HETEROGENEOUS INTEGRATION ROADMAP

2025 Edition

Chapter 13: Co-Design for Heterogeneous Integration

For updates, visit **the website**

The HIR is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

We acknowledge with gratitude the use of material and figures in this Roadmap that are excerpted from original sources. Figures & tables should be re-used only with the permission of the original source.

Contents

Executive Summary

Market focus on artificial intelligence, high-performance computing, data centers, health, automotive and communication is driving the quest for more creative integration methods. Future packaging technologies will exhibit high interconnect densities and large numbers of diverse components and devices. Such schemes are expected to offer relaxed transactions between components, increased bandwidth, and better energy efficiency. To achieve these potentials, the implementation of these systems will require assistance from design tools that can handle unmatched levels of complexity and highly complex computational challenges. The tools will need in addition to concurrently address problems related to signal integrity, delay, bandwidth, heat removal, electromigration, placement, routing, reliability, testing, resilience, and security which are increasingly intertwined and must be addressed concurrently, thus the need for codesign. Multi-domain and multi-physics challenges have given rise to the need for co-design tools and methodologies. Interdependence between different domains in the design flow dictates that analysis and design optimization in these domains must be performed concurrently. For heterogeneous integration, this implies the replacement of separate approaches with a parallel and integrated flow for design, analysis and optimization. Solutions are needed to address energy efficiency and high bandwidth density requirements necessary for high-performance compute and storage while providing an optimum solution for power, performance, area, cost, and reliability (PPACR).

Figure 13.0.1 captures the essentials of an advanced packaging design flow. The abstraction layers representing various stages of the design process are software point tools that may exhibit different levels of interaction. The constraints must be considered as concurrent drivers of the design flow that ensure that performance objectives are attained.

Figure 13.0.1: Constraint-driven co-design flow for advanced packaging.

Co-design addresses the discontinuities that exist between the various levels of integration. It manifests itself in the form of different formats or physics employed. The objective is to streamline the process of assembling and optimizing the system while applying constraints pertaining to the physical and logical interactions between the various domains.

This chapter explores how co-design practices need to be defined in the context of heterogeneous integration. It addresses the current capabilities and future challenges in the design flow for advanced packaging. The vision for co-design is expected to create an environment where design closure is achieved with a minimum number of iterations meeting all requirements for PPACR. This environment must leverage from currently available technologies, namely computing power, algorithms and artificial intelligence. The TWG explored challenges and potential solutions associated with:

1. Signal integrity/chiplets
2. Power delivery
3. Thermal analysis
4. Mechanical stress and warpage
5. Testability, Resilience, Reliability and Security
6. System Architecture / Partitioning / NOC Design
7. Chip-to-Chip Interface
8. Floorplanning / Layout / Routing / Assembly
9. HI Signoff (Extraction / Verification / System analysis / Signoff)
10. Multiphysics Co-Design
11. AI for Co-Design

It is imperative that designers work together to assess early on the system-wide impact of various packaging scenarios. Design teams need to bridge the process between optimization and production design. Experience has shown that tools must and will mature with customers. It is a very beneficial model, because software can be re-configured and updated continuously depending on the needs and specific applications. New methodologies need to be introduced and can only succeed with support from the major EDA players.

Send corrections, comments and suggested updates to the TWG chair, using our HIR SmartSheet:

https://rebrand.ly/HIR-feedback

13.1 Co-Design and Analysis of Chiplets, Interposers and Packages for Multi-Die Systems

Competition in the process technology race will continue for the next decade. However, this will not be the only factor for competitiveness. Cost and bandwidth have been the driving forces for the computer industry over the past 60 years. These relied on the assumption that advances in technological manufacturing would lead to advances in integration. While scaling occurred for many component parameters, it did not occur uniformly. Irregular scaling is exemplified by analog circuits for which certain components scale with the wavelength rather than technology node. Thus, the difficulty of interfacing digital and analog components has become more serious in the lower nodes. At this point additional scaling does not necessarily result to a proportional increase in performance on a per-node basis.

Chiplets offer manufacturers with a more efficient means of achieving die scaling on those portions of the chip that would yield more substantial gain in performance. With such an approach, an entire IC does not need to shrink. In addition, smaller die leads to less wafer waste, higher number of cores per wafer, improves yield and facilitates testing. Finally, chiplets can facilitate a heterogeneous integration by combining different types of technology (e.g. silicon and gallium arsenide).

In addition, scaling has become more difficult and expensive at each node, and the power and performance improvements to be gained have diminished. While scaling alone can improve performance for each new node below 7nm, changes related to advanced packaging, heterogeneous integration and hardware-software co-design can boost performance by several orders of magnitude.

13.1.1 Co-Design of Chiplets

In current 2.5D chiplet systems, each chiplet is often designed independently as a single unit and then the chiplets are mounted on an interposer or a package to form a heterogeneous system. There is very little co-optimization of chip-interposer-package design. This design approach is not optimal when the interactions between chips, interposer and package is critical to the overall system performance, power, chip-to-chip communications and readability as well as cost. The RDL routing, interposer design and package layout need to be implemented with constraints to improve the overall system.

The design flow that considers the Chip-Package Co-Design goals in 2.5D integration technology is shown in Figure 13.1.1. In this flow, the design of the 2.5D package together with the chiplets is implemented in the same design environment. This enables exchanging design information between the chiplets and the package during the optimization steps, which is essential to achieve an optimized heterogeneous system [13.1.1]. It starts with partitioning scheme that is compatible with the 2.5D package routing layers together in a chip design process. Then the floor planning of chip and package routing analyzed to reach an optimized heterogeneous system.

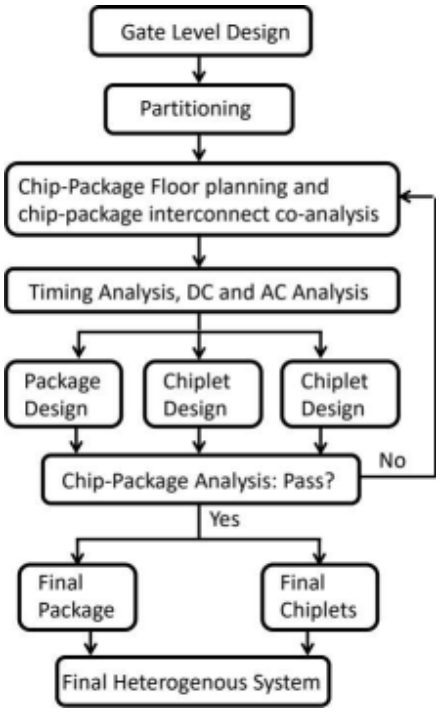


Figure 13.1.1: Chiplet-package co-design flow.

The common implementations of heterogeneous integration systems are using organic substrate and silicon substrate with and without Through-Silicon Via (TSV) are reviewed. The performance, density and cost of the die-to-die interconnects using these three approaches can be significantly different. A conceptual drawing of the cross-section of a heterogeneous system with organic or silicon interposer, chip-on-wafer-on-substrate (CoWoS), and Embedded Interface Bridges (EMIB) are shown in Figure 13.1.2. Most common implementations of 3D system-in-package (SiP) are designed and manufactured with a large organic interposer (substrate) with fine-pitch and fine-line interconnections [13.1.2]. Silicon interposer and EMIB are also used for high-end and high-performance systems [13.1.3]–[13.1.4]. Figure 13.1.3 summarizes different characteristics of EMIB and silicon interposers with respect to organic substrates. Due to the higher interconnect density, silicon interposer and EMIB more effective in high-end heterogeneous systems.

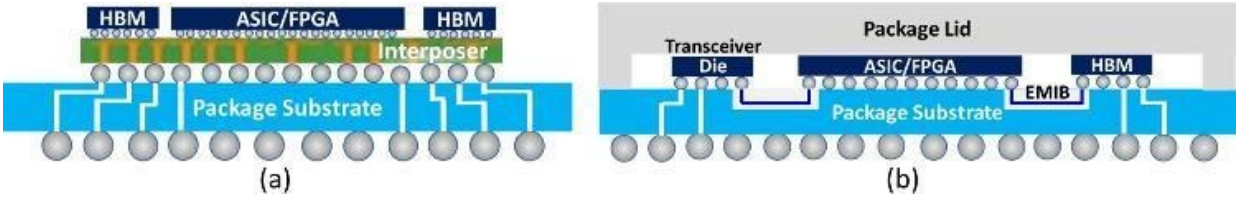


Figure 13.1.2: Heterogeneous integration using (a) organic or silicon interposer and (b) EMIB technology.

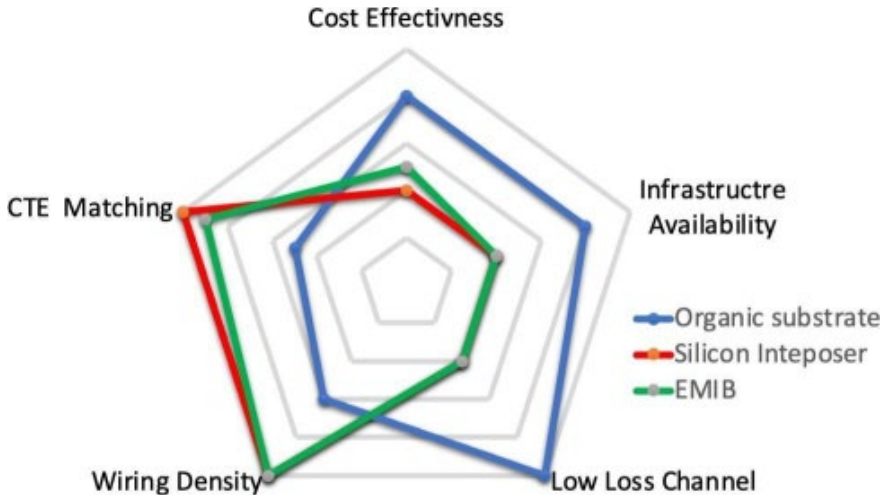


Figure 13.1.3: Comparison EMIB and silicon interposers with respect to organic substrates.

Although the scale of interconnect dimensions achievable in silicon interposer is finer than what is realizable in organic packages, the signal conductor and dielectric losses in organic substrate are significantly less than that in silicon interconnect. The dimension of interconnects used in silicon interposer or EMIB is similar to that of silicon metallization, thus the resistive loss and crosstalk between signal routings can be one of the dominant sources of noise and timing error when compared to package and board. While traces in package and board may require only 2D-based crosstalk analysis, an interposer and silicon bridge need 3D crosstalk analysis including vertical paths such as via, bump and micro-via.

As shown in Figure 13.1.4, a typical silicon interposer often uses one-sided 3 or 4 layers of redistribution layer (RDL) and through-silicon via (TSV). The physical dimensions and material properties shown in Figure 13.1.4(a) and Table 13.1 are used in the analysis. The metal configuration of the three copper conductor layers with signal layer and power mesh are also shown in Figure 13.5(b). The diameter of the TSV is 10 µm and the insulation layer thickness is 0.5 µm.

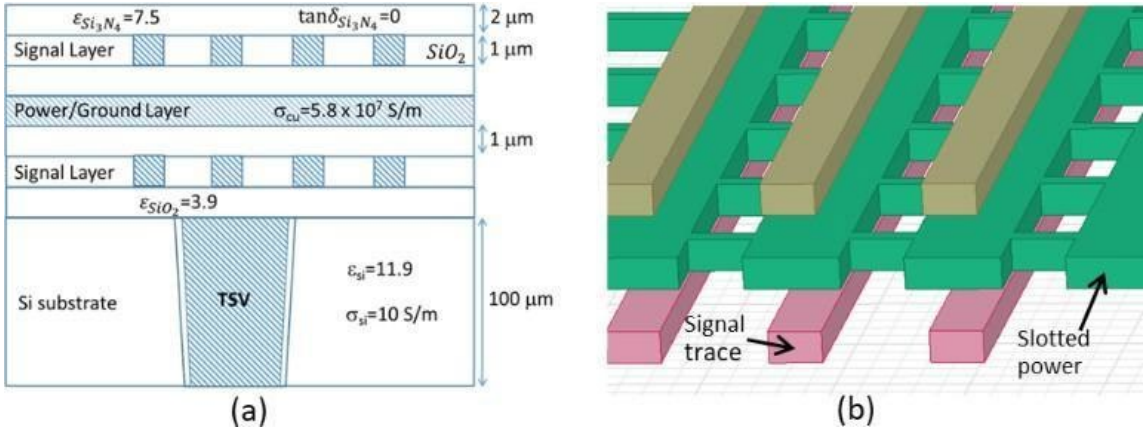


Figure 13.1.4: (a) A typical physical dimension and material property and (b) the signal and power/ground interposer design.

The fine pitch offered by silicon interposer and EMIB allows high density routing among multiple processors, transceivers and HBM dies. Although the channels between the dies on the same package are relatively short, the transceiver dies can communicate with another processor placed on another organic package using TSV or micro-via, package, board traces and via's, thus signal integrity can be a concern. The effect of TSV or interposer via with optimized design, using single or multiple TSV's, in off-chip high-speed channels can be negligible.

Unlike silicon interposer, the silicon bridge (i.e., EMIB) does not require TSV. EMIB is embedded in the package and uses micro-bumps to connect signals as well as power/ground planes between two silicon dies as shown in Figure 13.1.4(b).

13.1.2 Signal Integrity Analysis of Die-to-Die Interfaces

Since hundreds of signal interconnects are densely integrated in a small space in either organic or silicon interposer, there exist several coupling mechanisms. Since all signal micro-bumps are very crowded in a limited space, only a limited number of ground signals often provide the reference, it is necessary to analyze the coupling effect between power and signal nets. Power-to-signal coupling can be a major source noise and jitter in both interposer and EMIB.

The signal integrity of the three types of second level interconnects: organic and silicon interposers and EMIB, are analyzed using the dimensions and material properties summarized in Table 13.1.1. All implementations have two signals surrounded by power and ground traces. The insertion loss of the three implementations are shown in Figure 13.1.5(a). The attenuation of the silicon interposer is higher due to the aggressive design rules shown in Table 13.1.1. The DC loss is higher due to the smaller cross section of the silicon interposer traces. As shown in Figure 13.1.5(b), the crosstalk for the silicon interposer is minimized by using ground and power traces for shielding as shown in Figure 13.1.4(b), even though the ground and power planes are hatched.

Table 13.1.1: Cross-Section and Material Properties of the Three Channels

Interposer Type	H (μm)	T (μm)	W (μm)	S (μm)	ε_r	tan (δ)
Organic	10	10	7	7	4.6	0.02
EMIB	2	1	2	2	3.9	0.001
Silicon	1	1	1	1	3.9	0.001

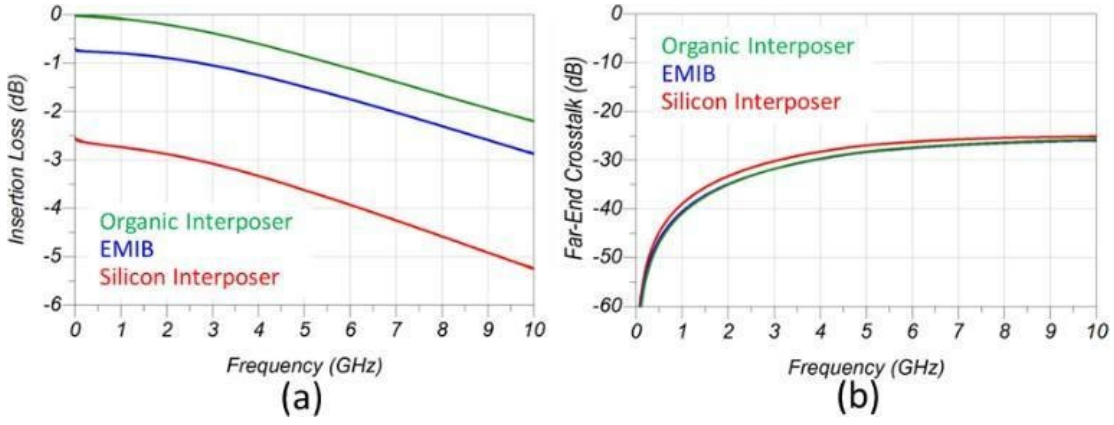


Figure 13.1.5: Channel characteristics of the organic, silicon interposer and EMIB: (a) insertion loss and (b) far-end crosstalk.

13.1.3 Power Integrity Analysis

Power distribution network (PDN) of the heterogeneous integrated system includes the interposer or EMIB power and ground routings connected to the PDN of the rest of the system that consists of PCB, package and micro-bump routing, as shown in Figure 13.1.6. If a PDN connection exists in the interposer or silicon bridge, then the PDN impedance or resonance of the dies with smaller on-chip decoupling capacitor can significantly be improved through charge sharing from the dies with large decoupling capacitor. Since the EMIB technology does not use TSV, power connections among multiple dies are made through package surface or horizontal silicon routing, thus the power delivery networks of IPs must be planned during the early design phase. For example, if there are power network connections between die 1 and die 2 through the interposer and/or package routings, even though die 1 has very small on-chip decoupling capacitor and die 2 has a huge amount of decoupling capacitor, because the power supplies are shared, die 1 can take advantage of available capacitance provided by die 2. A typical example of capacitor or charge sharing often occurs in a multi-die system with HBM devices. Therefore, without the interposer or EMIB connection, PDN peak in die 1 may have a very high peak exceeding 1.0 Ohm while die 2 PDN peaks to only 0.1 Ohm. Such shared PDN impedances for digital and analog supplies are shown in Figure 13.1.7. Therefore, the power network connections between multiple dies through interposer or EMIB routings allow sharing of decoupling capacitors to lower self-generated noise. However, the proper tradeoff has to be made as the noise coupling may increase among the charge sharing silicon dies leading into elevated coupled power supply noises.

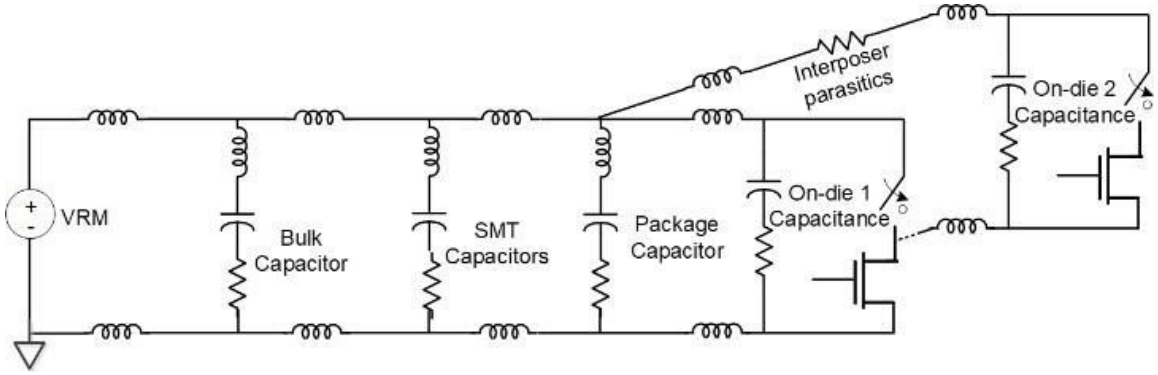


Figure 13.1.6: PDN model for a heterogeneous integrated system including board, package, interposer, and parasitic and die decoupling capacitance.

