White Paper

Intel Delivers Cutting-Edge Process Technologies to the Data Center with Intel 18A and Advanced Chiplet Packaging

Authors

Intel Design Engineering Group

Pushkar Ranade
Senior Principal Engineer and Lead Technologist, Intel

Mondira Pant
Technical Assistant to SVP and GM of Design Engineering Group and Lead Technologist, Intel

Srikanth Nimmagadda
Senior Principal Engineer and Lead Technologist, Intel

Eric Fetzer
Fellow and Lead Technologist, Intel

Intel is delivering several advanced logic, packaging, and systems capabilities as part of its new systems foundry for the AI era. These technologies enable pioneering new approaches for customers to develop architectures, products, and high-performance, efficient systems to support demanding applications like AI. Intel sees these technologies as critical building blocks for future silicon-based computing systems. These groundbreaking features are ready for design by Intel Foundry customers and will debut in a future Intel® Xeon® processor (codenamed Clearwater Forest) in 2025 using Intel 18A technology.

Best-in-Class Power Efficiency for Throughput Computing

Increasingly, a variety of modern computing workloads are better served with flexible CPU systems that can scale compute performance through improved core performance or higher core density. In addition, power efficiency is becoming a more central aspect of data center server architecture and design. A state of the art many-core CPU implementation today requires more silicon area (span) than a single lithography reticle field (~800mm^2). This in turn necessitates a disaggregated architecture and drives the need for advanced packaging technologies to maximize die-to-die communication bandwidth while minimizing any latency penalty. In order to meet these needs, Intel has pioneered a number of new technologies in its Intel 18A process node as well as its advanced packaging and assembly techniques.

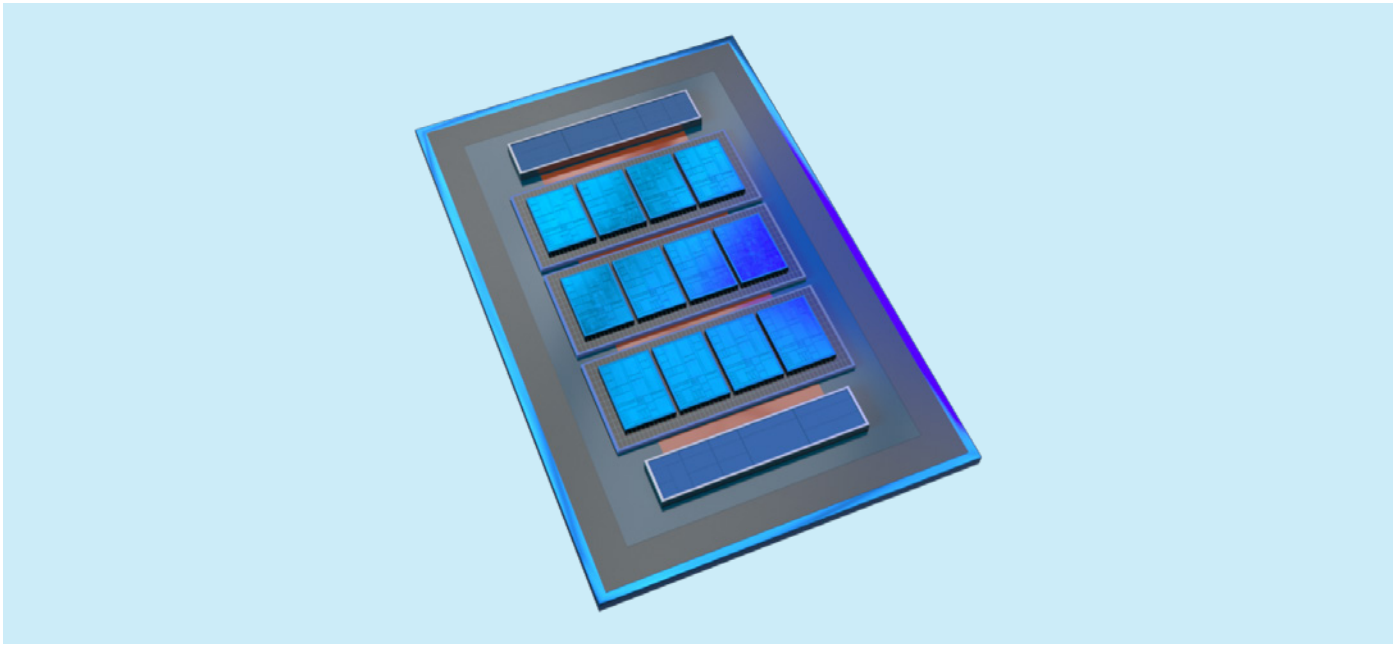

Figure 1. A rendering shows multiple chiplets connected with a combination of 2D and 3D advanced packaging techniques to create a complex system in a package.

