Techstrong | Semi

THE CENTER OF GRAVITY FOR THE SILICON ECONOMY

FROM TECHSTRONG GROUP, A FUTURUM COMPANY · TECHSTRONGSEMI.COM

SPECIAL REPORT
VOLUME 01 · 2026

THE SILICON ECONOMY · ADVANCED PACKAGING

Beyond the *Die*

The definitive guide to advanced packaging in the AI era — how chiplets, HBM, hybrid bonding, die-to-die interconnect and CoWoS rewrote the rules of the semiconductor industry.

BY

Alan Shimel

CEO & Founder, Techstrong Group

With **Brendan Burke**, Research Director, Futurum Research

RESEARCH IN PARTNERSHIP WITH

Techstrong a Futurum company

Futurum®

01 · EXECUTIVE SUMMARY

The Package Has Become the Computer

> For decades, progress in semiconductors was explained mainly through the chip itself. Shrink the transistors, fit more of them onto a die, and each generation becomes faster, cheaper or more efficient. That model still matters. It no longer explains how the most advanced AI systems improve.

Leading AI accelerators are systems assembled from logic dies, high-bandwidth memory, dense interconnect, substrates, power delivery and thermal engineering inside one enormous package. The consequential engineering has moved around the die as well as inside it. Packaging is no longer a protective enclosure or a final manufacturing step. It is part of the architecture, part of the yield model and an increasingly important part of the bill of materials.

Calling advanced packaging "the bottleneck" is directionally correct but incomplete. Accelerator supply is governed by a coupled chain. At different times the binding constraint may be silicon-interposer capacity, high-bandwidth memory, advanced substrates, hybrid-bonding yield, final test, power, cooling or network availability. Adding capacity at one point does not necessarily increase finished-system output if the rest of the chain cannot expand with it.

KEY TAKEAWAYS

- **Leading AI accelerators are systems, not chips.** Logic, HBM, interconnect, substrates, power and cooling must work as one architecture.
- **Packaging keeps Moore's Law moving through system-level scaling** when transistor density alone cannot deliver the required gains.
- **Yield is becoming an economic variable of the first order.** In a high-loss scenario, advanced packaging can cost more than memory and directly shape gross margin and price performance.
- **CoWoS- and HDFO-class capacity is a persistent bottleneck** that Futurum Research expects to extend through 2028.
- **The next technical frontier is not CPO alone.** Hybrid bonding and chip-to-chip interconnect determine how much of a system can be integrated, how efficiently its components communicate and whether modularity can become a commercial ecosystem.
- **The winners will coordinate the full stack:** logic, memory, interconnect, substrates, assembly, test, power, cooling and networking.

The numbers that frame the market

$1.21 T

PROJECTED 2030 DATA-CENTER SEMICONDUCTOR MARKET (BASE CASE)

Futurum Intelligence, 1H 2026 forecast

$66.9 B

COWOS AI WAFER MARKET BY 2030 (46% CAGR)

Futurum Intelligence, 1H 2026 forecast

Through '28

COWOS & HDFO-CLASS CAPACITY CONSTRAINT WINDOW

Futurum Research outlook

> *"Advanced packaging is becoming a standalone industry because the package, not the die, is now the unit of integration for AI accelerators, giving it distinct capacity constraints, capital cycles and pricing power. In high yield-loss scenarios, packaging can cost more than memory or logic, since one scrapped assembly destroys multiple known-good dies and the wafers behind them. That economic weight positions AI-accelerator packaging to reach $67 billion by 2030, a significant share of the $1.2 trillion data-center semiconductor market.*
>
> **BRENDAN BURKE** Research Director, Futurum Research

The two Futurum market metrics above frame the same story from different angles. The $66.9 billion CoWoS AI wafer figure measures the value of AI wafers flowing through TSMC's dominant packaging platform — not a forecast of TSMC packaging revenue. The $1.21 trillion 2030 data-center semiconductor forecast is the wider market that this concentrated flow increasingly anchors. Read together, they say a very large compute market is becoming dependent on a small set of packaging flows, while the value of the silicon committed to those flows rises sharply.

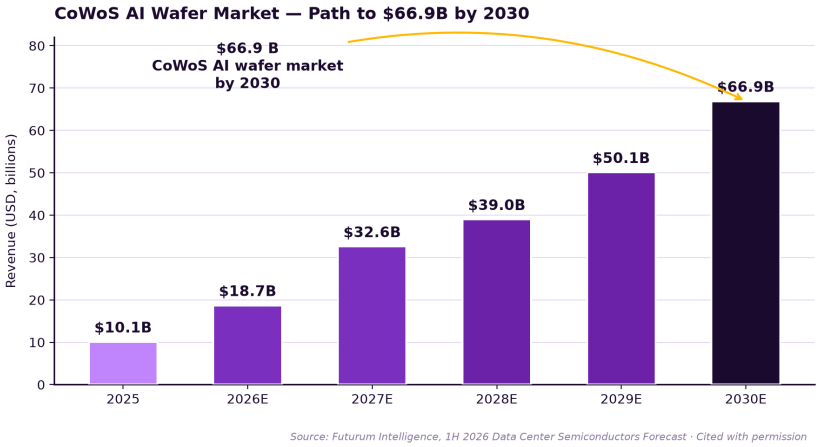


Figure 1 · Futurum Intelligence's CoWoS AI wafer market trajectory, from $10.1 billion in 2025 to $66.9 billion in 2030. Figures are proprietary Futurum forecast values cited with permission and are not reproduced from a proprietary Futurum table or chart. Visualization is by Techstrong Semi.[1]

02 · THE NEW ARCHITECTURE

Packaging Becomes the Architecture

Before going further, five basic terms establish the physical picture.

FOUNDATIONAL VOCABULARY

Die

The piece of silicon containing the transistors and circuits that perform a function. It is what most people mean when they casually say "the chip."

Package

The physical assembly that holds one or more dies, connects them to the rest of the computer, supplies power and helps remove heat.

Substrate

The layered platform beneath the dies that carries electrical connections between the package and the system board.

