Futurum®

intel foundry

Packaging the AI Frontier:

Intel Foundry's Advanced Packaging Alignment with the XPU Industry Roadmap

APRIL 2026

Packaging the AI Frontier: Intel Foundry's Advanced Packaging Alignment with the XPU Industry Roadmap

Executive Summary

AI infrastructure scaling is increasingly limited by packaging architecture. As monolithic silicon approaches reticle limits and power-density ceilings, the ability to disaggregate compute, memory, and I/O into chiplet-based systems connected via advanced packaging determines which AI accelerators can meet hyperscale performance requirements.

Intel Foundry's advanced packaging portfolio aligns with the roadmap for next-generation AI systems, reaching 12x the reticle size by 2028. With most XPU chip sizes projected to multiply 2–3x in the next generation and continue growing annually through the end of the decade, proactively supporting superchip architectures will differentiate foundries.

High-yield cost-efficient packaging can achieve up to 21.6% cost savings in AI accelerator manufacturing by avoiding packaging yield loss – a highly variable component of manufacturing costs (see Figure 1). Advanced packaging can surpass memory as the leading cost driver for a leading-edge accelerator, making gross margins and price performance dependent on wafer efficiency.

Figure 1. Cost Savings from High-Yield Packaging in AI Manufacturing

Up to
21.6%
Cost Savings

Optimized Chip Design
Reducing Die Interconnect Complexity

Improved Yield
Enhancing Manufacturing Efficiency to Minimize Defects

Streamlined Manufacturing
Applying Advanced Process Control

Logistics & Supply Chain
Effective Cost and Resource Management

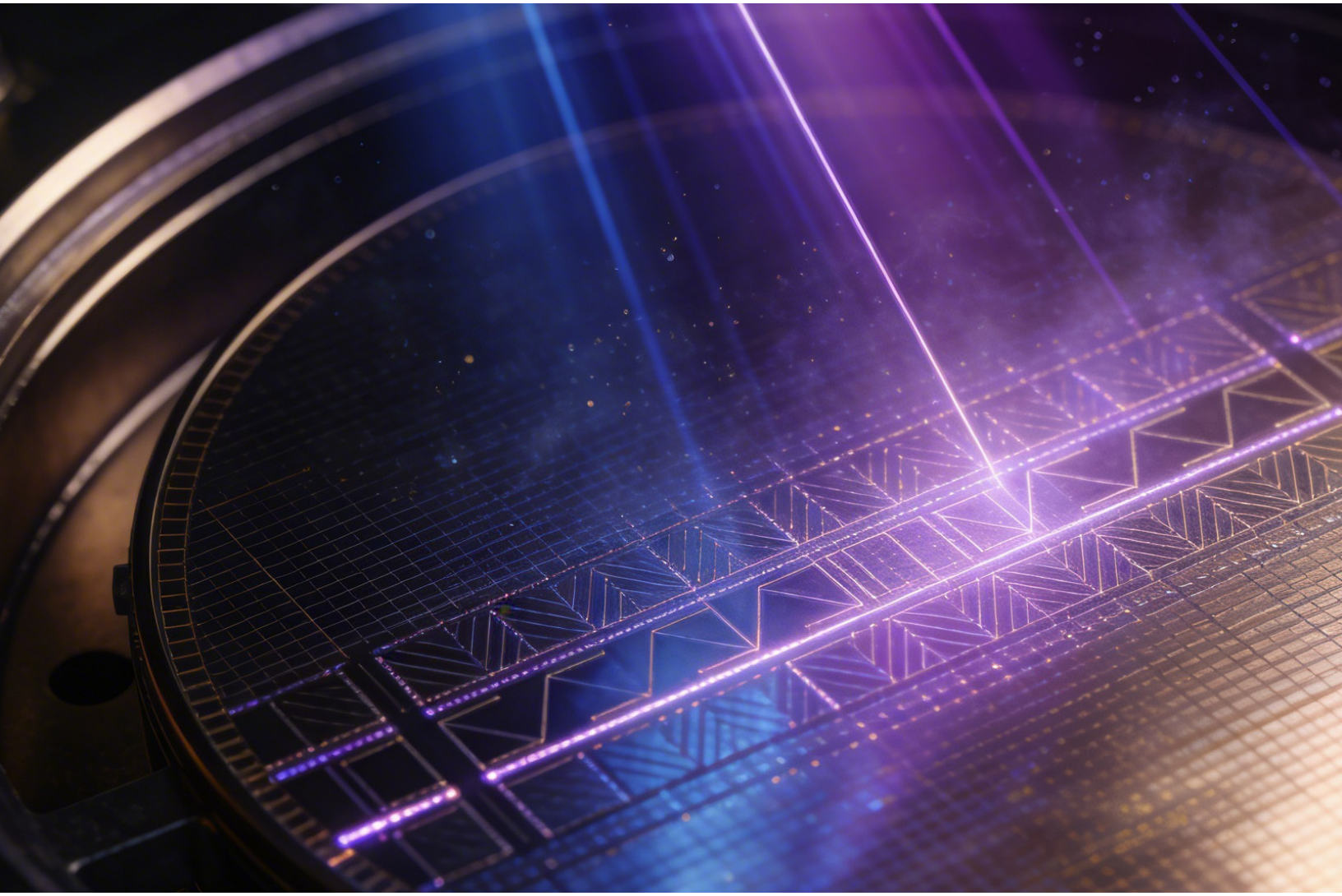
Source: Intel Foundry

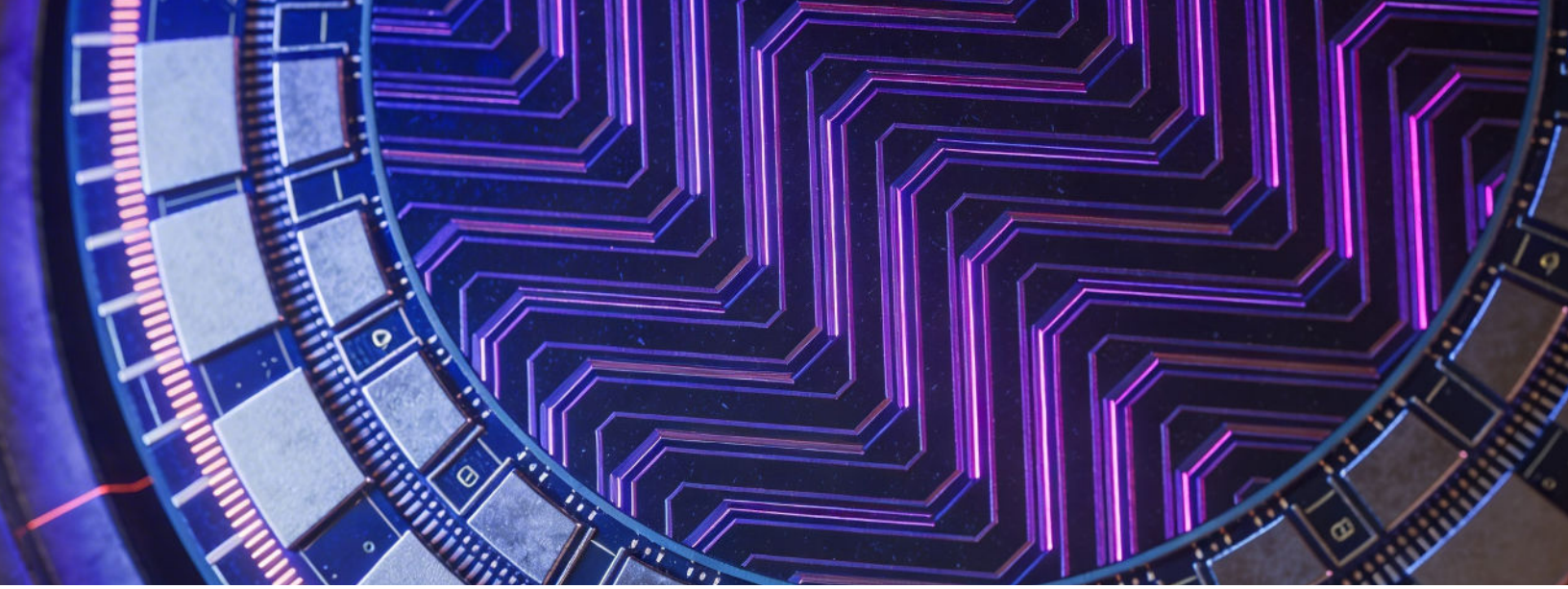
Intel's pivot to a Systems Foundry model with unbundled services for fabless customers, including advanced packaging and heterogeneous integration, can transform the company from a vertically integrated chipmaker into a catalyst for the broader AI accelerator ecosystem.