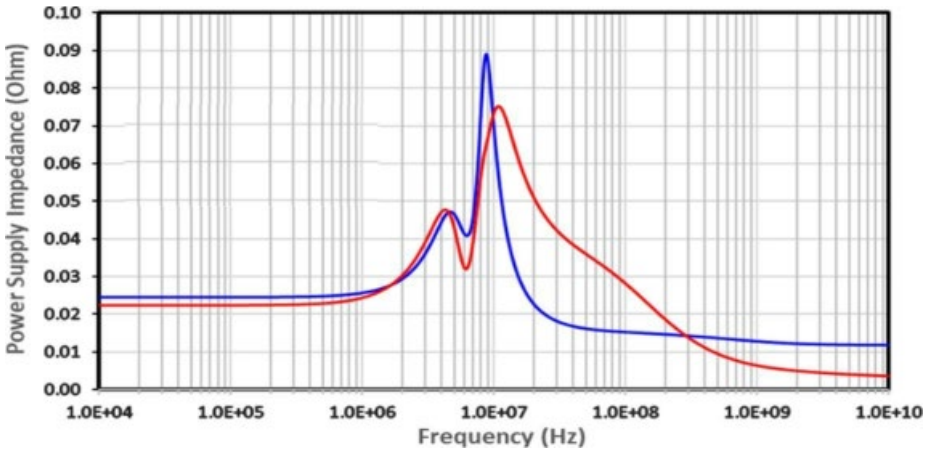


Figure 13.1.7: Typical PDN profiles of two supplies.

13.1.4 Analysis of Chiplet Interface

A Die-to-die interface, shown in Figure 13.1.8, is very different from traditional memory interfaces or SerDes (Serializer/Deserializer) links in many respects. The massive I/O often present in die-to-die interconnects is not impedance controlled and designed to efficiently operate strongly coupled to the channel. The interface operates power efficiently in few Gbps data range. The I/O often use HSUL (high swing unterminated logic) where the transmitter is not designed to be linear. Some I/Os use LVSTL (low-voltage swing termination logic) to save power with reduced swing.

Although the operating data rates of die-to-die I/O are often lower compared to the latest modern links, there are unique challenges to optimize the performance. Often, the I/O can be partitioned into 'Analog' and 'Digital' sections where only the jitter of the 'Digital' section is calculated and the voltage noise and timing jitter for the 'Analog' section of the link is analyzed. The jitter and noise in the 'Analog' section of link is summarized in Table 13.1.2.

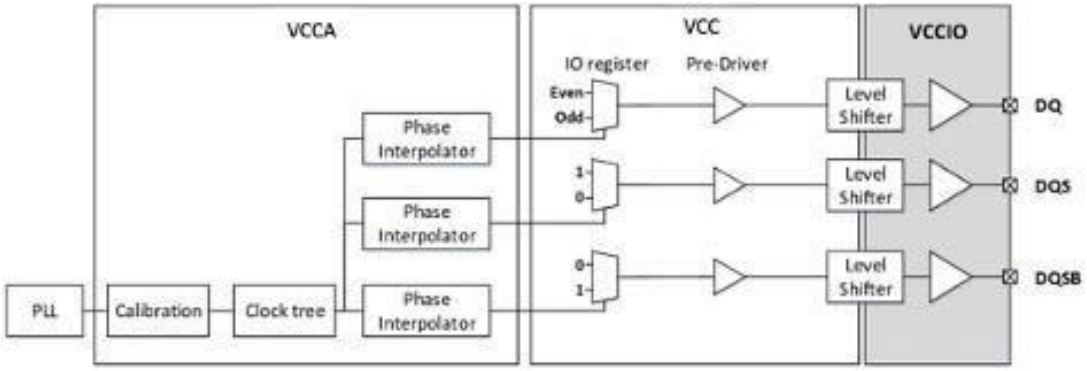


Figure 13.1.8: Block diagram and power supply partition of a typical I/O PHY TX for source- synchronous parallel.

Table 13.1.2: Timing and Noise Contributions of the Analog Link

Channel Jitter/Noise Sources	Contribution
ISI Jitter [UI]	0.07 – 0.10
ISI+ Crosstalk Jitter [UI]	0.12 – 0.19
ISI + Crosstalk + SSN jitter[UI]	0.19 – 0.25
Power Supply Noise [mV]	60 – 85

13.2 Power Delivery

13.2.1 Power Delivery Challenges

As heterogeneous integration and chiplet-based architectures become foundational for high- performance computing, ensuring robust power delivery across advanced packaging platforms has emerged as one of the most critical and complex challenges, driven by increasing power density, signaling speeds, and multi-die integration.

Non-uniform power and ground bump maps present a significant challenge in heterogeneous chiplet-based systems. Advanced packaging techniques used for high-performance computing (HPC) often combine chiplets utilizing fine-pitch microbumps, such as those found in UCIe and HBM, with bump pitches ranging 25–73 µm, with standard interfaces like C4s that employ coarser 100–150 µm pitch. This creates spatial non-uniformity in power and ground accessibility across the die surface and package boundaries. As a result, localized circuits, especially in high-speed PHY regions or dense beachfronts, may experience uneven power distribution. This can lead to localized IR drop variations and elevated power supply noise that directly impact circuit timing margins and signal integrity.

Smaller footprint PHYs in such advanced architectures exacerbate power delivery challenges due to higher current densities. As chiplets integrate more transceivers and logic into smaller die regions, each power bump or copper pillar must carry greater current while constrained by electromigration and thermal limits. The dense placement of signal bumps in high-speed interfaces like UCIe further restricts the number of power and ground bumps that can be allocated within beachfront regions. This heightened power density can make it difficult to meet current delivery requirements without violating layout or mechanical constraints, especially in transient or high-activity scenarios.

These physical challenges contribute to more complex modeling and analysis of the power delivery network (PDN), especially as power must traverse multiple hierarchical levels: on-chip, package-level, interposers or bridges, PCB stackups, and voltage regulator modules. Each level introduces impedance characteristics that must be factored into system-level power integrity modeling. Designers are tasked with ensuring the aggregate PDN impedance remains below target thresholds over relevant frequency ranges while considering the effects of dynamic load transients such as clock gating, lane wake-up, or high-speed link bursts. The interconnected nature of advanced packaging means voltage droop, power supply noise, and PDN resonances can propagate across chiplet boundaries, making full-system simulation and hierarchical modeling critical for achieving reliable operation at increasing signaling speeds and integration densities.

In time-domain simulations, the impact of rapid current changes on supply integrity becomes especially pronounced, as supply noise can create timing errors in sensitive high-speed circuits. Yet accurate modeling of these effects requires detailed extraction and co-analysis across electrical and physical domains, which increases both design complexity and verification time. Moreover, constraints in bump placement and metal routing limit designers' flexibility in uniformly distributing current paths, amplifying the importance of early-stage, multi-physics modeling to anticipate and mitigate worst-case power integrity scenarios.

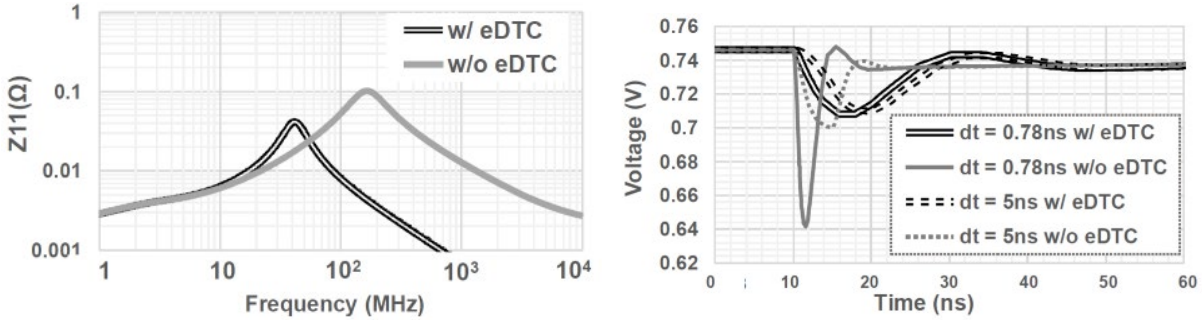


Figure 13.2.1: Power integrity modeling for UCIe-compliant chiplet interconnect design [13.2.1].

13.2.2 State-of-the-Art Approaches to Address Power Integrity Constraints

Staggered Activation

A current approach to relaxing supply voltage ripple during high-speed chiplet interface testing is staggered activation, as demonstrated in the referenced UCIe implementation [13.2.1]. Because the small ubump pitch inherent to advanced packaging designs restricts power and ground access, simultaneous activation of all transceivers during at-speed testing can cause significant IR drop and supply voltage transients. To address this, the design divides the full module into nine segments, each containing up to eight transceivers, and activates only one segment at a time during bit-error rate testing. This staggered activation helps distribute the current load over time and reduces peak power demand, thereby minimizing voltage droop and ripple on the shared power delivery network. Such architectural techniques illustrate how decoupled activation strategies play a critical role in ensuring power integrity in high-density die-to-die systems, especially under the strict constraints posed by chiplet-based advanced packaging.

Vertical Power Delivery

Another promising approach to these challenges is the use of integrated voltage regulators (IVRs). Instead of distributing high currents at low voltage (e.g., 1 V), a higher-voltage, lower-current supply such as 48 V is delivered into the package and converted near the point of load. IVRs can be integrated in multiple locations, including: 1) within the silicon interposer, 2) directly beneath die bumps, or 3) inside a cavity in the package substrate (similar to CoWoS-L or EMIB approaches).

By performing voltage regulation as close as possible to the die, IVRs significantly reduce current requirements through the external PDN, improve efficiency, and simplify power routing in advanced packaging.

Most of the challenges discussed so far are associated with DC power distribution. To ensure proper AC and transient behavior of the power-delivery network (PDN), adequate decoupling is required to minimize supply ripple and reduce transfer impedance (i.e., crosstalk) between supply domains.

Decoupling Capacitors

Decoupling capacitance is most effective closest to the circuits that need them. On-die MOS cap is the first layer of decoupling along with the intrinsic capacitance of the gates, drain/source/well capacitance. Next level capacitance available to the designer is MiM cap (Metal-Insulator-Metal) capacitance which has a capacitance density of ~200fF/um^2. For larger than reticle size partitioned die or die interfacing with High Bandwidth Memory (HBM) the use of silicon interposer includes additional capacitance in the form of Deep Trench Capacitors (DTC) which have a capacitance density of 300fF/um2. All of these capacitors are very close to the chip and have very low parasitic effective series resistance (ESR) and inductance (ESL). The next level of decoupling, which is much less effective, are ceramic chip capacitors which have very high capacitance density (~40,000 fF/um2) however they are placed much further away from the die edge (~2-3mm) to accommodate the underfill process. Connectivity to the substrate capacitors may require additional layers or compete with other resources to connect. These capacitors have self-inductances on the order of 80-400pH depending on the component in addition to the lateral mounting inductance of the substrate which is on the order of 30pH-per-square. With a vertical loop-inductance of the PDN on the order of ~1-2pH to the BGA balls for the core supply the substrate capacitors are significantly less effective. The roadmap for the decoupling capacitance density is provided in the following table.

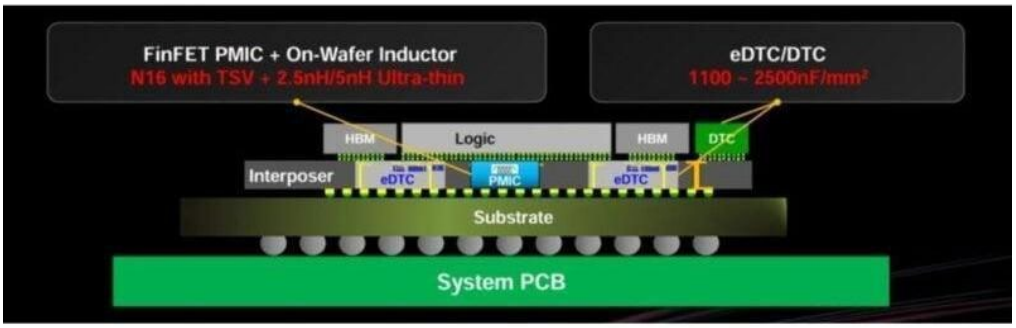


Figure 13.2.2: Staggered activation approach to managing supply voltage ripple in UCIe implementations. Courtesy: TSMC CoWoS [13.2.2]

Table 13.2.1: Metrics of interest in power delivery for advanced packaging in high performance computing

	2026	2028	2030	2032	2034
Decoupling capacitance density (μF/mm²)	2	3	4	5	6
Power density (W/mm²)	1-3	3–6	6–10	10–20	20–30
Slew rate: dI/dt (kA/μs)	1	3	5	7	>10

13.3 Thermal Analysis

13.3.1 Thermal Management Challenges

The longevity of Moore's law has been under constant debate [13.3.1]. Many packaging experts still believe the Moore's law would hold good for at least the next 10 years due to recent technological efforts to combine functionalities horizontally and vertically into a single chip module, commonly referred to as heterogeneous integration (HI). Heterogeneous integration (HI) is the basis of the future generation computer systems and refers to the assembly and placing of multiple separately manufactured components onto a single module to improve functionality and enhance operating characteristics. In such application, heat sinks are typically designed to cool multiple dies with different powers, sizes, heights, and allowable maximum temperatures in specified arrangements on electronic boards simultaneously. For several decades, electronic industries and relevant Academic research communities globally have made constant effort to develop and commercialize the Moore's law [13.3.1, 13.3.2], leading to many technological breakthroughs and revolution in packaging technologies and skyrocketing thermal management challenges. The thermal management challenges are attributed to a decrease in device size and increases power consumption. With simultaneous shrinkage in size and more considerable heat dissipation, the flux values have increased exponentially, yet the required temperature difference coined as the thermal budget for thermal management is also reduced or unchanged making it even more challenging for effective thermal management. Thermal management of electronic devices is important to improve the reliability of devices by maintaining silicon junction temperatures below critical safe operating temperatures. While the academia efforts on optimization of thermal solutions have been predominantly streamlined to theoretical studies, the constraints on manufacturability and reliability are not widely considered part of the optimization problem. However, incorporating the commercially available cold plates into the optimization process makes the study more pragmatic.

A typical thermal solution is by having a remote separable cold plate/heat sink attached to the chip with a thermal interface material (TIM). The thermal solutions for heterogeneously integrated packages, in general, can be categorized into:

- Technology 1 (T1) – Embedding cooling solution onto the device (Two phase)
- Technology 2 (T2) – Embedding cooling solution onto the chip lid/Integrated heat spreader (Single- and two-phase liquid cooling)
- Technology 3 (T3) – Remote separable heat sink/cold plate thermal solution (Air and single-phase liquid cooling)

The heat transfer limit of each of the stated technologies are as shown in Figure 13.3.1.

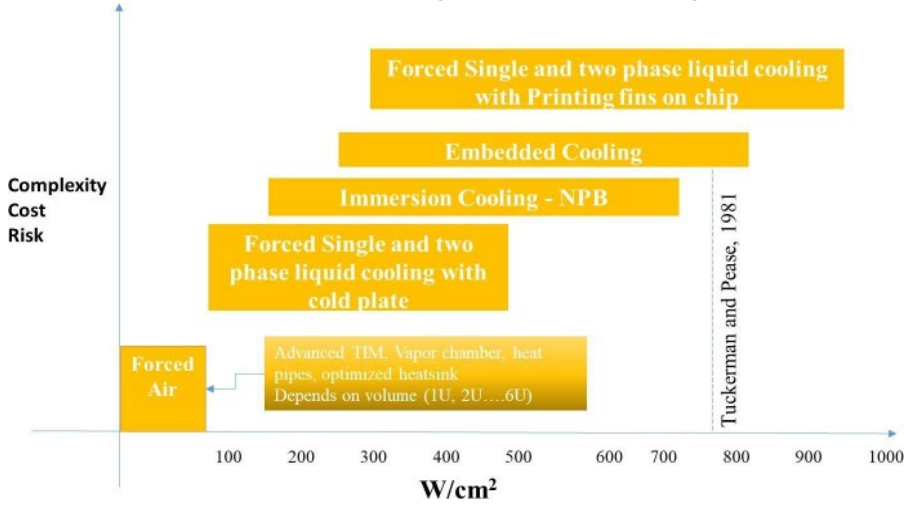


Figure 13.3.1: Heat transfer limits of different thermal management technologies.

The associated complexity and risk in implementing each of these technologies are marked on the ordinate of Fig 13.3.1. In practice for today's ongoing data center power demand, the most common cooling method takes the

form of air-cooling [13.3.3]. Air-cooling is often claimed to be the most reliable technology that is of practical interest. However, in practice air-cooling suffers from low heat transfer performance and acoustic-related problems. The low heat transfer performance is attributed to low heat transfer coefficient, high Inlet air temperature and low thermal spreading owing to requirement of larger heat sink base area. The noise issue is addressed by imposing an engineering constraint on the maximum fan velocity [13.3.4]. To address the thermal spreading issue, technologies like vapor chamber integrated with heat pipes is employed. The acoustic constraints and the allowable pressure drop dictate the limit on the maximum airflow in the system. The constraint on minimum inlet air temperature is dictated by the Telcordia GR63 standard and ASHRAE for telecom application (55°C) and data center application (45 °C) respectively. With regard to easier manufacturing, the plate fin parallel channel heat sinks are still the most widely adopted fin design. With most of the stated factors associated with air cooling have reached standard limits, only the Geometric optimization of air-cooled heat sinks could help the research community define the limits of the technology under practice. Unlike the air-cooled heat sink modules, the liquid cooled heat sinks (commonly referred to as cold plates) can perform extremely well due to superior thermal properties of liquid water. However, the superior thermal performance trades with the reliability and technological risks associated with using water closer to the circuit (e.g. unfavorable dielectric strength). In a liquid-cooled module, the heat sink resistance is 10-fold lower than the spreading and the TIM counterparts. Tuckerman and Pease [13.3.5] dissipated heat flux up to 790 W/cm2 for a substrate temperature rise of 71 °C by etching a microchannel on the backside of the active device. However, there are practical limitation, challenges, and risks in implementing the work in practice and requires a massive change in the existing infrastructure. Understanding the scaling limits of air and liquid cooling is essential to understand better the potential of embedding the cooling solutions onto the silicon die. Fig. 13.3.2(a) and Fig. 13.3.2(b) show the limit of air and liquid cooling with increasing area ratio (AR= Ratio of area of the spreader to the area of the chip). AR=1 translates to a scenario of embedded cooling. The results from the scaling analysis make it evident that the liquid cooling with AR=1 has the minimum total external resistance. Hence liquid cooling with printed fins on-chip with no additional spreader area ideally can extend the heat transfer limit of single-phase liquid cooling. However, air cooling requires a large area for effective heat dissipation and does not perform well when embedded.