Interposer

A high-density connection layer placed between dies and the substrate, allowing components to communicate across far more connections than a conventional board can provide.

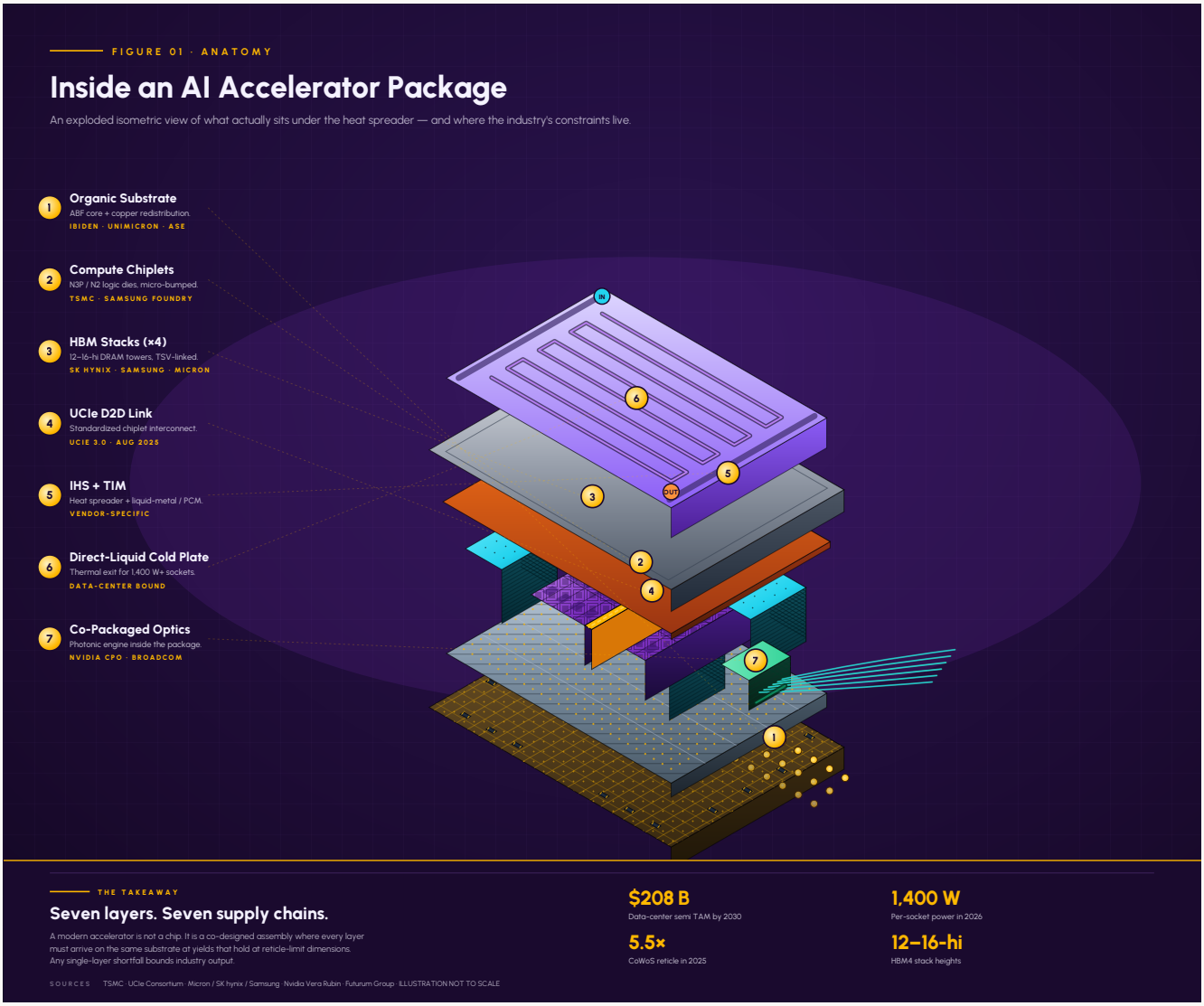
High-Bandwidth Memory (HBM)

Memory built from vertically stacked DRAM dies and placed close to a processor so data can move between memory and compute at extremely high rates.

For most of the industry's history, the package was described as the protective enclosure around the valuable part. The die performed the computation. The package connected it to a board, carried power and signals, removed heat and protected the silicon from physical damage. Packaging mattered, but it sat downstream from architecture.

An advanced AI accelerator overturns that model. The logic may be divided among several dies. Multiple HBM stacks sit beside it. An interposer, redistribution layer or set of local silicon bridges carries thousands of connections among those components. I/O dies handle external communication. The entire assembly is attached to an organic substrate that must remain flat and reliable despite increasing size, power density and thermal cycling. Co-packaged optics may join this assembly in selected designs, but it is an emerging option, not a universal package layer.

No single die contains the complete architecture. The architecture exists in the relationship among the logic, memory, I/O, interconnect and package.



Infographic 1 · *Vocabulary in context. Each numbered element in the cross-section corresponds to a coordinated technology described in the legend — the layers that together make a modern AI accelerator package function as a single system.*

Why one giant chip is no longer enough

1. **Yield economics.** Manufacturing yield generally becomes harder as a die grows. A defect that might ruin one small region can make an entire large die unusable. Dividing a design into smaller dies can improve wafer economics, provided assembly yield remains high.
2. **Right-sized process nodes.** Not every function benefits equally from the newest and most expensive process technology. Leading-edge transistors may be valuable for compute, while I/O, analog and control functions can often be built efficiently on older nodes.
3. **Memory bandwidth.** AI processors need access to far more memory bandwidth than conventional off-package channels can deliver within a practical power envelope.
4. **Geometry.** The reticle limits the practical area of one lithographic exposure. Frontier systems need more compute, cache, memory and I/O than one reticle-sized die can economically contain. Packaging lets the architecture expand beyond that boundary.