The Evolution of the AI XPU from Monolithic to Modular

Modern AI accelerators face a geometric constraint that no process node advancement alone can solve. The compute and memory complexes required for frontier model training and large-scale inference exceed the maximum size of a single lithography reticle (approximately 858mm²). NVIDIA's Vera Rubin platform uses a 2x GPU reticle design integrating two massive GPU dies with supplemental dies including HBM4 memory that stretch the overall die complex to approximately 5.3x reticles, and the subsequent Rubin Ultra is likely to push toward 11x reticles, requiring multiple interposers. AMD's MI400 series will rely on multi-die packaging to combine CPU and GPU chiplets. Beyond these platforms, we expect XPU chip sizes to continue increasing by 2–3x per generation throughout the decade due to aggressive scaling of frontier model sizes and demand for inference performance gains.

AI workloads demand heterogeneous compute that can benefit from wafer-efficient packaging, including matrix math engines, vector processors, SRAM, and high-bandwidth memory controllers. Most of these use different process nodes and all share a need for die-to-die interconnect with unique power delivery profiles. Manufacturing an entire monolithic accelerator on a leading-edge node can waste expensive transistor density on components that perform identically on mature nodes. Disaggregated chiplet architectures solve this by manufacturing each functional block on its optimal process node, and then integrating them through 3D stacking to make a disaggregated system perform as a unified accelerator.





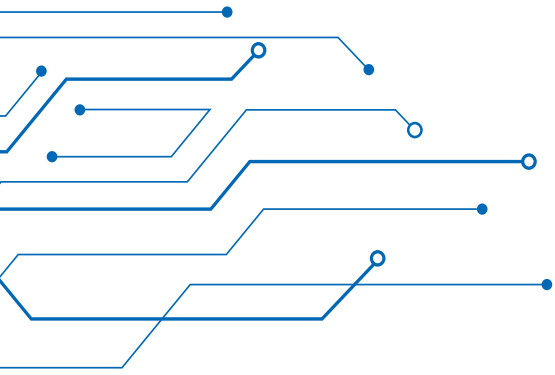
Winning the Race to Superchip Packaging

Intel anticipated the future need for die disaggregation in 2016 with the first paper on EMIB, "Embedded Multi-Die Interconnect Bridge (EMIB) – A High Density High Bandwidth Packaging Interconnect," published in IEEE. Since then, the company has gained multi-generational manufacturing experience with EMIB-based Xeon CPUs that function as 'logically monolithic'. This legacy manufacturing scale creates supply chain resilience that new entrants and competitors without equivalent packaging volume cannot replicate, particularly as AI accelerator demand strains global advanced packaging capacity. Intel Foundry's existing infrastructure provides a buffer that fabless customers cannot easily source elsewhere.