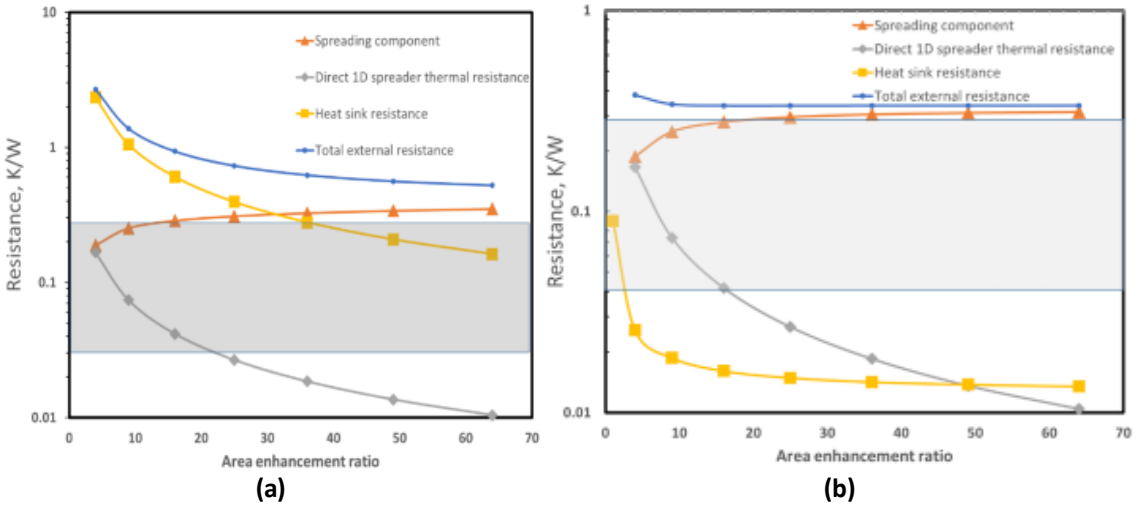


Figure 13.3.2: (a) Tradeoff analysis for air cooling with variable area enhancement ratios, assuming the chip size Ac = 1cm2, metal conductivity, k = 150W/mK, spreader thickness t = 10mm, wch = 1mm, hch=25mm, Vair= 10 m/s. This figure excludes the TIM resistance. (b) Tradeoff analysis for a single phase liquid cooling scenario for a chip size Ac = 1 cm2, metal conductivity k = 150 W/mK, Spreader thickness t = 10 mm, wch = 0.05 mm, hch=0.3 mm, Q= 0.5 liters/min. The plot excludes the TIM resistance.

With growing heterogeneous integration of multiple functionalities on to a single chip, there are growing regions of intense hotspots. 3D printing of fins on to the chip [13.3.6][13.3.7] is seen to be a potential solution to mitigate hotspots and achieve minimum of the maximum chip temperature. Vahideh et al [13.3.7] numerically demonstrated single-phase heat transfer enhancement of a liquid water-based cooling system employing printed fins directly on chip. The fin material was pure silver. The authors demonstrated that a careful optimization of the pin fin profile and shape could lead to a total chip to ambient thermal resistance of 0.26 ℃- cm2/W (compared to separable cold plate of 0.58 ℃- cm2/W) for a 4cm2 uniformly heated chip under a constrained pressure drop [13.3.7]. The same technique was extended by Vahideh et al [13.3.8] to chip with simultaneous hotspots and background heating conditions. A minimum total case to fluid resistance of 0.21 ℃- cm2/W for a four 4cm2 chips under similar flow conditions. The rationale behind the low resistance was attributed to the fact that each core/hotspot had a dedicated inlet on top, thus minimizing the total resistance's caloric resistance component.

While recent efforts on exploring two-phase liquid cooling using separable cold plates for data center applications have shown remarkable performance [13.3.9] [13.3.10] in mitigating high heat fluxes with significant energy savings, however, two-phase cooling is essentially recommended for direct chip cooling using dielectric coolant (Technology 1) [13.3.11].

Two-phase Immersion cooling could be a potential solution for highly intense and non-uniform heated devices. Boiling enhancement coatings on top of the heated surface is seen to increase the critical heat flux (CHF) and the boiling heat transfer coefficient(hb) simultaneously [13.3.12]. 3D printing of fin and microporous structures can be beneficial in promoting nucleation and enable rewetting of the heated surface. The combination of fin and microstructures has been highly beneficial in the simultaneous improvement of CHF and hb[13.3.13]. Wong et al [13.3.14] printed lattice structures on top of a heat surface, performed pool boiling experiments, and demonstrated a critical heat flux of 107 W/cm2 correspondings to a boiling heat transfer coefficient of 1.5 W/cm2K.

For all of the listed technologies, the optimization process of heat sinks/ cold plates is as shown in Figure 13.3.3.

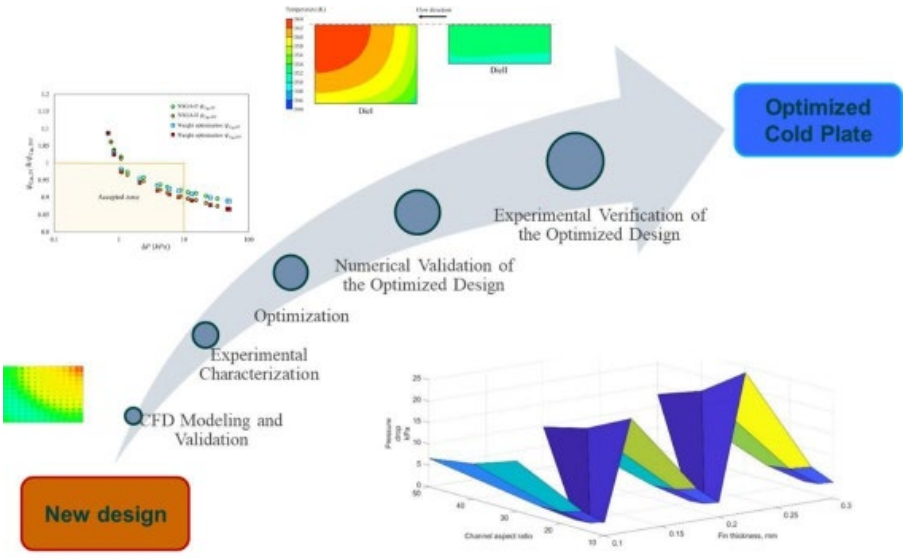


Figure 13.3.3: Optimization approach for the cold plate/ heat sink thermal solution

13.4 Mechanical Stress and Warpage

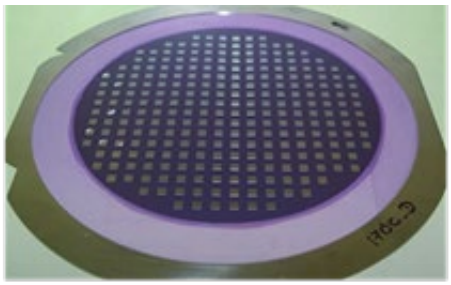
13.4. 1 Warpage and stress challenges

Mechanical warpage - defined as the out-of-plane deformation (bending, bowing, or twisting) of a package, substrate, wafer, or panel from its nominally flat configuration - and the resulting stress states in advanced semiconductor packages arises from a combination of intrinsic material behavior and extrinsic loading conditions. Key material drivers include the coefficient of thermal expansion (CTE), elastic modulus, cure-induced shrinkage, and nonlinear material constitutive responses such as viscoelasticity. These factors interact with process-dependent and operational environments, including manufacturing thermal cycles and in-field service conditions. Accordingly, a rigorous co-design framework must explicitly incorporate both warpage and stress evolution at the earliest stages of design. Such an approach is essential to proactively mitigate quality degradation and long-term reliability risks in advanced packaging systems.

13.4.2 Warpage and stress challenges in packaging build-up processes

Package build-up processes - including die placement, molding, redistribution layer (RDL) formation, interconnect integration (e.g., C4, thermo-compression bonding, and hybrid bonding), and reflow - constitute primary sources of warpage and stress accumulation within advanced package architectures [13.4.1], [13.4.2]. During manufacturing, maintaining surface co-planarity is a critical requirement for ensuring the integrity and yield of subsequent process steps. Thermally induced stresses, arising from coefficient of thermal expansion (CTE) mismatch, cure shrinkage, and chip–package mechanical interactions, can drive defect formation mechanisms such as low-k BEOL and RDL cracking, interfacial delamination, and degraded bonding quality due to non-planar or warped surfaces.

To address these challenges, physics-based modeling and simulation - particularly finite element analysis (FEA) - have become essential tools for predicting warpage and stress evolution throughout the build-up process. For example, as illustrated in Figure 13.4.1 for fan-out wafer-level packaging, variations in fan-out ratio alter the effective structural stiffness, thereby modulating wafer-level warpage responses [13.4.3] Furthermore, Figure 13.4.2 highlights the critical importance of accurate material characterization in such analyses; simplified assumptions, such as purely elastic behavior, can lead to significant deviations in predicted warpage, underscoring the necessity of incorporating realistic, nonlinear material models [13.4.4].



Structure	After Molding (150C to 22C)	At Debonding (At 180C)	Post Debonding (180C to 22C)
Homogeneous Mold (No VIB)	0.85 mm	-2.27 mm	-1.48 mm
Heterogeneous Mold FOR = 2.5	0.5 mm	-0.5 mm	-0.47 mm
Heterogeneous Mold FOR = 1.5	-0.91mm	-0.33 mm	-0.32 mm

Figure 13.4.1: Analysis of wafer warpage as a function of fan-out-ratio (Ref 13.4.3)

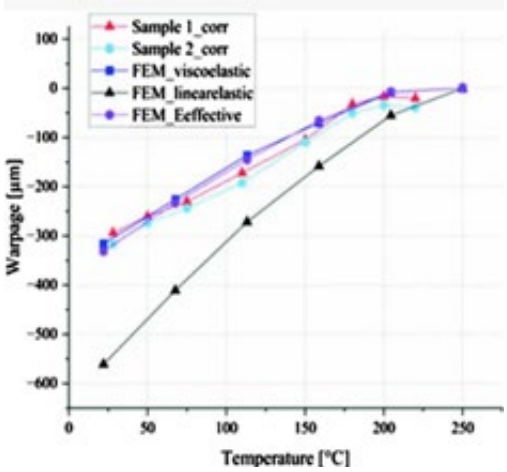


Figure 13.4.2: Warpage analysis - need for accurate materials data (Ref 13.4.4)

For HPC and AI applications, the trend to use 3D-die staking and hybrid bonding are key co-design challenges. Warpage and stress control must be assessed at the early design stages for challenges for 3D hybrid bonding as interconnect pitches scale and alignment tolerances tighten. Even small out-of-plane deformation can degrade bonding uniformity, leading to voids, incomplete contact, and reduced yield [13.4.5].

These issues stem from coupled thermo-mechanical effects across heterogeneous materials, including CTE mismatch and copper pad topography. Thermal steps during bonding further amplify stress, increasing the risk of delamination, cracking, and interfacial defects. Effective mitigation requires integrated co-design approaches combining accurate material modeling, process optimization, and layout strategies to control both global wafer warpage and local surface planarity.

13.4.3 Warpage and stress challenges for packages in the field

For AI and high-performance computing (HPC) systems incorporating large SoC dies and chiplet-based architectures, elevated power densities and the resulting thermal loads drive significant thermo-mechanical stress. These stresses propagate across the chip–package–board hierarchy and can induce multiple failure modes, including solder bump fatigue, RDL cracking, interfacial delamination, electro-migration, and device-level performance degradation due to stress-induced carrier mobility shifts within the silicon.

Addressing these challenges necessitates a holistic chip–package co-design paradigm that concurrently considers electrical, thermal, and mechanical interactions. Finite element–based modeling has been widely employed to evaluate thermo-mechanical response under thermal excursions (e.g. JEDEC-JESD22 test standards such as thermal cycling, etc); however, many existing approaches assume stress-free initial conditions following package fabrication. This assumption is increasingly untenable for advanced architectures, where residual stress from prior processing steps can significantly influence system behavior. Accordingly, future methodologies must incorporate process-induced stress histories and rely on accurate, temperature-dependent material properties to enable predictive and reliable assessment.

13.4.4 Co-Design needs and potential solutions

Advanced packaging requires modeling frameworks - spanning both physics-based and AI/ML-driven approaches - to enable integration with electronic design automation (EDA) flows, ensuring that early-stage design is simultaneously informed by electrical, thermal, and mechanical considerations. Such capabilities can also inform the development of Assembly Process Design Kits (APDKs), providing a critical link between design intent and manufacturing execution. Recent demonstrations in fan-out wafer-level packaging have illustrated this paradigm, where die shift is monitored during processing and fed back into EDA environments to enable accurate redistribution layer (RDL) formation [13.4.6], [13.4.7].

In-field performance of AI and HPC systems will increasingly depend on the availability of fast, physics-informed simulation frameworks capable of diagnosing and prognosticating thermally induced degradation in advanced packages. These tools must account for the coupled effects of operational loads and environmental conditions encountered during service. Such capabilities are essential to enable predictive reliability assessment, real-time health monitoring, and informed maintenance strategies at the package and system levels.

Looking forward (see Table 13.4.1), there is a clear need to extend these capabilities toward rapid, layout-aware assessment of multiphysics performance. Physics-informed machine learning [13.4.8] offers a promising pathway to accelerate such evaluations, enabling efficient exploration of design spaces while retaining physical fidelity. These approaches can further underpin the development of digital twins, supporting predictive control, process optimization, and assurance of package- and system-level quality and reliability.

Table 13.4.1: Estimated schedule for mechanical workflow components.

Key attribute	Near term	Mid term	Long term
EDA Co-design workflows and APDK's	Electrical and thermal informed package layouts	Electrical, thermal, and mechanical informed package layouts	Fast holistic chiplet-package co-design across manufacturing processes, materials, architecture, cooling, and workloads

13.5 Testability, Resilience, Reliability and Security

As the semiconductor industry transitions from monolithic System-on-Chip (SoC) designs toward heterogeneous integration (HI) using 2.5D and 3D architectures, the traditional boundaries of Design-for-Test (DFT) are being fundamentally challenged. While HI enables unprecedented bandwidth, energy efficiency, and functional specialization through chiplets, it also introduces a severe test wall. This test wall arises from the growing mismatch between the physical density of interconnects and our ability to electrically access, stress, and diagnose them using conventional test infrastructures. In chiplet-based ecosystems, failures can originate at multiple abstraction levels, such as front-end-of-line (FEOL) devices, back- end-of-line (BEOL) routing, micro-bumps, hybrid bonds, TSVs, or even optical interfaces, yet the cost of isolating these failures grows superlinearly with integration density. As a result, test can no longer be treated as a post-design activity. Instead, testability must co-evolve with packaging technology, power delivery, thermal management, and system-level redundancy strategies.

13.5.1 Physical Access, Probing Limits, and PDN Constraints

The scaling of micro-bump and hybrid-bond pitches is significantly outpacing advances in mechanical probe technology. While interconnect pitches are rapidly approaching the sub-µm regime [13.5.1, 13.5.2], probe card capabilities remain fundamentally constrained by pitch, mechanical tolerances, probe wear, and contact force limits [13.5.3]. Table 13.5.1 summarizes the reported evolution of probe tip radius and probe pitch in state-of-the-art wafer probing technologies. As a consequence, full probing of all signal, power, and ground bumps during wafer sort or Known Good Die (KGD) testing is no longer feasible. Therefore, unless probe pitch scales commensurately with micro-bump pitch, achievable fault coverage will degrade, directly undermining the reliability and validity of KGD qualification. To mitigate this, industry practice is shifting toward selective probing strategies in which only a subset of bumps is contacted and exercised, as probing every micro-bump becomes increasingly infeasible with shrinking interconnect pitches and escalating probe design and cost constraints [13.5.4]. However, this selective probing approach introduces additional challenges and trade-offs that must be carefully considered.

Table 13.5.1: Evolution of Probe Tip Radius and Probe Pitch Over Time

Year	Probe tip radius	Probe pitch	Source
2018	10 μm	—	[13.5.5]
2019	4.5 μm	45 μm	[13.5.6]
2023	4.3 μm	18.5 μm	[13.5.7]
2025	1.2 μm	—	[13.5.8]
2025	500 nm	2 μm	[13.5.9]

A particularly severe consequence of partial probing is the degradation of Power Delivery Network (PDN) integrity during test. When only a fraction of power and ground bumps are contacted, the effective PDN impedance seen by the die is significantly higher than in the final packaged configuration. This prevents the application of high-current workloads and full-speed switching activity during test, masking IR-drop, droop-induced timing failures, and marginal PDN designs. As a result, a die may pass wafer-level tests under benign conditions but fail catastrophically once integrated into a full 2.5D or 3D stack. This probing scaling gap highlights the need for alternative access mechanisms, including embedded test infrastructures, current-stressing techniques that do not rely on full probing, and architectural assumptions that explicitly account for limited external access.

13.5.2 Test Logic Scalability, Routing Mismatch and Thermal Logistics

Embedding test logic in HI designs faces severe scalability constraints. Standards such as IEEE 1838 [13.5.10], ATE-based schemes [13.5.11, 13.5.12], and conventional BIST architectures [13.5.13–13.5.15] implicitly assume that test circuitry can be placed behind each interface with manageable area overhead. In modern HI designs with millions of interconnects, this assumption no longer holds. Placing dedicated test cells per bump or per lane can consume an unacceptable fraction of FEOL area. As routing distances from test logic to bumps vary across the die, non-uniform parasitic delays are introduced. In ultra-high-speed interfaces, picosecond-scale path mismatches can obscure the distinction between true physical defects and process-induced delay variations, complicating both test interpretation and diagnosis.

These challenges are further exacerbated at the KGD stage, particularly in multi-vendor HI ecosystems. Once wafers are singulated, controlling the temperature of individual dies in a packaged stack becomes significantly more difficult than wafer-level thermal management on a chuck [13.5.16]. Thermal gradients and uncontrolled die temperatures during test can lead to false failures, excessive guardbanding, or yield loss. Moreover, insufficient screening at KGD dramatically increases diagnostic complexity at the package level, where isolating the faulty die or interface in a dense 3D stack becomes prohibitively expensive.

13.5.3 Multi-Domain Integration, Reliability Stressing, and Test Economics

Heterogeneous integration increasingly spans multiple physical domains, including logic, memory, analog, RF, and silicon photonics (SiPh). SiPh introduces optical interfaces whose test requirements differ fundamentally from electrical DFT flows. Performing optical alignment, calibration, and characterization at the KGD stage significantly increases test cost and complexity yet deferring these tests until package assembly risks catastrophic yield loss. This tension demands new cross-domain test methodologies that minimize total cost of test while preserving confidence in incoming die quality.

Beyond functional correctness, HI introduces unique reliability risks associated with mechanical and thermal stress. Structures such as micro-bumps, hybrid bonds, and TSVs are susceptible to latent defects driven by electromigration, thermo-mechanical fatigue, and stress-induced voiding. Conventional logic-oriented test patterns, dominated by simple 0-to-1 transitions, are insufficient to excite these failure mechanisms. Instead, HI demands new stress-oriented test patterns capable of intentionally hammering interconnect structures through controlled current surges, thermal cycling, and activity shaping.