> “
>
> ***Modern AI accelerators face a geometric constraint that no process node advancement alone can solve. We expect XPU chip sizes to continue increasing by 2–3× per generation throughout the decade due to aggressive scaling of frontier model sizes and demand for inference performance gains.***
>
> **BRENDAN BURKE** Research Director, Futurum Research

Chiplets address yield, process-node and reuse problems by dividing a system into functional dies. HBM addresses the memory problem by bringing wide, stacked memory close to compute. Advanced packaging turns those components back into one functioning product. It does not remove complexity. It relocates complexity from the monolithic die into the interfaces, assembly flow, test strategy and operating software.

03 · THE TAXONOMY

A Practical Guide to 2.5D, 3D and Fan-Out

> Advanced packaging terminology can make different technologies sound like mutually exclusive choices. In practice, leading systems combine them. The useful distinction is what each method places next to or above another component, and what medium carries the connections.

2.5D integration

In 2.5D integration, logic dies and memory stacks sit side by side on a high-density interconnect layer. In the classic form, that layer is a silicon interposer. It provides much finer wiring than an organic substrate and can connect a large compute device to several HBM stacks through very wide interfaces. TSMC's CoWoS is the defining commercial platform. Intel's EMIB embeds smaller silicon bridges in the substrate only where dense connections are required.

3D integration

Three-dimensional integration stacks dies vertically. Connections may use through-silicon vias, microbumps or direct copper hybrid bonding. Short vertical links can provide very high bandwidth density with less energy per bit than longer package traces. HBM is itself a 3D structure: multiple DRAM dies are thinned, stacked and connected vertically. Logic-on-logic and cache-on-logic stacking push the concept further.

Fan-out

Fan-out packaging embeds dies in molded material and creates redistribution layers that spread their connections beyond the original die footprint. Some implementations eliminate a conventional laminate substrate, producing a thinner package with favorable electrical characteristics. Fan-out is established in mobile devices and is being extended to larger, higher-performance assemblies through wafer-level and panel-level manufacturing. Futurum uses **HDFO** (high-density fan-out) as a class-level reference for these approaches.

TABLE 1 · COMPARING INTEGRATION METHODS

METHOD	BASIC CONSTRUCTION	PRIMARY ADVANTAGE	PRIMARY CHALLENGE	BEST-ESTABLISHED USE
2.5D	Dies placed side by side on an interposer or dense bridge layer	High bandwidth between logic and HBM; heterogeneous process nodes	Large-package yield, cost, substrates and warpage	AI accelerators & HPC
3D	Dies stacked vertically with TSVs, microbumps or hybrid bonding	Highest interconnect density; short, energy-efficient links	Heat, compounded yield, test and repair	HBM; increasingly logic & cache stacking
Fan-Out / HDFO	Dies embedded in mold with redistribution layers	Thin form factor; potential cost advantages at scale	RDL yield, warpage, power & thermals at large sizes	Mobile & consumer; expanding toward HPC

The next generation will increasingly mix these constructions. A system can include vertically stacked HBM, logic chiplets connected through local bridges, redistribution layers spanning a large package and conventional substrate connections to the board. ***The taxonomy matters, but the integration flow matters more.***

04 · COWOS IN DEPTH

Inside CoWoS: S, R and L

CoWoS stands for Chip-on-Wafer-on-Substrate. In simplified terms, logic and memory dies are attached to an interconnect wafer or layer, and the completed structure is mounted on an organic substrate. TSMC offers three principal variants.

TABLE 2 · TSMC COWOS VARIANTS

VARIANT	INTERCONNECT STRUCTURE	CORE STRENGTH	LIKELY ROLE
CoWoS-S	Full silicon interposer	Maximum mature routing density for logic + HBM	Established high-performance accelerator packages
CoWoS-R	Redistribution-layer interposer	Greater design and cost flexibility	Large heterogeneous packages without full silicon density
CoWoS-L	RDL interposer + local silicon bridges	Extends package size while concentrating dense links where needed	Reticle-busting multi-die AI systems

A **reticle** is the area patterned in one exposure by a lithography machine. Traditionally, it helped define how large a single leading-edge die could be. When a package is described as several times reticle size, the assembly has grown far beyond the practical footprint of one monolithic die. Reticle-busting is more than industry slang — it marks the transition from one large processor to a system assembled across a much larger surface.

> “
>
> ***When the reticle stops being a constraint, packaging becomes the architecture of intelligence itself.***
>
> **BRENDAN BURKE** Research Director, Futurum Research

CoWoS-L combines a redistribution-layer interposer with embedded local silicon bridges. TSMC says the platform entered volume production in 2024. It certified a **5.5-reticle-size solution in 2025** and scheduled volume production for 2026. The architecture illustrates why packaging is becoming the scaling vehicle: fine wiring is placed only where the design needs it, while the RDL structure carries the assembly across an area that a full silicon interposer would make more difficult and expensive.

Reading capacity numbers with care

The most important capacity number is ***effective finished output***. TSMC does not disclose customer allocations or a clean monthly series. Industry estimates place potential 2026 capacity around 120,000 to 140,000 wafer starts per month, with additional downstream OSAT capacity. Brokerage estimates suggest NVIDIA represents roughly 60 percent of demand. Neither figure is a TSMC disclosure, and neither tells the reader how many good accelerator packages leave the line.

Installed equipment is not qualified output. Qualified output is not good packages shipped. Product mix changes the translation from wafers to accelerators because larger products consume more area. Futurum's AI Chipsets Chip Library estimates the Rubin-generation CoWoS package at about 3,735 square millimeters, compared with roughly 2,905 square millimeters for Blackwell — an increase of approximately **29 percent**. The implication matters even if roadmap dimensions move: nominal wafer capacity can rise while finished accelerator output grows more slowly.