The new technology components include:

1. **RibbonFET –** the latest advancement in transistor architecture.
2. **PowerVia –** the latest advancement in power delivery technology.
3. **Foveros Direct 3D –** hybrid bonding to enable high-density direct stacking of active chips.
4. **Embedded Multi-die Interconnect bridge (EMIB) 3.5D –** EMIB 2.5D technology combined with Foveros Direct 3D.
5. **Intel Foundry FCBGA 2D+ –** high performance, multi-die, cost-effective, high pin-count packaging.

RibbonFET

RibbonFET is the most significant change to transistor architecture after today's FinFET transistor. The FinFET architecture was refined and optimized over the last 15 years to improve performance and power efficiency. But at today's geometries, FinFET has reached its limits and is no longer able to provide additional gain in performance or power. The RibbonFET transistor further improves the electrostatics of the FinFET by wrapping the transistor gate all around the channel which takes the shape of narrow ribbons of silicon. Intel Xeon processors (codenamed Clearwater Forest) will leverage Intel's second generation RibbonFET Technology (Intel 18A) to build the primary compute CPU chiplets. RibbonFET is expected to deliver exceptional energy efficiency improvement over today's FinFET transistor.

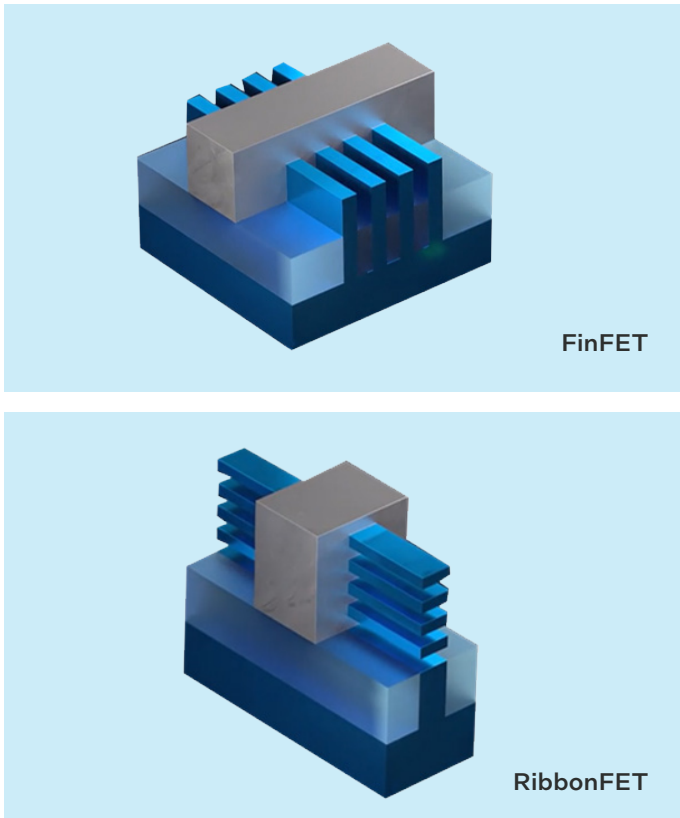

Figure 2. RibbonFET represents a generational shift in transistor architecture after the FinFET. Superior electrostatic control of the channel region allows for supply voltage reduction and improved power efficiency.

PowerVia

Since the very first integrated circuit nearly 5 decades ago, metal wires to interconnect transistors have always been on the top of the transistor layer (front-side interconnects), while the substrate under the transistors has always been primarily a structural support layer. Starting with its Intel 20A process node, Intel is changing this paradigm to introduce metal interconnects below the transistor layer (back-side interconnects). In the old paradigm, the front-side interconnect architecture was shared among wires to route electrical signals between transistors and wires to deliver power to the transistors. With the introduction of PowerVia technology on Intel 20A, signal routing and power delivery are decoupled for the first time. This enables the front-side interconnect architecture to be optimized for signal routing while a new back-side interconnect architecture can be independently optimized for power delivery. This decoupling enables improved routability (thus saving chip area and power) and also lower voltage droop (thus enabling more performance at a given supply voltage).