Finally, the economics of test and repair are shifting. As interconnect counts reach the millions, the traditional test-diagnose-repair paradigm becomes increasingly impractical. For ultra-fine pitches, it is often more cost-effective to embrace redundancy rather than precise diagnosis. Strategies such as spare bumps, double bumping, or over-provisioned interconnect fabrics can trade modest area overhead for dramatic improvements in yield and test simplicity. When bandwidth requirements remain fixed but connection density increases, redundancy-based design becomes a powerful lever to relax test requirements and mitigate defect sensitivity. Collectively, these trends signal a fundamental shift in how test, reliability, and economics must be co-optimized in future HI systems.

Table 13.5.2: Testability, Resilience, and Reliability Metrics Roadmap

Metric	2026	2028	2030	2032	2034
Latent Defect Detection	Incidental	Opportunistic	Systematic	Proactive	Predictive (Probabilistic)
Thermo-mechanical Awareness	None	Package-level	Chiplet and package	Runtime - Aware	Feedback-Driven (Self-Adaptive)
Fault Tolerance Approach	Within Chiplet	Chiplet-to-Chiplet Interface	Chiplet-Level	Multi-Chiplet Domain	System-Level
AI's influence on DFT (%)	20	30	50	60	70
Reliability Margining	Fixed Guard-bands	Adaptive	Workload-Aware	Aging-Aware	Self-calibrating

13.5.4 Security

System Architecture/Chip Partitioning

Modern SoCs increasingly rely on multi-die integration, where compute, I/O and memory interface dies are integrated within a single package to improve system level efficiency and performance. This integration model introduces notable security challenges, since the added interfaces and protocol layers broaden the potential attack surface without robust protection. In response, the industry is incorporating security mechanisms directly into die-to-die interconnects. Technologies such as External Global Memory Interconnect (XGMI) introduce authenticated link encryption [13.5.18]. Platform security engines exemplified by architectures such as Converged Security and Management Engine (CSME) [13.5.19] and AMD Secure Processor (ASP) [13.5.20] provide hardware-based trust anchors, verify the integrity of firmware and manage the cryptographic keys required to authenticate participating components in inter-die interactions. Complementing these measures, security controllers and security-focused intellectual property such as cryptographic accelerators, attestation modules, authentication controllers and lifecycle management units are incorporated into the system architecture to reinforce the trust model and support hardware-enforced access control, secure provisioning and continuous integrity validation.

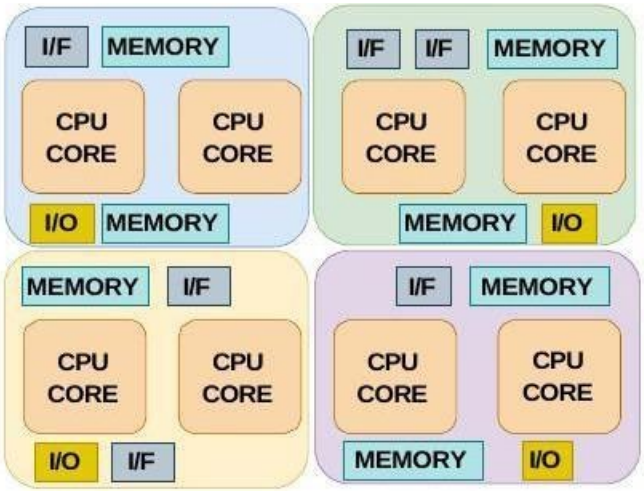


Figure 13.5.1: Multi-die system showing distribution of compute, memory, I/O and interface blocks across dies (adapted from [13.5.17]).

Front-End Design/NoC Design

Ensuring security in the front-end design stage is increasingly supported by industry adoption of shift left practices, in which security requirements and threat analyses are introduced at the earliest phases of architectural definition and RTL development. Simulation- based security checks are used to identify unauthorized information flows, privilege escalation paths and functional conditions that may result in exploitable behavior before synthesis and physical implementation. Tools such as the Codax platform [13.5.21] from Caspia Technologies support this process by detecting security-relevant defects directly in RTL, correlating them with known vulnerability classes defined in CWE [13.5.22] and documented instances in CVE [13.5.23] databases, and providing actionable guidance for remediation. Additional security assurance technologies, including formal verification engines, information flow analysis, structural code assessment tools and automated review frameworks, further strengthen the ability to identify and resolve weaknesses during the design phase, thereby reducing the likelihood of defects propagating into later stages of the SoC development lifecycle.

Interfaces/D2D Interfaces

Modern die-to-die interfaces are being extended with explicit security primitives to protect package-level communication and management. The Universal Chiplet Interconnect Express (UCIe) [13.5.24] specification builds on Peripheral Component Interconnect Express (PCIe) [13.5.25] and Compute Express Link (CXL) [13.5.26] foundations and includes manageability and security constructs for package-level links, with successive releases expanding those capabilities. PCIe has also introduced link security enhancements and related engineering change notices that add protections such as selective header encryption and mechanisms to reduce exposure from downstream devices, and platform DMA remapping and IOMMU usage remain primary defenses against malicious DMA. Compute Express Link provides an explicit integrity and data encryption capability for its cache and memory channels, with CXL 2.0 introducing Integrity and Data Encryption (IDE) style protection for Flow Control Unit (FLIT) level confidentiality, integrity and replay resistance. New interconnects targeted at accelerator fabrics such as UALink [13.5.27] include a security layer in their initial specifications, for example UALinkSec in the UALink 1.0 specification, reflecting industry attention to authenticated encrypted die-to-die transport in high scale accelerator pods.

Table 13.5.3: Security Metrics Roadmap

Metric	2026	2028	2030	2032	2034
Telemetry Sensor Coverage	On-die sensors	Expanded on-chip sensing	Cross-chiplet visibility	Full-system coverage	Pervasive, fabric-wide

Chiplet Trust Domain Enforcement for Multi-tenant Systems	Basic access controls	Stronger isolation mecha- nisms	Coordinated cross-chiplet controls	Automated enforcement	Predictive, adaptive controls
Fault vs. Attack Classification	Heuristic rules -assisted	ML ana lysi s	Context- aware correlation	Real-time automated decisions	Predictive and AI-assisted
Physical Tampering Detection	Voltage/clock glitch detection	Multi-vector sensing	Correlated tamper signals	Proactive detection	Predictive tamper modeling

13.6 Systems Architecture, Partitioning, and Network- on-Chips

Systems architecture is master planning a product to connect its business needs with technical implementation. Figure 13.1.1 below describes multiple steps involved in architecting a system from its specification to the hand-off for design implementation. Modeling and model-order-reduction are critical to achieving this.

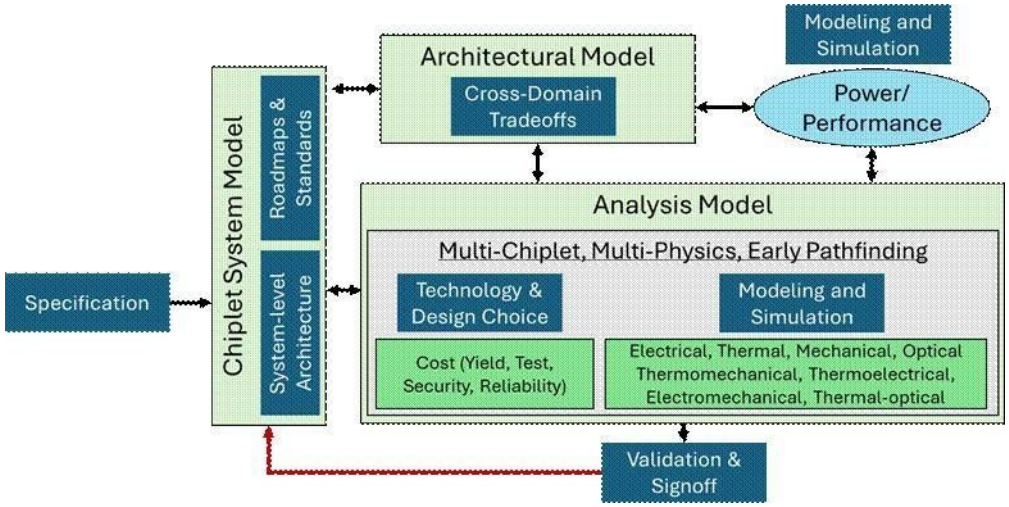


Figure 13.6.1: Systems Architecture—Specification to Implementation Handoff.

Heterogeneous integration and advanced packaging of multiple silicon dies, whether side- by-side, vertically stacked, or a combination thereof, is driving the need for new layers of abstractions and model representations of designs to support shift-left strategies. New abstraction layers relating to the systems architecture tasks must allow order reduction, be able to generate downstream models, and be able to act upon actionable inputs back from downstream analyses.

Table 13.6.1: System Architecture Codesign Roadmap

Metric	2026	2028	2030	2032	2034
Early system-scale multiphysics analysis	Power, thermal, thermo-mechanical stress, electro-thermal (electromigration, etc). Separate tools. O(days) full system runtime.	Single analyses cockpit. O(hours) full package runtime	Scalable multiphysics analyses for multi-layer 3D stacks, complex silicon and non-silicon substrates	Accurate, abstracted compact multiphysics models for O(minutes) full package runtime	Scalability to waferscale and panel-scale systems
Multi-Chiplet Specification Standards	Physical-only Assembly and Chiplet Design Kits	Materials, Thermal, Power added for multiphysics analyses	Test and repair added	Yield, Reliability added	Security added
Network in package	Point-to-point, mesh	Fault tolerance (bonding/chiplet failures)	Complex topologies (taurus, switched, hierarchical etc)	Heterogeneous within package interconnect support: photonic, serial/parallel electrical	Software-defined reconfigurable networks
System Partitioning and early pathfinding	Manual	Visualization, editing and analyses in a single cockpit with multiphysics analyses	Yield, cost, reliability awareness; semi-automatic design space exploration including assessment of impact of packaging/cooling/PDN technologies	Semi-automated block/chiplet fault tolerance/redundancy insertion. Scalability to multi-layer 3D stacks and ~50 chiplets per package.	Automated partitioning integrated with (third party) system performance models

13.7 Die-to-Die Interface

The rapid evolution of Artificial Intelligence (AI) applications is driving the need for disaggregating System-on-Chip (SoC) architectures as monolithic die hit reticle limit. The next generation of AI and HPC systems will require low latency, high bandwidth, power efficient die-to-die interconnects to facilitate large amounts of data transfer between chiplets.

13.7.1 Die-to-Die Interconnect Use Cases

System disaggregation into chiplets introduces new communication requirements across compute, memory, and I/O domains. Each of these use cases imposes distinct latency, bandwidth, and protocol requirements that must be addressed by a die-to-die (D2D) interconnect architecture.

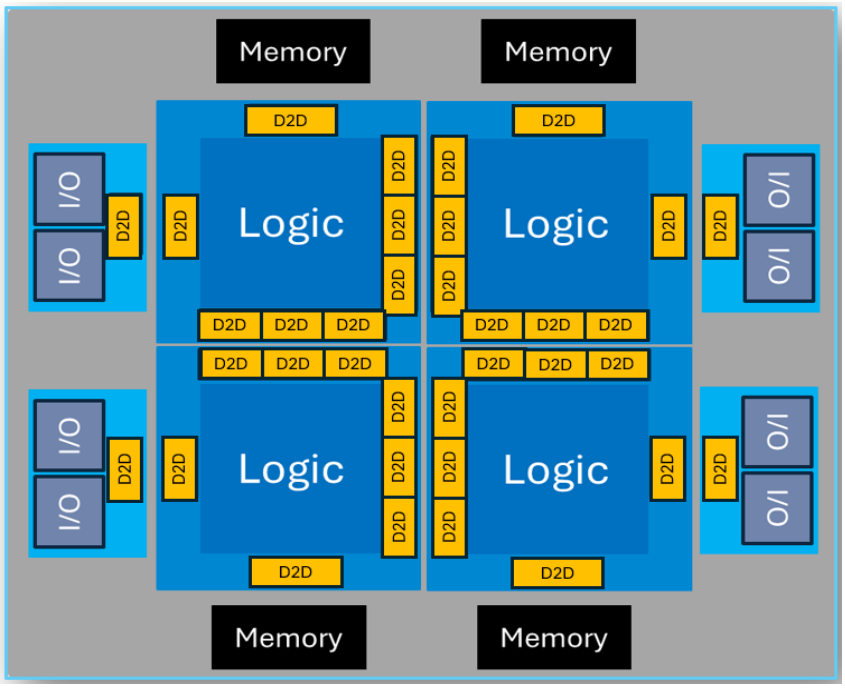


Figure 13.7.1: Disaggregated SoC Architecture.

13.7.2 Compute-to-Compute

Disaggregated compute architectures require high-bandwidth, low-latency links to preserve the performance of a unified fabric. Traffic that once traversed an on-die NoC now crosses die boundaries through D2D interfaces, driving bandwidth scaling from 32 Gbps to 64 Gbps and beyond. To maintain coherency and QoS, D2D links typically employ a streaming layer that transports standard protocols such as AMBA AXI, ARM CHI C2C, or proprietary interfaces, enabling seamless cache and memory sharing across compute tiles.

13.7.3 Compute-to-Memory

When disaggregating memory interfaces like HBM or DDR, the D2D interconnect must provide bandwidth density comparable to direct attachment. A D2D subsystem can be configured for HBM-class throughput in Advanced Package (AP) implementations or DDR- class bandwidth in Standard Package (SP) designs. Reusing the same D2D shoreline for memory connectivity eliminates large on-die PHYs, improving bandwidth-per-millimeter (BW/mm), silicon efficiency, and total energy per bit.

13.7.4 Compute-to-I/O

AI and HPC systems integrate multi-terabit I/O for both scale-up (intra-system) and scale-out (inter-system) communication. Scale-up paths within a package currently demand 64 Gbps-class D2D links implemented over advanced or organic packages, depending on routing density and loss constraints. Scale-out I/O connects to external hosts or switch fabrics using similar signaling and protocol stacks. A unified D2D architecture therefore enables scalable compute, memory, and I/O integration for heterogeneous AI platforms.

13.7.5 Evolution from 2.5D to 3D Packaging for Die-to-Die Connectivity

The transition from 2.5D to 3D integration signifies a major advancement in the design and optimization of die-to-die (D2D) interconnects, enabling higher performance, greater interconnect density, and improved energy efficiency. While both enable heterogeneous integration across multiple dies, they rely on distinct interconnect structures and design trade-offs. In practice, 2.5D microbump-based links provide millimeter-scale lateral routing for flexible chiplet placement, whereas 3D hybrid-bonded links offer <10 µm hybrid copper bond vertical connections with significantly higher bandwidth efficiency. These two classes will coexist, each serving different latency, density, and packaging requirements.

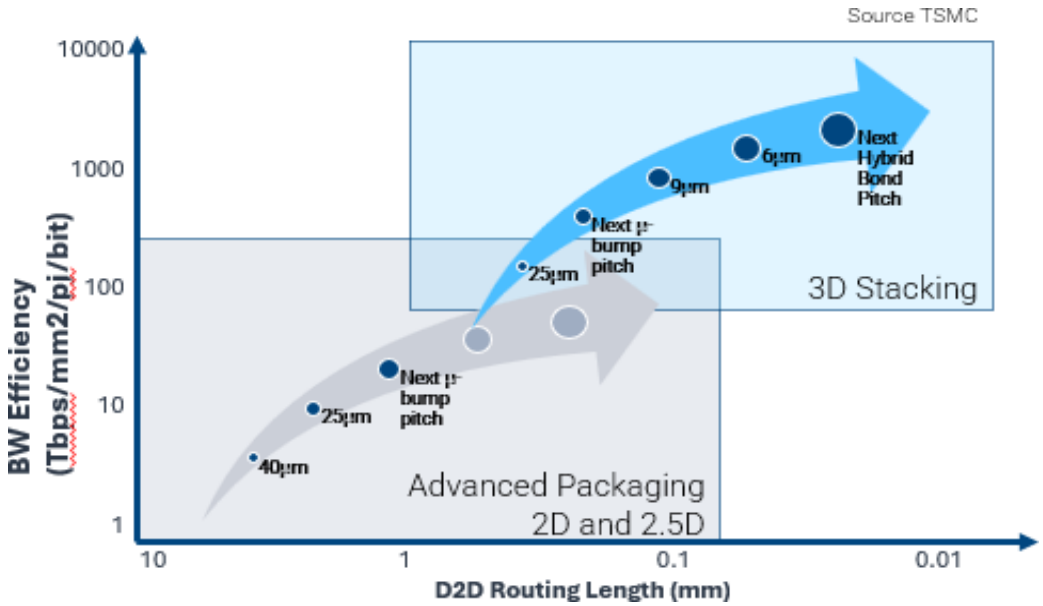


Figure 13.7.2: Illustrative Bump Pitch Scaling Trend in 2.5D and 3D Packages.

2.5D packaging leverages a silicon or organic interposer to provide lateral interconnects between chiplets placed side by side. The interposer serves as a high-density redistribution layer (RDL) platform that enables fine-pitch wiring (typically 1–5 µm line/space for silicon interposers and 10–15 µm for organic substrates). Through-Silicon-Vias (TSVs) in the interposer deliver vertical connectivity to package-level power and I/O planes, while microbumps at the top surface provide die-to-die interconnect terminations.

To address bandwidth density and latency limits of lateral 2.5D interconnects, 3D stacking introduces vertical die-to-die connectivity through hybrid bonding technologies. In this scheme, multiple active dies are directly bonded face-to-face (F2F) or face-to-back (F2B) with sub-10 μm pitch interconnects, eliminating the need for microbumps.

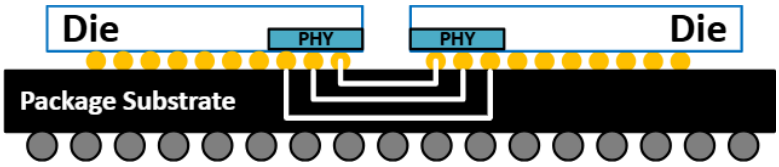


Figure 13.7.3: Standard Package (Organic Interposer) Die-to-Die Interconnect.

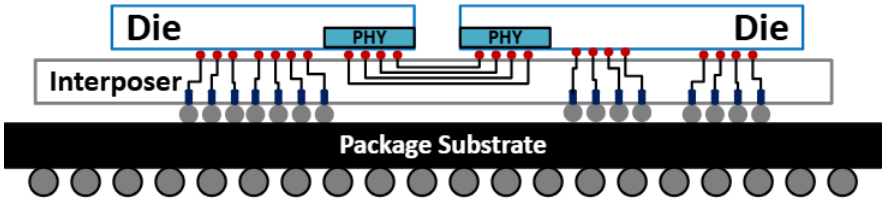


Figure 13.7.4: Advanced Package (Silicon Interposer) Die-to-Die Interconnect.

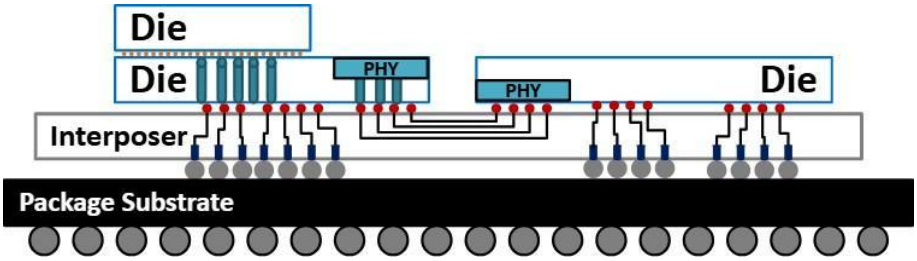


Figure 13.7.5: 3D Stacked Die-to-Die Interconnect.