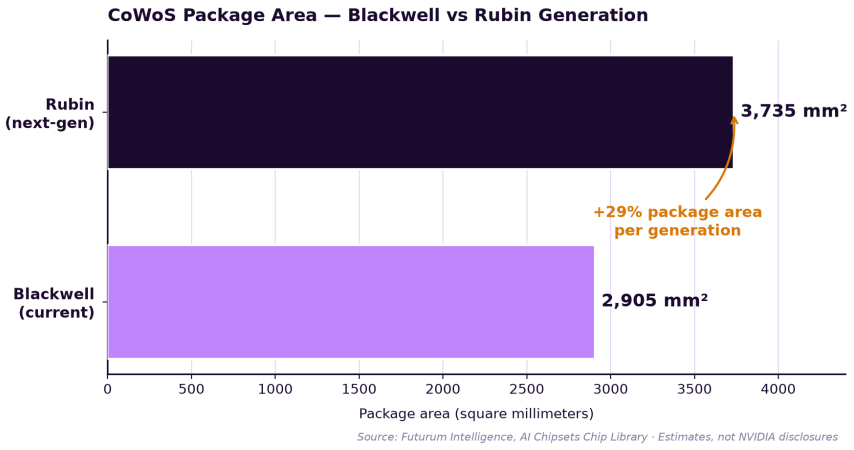


Figure 2 · Rubin-generation CoWoS packages are estimated at roughly 3,735 mm², a **29% increase** over Blackwell's 2,905 mm². Larger packages consume more interposer, redistribution and substrate area per device — so nominal wafer capacity can rise while finished accelerator output grows more slowly. Package specifications are Futurum estimates, not NVIDIA disclosures.[2]

Futurum estimates CoWoS AI wafer demand will grow from $10 billion in 2025 to $67 billion in 2030 — a **46 percent compound annual growth rate**. The concentration is the story. A single TSMC packaging platform tied to AI accelerators is on a path to approach the size of the entire advanced-packaging market as measured in 2024. This is proprietary Futurum forecast data cited as narrative metrics with permission; it is not reproduced as a proprietary chart or table.

HBM Becomes a Co-Designed Component

> An AI accelerator is valuable only when it can keep its compute engines supplied with data. Conventional off-package memory channels cannot provide enough bandwidth within a reasonable power envelope. HBM addresses that problem by stacking DRAM dies and connecting each stack to the accelerator through an exceptionally wide interface.

The simplest analogy is a roadway. Compute engines are destinations, data is traffic and memory bandwidth is the number and speed of available lanes. Adding more compute without widening the route leaves expensive processing units waiting for data.

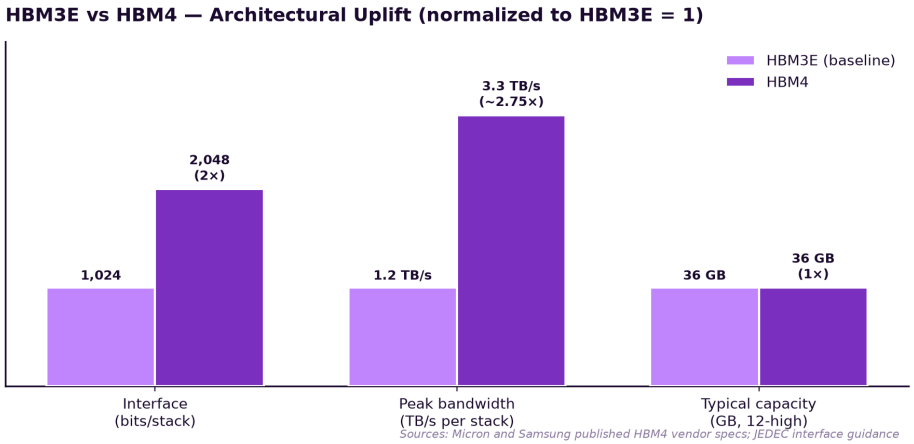


Figure 3 · HBM4 is not simply the next memory product with a higher speed printed on its spec sheet. Its physical interface **doubles** in width, its package routing becomes more demanding, and its logic base can be tailored to the processor it serves.

TABLE 3 · HBM3E VS HBM4 — ARCHITECTURAL CONSEQUENCES

CHARACTERISTIC	HBM3E	HBM4	ARCHITECTURAL CONSEQUENCE
Interface width	1,024 bits	2,048 bits	More routing density; deeper package co-design
Stack height	8-high or 12-high	12-high, 16-high in dev	More capacity in the same footprint; harder yield and thermals
Capacity examples	24 GB (8-high) / 36 GB (12-high)	36 GB at 12-high	System capacity grows via more stacks and taller ones
Peak bandwidth	~1.2 TB/s per stack	2.8–3.3 TB/s per stack	Memory increasingly sets useful accelerator throughput
Base-die strategy	Memory-interface functions	Advanced logic base + customization	Tighter coupling among memory, foundry and designer

HBM4 is qualified. The watch item is no longer qualification in the abstract. It is the scale and portability of qualified implementations. The base die is the architectural pivot. A custom advanced-node logic base can improve efficiency, signal behavior and memory management, but it can also make second-sourcing harder. The JEDEC interface may be standardized while the implementation, base logic, package routing and customer qualification remain platform-specific.

The three suppliers, three positions

SK hynix

VOLUME 12-HIGH HBM3E IN 2024;
16-LAYER HBM4 SHOWN 2026

MR-MUF packaging differentiator

Micron

VOLUME 36GB 12-HIGH HBM4 FOR NVIDIA VERA RUBIN, Q1 2026

>$1B HBM4 revenue by fiscal Q3

Samsung

HBM4 MASS PRODUCTION ANNOUNCED FEBRUARY 2026

1c DRAM + 4nm logic base die

SK hynix, Micron and Samsung each enter this phase from a different position. Market-share snapshots move rapidly because revenue, bits shipped, stacks, contracted capacity and qualified platforms measure different things. The strategic question is not which supplier wins one quarter. It is which suppliers can qualify custom HBM4 at high yields, across enough platforms, without letting the customization that improves performance destroy supply flexibility.