Foveros Direct 3D

Foveros Direct 3D is an Intel technology that enables direct attach of one or more chiplets to an active base tile to create complex system modules. "Direct" attach is achieved by thermocompression bonding of copper vias on individual chiplets to those on a wafer or even direct bonding of entire wafers stacked atop each other. The attachment can be "Face-to-Face" or "Face-to-Back" and can include chips or wafers from different source foundries, offering more flexibility in product architecture. The connection bandwidth is determined by the copper via pitch (and resulting density). The first generation of Foveros Direct 3D will use copper bonding at a pitch of 9um while the second generation will shrink the pitch to just 3um.

This unit of CPU chiplets sitting atop a large "local" cache becomes a complete compute module, which can then be replicated to scale up compute capability and create a SKU stack based on core count and cache requirements.

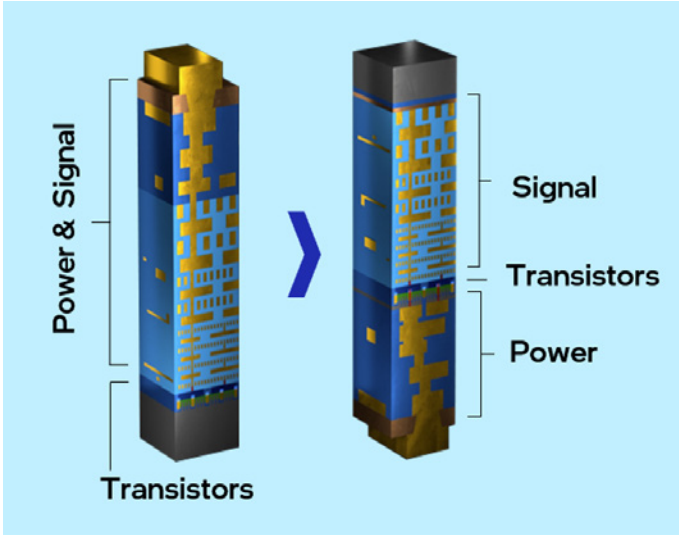

Figure 3. PowerVia enables up to 90% chip area utilization along with a 30% reduction in voltage droop and a 6% performance improvement. These benefits, already proven on test chips and published in June 2023, are expected to translate to the product level too.

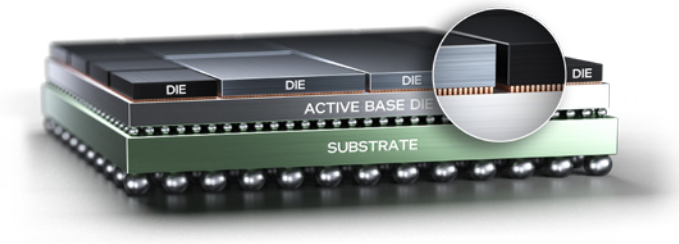

Figure 4. Foveros Direct 3D enables high-bandwidth and low latency interconnects between stacked chips.

EMIB 3.5D

Embedded Multi-die Integrated Bridge (EMIB) is a proven Intel technology that enables high bandwidth connectivity between multiple large chiplets without using a silicon interposer. EMIB technology can also be used to connect multiple compute modules constructed using Foveros Direct 3D technology as described earlier. This combination of EMIB and Foveros in a single package is called EMIB 3.5D and enables the creation of flexible, heterogeneous computing systems. Individual tiles or modules can be identical (e.g., to create a scalable compute architecture) or they can be disparate (e.g., to connect compute modules with I/O tiles or with DRAM modules). The scalability and flexibility enabled by EMIB 3.5D allows the creation of systems in package with total silicon surface area far greater than that achieved by silicon interposers alone. Intel Foundry customers can leverage 2nd generation EMIB technology (bump pitch scaled from 55 micron to 45 micron) to achieve high bandwidth connectivity with either Foveros Direct 3D chiplets or multiple I/O chiplets.

Figure 5. A combination of EMIB and Foveros enables the creation of flexible and heterogeneous systems with significantly larger total silicon surface area within a single package.

Intel Foundry FCBGA 2D+: Cost-aware packaging solutions

In addition to the wide capabilities of advanced 3D packaging, Intel also has specific architectures and design techniques to deliver cost-optimized packaging. One such architecture is called Intel Foundry FCBGA 2D+ (Flip Chip Ball Grid Array 2D+). The schematic rendering below shows the high-level concept of Intel Foundry FCBGA 2D+.