Table 13.7.1: 2.5D Die-to-Die Interconnect Metrics Projection

	Unit	2025		2028		2031		2034	
		D2D-SP	D2D-AP	D2D-SP	D2D-AP	D2D-SP	D2D-AP	D2D-SP	D2D-AP
Per-wire Data Rate	Gbps	32	32	64	64	128	128	128	128
Tx+Rx (# of IOs)		64	128	64	128	64	128	64	128
Number of Modules		2	2	2	2	2	2	2	2
Bump Pitch	μm	110	45	110	45	110	45	90	25
Bandwidth Density	Tbps/mm^2	1.03	6.31	2.06	12.61	4.13	25.22	4.13	25.22
Beachfront BW Density	Tbps/mm	1.70	9.14	3.40	18.29	6.81	36.57	6.81	36.57
Energy/bit	pJ/b	0.4	0.3	0.5	0.4	0.4	0.3	0.3	0.2

Table 13.7.2: 3D Die-to-Die Interconnect Metrics Projection

	2025	2028	2031	2034
	D2D-3D	D2D-3D	D2D-3D	D2D-3D
Per-wire Data Rate (Gbps)	4	4	6	8
Cluster Width	80	80	80	80
Bump Pitch (μm)	6	4	4	2
Bandwidth Density (Tbps/mm^2)	8.89	20.00	30.00	160.00
Energy/bit (pJ/b)	0.07	0.05	0.03	0.02

For AI workloads the compute-to-memory path is often the most performance-critical connection in the entire system. AI inference and training workloads place extremely high pressure on memory bandwidth, latency, and efficiency. And that has led many AI and HPC teams to develop custom HBM architectures that go beyond what the standard HBM sub-system provides. One of the major drivers for customization is that traditional HBM integration is limited by die shoreline. You only have so much perimeter to expose high- bandwidth connections. As AI compute dies scale in size, that shoreline becomes a bottleneck. This is where combining die-to-die links with custom HBM becomes really powerful. You can reuse the same die edge that is used for D2D connectivity to also attach memory thus allowing you to scale memory bandwidth without growing the compute die or increasing PHY overhead. Importantly, the custom memory interface roadmap will closely follow D2D scaling in lane speed and bandwidth density to sustain compute-to-memory performance.

Table 13.7.3: Standard HBM4 vs. Custom HBM4 Using 64G D2D Link

Parameter	Standard HBM4	Custom HBM4 using 64G D2D Link
Data Width [#]	2048	64
Data Rate [T/s]	10G	64G
BW/Density [Tbs/mm]	2.6 Tbs/mm	21.1 Tbs/mm (~8X)
Power [pJ/b]	0.36 pJ/b	0.35 pJ/b
Reliability (BER)	Comparable	E-15

As we move from compute-to-memory into compute-to-IO, the challenge becomes scaling bandwidth beyond what traditional electrical links can support, especially in large AI clusters. Scale-up architectures increasingly rely on a combination of high-speed electrical and optical interfaces. Passive copper cables are the most energy-efficient choice, but their reach drops significantly at high symbol rates of 224G. Larger AI pod architectures require longer reach, forcing designers to insert more retimers. That solves the reach problem but increases power and system cost. This is where optical begins to play a major role.

We are seeing an increased adoption of slow-wide interfaces for optical connectivity in scale-up networks, which leverage a wide, lower-speed electrical interface feeding an integrated optical DSP placed inside the transceiver. In this model, the traditional PHY effectively migrates into the optical module, making D2D connectivity essential. This architecture is foundational for next-generation 2D and 3D co-packaged optics (CPO). Advanced CPO solutions bring optics directly adjacent to the compute die, enabling photonic integration at the package level.

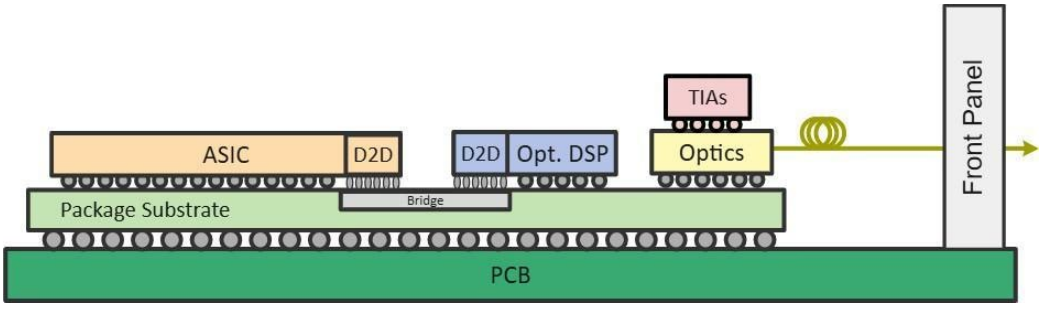


Figure 13.7.6: Co-packaged optics and scale-up connectivity for AI platforms.

13.8 Physical Design for Heterogeneous Integration

13.8.1 Technology and Systems Focus

Physical design for heterogeneous integration provides the layout foundation that enables multiple dies, technologies, and materials to operate as a unified system, translating architectural intent into manufacturable and reliable layouts [13.8.1, 13.8.2, 13.8.18]. As monolithic scaling slows, heterogeneous integration has emerged as a primary path to sustain performance, bandwidth, and energy-efficiency improvements across computing platforms [13.8.1, 13.8.8, 13.8.18]. Within this landscape, 2.5D and 3D ICs play complementary but distinct roles, each imposing different physical-design requirements and aligning with different system objectives [13.8.4, 13.8.20].

2.5D IC physical design primarily targets AI and high-performance computing (HPC) applications, where performance scaling is driven by massive parallelism, high memory bandwidth, and aggressive power delivery [13.8.6, 13.8.10, 13.8.11]. In this context, "IC" refers not to a monolithic die, but to a heterogeneous system integrating multiple chiplets, memory stacks, and supporting dies through advanced packaging. Large silicon or glass interposers enable wide, low-latency lateral connectivity among many chiplets, supporting hundreds of chiplets and millions of inter-chiplet links in future systems [13.8.7, 13.8.18]. Physical design in this context emphasizes interposer-scale floorplanning, high-density redistribution layer (RDL) routing, and robust power delivery network (PDN) construction to maximize bandwidth density while maintaining signal integrity, power integrity, and thermal scalability [13.8.7, 13.8.12, 13.8.15, 13.8.21].

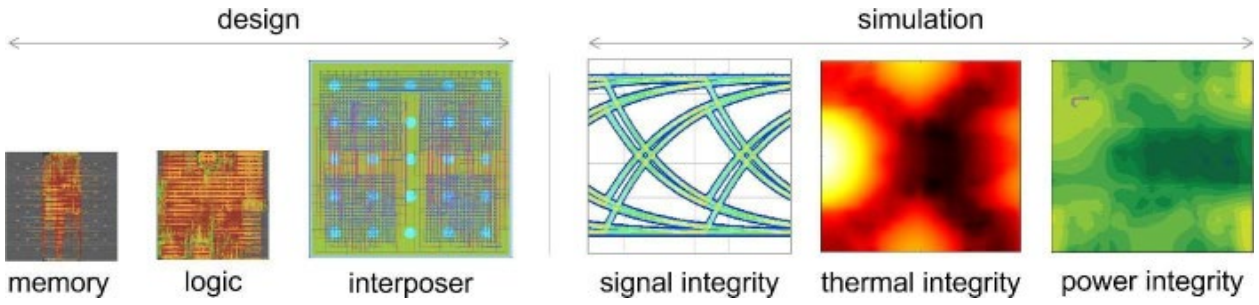


Figure 13.8.1: Representative 2.5D IC physical-design, illustrating chiplet design, interposer design, and associated multi-physics analysis [13.8.21].

In contrast, 3D ICs take the form of vertically integrated chips that are either integrated into a larger 2.5D platform or deployed as standalone devices [13.8.4]. In this context, 3D IC physical design emphasizes tier-aware placement, dense vertical interconnects such as TSVs and hybrid bonds, and rigorous thermal and mechanical control to realize compact, energy-efficient building blocks [13.8.5, 13.8.8, 13.8.22]. As standalone devices, these vertically integrated ICs are well suited for mobile and energy-constrained applications, including logic-memory stacking, mixed-signal integration, and form-factor-limited subsystems [13.8.9, 13.8.14, 13.8.22]. When incorporated into a 2.5D interposer-based system, 3D chiplets combine the energy efficiency and compactness of vertical integration with the scalability and bandwidth advantages of 2.5D integration, as exemplified by modern NVIDIA GPU platforms in which a large 2.5D GPU system is co-packaged with 3D-stacked HBM chiplets on a shared interposer [13.8.6, 13.8.21].

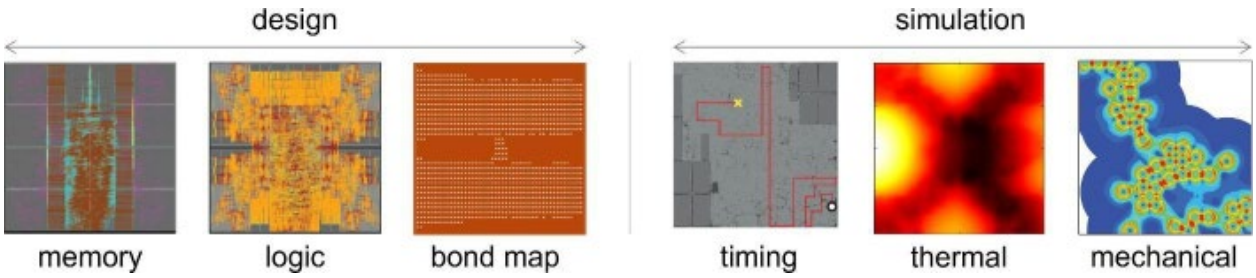


Figure 13.8.2: Representative 3D IC physical-design, illustrating memory tier design, logic tier design, inter-tier routing, and associated multi-physics analysis [13.8.19].

13.8.2 Technology and Systems Attributes

2.5D IC Physical Design

Physical design for 2.5D ICs focuses on optimization in the (x, y) plane and is dominated by large-area lateral integration on an interposer, where chiplet floorplanning, micro-bump placement, and high-density RDL routing are the primary challenges [13.8.6]. As interposer dimensions grow from tens to hundreds of millimeters and chiplet counts scale from single digits to hundreds, electrical routing congestion across wide interposer fabrics becomes a first-order concern [13.8.12]. Long lateral interconnects introduce parasitics and crosstalk that directly impact bandwidth and energy efficiency, requiring accurate extraction models and routing-aware optimization [13.8.7, 13.8.10, 13.8.12].

Robust PDN design is equally critical, as large AI and HPC systems demand high current delivery with tight IR-drop and electromigration constraints [13.8.11]. Thermal spreading across the interposer must be carefully managed to avoid hotspots and performance throttling [13.8.7, 13.8.15], while mechanical reliability issues such as die-to-interposer coefficient-of-thermal-expansion (CTE) mismatch and interposer warpage must be addressed during physical signoff [13.8.2].

3D IC Physical Design

Physical design for 3D ICs introduces a vertical dimension, fundamentally changing the EDA optimization problem [13.8.4, 13.8.5, 13.8.8]. Instead of planar (x,y) optimization, 3D ICs require simultaneous optimization across (x,y,z), where lateral and vertical decisions are tightly coupled. Tier-aware floorplanning and co-design of vertical interconnects such as TSVs and hybrid bonds are critical for performance and energy efficiency [13.8.3, 13.8.9]. Compared to 2.5D ICs, 3D ICs exhibit stronger coupling among routing, power delivery, thermal behavior, and mechanical stress, with limited vertical heat removal and new IR-drop and reliability challenges [13.8.11, 13.8.12].

Thermo-mechanical stress accumulation across bonding interfaces and TSVs further complicates physical design, necessitating early awareness of mechanical constraints rather than late-stage signoff checks [13.8.14, 13.8.22]. In addition, physical-design-driven DFT becomes significantly more complex, as test access and observability must be ensured across tiers, often requiring pre-bond test strategies and tier-aware test routing [13.8.14].

13.8.3 Future Projections

2.5D IC Physical Design

As summarized in Table 13.8.1, the evolution of 2.5D heterogeneous integration from 2026 to 2034 is characterized by aggressive scaling of both technology and EDA capabilities [13.8.18]. Interposer interconnect pitch is projected to shrink from 2–5 µm to 0.1–0.25 µm, while chiplet count grows from 8 to 512 and total inter-chiplet links increase from ~1K to ~20M [13.8.18]. Interposer sizes expand from 50×50 mm2 to 300×300 mm2, accompanied by a transition from silicon to glass substrates and from micro-bumping to hybrid bonding [13.8.2, 13.8.18].
These trends make EDA scalability the dominant enabler. Chiplet heterogeneity support increases from ~20% to 100%, while chiplet partitioning, placement, and chiplet–interposer routing evolve from manual or separate flows to fully automatic, concurrent optimization by 2030–2034 [13.8.12, 13.8.13]. Awareness of chiplet NoCs and clocking similarly rises toward 100%, enabling timing- and clock-correct closure across hundreds of chiplets [13.8.6, 13.8.17].
Multi-physics shift-left adoption grows from ~20% to ~60%, allowing thermal, PDN, and reliability constraints to be enforced during early physical design rather than at late signoff [13.8.7, 13.8.11]. These advances are essential to achieving the targeted reduction in constraint closure time from months to days by 2034. Design-for-Test evolves from interposer-only testing in 2026 to die-to-die BIST with repair and ultimately full-system DFT with on-chip monitors by 2034 [13.8.14]. AI-in-the-loop optimization, cross-vendor tool interoperability, full-stack database unification, and chiplet-marketplace interoperability all approach 80–100% maturity, positioning EDA as both the bottleneck and the key enabler for large-scale 2.5D systems targeting AI and HPC workloads [13.8.6, 13.8.13, 13.8.17].

2.5D IC Physical Design

As shown in Table 13.8.2, the evolution of 3D IC technology and EDA from 2026 to 2034 is driven by increasingly fine-grained vertical integration [13.8.3, 13.8.8, 13.8.18]. 3D interconnect pitch scales from ~5–10 µm to 0.1–0.5 µm, while the maximum number of vertical interconnects grows from ~1K to ~20M and tier count increases from 2 to 6 [13.8.18]. In parallel, device technology transitions from GAA to CFET and eventually atomically thin channel transistors, and bonding evolves from micro-bumping to hybrid bonding and monolithic integration [13.8.9, 13.8.18].

These trends initially exacerbate thermal resistance, peaking at ~1.5 °C/W around 2030, before improved PDN and thermal design reduce it back toward ~1.0 °C/W, with PDN impedance improving from ~5 mΩ to ~1 mΩ [13.8.11, 13.8.15]. On the EDA side, flows transition from die-by-die, memory-on-logic integration to fully concurrent, logic-on-logic co-design, with partitioning granularity tightening from die and module level to gate- and transistor-level [13.8.5, 13.8.8].

Placement, routing, and design closure evolve from die-by-die execution to concurrent multi-tier optimization by 2030 [13.8.5]. In parallel, multi-physics shift-left adoption increases from ~20% to ~60%, enabling early enforcement of PPA and reliability constraints and supporting the targeted reduction in constraint closure time from months to days. Die-to-die variation awareness grows from 0% to ~50%, reflecting the growing importance of vertical process and alignment variability [13.8.4, 13.8.14, 13.8.15]. Design-for-Test moves from post-bond to pre-bond strategies, and advanced capabilities such as AI-in-the-loop optimization, cross-vendor interoperability, full-stack database unification. 3D DRC/LVS maturity all rise toward 80–100%, underscoring that scalable 3D IC adoption is fundamentally gated by EDA's ability to manage fine-grained, vertically coupled physical design [13.8.4, 13.8.13, 13.8.19].

Table 13.8.1: Roadmap of 2.5D heterogeneous integration technology and EDA evolution from 2026 to 2034, showing scaling in interposer pitch, chiplet count, interposer size and materials, bonding technologies, physical performance metrics, and the increasing maturity of automated, multi-physics, and interoperable EDA flows. The HIR column refers to the corresponding chapters in the 2024 edition of the Heterogeneous Integration Roadmap [13.8.18].

	metric	HIR	2026	2028	2030	2032	2034
TECHNOLOGY	Interposer lateral interconnect pitch (micron)	Ch8, Ch22	[2, 5]	[1, 2]	[0.5, 1]	[0.25, 0.5]	[0.1, 0.25]
	Chip-to-interposer bump pitch (micron)	Ch8, Ch22	[40, 55]	[25, 40]	[10, 25]	[5, 10]	[2, 5]
	Chiplet count	-	8	32	128	256	512
	Total inter-chiplet link count	-	1K	50K	1M	5M	20M
	Interposer dimension (mm)	Ch8, Ch23	50x50	100x100	150x150	200x200	300x300
	Interposer material	-	silicon	silicon	silicon + glass	glass	glass
	Chiplet bonding technology	-	micro bumping	micro bumping	hybrid bonding	hybrid bonding	mix
	2.5D ADK accuracy	-	0.6	0.7	0.8	0.9	0.95
	2.5D Thermal resistance (°C/W)	Ch20	1.0	0.8	0.6	0.5	0.4
	2.5D PDN impedance (mΩ)	Ch22	5.0	3.0	2.0	1.0	0.5
EDA	Chiplet heterogeneity (%)	-	20	40	60	80	100
	Chiplet partitioning	-	manual	semi	auto	auto	auto
	Chiplet NoC/protocol awareness (%)	-	20	40	60	80	100
	Chiplet floorplanning	-	manual	semi	semi	auto	auto
	Chiplet/interposer routing	-	separate	separate	together	together	together
	Clocking across chiplets (%)	-	20	40	60	80	100
	Chiplet/interposer co-design closure	-	separate	separate	together	together	together
	Multi-physics shift-left (%)	-	20	30	40	50	60
	Constraint closure time	-	months	weeks	weeks	days	days
	Design-for-Test	-	interposer only	KGD + interposer	die-to-die BIST	BIST + repair	full-chip DFT + monitors
	AI-in-the-loop (%)	-	20	40	60	80	100
	Cross-vendor tool interoperability (%)	-	20	40	60	80	100
	Full-stack DB unification (%)	-	20	40	60	80	100
	Chiplet interoperability (%)	-	20	40	60	80	100
	2.5D DRC/LVS maturity (%)	-	20	40	60	80	100

Table 13.8.2: Roadmap of 3D IC technology and EDA evolution from 2026 to 2034, illustrating scaling in vertical interconnect pitch and count, tier depth, device and bonding technologies, and physical metrics, together with the transition of EDA from die-by-die flows to concurrent, fine-grained, multi-physics-aware 3D physical design and signoff. The HIR column refers to the corresponding chapters in the 2024 edition of the Heterogeneous Integration Roadmap [13.8.18].

	metric	HIR	2026	2028	2030	2032	2034
TECHNOLOGY	3D vertical interconnect pitch (micron)	Ch22	[5, 10]	[1, 5]	[0.5, 1]	[0.25, 0.5]	[0.1, 0.25]
	3D interconnect count (max)	Ch22	1K	50K	500K	5M	20M
	Number of logic tiers	Ch20	2	2	4	4	6
	Transistor type	-	GAA	GAA	CFET	CFET	atomically thin
	Bonding technology	-	micro bumping	micro bumping	hybrid bonding	hybrid bonding	monolithic
	3D ADK accuracy	-	0.6	0.7	0.8	0.9	0.95
	3D Thermal resistance (°C/W)	Ch20	1.0	1.2	1.5	1.3	1.0
	3D PDN impedance (mΩ)	Ch22	5.0	3.0	2.0	1.5	1.0
EDA	3D Architecture	-	memory-on-logic	memory-on-logic	logic-on-logic	logic-on-logic	logic-on-logic
	3D Partitioning granularity	-	die	module	module	gate	transistor
	3D Floorplanning		manual	semi	semi	auto	auto
	3D Placement	-	die-by-die	die-by-die	together	together	together
	3D Routing	-	die-by-die	die-by-die	together	together	together
	3D Design closure	-	die-by-die	die-by-die	together	together	together
	Multi-physics shift-left (%)	-	20	30	40	50	60
	Constraint closure time		months	weeks	weeks	days	days
	Die-to-die variation awareness (%)	-	0	20	30	40	50
	3D Design-for-Test	-	post-bond	post-bond	pre-bond	pre-bond	pre-bond
	AI-in-the-loop (%)	-	20	40	60	80	100
	Cross-vendor tool interoperability (%)	-	20	40	60	80	100
	Full-stack DB unification (%)	-	20	40	60	80	100
	3D DRC/LVS maturity (%)	-	20	40	60	80	100

13.8.4 Research and Development Directions

Despite rapid progress, current 2.5D IC physical design flows lack robust support for system- level chiplet–interposer co-design at scale [13.8.16, 13.8.20, 13.8.21]. A key research direction is the development of standardized physical abstractions for chiplets and interposers, including unified ADK schemas, interoperable layout and parasitic models, and consistent representations of power, thermal, and mechanical constraints across vendors [13.8.13, 13.8.18]. In addition, shift-left integration of multi-physics modeling remains limited, motivating research into early-stage, physically informed optimization frameworks that reduce late-stage iterations [13.8.7, 13.8.15]. Finally, physically-driven DFT at interposer scale requires standardized test-access and modeling infrastructures to support manufacturability and yield analysis [13.8.14]. Advancing open, tool-agnostic interfaces and standardized representations is essential for scalable and vendor-neutral 2.5D IC physical design [13.8.13, 13.8.21].

