD270-4 and HBM thermal-density reference; IEEE ECTC 2026; SK hynix iHBM release (May 2026); Hot Chips 2026. Details on earlier slides.

© 2026 Marvell. All rights reserved.

Call to action

Custom HBM scales only as a co-designed:

The base-die interface

Close collaboration on custom base-die interface that works across memory vendors.

Converge on one low-power D2D

A D2D spec customized built for HBM, from 2.5D to 3D hybrid bonding.

Co-design package thermals

Memory makers, foundries, OSATs, and XPU designers plan TIM, lid, and cooling together

Publish models and data

Thermal models, stackup data, and interop results every vendor can build on



OCP GLOBAL SUMMIT | 2026

© 2026 Marvell. All rights reserved.