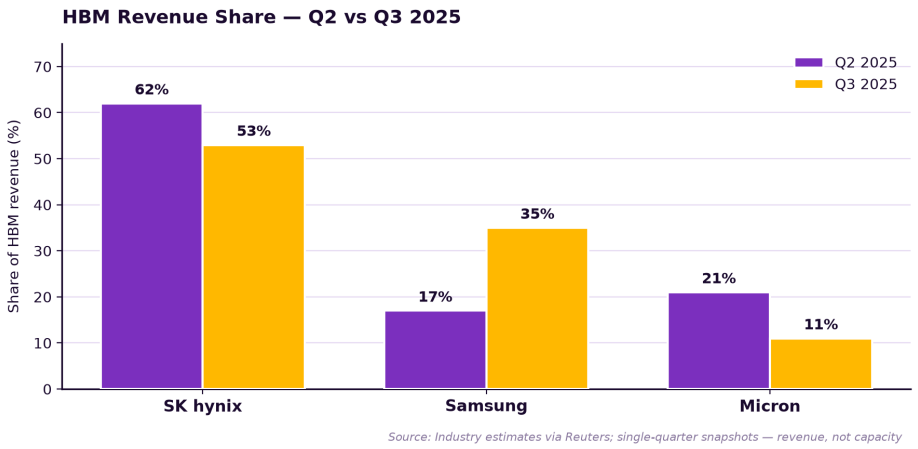


Figure 4 · Market-share figures should be read as dated snapshots. A move this large in one quarter illustrates the danger of calling any one figure permanent "capacity share."

06 · THE BONDING FRONTIER

Hybrid Bonding Turns Assembly Into a Front-End Discipline

> Hybrid bonding deserves more attention than CPO in an advanced-packaging report because it sits closer to the central scaling problem. Conventional microbumps use tiny solder connections to join dies. They are proven and repair the geometric break between separate pieces of silicon, but their pitch, electrical resistance and parasitic effects limit how densely and efficiently dies can communicate.

Copper-to-copper hybrid bonding removes the solder bump. Precisely planarized copper pads and surrounding dielectric surfaces are aligned and bonded directly, combining mechanical attachment and electrical connection in one interface. TSMC describes SoIC as a sub-10-micrometer-pitch platform and says its 3-nanometer stacking technology entered volume production in 2025. Intel says Foveros Direct achieves sub-10-micrometer pitch, with a roadmap to below 5 micrometers. Imec research has demonstrated die-to-wafer bonding at 2 micrometers and wafer-to-wafer work at still finer pitch. These are not equivalent maturity claims, but together they show where interconnect density is headed.

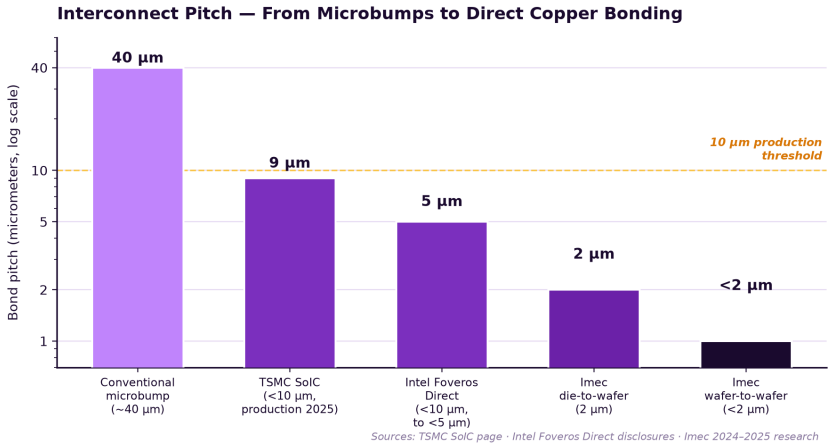


Figure 5 · From conventional microbumps at ~40 µm to hybrid-bonding platforms in volume production below 10 µm, with research demonstrations at 2 µm and finer. Log scale. The 10 µm line marks the current production threshold for direct copper bonding.

Why it matters

- **Bandwidth density rises** because far more connections fit in the same area.
- **Energy per bit can fall** because the connection is short and low resistance.
- **Logic, cache and memory can be partitioned** with finer granularity than a microbump interface permits.
- **The electrical boundary between dies** starts to look less like a package trace and more like on-chip wiring.

The advantages carry manufacturing consequences. Surfaces must be exceptionally flat and clean. Alignment tolerances tighten. A particle, void or overlay error can destroy a dense field of valuable connections. Die-to-wafer flows preserve some known-good-die flexibility but require precise placement of many individual dies. Wafer-to-wafer bonding can deliver throughput and alignment advantages when the top and bottom wafers match, but it can compound yield if one wafer contains defective die locations. Repair after bonding is limited.

> *“*
>
> ***Hybrid bonding is not merely an interconnect improvement. It moves packaging toward front-end process discipline. Cleaning, chemical-mechanical planarization, metrology, alignment and defect control become as decisive as the bonding tool itself.***

The commercial question is not whether direct copper bonding works. It is whether manufacturers can sustain its promised density at production yields across larger, hotter and more heterogeneous products.

07 · THE MODULAR IDEAL

Chiplets and Die-to-Die Interconnect

> Chiplets are already a commercial success. AMD's CPU and accelerator families use multi-die construction extensively. Intel has shipped products using advanced die integration for years. AI accelerators increasingly separate compute, I/O, cache and specialized functions into dies optimized for different purposes and process nodes.

The appeal begins with yield and reuse. A company can replace one enormous monolithic die with several smaller dies, use an advanced process only where it creates value and carry proven I/O tiles across products. The complexity moves into the package and system design. Die-to-die links consume power and add latency. The package must distribute power and clocks, preserve signal integrity and remove heat from components with different thermal profiles. Every chiplet must be known good before assembly. Firmware must initialize and manage the system. Security boundaries and failure diagnosis become more difficult when functions span several dies.

The physical and logical link are different problems

Hybrid bonding, microbumps and silicon bridges describe *how* dies are physically connected. UCIe describes a standardized logical and electrical interface that can run across supported package channels. Conflating the two hides the real work. **A dense physical interface does not create interoperability by itself, and a protocol standard cannot rescue a package whose signal integrity, clocking, power delivery or thermal behavior is inadequate.**

The UCIe standard

UCIe addresses the die-to-die connection. Version 2.0 added manageability, testing and support for 3D packaging. **UCIe 3.0, released in August 2025**, supports 48 and 64 gigatransfers per second, extended sideband reach and runtime recalibration. Those features matter because chiplet systems must be tested from wafer sort through field operation, not merely shown to exchange bits in a lab.