For 3D ICs, research directions must address the lack of standardized representations for multi-tier structures, including tiers, vertical interconnects, and bonding interfaces [13.8.5, 13.8.8]. New 3D ADK frameworks are needed to support concurrent multi-tier optimization and early evaluation of thermal, mechanical, and electrical interactions [13.8.4]. Equally important is the development of pre/during/post-bond DFT methodologies and open interfaces that enable multi-physics-aware verification and signoff across tools and vendors [13.8.14, 13.8.19].

13.8.5 Difficult Challenges

The fundamental challenges in 2.5D IC physical design arise from scaling lateral integration to unprecedented system sizes while maintaining closure across signal, power, thermal, and mechanical domains [13.8.11, 13.8.20]. Terabyte-scale layouts, dense interposer routing, and massive chiplet counts significantly strain existing algorithms and data models, making concurrent chiplet–interposer optimization and accurate early-stage multi-physics prediction intrinsically difficult [13.8.12, 13.8.13]. Ensuring cross-vendor interoperability for DRC/LVS and signoff further compounds these challenges as system complexity grows.

For 3D IC physical design, the challenges are driven by extreme vertical coupling and fine-grained partitioning, which fundamentally complicate concurrent multi-tier placement, routing, and closure [13.8.4, 13.8.5, 13.8.8]. Thermal constraints are particularly severe due to limited vertical heat removal, while accumulated thermo-mechanical stress across bonding interfaces directly impacts reliability [13.8.22]. Additional uncertainty from die-to-die variation and alignment errors introduces stochastic behavior that challenges deterministic optimization methods, necessitating probabilistic and variation-aware design techniques [13.8.14]. Without robust abstractions and scalable methodologies, these intrinsic difficulties will continue to limit large-scale 3D IC adoption [13.8.6, 13.8.14].

13.9 HI System Signoff

Verification and signoff for heterogeneously integrated (HI) designs present a number of major challenges. Many of these stems from the fact that these are indeed miniature systems and are often comprised of multiple physical layouts (ex. interposer and package) that may come from different sources. This section is broken down into the following categories:

- Layout vs. schematic checking
- Thermal integrity
- Electromagnetic extraction
- Power integrity, Signal integrity
- Manufacturing tolerance impact.

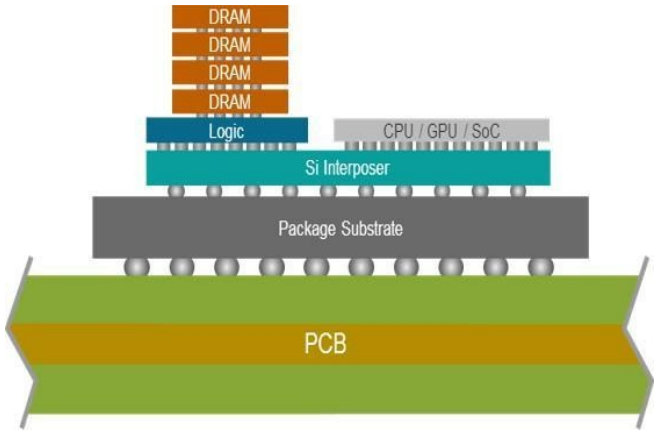


Figure 13.9.1: Example HI system.

13.9.1 Layout vs. Schematic Checking

Layout vs. Schematic (LVS) checking is usually the first verification activity, checking to see that the logical design data is in sync with the physical layout implementation. The key aspect of this is how the logic is captured. This can range from simple spreadsheets to a system-level schematic, to system planning tools that contain abstracts of the physical layouts. As engineers move into HI designs, it implies multiple die are present in the design. With multi-die designs, the complexity of the logical netlists increases rapidly, and it is highly recommended to move to commercial tools such as schematics or system planners and utilize the automated LVS capabilities available with those flows.

13.9.2 Thermal Integrity

One of the benefits of HI designs is the ability to pack multiple chiplets closely together. One byproduct of this is increased thermal challenges, making thermal analysis a necessity. The main objectives are to verify that the maximum junction temperatures of the bare die are not exceeded, and that the warpage specifications of the design are met. The general approach is to merge the multiple layout fabrics (ex. interposer, package) together into a single 3D model, then mesh and simulate. This is expected to generate heat maps, junction temperature reports, and warpage reports.

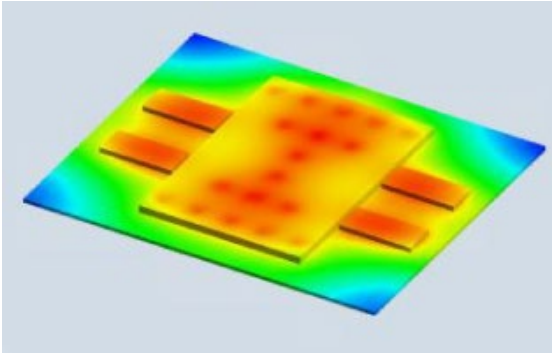


Figure 13.9.2: Example thermal map.

Some associated metrics are shown below:

Table 13.9.1: Thermal integrity metrics.

	2026	2028	2030	2032	2034	Notes
Package TDP	1200 W	2100 W	3200 W	6000 W	8500 W	Compute to go from 1 chip to multiple stacked dies and chiplets, HBM3 to HBM4 to HBM5 to increase stacks from 12 today to 20 by 2032
HBM power/stack	80 W	90 W	100 W	110 W	120 W or more	
Die Power Density	125 W/cm2	160 W /cm2	200 W/cm2	270 W/cm2	300 W/cm2	3nm to 2nm to 1 nm process; increased transistor density
Server Power	3200 W	5500 W	8000 W	14000 W	22000 W	
Cooling mechanisms	D2C	microfluidics	immersion	laser cooling	cryogenic	can be used in combination, EDA tools will need to model all these techniques
Theta JC	0.12 K/W	0.09 K/W	0.07 K/W	0.05 K/W	0.03 K/W	Improved cooling expected with D2C, TIM improvements (ex. graphene and indium)
Package Area	1000 mm2	1100 mm2	1300 mm2	1500 mm2	1600 mm2	
Warpage *	0.7 mm	0.8 mm	1.2 mm	1.6 mm	2 mm	If thermal concerns are not addressed, warpage will increase to the levels of causing solder joint failure in fewer cycles and delamination - greatly increasing cost per good product
Methodology	FEA of multi-fabric system + system level CFD	prev + MLMs for faster prediction	In-design optimization looking at PI + thermal + mechanical tradeoffs	AI agents directly working inside design tools to find new designs	AI agents working from top-level specs	How thermal analysis methodologies may evolve

Definition of terms:

- CFD; computational fluid dynamics
- D2C; direct-to-chip cooling, incl. cold plates
- FEA; finite element analysis
- MLM; machine learning model
- TDP; thermal design power
- TIM; thermal interface material

Looking at High Bandwidth Memory (HBM) as a driving factor, package TDP and power densities are all expected to increase significantly. Multiple cooling mechanisms are expected, including direct-to-chip cooling, increased use of microfluidics, full immersion, laser, and cryogenic cooling. Methodology is also expected to evolve from today's FEA-based approaches for the merged physical system to leverage machine learning, in-design optimization, and increased use of artificial intelligence (AI) agents.

13.9.3 Electromagnetic Extraction

Electromagnetic (EM) extraction is a key requirement for signoff, as it not only allows the engineer to assess the characteristics of the passive interconnect in the frequency domain, but it is also the enabler for signal and power integrity. Whereas the active chiplets are modeled with either IBIS (for SI) or die power models (PI), the passive interconnect between them is usually modeled with S-parameters generated from EM extraction.

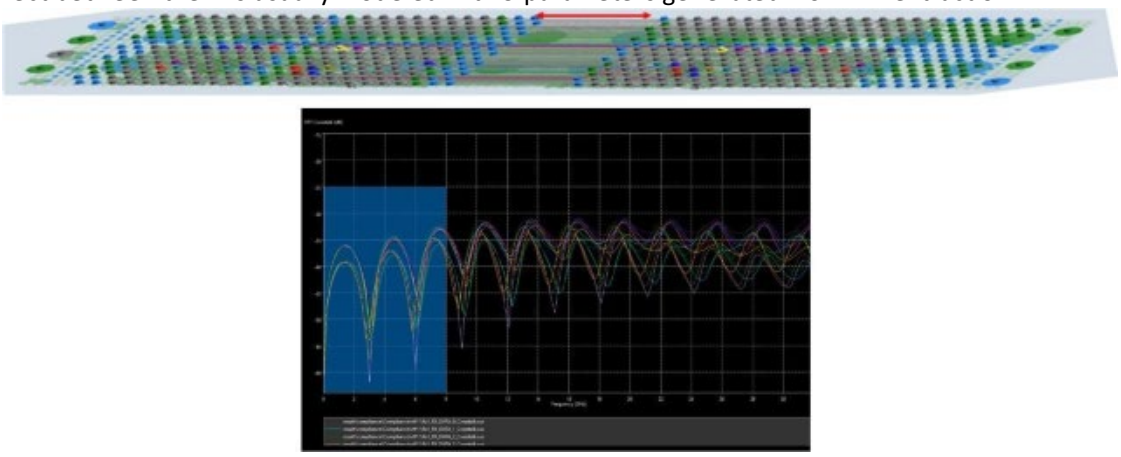


Figure 13.9.3: – UCIe interposer example with VTF crosstalk results

HI design has a huge impact on EM extraction. As seen in the table below, physical layout database size increases significantly with the number of chiplets increasing, the line and space geometries shrinking, and the power distribution networks (PDNs) becoming more complex with many degassing holes and cross hatch patterns. This all drives up the number of tetrahedral mesh elements to solve and overall computation times. Line data rate will also continue to increase, with the associated rise and fall times shrinking to fit within the smaller bit periods. To cover the bandwidth associated with these edge rates, EM extractions will be pushed to go higher and higher in frequency, again becoming more computationally expensive. The number of signals in die-to-die parallel buses is also expected to increase, meaning that to extract the S-parameters of an entire interface to enable signoff, the port count will be increasing significantly as well. It is predicted that this will have a major impact on the hardware requirements for full wave 3D solvers, moving from fairly large on-prem machines with hundreds of cores to thousands of cores in the cloud or compute farms, and eventually to GPU-based hardware.

Some associated metrics for EM extraction of single-ended parallel buses are shown below:

Table 13.9.2: EM extraction metrics for single-ended parallel bus interfaces

	2026	2028	2030	2032	2034	Notes
Layout database size (GB)	2	4	8	16	32	assume doubling every 2 years
SE line data rate (Gbps)	32	64	128	128	128	Based loosely on UCIe and projections
Bit period (ps)	31.3	15.6	7.8	7.8	7.8	1/data rate
Rise / fall time (ps)	8.3	4.2	2.1	2.1	2.1	Assume Trise is 0.2 UI, convert 20/80 to 10/90
Extraction bandwidth (GHz)	42.0	84.0	168.0	168.0	168.0	Estimate from rise time
# UCIe modules	4	2	2	2	2	
# Signals (64 / module)	256	128	128	128	128	assuming advanced package, 4 modules in 2026, then down to 2 as data rates increase
Port count	512	256	256	256	256	Lane is pair of UCIe signals, 1 unidirectional Tx signal and Rx signal
# of mesh elements	20M	40M	80M	160M	320M	assume tracks with database size, driven by PI extraction
Compute platform	on-prem, 128 CPU cores	on-prem, 512 CPU cores	cloud, CPU, 1028 cores	on-prem, GPU	cloud, GPU	assumptions based on customer observations

Serial link interfaces are expected to have a similar impact. With Pulse Amplitude Modulation (PAM) signaling, it becomes less straightforward to look at simple rise and fall times to determine extraction bandwidth requirements. One common approach to determine bandwidth extraction requirements is to look at the Nyquist frequency of the PAM4/8 signals and use a simple multiplier.

Some associated metrics for EM extraction of serial link interfaces are shown below:

Table 13.9.3: EM extraction metrics for serial link interfaces

	2026	2028	2030	2032	2034	Notes
Lane data rate (GT/s)	64	128	256	512	?	Based loosely on PCIe and projections
Signaling scheme	PCIe6 (PAM4)	PCIe7 (PAM4)	PCIe8 (PAM4)	PCIe9 (PAM8?)	optical?	Guesstimates for 2032 and 2034
Nyquist frequency (GHz)	16.0	32.0	64.0	85.3	?	Half the BW needed because 2 bits per symbol for PAM4
Extraction bandwidth (GHz)	48.0	96.0	192.0	256.0	?	Estimate 3x Nyquist

13.9.4 Power Integrity

From a PI perspective, nominal supply voltages are dropping further from the sub-1V levels in the 700 mV range today down to 500 mV in the future. As supply voltage drops, the max IR drop and power ripple requirements also scale down, making PI analysis more and more critical. Power consumption is growing significantly even though supply voltage is decreasing, meaning that transient current will increase, increasing current density and stressing the power distribution network (PDN).

Regarding methodology, it is expected to evolve from the single-fabric analyses seen today to looking at entire power rails at the system level, spanning from the voltage regulator module (VRM) on the printed circuit board (PCB) through the PCB, package, interposer, all the way to the on-die PDN. A shift in methodology is anticipated to emphasize up-front, pre-layout analysis, generating constraints that will be used to drive the implementation of the various physical layouts downstream. Machine learning is also expected to have an increasing role, accelerating the turnaround time for these analyses.

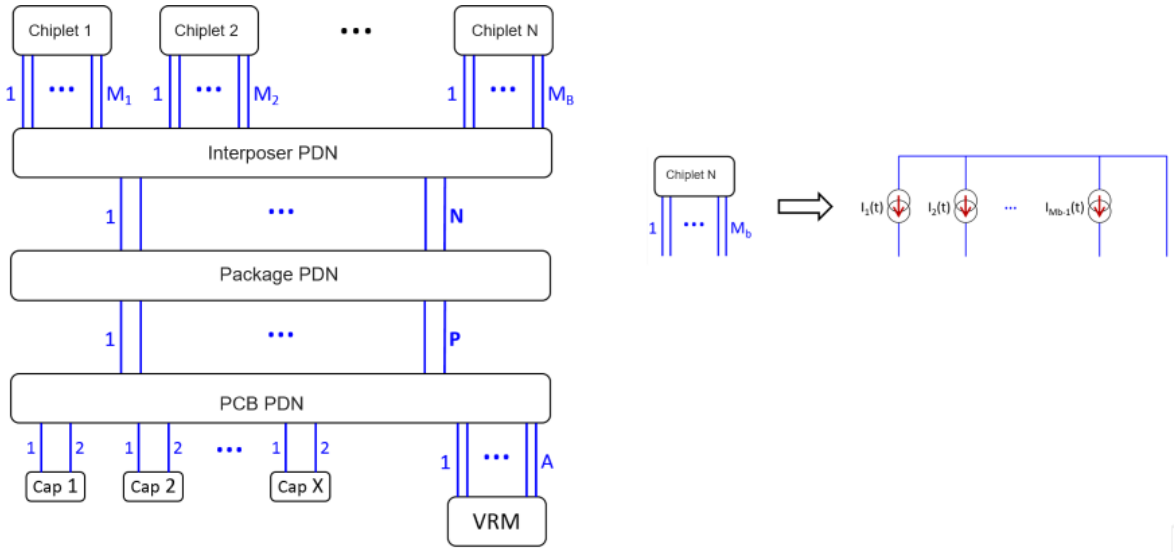


Figure 13.9.4: Simple representation of a HI system PDN.

Some associated metrics with power integrity signoff are shown below:

Table 13.9.4: PI metrics for HI design

	2026	2028	2030	2032	2034	Notes
Nominal Supply Voltage (V)	0.7	0.65	0.6	0.6	0.5	taking some judgement on IRDS doc
Max IR Drop (V)	0.014	0.013	0.006	0.006	0.005	assume 2% until 2030
Max Power Ripple (V)	0.035	0.0325	0.030	0.030	0.025	assume 5% of supply
Target Impedance (microohms)	50.0	40.0	30.0	20.0	10.0	The "Novak prediction"
Max Transient I (A)	700.0	812.5	1000.0	1500.0	2500.0	derive from target Z
DC Methodology	IR drop analysis on individual fabrics	IR drop analysis of entire power rail (PCB VRM to chiplet)	prev + top-down IR drop constraint partitioning across layouts w/ IDA	prev + machine learning	same	
AC Methodology	fabric-level decap analysis	prev + TD circuit sim of entire power rail (PCB VRM to chiplet)	prev + top-down target Z constraint partitioning across layouts w/ IDA	prev + machine learning	same	

13.9.5 Signal Integrity

The focus with SI is typically to check signals at the receiver die or chiplet against compliance criteria defined for a specific interface standard, such as Universal Chiplet Interconnect Express (UCIe) or PCI Express (PCIe). This is done by building out a full topology of the complete chiplet-to-chiplet signal path of the interface, which can traverse chiplet, interposer, package, and PCB as applicable. The topology is then simulated in the time domain with transient circuit and channel simulation, and the resulting waveforms and eye diagrams are checked against the interface-specific compliance criteria. This process is largely automated in most of today's SI tools.