EMIB: The Bandwidth Engine for 2.5D Integration

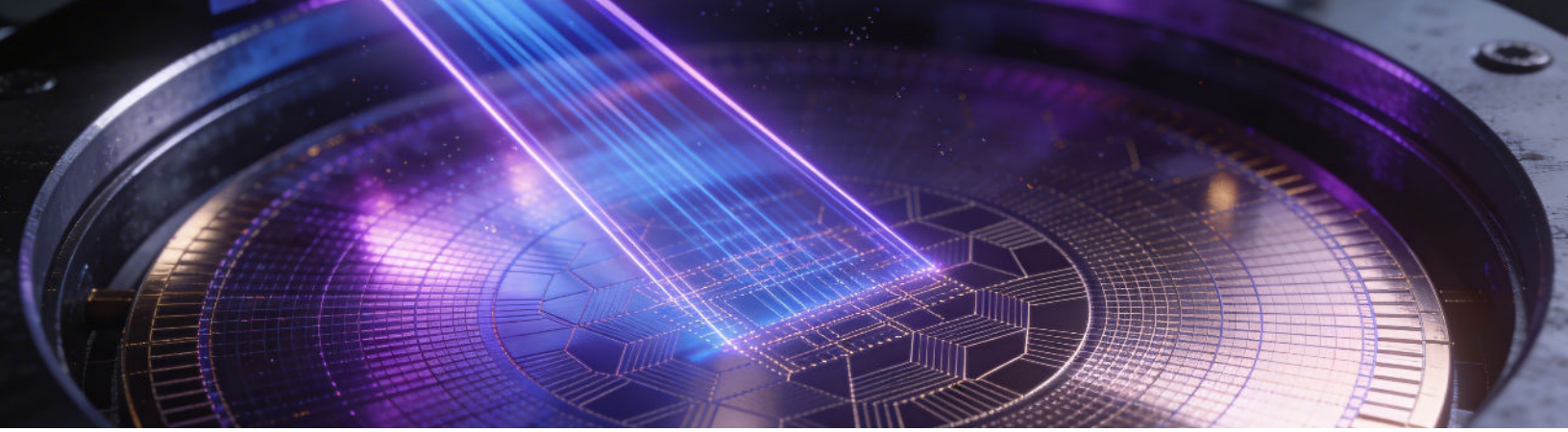
Quantifying the EMIB Wafer Utilization Advantage

Intel Foundry's EMIB technology diverges from the silicon interposer approach used by XPU packaging competitors with fundamentally different process flows. Silicon interposers require manufacturing a single large interposer and attaching to a substrate to connect chiplets, consuming significant wafer area and introducing yield challenges at scale. In contrast, EMIB embeds small silicon bridges directly into the organic substrate only where high-bandwidth die-to-die connections are needed. These bridges are then embedded into large-format organic rectangular panels, which are separated into individual rectangular substrates, leading to overall high material utilization. Compute die, HBM stacks, and I/O components are subsequently attached to these substrates. Non-critical areas use standard substrate routing, dramatically improving wafer utilization. As a result, EMIB-T delivers not only superior silicon efficiency but also a packaging flow inherently aligned with panel-based manufacturing.



"Every part of the EMIB platform is tuned towards better yield, lower cost, higher utilization."

Jan Krajniak, Strategic Business Development Manager for Advanced Packaging, Intel Foundry



The economic implications of wafer utilization are significant for AI accelerator manufacturing economics. EMIB's highly dense interconnects achieve approximately 90% wafer utilization, compared to 60% with alternative packaging based on an 8x reticle RDL interposer. Advanced packaging can be the highest contributor to XPU costs, ranging up to 42.7% of the bill of materials in a scenario of high packaging yield loss. In this scenario, packaging can cost more than memory or logic dies, according to Epoch AI's NVIDIA B200 Bill of Materials Model. Packaging yield loss is a highly variable contributor to system costs and can surpass the advanced packaging components themselves.

Minimizing packaging yield loss can reduce AI accelerator production costs by as much as 21.6%. For a leading-edge GPU, this improvement translates to approximately 4.9 gross margin percentage points. At industry scale, this would have contributed $10 billion in additional gross profit across the AI Accelerator industry in 2025 alone, based on Futurum's Data Center Semiconductors Forecast.