In the Intel Foundry FCBGA 2D+ architecture, the finer feature (expensive) capabilities of organic substrate technology are invoked in a smaller footprint (a high-density "patch" substrate) and assembled onto an interposer (larger footprint) which invokes "printed circuit board" or PCB-like capabilities at a lower cost. This composite (package-on-package) is then assembled onto a board. Overall cost reduction gains using such an architecture for Intel Xeon processors can easily be in the hundreds of millions of dollars. Intel has been successfully deploying this technology in its Intel Xeon product line for several generations. More recently, as the interconnect speeds continue to rise and electrical margins make it challenging to overcome the margin loss implications (discontinuity in the electrical path), material advancements and design techniques have been developed which can help realize PCIe Gen6, DDR5, and MR DIMM like speeds.

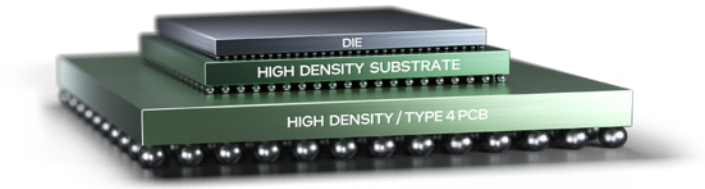

Figure 6. A high density "patch" with finer features is sandwiched between an active chip (top) and a PCB-like interposer (bottom).

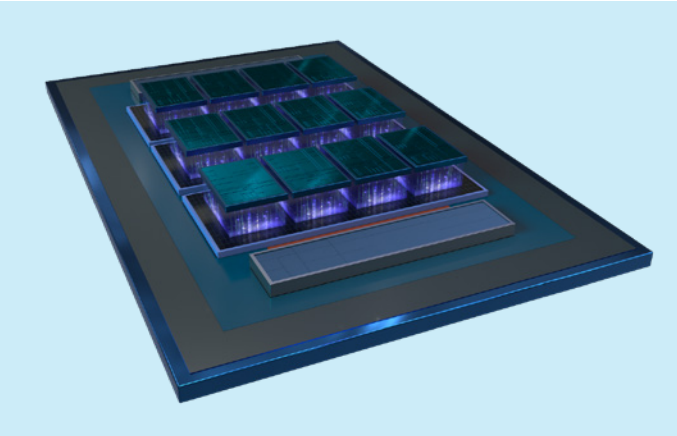

Figure 7. Innovative construction enables mix and match of multiple different process technologies to optimize fab yield, unit cost, and design turnaround time.

Mix and match: Multiple process nodes in a package

Best-in-class high performance computing products require a significant silicon span (total silicon surface area) within a package. This is driven by increasing core counts, increasing I/O and connectivity requirements, increasing accelerator IP content and other functionality. This requirement makes disaggregation a necessity for high performance computing products today and even more so in the future. As is well documented, smaller chiplets are easier to yield than large, near-reticle sized chips. Advanced packaging technologies like Foveros Direct 3D and EMIB 3.5D enable larger than reticle silicon spans as described earlier—but they also enable significantly more choices and flexibility in product architecture. Leveraging this flexibility, not only can architects now break up large monolithic chips into identical tiny chiplets to improve yield (and hence cost), but they can also disaggregate functional blocks into unique chiplets. This enables disaggregation by process node—allowing less scalable IP (e.g., analog and SRAM) to be retained on trailing edge geometries, while only migrating more scalable IP (e.g., digital logic) to leading edge geometries. Technologies like Foveros Direct 3D also enable the combination of chiplets from disparate sources (foundries) which adds even more flexibility to product architecture.

Compute chiplets benefit the most from geometry scaling and will leverage Intel 18A technology for best-in-class performance-power-area (PPA). The size of an individual compute chiplet is chosen to optimize process yield while also enabling modularity in product architecture. Compute chiplets are stacked atop an active base tile using Foveros Direct 3D as described earlier. The base tile can contain logic and memory IP for data caching and routing from I/O to cores and between cores. The base tile can leverage previous designs using a previous generation process node to lower R&D costs while providing adequate functionality. The I/O tiles can also reuse investments from previous products speeding development turnaround time (TAT) and provides a significant product cost advantage. These ingredients can be mixed and matched in future products as needs arise for different processor core IPs and/or I/O functionality enabling derivative products relatively quickly, while retaining the existing system architecture.

Bringing these flexible architecture advancements to market represents Intel's vision for computing systems of the future, and a moment where these innovative technologies come together in a package that will significantly advance data center computing. Intel 18A, Foveros Direct 3D and EMIB 3.5D are ready for design by Intel Foundry customers and will debut in the market in 2025 in a future Intel Xeon processor, codenamed Clearwater Forest.

Further Reading

In 2024, Intel Hopes to Leapfrog Its Chipmaking Competitors - IEEE Spectrum

With PowerVia, Intel Achieves a Chipmaking Breakthrough