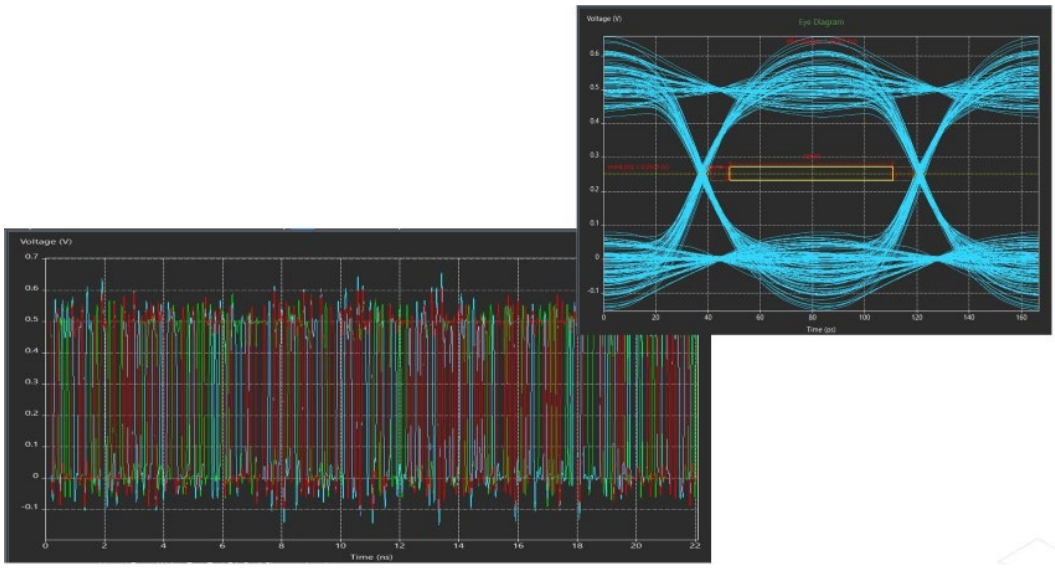


Figure 13.9.5: UCIe waveforms and eye diagram.

Some associated metrics for SI analysis of parallel and serial link interfaces are shown below:

Table 13.9.5: SI metrics

	2026	2028	2030	2032	2034
SE / SL Data Rate (Gbps)	32 / 64	54 / 128	128 / 256	128 / 512	128 / ?
EQ for SE	none	optional	standard	standard	standard
Methodology	Freq domain masks (IL, RL, etc.)	prev + TD eye diagram masks + BER	prev + MLMs derived from early exploration	same as prev	prev + MLMs provided by foundry
Notes	Values will be driven by specific interface standard	same as prev	Early sweeps / optimization data mined and capture for use downstream	Expect this to evolve and propagate over a few years	Eventually some MLMs will be provided as part of ADK from foundry

Definition of terms:

- ADK; Assembly Design Kit
- BER; bit error rate
- EQ; equalization
- IL; insertion loss
- MLM; machine learning model
- RL; return loss
- SE; single-ended
- SL; serial link

Taking some guidance from the commonly used UCIe standard, it is expected for line data rates to continue to climb, with equalization options being first optional, then standard practice after the 128Gbps point. Initial methodologies will likely be focused on frequency domain requirements for the passive interconnect, but quickly incorporating full time domain simulation, with eye diagram, and bathtub / bit error rate (BER) analysis. Early pre-layout sweeping and optimization will become commonplace to generate constraints for physical implementation, and analysis will be augmented with machine learning models to capture and reuse computationally intensive data and accelerate the signoff process.

13.9.6 Manufacturing Tolerance Impact

As pre-layout feasibility/trade-off analysis, constraint generation, and in-design analysis enforcement functionalities and methodologies become more robust, it is postulated that the main focus of post-layout analysis and signoff activity will be increasingly focused on the impact of manufacturing tolerances on the final thermal, power, and signal integrity margins. Accelerated by machine learning techniques, this aspect of analysis is expected to develop quickly over the next few years, in concert with tight partnerships with the foundries and manufacturing facilities that produce the various components that make up HI systems.

13.10 Multiphysics Co-Design

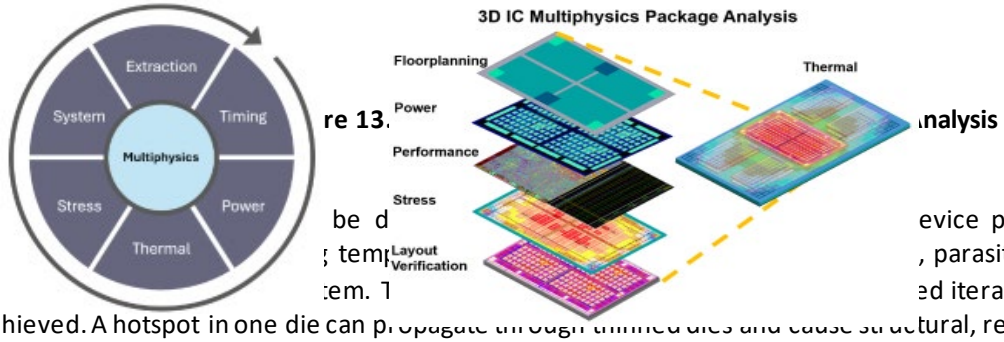
13.10.1 Multiphysics Analysis Enables Heterogeneous Integration

Imagine turning on an accident avoidance system in your car just milliseconds before a high-speed collision. That is precisely what it would be like to wait until sign-off to address power, timing, structural, or thermal issues in a heterogeneous 3D IC. Multiphysics analyses are no longer a luxury; they are a core requirement that must be integrated into your entire design flow to avoid catastrophic, last-minute surprises.

Characterizing silicon over power, timing, voltage, thermal, and process variations has been a core component of successful IC design for decades. Moving from traditional 2D implementations to the 3D IC space represents a much larger undertaking than simply scaling from one technology node to another. There are far more degrees of freedom with 3D packages, and each design decision can have potentially unpredictable consequences for the performance, reliability, and feasibility of your packaged product.

Unlike well-established 2D design processes, where designs typically move through each stage along a largely sequential path with well-defined constraints, advanced heterogeneous 3D ICs require the power, timing, temperature, and stress behavior to be modeled continuously in an iterative cycle, from design concept all the way through to final sign-off. Engineers must have concrete data on how design decisions impact other components within the package, far beyond their traditional area of expertise. This means increased communication between the

different die designers, the package architect, and the manufacturer. These new challenges represent the core problem that integrated multiphysics tools must address.



re 13. nalysis

be d evice power characteristics and w(; temp , parasitics, stress, and electrical pe tem. T ed iteratively until steady-state is achieved. A hotspot in one die can p ctural, reliability, and performance issues in other dies if not properly accounted for. Multiple overlapping hotspots could cause increased power consumption that drives a thermal runaway situation. Clearly, attempting to fix these issues after the die layouts are complete and the package floorplan chosen would be nearly impossible. This is why multiphysics simulations must start all the way at the design concept stage, where you have the degrees of freedom to adjust your floorplan, stacking arrangement, materials, and cooling technologies to avoid these issues entirely.

13.10.2 Multiphysics Tool Challenges

A key challenge in incorporating multiphysics simulations is the risk of exceeding current compute resources and realistic development cycles. Traditional stand-alone power, timing, and physical verification runs on large IC designs can consume hundreds of CPUs and hours of wall time. Now, these runs are required iteratively, per die in the package, until steady-state is achieved. For example, attempting a multi-corner, multi-mode static timing analysis across all dies in a package can render the number of permutations intractable. Furthermore, each multiphysics flow demands diverse inputs in various formats from different vendor tools and standards. Each run also necessitates detailed scripts and configurations tailored specifically to each design under test. Collectively, these factors make integrated multiphysics simulations highly complex to manage and challenges vendors to develop more standardized solutions for data input and flow management.

13.10.3 Emerging EDA Solutions for Multiphysics

EDA companies have traditionally provided a collection of point tools that plug into various stages of a development flow. Just as multiple dies are integrated into a single 3D IC package to scale performance, there is a parallel and equally strong demand to transition these point tools into cohesive, integrated flows. Major EDA companies are converging on ‘cockpit’ solutions for 3D-ICs, enabling the singular input of the overall design and collaterals for running complex physical analyses. Notable examples include Cadence Integrity, Siemens EDA Innovator3D IC, and Synopsys 3DIC compiler.

Reflecting this shift towards integrated solutions for heterogeneous 3D ICs, in 2025 alone, the three major EDA companies collectively invested close to 50 billion US dollars in multiphysics-centric software acquisitions. For instance, Synopsys acquired Ansys, Siemens purchased Altair, and Cadence acquired BETA CAE. Additionally, EDA, foundries, and OSATs are focusing on interoperability for 3D IC projects, with 3D IC languages like 3Dblox and 3DStack+ to define standard 3D IC assemblies that can be reused across multiple analyses and flows. Tool changes are expensive, and multi-vendor solutions provide the best flexibility for end-users.

13.10.4 Tracking Multiphysics as a Manufacturing Requirement

Until recently, the adoption of these tools was largely driven by physical characterization, simulation, and design and performance optimization, but manufacturing requirements were often not the primary focus. However, foundries are increasingly incorporating multiphysics analyses into their sign-off procedures to ensure these advanced 3D ICs can be manufactured to specifications. Thermal solutions from the various EDA vendors have achieved certification from numerous manufactures, but now these thermal tools are integrating with other sign-off flows. For example, the thermal impact on EMIR is already required for certain advanced nodes. As stacking technology matures,

operational and manufacturing thermal impacts on both static timing analysis (STA) and stress are likely to become sign-off requirements in the near future.

The degree to which multiphysics analyses are required to implement a product depends heavily on the design application, technology nodes of the dies, 3D IC manufacturing technology, and numerous other parameters. The following table provides a general summary of where true iterative multiphysics analyses are expected to apply. These flow categories assume high-resolution thermal results and mechanical properties are used (e.g. device-level thermal maps) and not uniform corners, as is applied in most designs today.

Table 13.10.1: Multiphysics Analysis Applicability by Design Category.

Multiphysics Flow	2D Low Power	2D HPC	2.5D Low Power	2.5D HPC	3D IC Low Power	3D IC HPC
Stand-alone Thermal Analysis	No	Yes	Limited	Yes	Limited	Yes
Thermal-Aware Floorplanning	No	No	No	Developing	Limited	Developing
Thermal-Electrical Co-simulation	No	Limited	Limited	Yes	Limited	Yes
Thermal-Mechanical (Stress) Co-simulation	No	No	Limited	Developing	Developing	Developing
Thermal-Reliability (Thermal-EM) Analysis	Limited	Yes	No	Yes	No	Yes
Thermal-Timing (STA) Co-simulation	No	No	No	Developing	No	Developing
Power Integrity (PI) and Thermal Co-simulation	No	Yes	No	Yes	No	Developing
Signal Integrity (SI) and Thermal Co-simulation	No	Yes	No	Yes	No	Yes
Electro-Thermal-Mechanical Co-simulation	No	No	No	Developing	Developing	Developing

13.10.5 Enabling Multiphysics Through Multiscale Flows

As mentioned, multiphysics should commence at the concept and planning phase of a design. A key objective at this early phase is to determine an ideal arrangement of dies within the package, based on rough power and thermal estimations that ensure acceptable thermal windows and mechanical stresses under both manufacturing and operational conditions. This feasibility simulation can be performed with minimal inputs, such as RTL power estimates, die sizes, package components, and general material properties. At minimum, this enables the determination of whether certain stacking arrangements are physically possible and what configuration may deliver optimum thermal conditions to avoid hotspots. While the results are coarse at this stage, they provide excellent boundary conditions for subsequent package and die-level floorplanning and optimization that can be shared with different teams.

This approach introduces the concept of multi-scale multiphysics flows. It allows for the execution of fast simulations with minimal inputs, which in turn provide boundary conditions for highly detailed simulations. In Figure 13.10.2, the left image shows a coarse simulation result of thermal-induced stresses across an entire 3D IC package. These results can then be used to pinpoint areas of interest for performing compute-intensive, highly-detailed die-level stress simulations (as shown on the right). The detailed simulations provide the transistor-level data required to model stress impacts on electrical and reliability simulations. Multi-scale simulation is one methodology that can significantly reduce the intractability of multiphysics analyses.

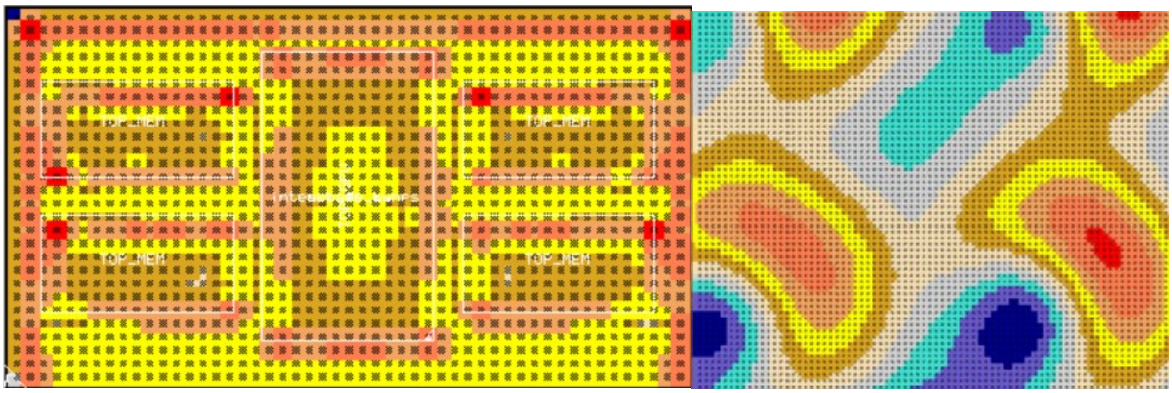


Figure 13.10.2: Coarse stress results at the package-level (left) and fine stress results at the device-level (right) generated from heating and cooling effects during the manufacture of a 3D IC Package

Ultimately, the success of multiphysics simulations hinges on the introduction of smarter, more capable EDA tools that integrate existing technology with AI/ML that boosts performance and usability. This integration is essential for reducing both the complexity of analyzing results and improving the accessibility of these tools for typical designers. What is clear in the space of multiphysics and 3D ICs is that the silos between development teams must come down and the traditional development processes must adapt to accommodate even more simulation and validation earlier in the flow.

13.11 AI for Co-Design

Recent advances in agentic AI and machine learning (ML) are poised to significantly accelerate the co-design process for 3D heterogeneous integration. With the rapid adoption of large language models (LLMs) and agentic AI frameworks, LLM-based agents can now be customized to act as co-designers or co-pilots alongside engineers.

By decomposing complex design workflows into hierarchical, multi-agent pipelines, each agent can take on specialized roles—whether handling routine but essential tasks, performing modeling and analysis, or executing debugging and diagnostic operations through EDA tools. Newly developed generative AI and ML models can further augment these agents, enhancing their ability to explore design spaces, predict performance, and synthesize solutions.

Below are several examples illustrating how a multi-agent co-piloting workflow can support heterogeneous integration design.

13.11.1 Multi-agent AI Workflow for Co-Design/Co-Pilot

A multi-agent AI workflow can significantly enhance the co-design process. Individual AI agents can function as *co-pilots*, autonomously executing many design tasks. In this framework, an orchestrator agent—with a human in the loop—creates a structured plan and delegates tasks to specialized agents. These agents coordinate through a shared memory buffer, enabling data exchange and seamless collaboration across the workflow.

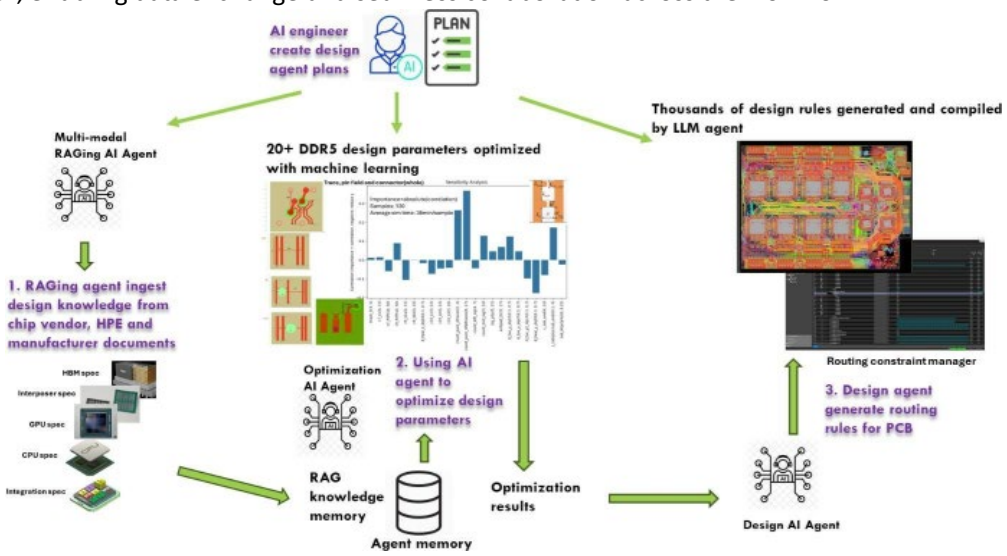


Figure 13.11.1: An example of a collaborative multi-agent system in which a planning agent, with human-in-the-loop oversight, creates executable plans that specialized agents carry out for the co-optimization of a DDR5 design.

Domain knowledge—such as information encoded in a knowledge graph—can be incorporated into the design flow through AI agents, stored in a shared memory space, and passed to other agents as needed. A retrieval-augmented generation (RAG) agent can ingest large volumes of design specifications, extract critical design parameters, and provide actionable inputs to downstream agents. Figure 13.11.1 illustrates an example of a collaborative multi-agent system in which a planning agent, with human-in-the-loop oversight, creates executable plans that specialized agents carry out for the co-optimization of a DDR5 design. Also multiple documents—across formats such as PDFs, videos, and text—can be pre-processed into embedding vectors, which then serve as inputs to a RAG agent for customer-facing Q&A or analysis input, as illustrated in Figure 13.11.2.

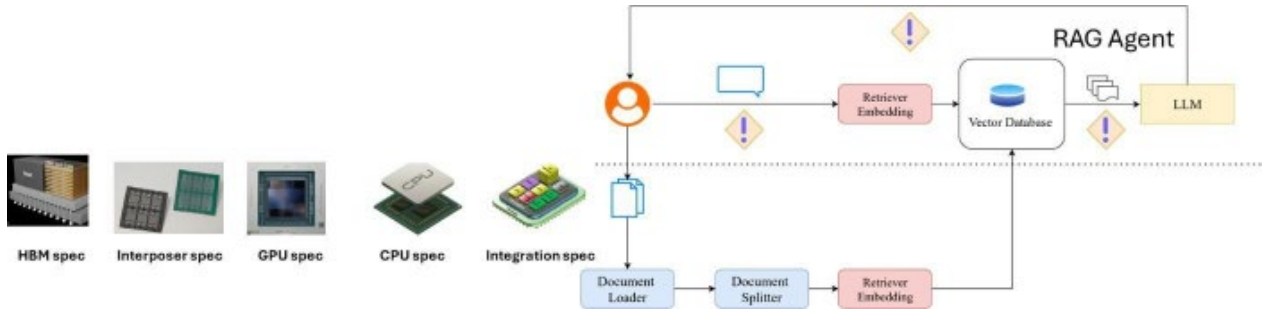


Figure 13.11.2: Multiple documents—across formats such as PDFs, videos, and text—can be pre-processed into embedding vectors, which then serve as inputs to a RAG agent for customer-facing Q&A or used by a simulation agent to initiate analysis by EDA tools [13.11.1].