In a case where EMIB only reduces packaging yield loss by 50%, cost savings could be lower than the maximum, yet remain highly material to overall system costs. For EMIB, additional time savings accrue from eliminating the wafer-level assembly process that silicon interposers require. While actual cost savings will vary by a design's wafer requirements and complexity, margin improvement will be crucial for price-performance and requires cost savings on both memory and packaging.

EMIB Scaling Roadmap for AI

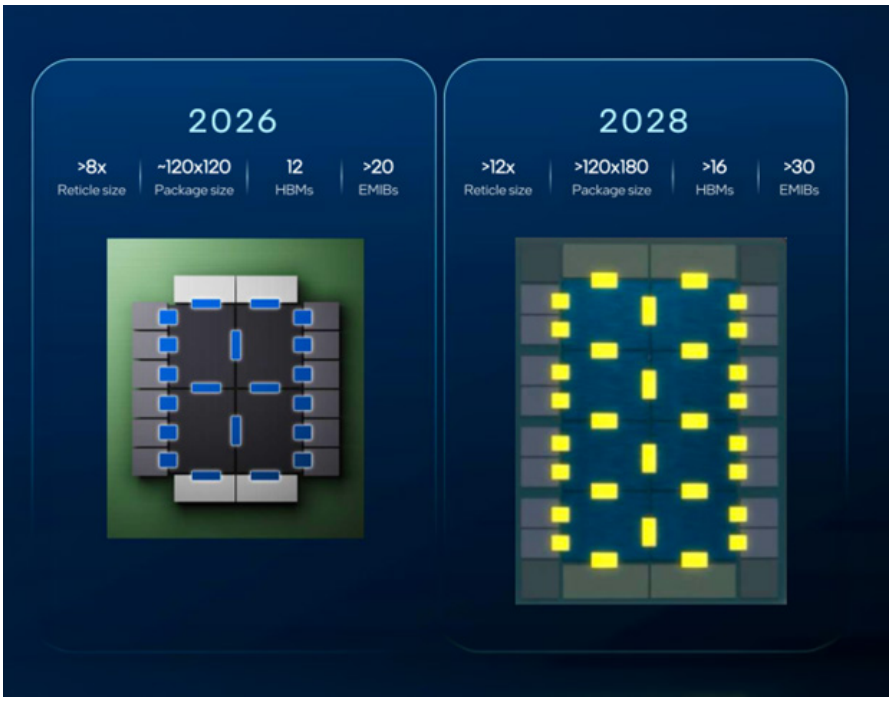
EMIB's technological roadmap directly addresses AI-specific requirements across three dimensions (see Figure 2):

1. **Bandwidth Without Power Penalty:** EMIB supports high-bandwidth, low-latency communication between chiplets without power-hungry (serializer/deserializer) SerDes circuits. By eliminating SerDes conversion at die boundaries, EMIB preserves the power budget for compute rather than data movement. As the industry considers upgrading die-to-die interconnects to co-packaged optics for power savings and signal quality, risking system reliability, this innovation extends the utility of highly reliable silicon bridges.

2. **Advanced Power Delivery Through EMIB-T:** The forthcoming EMIB-T variant introduces through-silicon vias (TSVs) directly into the bridge to supply power vertically from the bottom of the package, reducing power noise by approximately 50% compared to traditional cantilever paths.

3. **HBM Scaling for Frontier Models:** The EMIB roadmap supports attaching 16+ HBM modules per package, aligning with frontier XPU designs including Rubin Ultra. As HBM4 interfaces demand higher bandwidth per module, EMIB-T's TSV-enabled power delivery will become essential for maintaining signal integrity and thermal stability across dense memory-to-compute interconnects.