FOUR LEVELS OF CHIPLET MARKET MATURITY

1. **Proprietary multi-die products shipping at scale.** Industry status: mature.
2. **UCIe interface IP, packaging services and test tools commercially available.** Industry status: mature.
3. **Multi-company UCIe interoperability demonstrated in silicon.** Industry status: demonstrated.
4. **End products shipping with independently sourced chiplets connected through UCIe.** Industry status: *remains the exception.*

The fourth level is the test of a genuine merchant chiplet market. A system company must be able to source dies from different suppliers, know their thermal and electrical behavior, authenticate them, test them, integrate their firmware, assign liability and warrant the finished product. UCIe is necessary, but the standard is not the commercial structure. ***Test, telemetry, security, lifecycle support and business accountability will determine whether the modular ideal becomes an ecosystem rather than a collection of proprietary multi-die designs.***

08 · THE OPTICAL BOUNDARY

Co-Packaged Optics, Important but Selective

As AI clusters scale, data movement becomes a power and reach problem. Conventional pluggable optical modules sit at a switch faceplate. High-speed electrical signals travel from the switch ASIC across the board before conversion to light. At higher lane rates, those traces consume more power and become harder to engineer.

Co-packaged optics moves optical engines beside the switch or compute ASIC. The electrical path becomes shorter, and conversion to light happens inside or immediately adjacent to the package. The potential benefits are higher bandwidth density, lower electrical I/O power and longer optical reach. Broadcom is pursuing merchant switch silicon with partner systems, NVIDIA is integrating photonics into its networking platforms, and Marvell is developing optical chiplets for custom XPU and scale-up designs.

The tradeoff is serviceability and yield. A pluggable module can be replaced at the front of a switch. An optical engine beside a costly, hot ASIC is harder to service. Fiber attach, laser placement, contamination, thermal isolation and assembly yield become production issues. A failed optical component can threaten a far more expensive unit.

CPO therefore belongs in the report, but not as a universal seventh package layer. It will appear first where bandwidth density and power justify tighter integration. Linear pluggable optics and conventional pluggables will remain where serviceability, supply flexibility and field replacement carry more weight. Hybrid bonding and die-to-die interconnect have a broader claim on the architecture of nearly every advanced multi-die accelerator; CPO is a more selective frontier.

09 · THE BACK END RISES

The OSAT Landscape and the Redistribution of Value

Advanced packaging is erasing the boundary between front-end manufacturing and the back end. Foundries perform packaging steps. Memory suppliers differentiate through stacking and logic base dies. Equipment and materials vendors determine bonding, redistribution and substrate capability. OSAT companies are taking on more technically demanding work and larger capital programs.

TABLE 4 · THE OSAT LANDSCAPE

COMPANY	DISTINCTIVE POSITION	RELEVANT CAPABILITIES	STRATEGIC DEVELOPMENT
ASE / SPIL	Largest OSAT; scale in outsourced AI packaging and test	VIPack, FOCoS, 2.5D/3D, flip-chip, test	Forecast **$3.2B** leading-edge packaging & test revenue in 2026 (2× 2025); raised 2026 CapEx to ~$10.5B
Amkor	Global OSAT with major U.S. capacity build	Flip-chip, wafer-level, system-in-package, advanced test	Arizona campus first fab complete mid-2027; production early 2028; **$1.5B NVIDIA agreement** (July 2026); AMD advanced-packaging confirmed
JCET	China's leading global OSAT	XDFOI heterogeneous integration; broad assembly/test	RMB 38.87B 2025 revenue; ~$3.78B advanced-packaging revenue; strategic role amid export controls
PTI	Memory packaging/test specialist expanding into panels	Probe, packaging, final test, burn-in, modules, FOPLP	Panel-level fan-out capacity and qualification program; 67% of 2025 revenue from packaging

These companies are not interchangeable capacity. A foundry may create the interposer and perform die bonding while an OSAT completes assembly and test. A memory vendor may stack HBM but rely on another company for final integration. *Reported capacity can double-count different stages of one package.*

> *"Test is undervalued in public discussion. As component cost rises, known-good-die testing becomes more important. Because repairability is limited, the cost of a final-package defect rises with every chiplet and HBM stack. Hybrid bonding intensifies the point: the denser and less repairable the interface becomes, the more value must be protected before final assembly.*