Figure 2: EMIB Scaling Roadmap



Source: Intel Foundry

Foveros Direct 3D: Hybrid Bonding for 3D Density

Foveros Direct 3D represents Intel Foundry's entry into direct copper-to-copper hybrid bonding interconnect (HBI), enabling vertical 3D stacking of chiplets with interconnect density that solder-based approaches cannot match. Current Foveros Direct implementations achieve bump pitches at 9 microns, with an aggressive roadmap toward sub-5 microns. This density enables bandwidth between vertically stacked dies that approaches the interconnect density within a single die.

Foveros Direct's unique value for AI accelerators lies in its integration with EMIB in the EMIB 3.5D configuration. 3.5D's Lego-like approach allows multiple 3D-stacked chiplet assemblies to be laterally connected via EMIB bridges, enabling package architectures of unprecedented complexity without requiring full-custom packaging development for each product variant. This architectural flexibility is particularly valuable for the AI accelerator market, where customer requirements vary significantly between inference-optimized, training-optimized, and domain-specific architectures.

Strategic Implications for the AI Market

The Foundry Turnkey Pivot: Packaging as a Service

Intel is no longer building advanced packages solely for its own products. Since its launch of Intel Foundry, the company's foundry ambitions are transforming packaging capabilities into infrastructure available to the broader AI accelerator ecosystem. Customers can bring their own foundry silicon and have Intel Foundry package it with EMIB, Foveros, or both via EMIB 3.5D.

Intel Foundry's packaging services have thus become independent of front-end wafer fabrication competitiveness. Even as TSMC maintains leading-edge process node leadership, Intel's foundry business can capture packaging value from silicon manufactured by any other foundry. The packaging layer, not the process node, becomes Intel Foundry's initial value capture mechanism.

The Superchip Reality: Panel-Level Packaging Enables 12x Reticle Size Die Complexes

The roadmap to package sizes exceeding 120mm×180mm enables the creation of massive AI accelerator complexes that function logically as one chip but are manufactured as cost-effective chiplets. This capability directly serves the competitive dynamic in AI infrastructure where accelerator performance scales with die area, memory bandwidth, and interconnect density.

We approximate the die complex sizes of NVIDIA's Rubin GPU at 5.3x reticles and AMD MI455X at 7x reticles, representing the current frontier. Custom XPUs tend to lag based on their optimization for inference and lack of component supply. Intel Foundry's 12x reticle size target for 2028 will enable scaling of these roadmaps and competition with merchant leaders, supporting proportionally larger compute complexes, more HBM modules, and higher aggregate memory bandwidth.