10 · THE COUPLED CHAIN

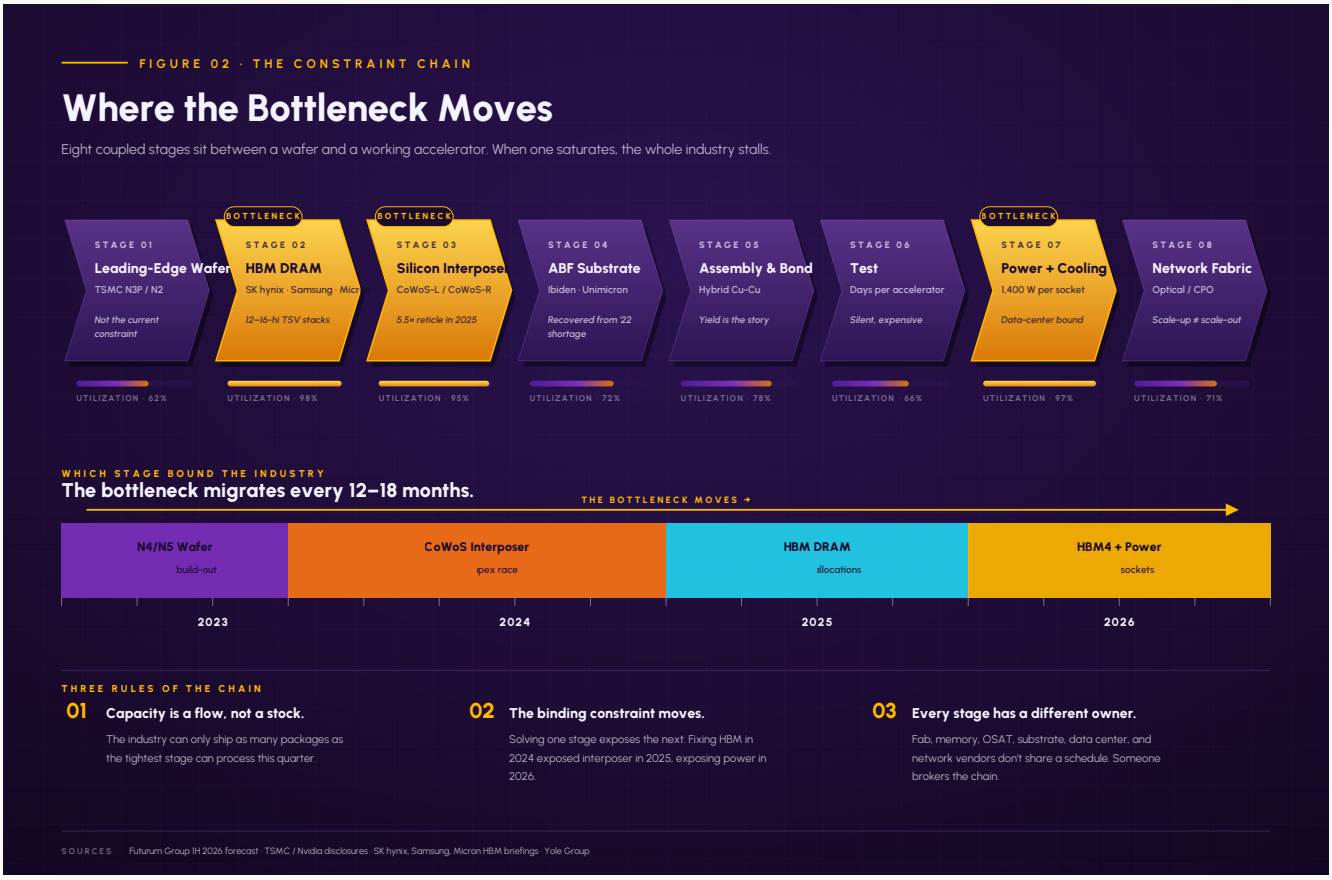
The Bottleneck Is a Moving Chain

The semiconductor industry likes singular bottlenecks because they imply singular solutions. Build more CoWoS capacity. Add another HBM supplier. Qualify a second substrate vendor. Each action is necessary; none operates in isolation.

Before an AI accelerator reaches a customer, logic wafers must yield acceptable dies; HBM stacks must be built, tested and qualified; interposers or redistribution structures must be manufactured; substrates must meet increasingly difficult size and warpage requirements; dies and memory must be bonded without sacrificing the value already embedded in them; and the completed package must pass test, receive power and cooling, connect to a network and enter a data center with electrical capacity.

Expansion at one stage moves the constraint to the next. More interposer starts do not create more accelerators if HBM4 supply is limited. More HBM does not help if substrate or hybrid-bonding yield falls as package size grows. More completed accelerators do not create deployed clusters when power interconnection or networking is late.

> “
>
> ***Futurum Research identifies advanced packaging capacity, specifically for CoWoS and HDFO-class technologies, as a persistent bottleneck that could extend through 2028.***
>
> **BRENDAN BURKE** Research Director, Futurum Research



Infographic 2 · *Eight coupled stages sit between a wafer start and a deployed AI cluster. The color band below shows how the binding constraint has migrated between 2023 and 2026 — from logic wafers to HBM, to CoWoS interposers, to substrates, to test and power, and toward network and CPO. Expansion at any single stage does not translate to more finished systems unless the entire chain expands with it.*

Allocation figures are unstable because effective output depends on product mix and package area. One generation may use a full silicon interposer; the next may shift to a larger CoWoS-L assembly. The same wafer starts can support a different number of packages. *Capacity is a flow, not a stock.*

Packaging yield is the economic hinge. A package can contain several leading-edge logic dies and multiple valuable HBM stacks before final test. A defect at that point does not waste one cheap component; it can strand an assembly whose input value is already enormous. **The cost of failure rises geometrically with system complexity**, and that is why wafer efficiency, known-good-die coverage, bonding yield and final test affect gross margin as directly as wafer pricing or memory contracts.

11 · THE WATCH LIST

What to Watch Through 2028

The next two years will show whether capacity is genuinely catching demand or whether expansion is being absorbed by larger and more complex products. The most useful signals are operational, not promotional.

01

Effective CoWoS output, not nominal wafer capacity

Track qualified packages shipped by type, package area and yield. CoWoS-L adoption and products beyond five reticles make simple wafer-start comparisons less meaningful.

02

Hybrid-bonding yield at production pitch

Watch die-to-wafer alignment, void and defect rates, known-good-die coverage and sustained yield below 10 micrometers. Demonstrations at 2 or 3 micrometers are important, but repeatable production economics are the milestone.

03

HBM4 shipment scale and second-source portability

HBM4 is qualified. The question is how quickly qualified custom-base-die implementations ramp, and whether a customer can move between suppliers without redesigning the package.

04

Named production products using independently sourced UCIe dies

A merchant chiplet ecosystem begins when a company can name dies from different suppliers, connect them through the standard, manage and secure them, and warrant the assembled system.

05

Package cost versus memory cost

Track whether packaging yield loss makes the package the leading cost driver. This is the point at which advanced packaging moves from an engineering constraint into a gross-margin constraint.

06 **HDFO and panel-level fan-out qualification**

If manufacturers can produce large AI packages on panels or high-density fan-out flows with acceptable routing density, warpage, yield and reliability, the cost structure could change.

07 **CPO field reliability**

Measure service procedures, laser lifetime, fiber attach and failure isolation, not aggregate bandwidth alone. CPO remains important, but it is one frontier among several.

08 **A complete U.S. advanced-packaging chain**

A domestic assembly building is not a domestic ecosystem. Substrates, materials, equipment, memory, metrology, test and specialized skills remain globally distributed.

12 · CONCLUSION

Beyond the Die

> Moore's Law is not over, and advanced nodes remain essential. But AI requires more compute, memory bandwidth, die-to-die communication and network capacity than one monolithic processor can deliver economically. Packaging keeps Moore's Law moving by changing the unit of scaling from the die to the system.

Advanced packaging allows designers to combine logic built on different processes, surround it with vertically stacked memory, connect chiplets through dense local links and, where the economics justify it, move optical I/O closer to the silicon. It also concentrates risk. Larger packages, taller memory stacks, direct copper bonding, customized base dies and tighter integration make yield, test, qualification and supply coordination harder.

The winners will not simply have the smallest transistors. They will coordinate logic, memory, interconnect, substrates, assembly, test, power, cooling and networking. **TSMC's CoWoS position** demonstrates the value of an integration platform. **HBM suppliers** show memory becoming part of accelerator strategy. **Hybrid bonding** turns assembly into a front-end discipline. **UCIe** shows the possibility, and the limits, of a standardized chiplet market. **OSATs** show value moving toward a back end that is no longer meaningfully back.

> *Advanced packaging is sometimes described as the bottleneck for AI. The larger conclusion is more accurate. It is where nearly all of AI computing's constraints meet, and where a growing share of the system's value can be lost or preserved. The package is no longer what surrounds the computer. Increasingly, the package is the computer.*

CONTINUE THE RESEARCH

Get the Latest Futurum Intelligence

The data-center semiconductor forecast, decision-maker survey, and vendor Signal scores cited throughout this report are produced by the analyst team at **Futurum Group**. Subscribers to the Futurum Intelligence Platform get continuous access to 50,000+ proprietary research items, five-year forecasts, and decision-maker survey data across 11 practice areas — including semiconductors, AI infrastructure, cloud, security, and DevOps.

Subscribe to Futurum Intelligence →

Or get the free newsletter

Source Notes & Research Citations

1. Futurum Group, *1H 2026 Data Center Semiconductors Market Sizing & Five-Year Forecast*, July 2026, Doc. #1H26MSDCSBB202607. Proprietary Futurum Intelligence subscriber research; selected metrics cited with permission.
2. Futurum Group, *AI Chipsets Chip Library*, accessed August 2026. Proprietary Futurum Intelligence subscriber research; package specifications are Futurum estimates, not NVIDIA disclosures.
3. Futurum Group, *1H 2026 Data Center Semiconductor Decision Maker Survey*, May 2026 (n=824). Public high-level findings; detailed cluster-expansion metrics are proprietary subscriber research cited with permission.
4. Futurum Group, *State of the Market: Semiconductors, Q2 2026 — The Age of Agentic Silicon Begins*, April 2026, Doc. #SOMSEMIBBQ2202604; and 1H 2026 Forecast, July 2026. Proprietary subscriber research.
5. Futurum Group, *1H 2026 Data Center Semiconductors Market Sizing & Five-Year Forecast*. Growth figures depend on Futurum's XPU/custom-accelerator market definition.
6. Yole Group, *Advanced Packaging Market Forecast 2024–2030*. See yolegroup.com.
7. TSMC, CoWoS technology page; TSMC 2025 5.5-reticle CoWoS certification disclosures.

8. Intel, EMIB and Foveros technology disclosures.
9. UCIe Consortium, UCIe 3.0 specification, August 2025.
10. Micron Technology, HBM4 volume shipment announcements for NVIDIA Vera Rubin, Q1 2026.
11. Samsung Semiconductor, HBM4 mass-production announcement, February 2026. Samsung HBM product page.
12. SK hynix HBM3E and HBM4 development milestones, 2024–2026.
13. Broadcom, *Tomahawk 6-Davisson* introduction, October 2025.
14. NVIDIA, *Spectrum-X Photonics* and *Quantum-X InfiniBand* platform disclosures.
15. ASE Technology Holding, 2026 leading-edge packaging revenue and CapEx guidance (July 2026).
16. Amkor Technology and NVIDIA joint statement on Arizona packaging investment, July 2026.
17. JCET Group, 2025 annual results and XDFOI platform disclosures.
18. Powertech Technology (PTI), 2025 revenue disclosures and panel-level fan-out program.
19. Reuters coverage of HBM revenue shares, Q2–Q3 2025.

ABOUT THIS REPORT

Beyond the Die: The Definitive Guide to Advanced Packaging in the AI Era is a Techstrong Semi Special Report by Alan Shimel, CEO and founder of Techstrong Group. It was produced in collaboration with the analyst team at Futurum Group, whose Data Center Semiconductor forecast and decision-maker survey provide the analytical backbone.

> **Contributing Analyst. Brendan Burke**, Research Director at Futurum Research, contributed the market sizing and forecast analysis that underpins the $66.9 billion CoWoS AI wafer-demand outlook, the Rubin-vs-Blackwell package-area estimates, and the 2028 capacity-constraint window discussed throughout this report. Four of his direct quotations appear in Sections 01, 02, 04 and 10 with full attribution. His analytical framing of the coupled-chain constraint model and the yield-driven gross-margin thesis shaped this report's structure and conclusions. Techstrong Semi thanks Brendan for his contributions.

Proprietary Futurum Intelligence metrics appear only as cited narrative figures; no proprietary tables or charts have been reproduced. Public estimates from Yole Group, TSMC, Intel, the UCIe Consortium, and vendor disclosures are cited where used.

Techstrong Semi at techstrongsemi.com is from Techstrong Group, A Futurum Company. It is *the center of gravity for the silicon economy*.

Stay ahead of the research. Subscribe to the Futurum Intelligence Platform for continuous access to the semiconductor forecasts, decision-maker surveys, and Signal scores that inform reports like this one — or start with the free Futurum newsletter.

Techstrong Semi · Techstrong a Futurum company · Futurum

Learn more: techstrongsemi.com · techstronggroup.com · futurumgroup.com

© 2026 Techstrong Group. All rights reserved. Special Report Volume 01.