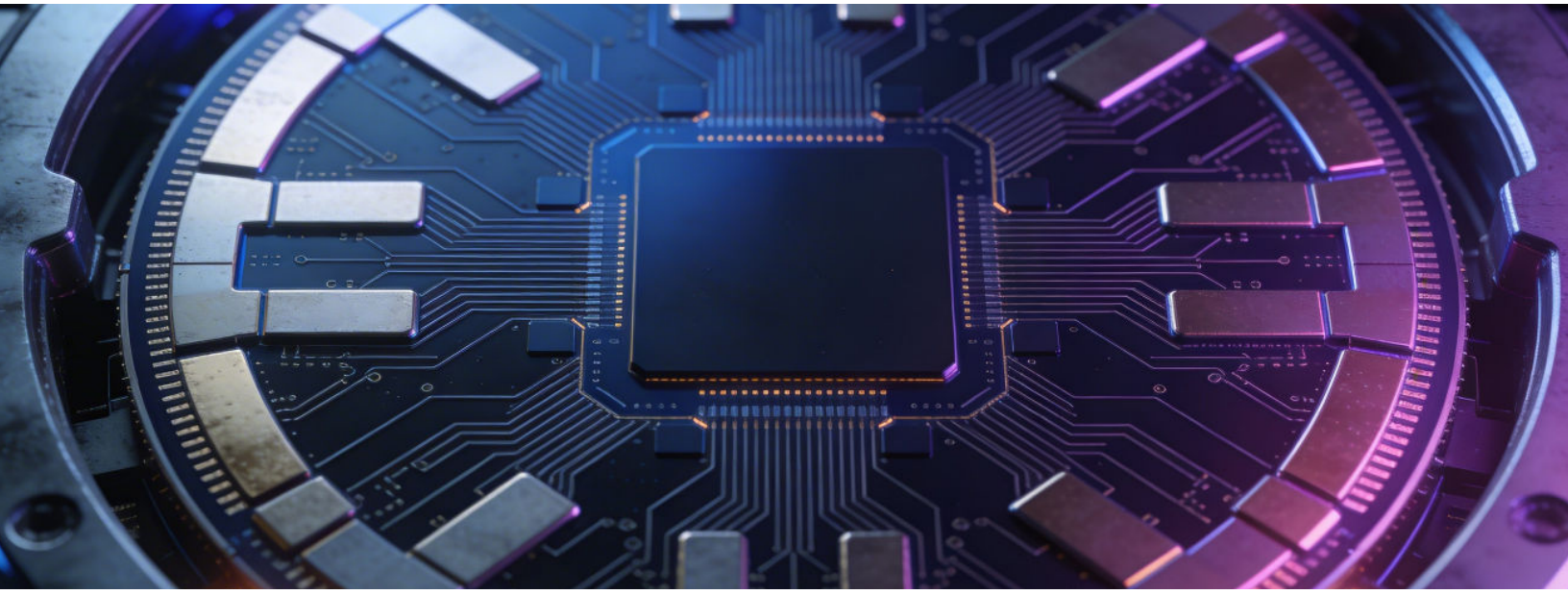
At this scale, AI accelerators can integrate compute, memory, and I/O at densities that enable new model architectures currently constrained by inter-chip communication bottlenecks. Intel Foundry's packaging roadmap has the potential to be the enabling infrastructure for the post-transformer era of AI model design.

The Post-Monolith Era: Packaging as the Scaling Vector

Intel Foundry's advanced packaging portfolio is the architectural foundation for AI systems that exceed what monolithic silicon can deliver. By solving the challenges of scaling silicon packages through EMIB bridge technology, Foveros Direct HBI, and panel-level packaging, Intel Foundry positions itself as the foundational infrastructure layer for post-monolithic computing where chiplet-based AI accelerators reaching 12x reticle size equivalents become the standard architecture for frontier AI systems.

The Systems Foundry model transforms this technological capability into a business strategy that serves the entire AI accelerator ecosystem, not just Intel's own products. Fabless AI chip companies require advanced packaging capacity that Intel's FCBGA manufacturing heritage, EMIB wafer efficiency advantages, and panel-level packaging roadmap are uniquely positioned to provide.

For AI architects and infrastructure decision-makers, Intel's packaging roadmap offers a pathway to build frontier accelerators by decoupling performance scaling from the slowing cadence of transistor density improvements. When the reticle stops being a constraint, packaging becomes the architecture of intelligence itself.



Important Information About This Report

AUTHORS

Brendan Burke
Research Director, Semiconductors, Supply Chain & Emerging Tech | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

intel foundry

ABOUT INTEL FOUNDRY

Intel Foundry is the contract manufacturing arm of Intel Corporation, created to deliver advanced semiconductor fabrication and packaging services to external customers. It combines Intel's leading-edge process technologies, including its roadmap of angstrom-era nodes, with differentiated capabilities in advanced packaging such as Foveros and EMIB. The business is designed to support a broad ecosystem of partners, from fabless chip designers to system companies, with a focus on performance, power efficiency, and supply chain resilience. Intel Foundry also integrates design services, IP, and software tooling to enable end-to-end silicon development and faster time to market. As geopolitical and capacity considerations reshape the semiconductor landscape, Intel Foundry positions itself as a strategic alternative for customers seeking geographically diverse, high-performance manufacturing.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum®

CONTACT INFORMATION: **The Futurum Group LLC | futurumgroup.com | (833) 722-5337**