Furthermore, the optimization agent can interact directly with EDA tools—potentially through an MCP (Model Context Protocol) server with embedded APIs—using parameters passed from the RAG agent to run simulations and determine optimal design solutions using appropriate optimization algorithms [13.8.2].

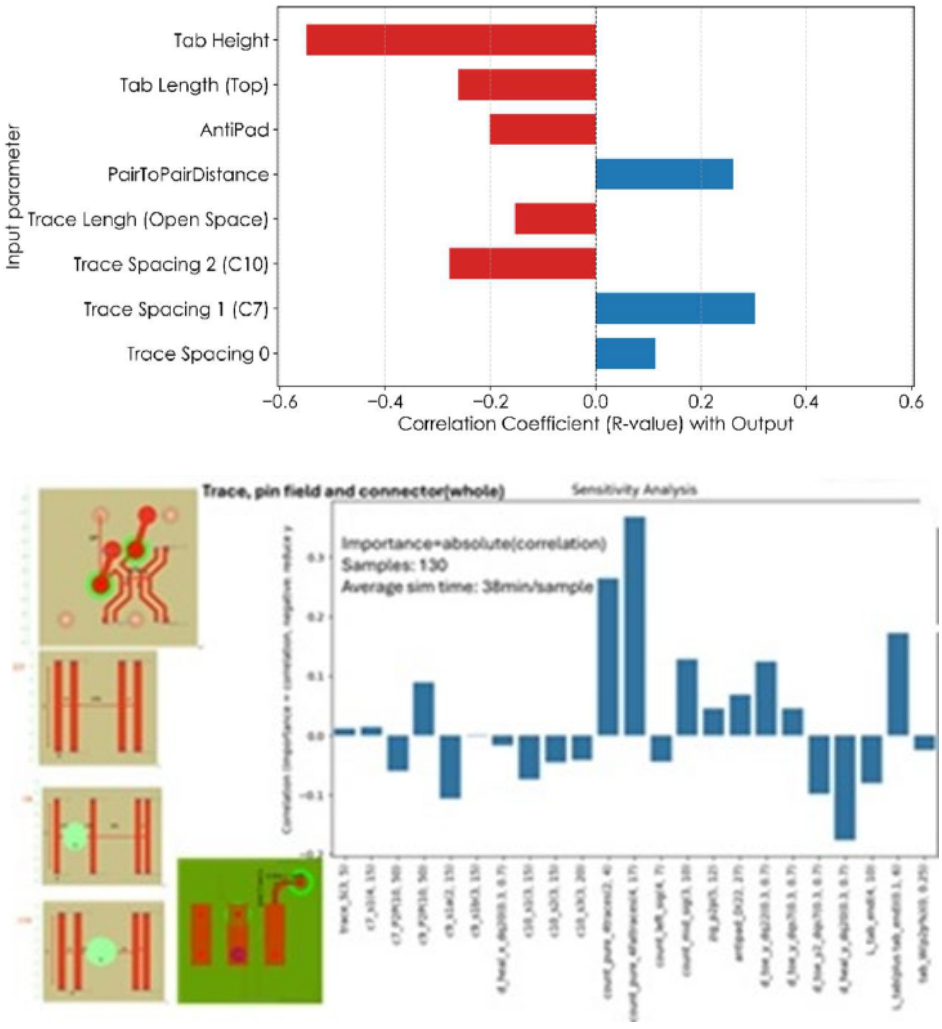


Figure 13.11.3. The optimization agent automatically determines and applies the most suitable optimization algorithm for the specific design problem.

Once the optimized results are obtained (Figure 13.11.3), the optimization agent passes the solution to a downstream agent to generate the corresponding routing constraints (Figure 13.11.4).

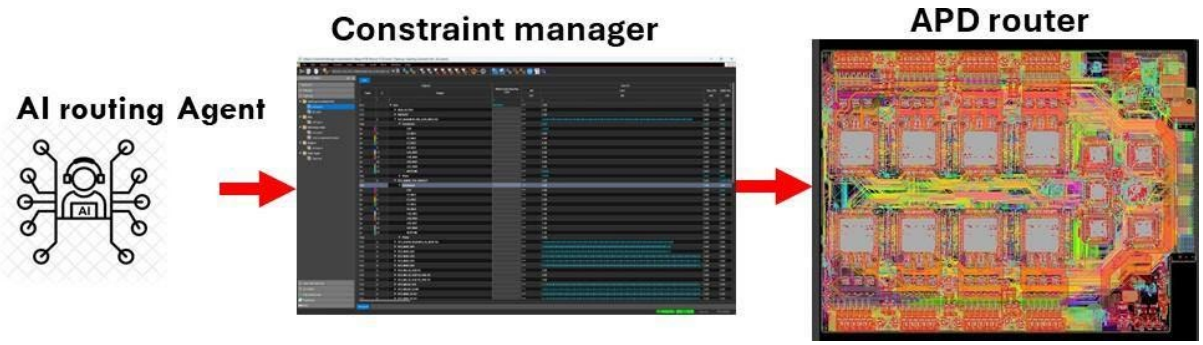


Figure 13.11.4: An optimized routing can automatically be done through an AI routing agent.

13.11.2 A Foundational Model Approach to Future-Proofing the Co-Design Process

One of AI's greatest promises is its ability to model highly complex problems using proven architectures such as GANs and transformer networks. With encoder–decoder transformers, high-dimensional challenges—such as complex electromagnetic channel characteristics or multiphysics power-to-thermal interactions—can be abstracted and accurately modeled through the transformer's encoding and decoding process as shown in Figure 13.11.5.

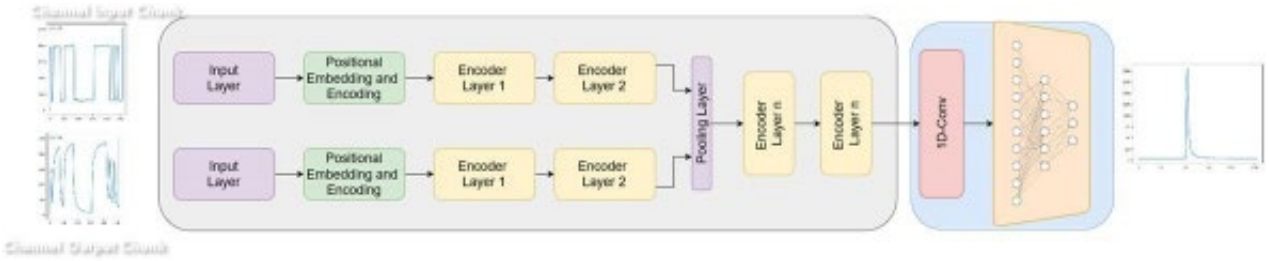


Figure 13.11.5: Using an encoder–decoder transformer architecture to model complex electro- magnetic channel characteristics and multiphysics power-to-thermal mappings.

Training a large AI model demands vast datasets and significant computational resources. In the fast-moving power and performance race, by the time enough data is collected for model training, design requirements and target technologies may have already evolved. Luckily, efficient adaptation techniques like fine-tuning and transfer learning enable us to update models without retraining the entire system, dramatically reducing computational needs as illustrated in Figure 13.11.6. These methods ensure that generative AI models remain future-proof compared to conventional EDA tools when design ranges or process technologies shift.

AI agents built on foundation models help abstract complex design tasks using generative or agentic AI, while fine-tuning and transfer learning extend the useful life of trained models as the underlying technology advances. Examples of foundation models include the channel model for signal integrity [13.11.2] or thermal models for advanced 3DIC design [13.11.3-13.11.4]. In summary, transfer learning and fine-tuning empower EDA workflows to quickly adapt to changing technology, maximizing model relevance and reducing total cost of ownership versus traditional tools.

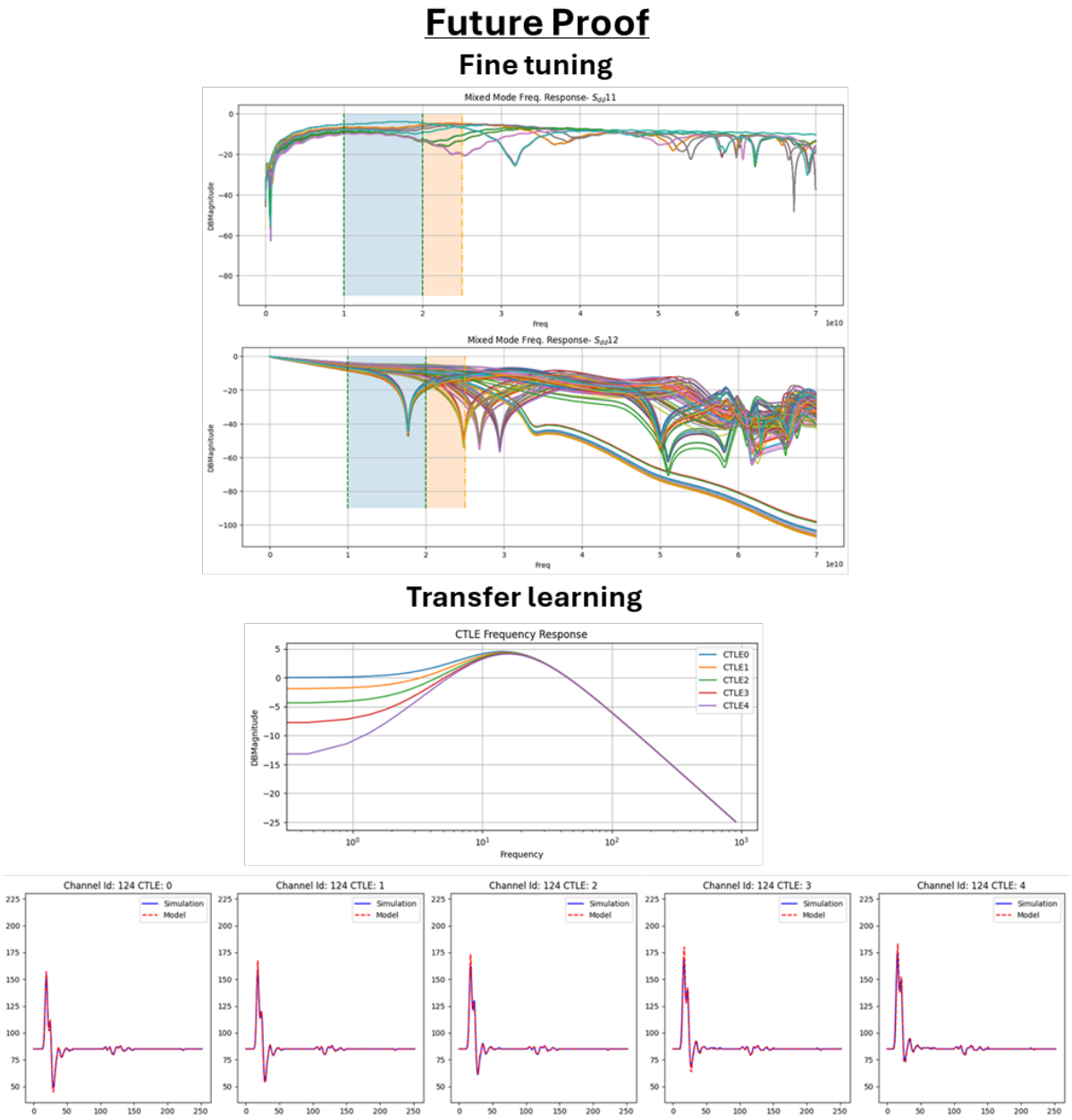


Figure 13.11.6: Future-proofing of the co-design process through fine tuning and transfer learning techniques.

13.11.3 AI Digital Twins for Real Time Multi-Physics System Technology Co-Optimization

Heterogeneous integration brings together electrical, mechanical, optical, and other physical domains within advanced system designs. When generative or agentic AI systems are trained with multiphysics-aware strategies, they enable the abstraction of complex multiphysics interactions—these AI-enhanced models are commonly known as digital twins.

Recent digital twin technologies are engineered for high computational efficiency, leveraging the latest GPUs or advanced CPU instruction sets like AVX-10, resulting in inference times measured in seconds or less. This paves the way for dynamic, real-time optimization of multiphysics challenges across heterogeneous environments—for example, rapidly tuning power, thermal, and cooling solutions, or optimizing high-speed SerDes links in operation. A hybrid digital twin approach can incorporate measurement data to improve simulation accuracy and capture residual physical effects not fully modeled [13.11.8].

Digital twins now allow co-optimization in the design and runtime phases, transforming the design process from slow, siloed simulations to holistic, adaptive systems capable of reflecting and responding to electrical, thermal, and mechanical changes in real time. As this field matures, the integration of AI-accelerated digital twins is becoming essential for the future of 3DIC, chiplet assemblies, and advanced packaging ecosystems.

Foundational models and digital twins represent distinct yet complementary applications of generative and agentic AI modeling. Foundational models are extremely large, high- complexity models designed to encode vast amounts of context—such as intricate design phenomena—within their embeddings, similar to the ever-increasing parameter sizes seen in leading large language models. This immense capacity makes those models highly adaptable and "future-proof," as emerging design situations and technical changes can be captured through targeted fine-tuning without the need to retrain the model from scratch.

By contrast, digital twins are typically focused, real-time virtual counterparts of specific physical systems such as a SerDes design. While digital twins can leverage foundational models for advanced predictive or optimization tasks, their primary purpose is mirroring and controlling actual system behavior, often in conjunction with real-time data feeds from measurements [13.11.6].

Digital twins are designed for speed and efficiency, enabling near real-time inference and operational feedback. Models such as U-Net and autoencoders excel at this by imposing a bottleneck structure, streamlining computation and minimizing complexity so the model remains lightweight and highly responsive to input data. See Figure 13.11.7 for a GAN-based digital twin that rapidly predicts thermal profiles from power distributions, enabling real-time optimization in electronic systems [13.11.7].

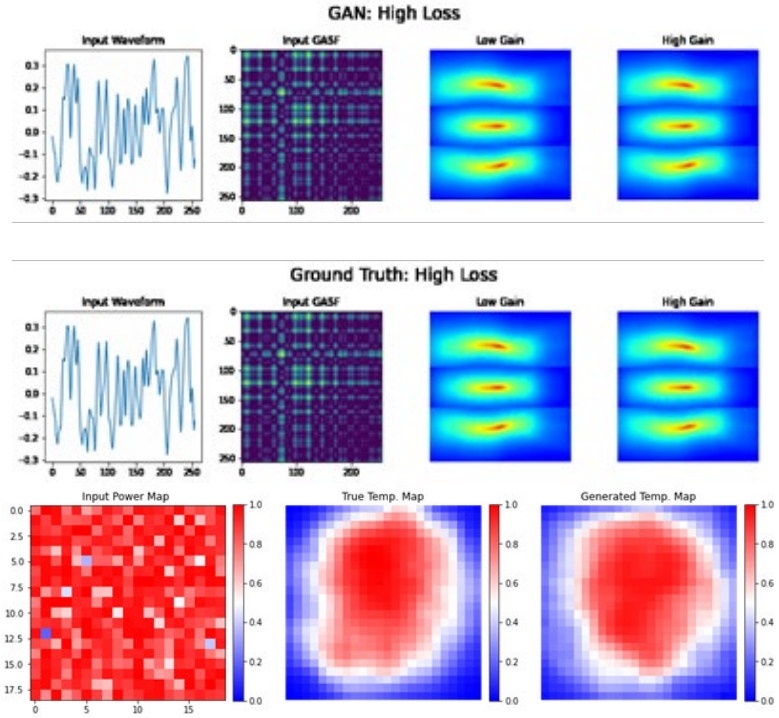


Figure 13.11.7: A GAN-based digital twin rapidly predicts thermal profiles from power distributions, enabling real-time optimization in electronic systems.

Conversely, foundational models take the opposite approach: they feature large embeddings between their encoder and decoder components. This architecture deliberately maximizes context capacity, allowing the models to flexibly represent diverse and growing design scenarios—similar to large-scale language models whose parameter sizes continually increase for future-proofing and adaptability.

In summary, digital twins optimize for rapid, targeted decision-making with compact model designs, while foundational models optimize for flexibility and context retention with very large, complex embedding spaces.

13.11.4 AI Agent as a Global Multi-Physics Optimizer

The capability of AI-driven frameworks to abstract and co-analyze multi-physics interactions— including electromagnetic signal integrity, power-delivery-network (PDN) impedance behavior, and steady-state/transient thermal distribution—across the full 3DIC hierarchy (chiplets, interposer, package-level PDN, and high-density interconnects) enables a unified global- optimization engine that operates on the entire heterogeneous stack. In the referenced study [13.11.9], AI compute throughput is jointly optimized by exploring chiplet-partitioning design spaces and evaluating the coupled effects of PDN droop, thermal hotspots, and inter- connect latency/bandwidth constraints. The highest throughput configuration emerged from a policy optimized using imitation-learning-based reinforcement learning, which more efficiently captured system-level performance gradients compared to conventional RL methods.

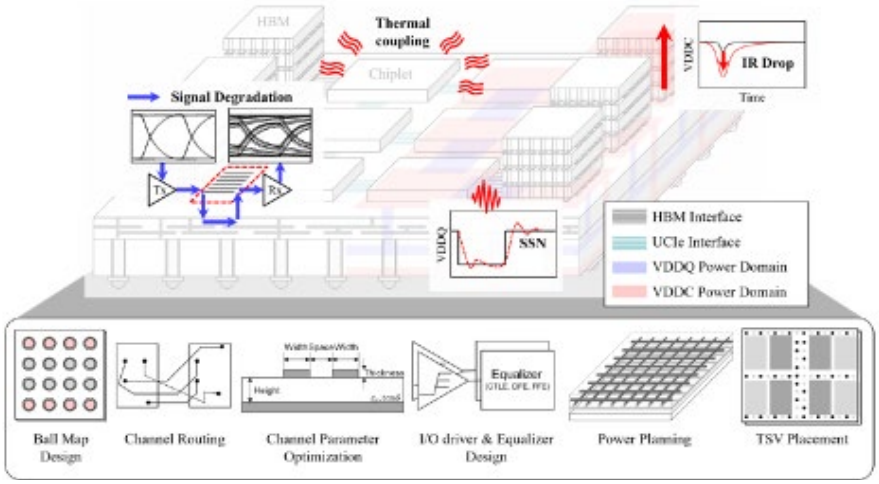


Figure 13.11.8: Multidisciplinary nature of chiplet design with signal routing, power integrity and thermal analysis.

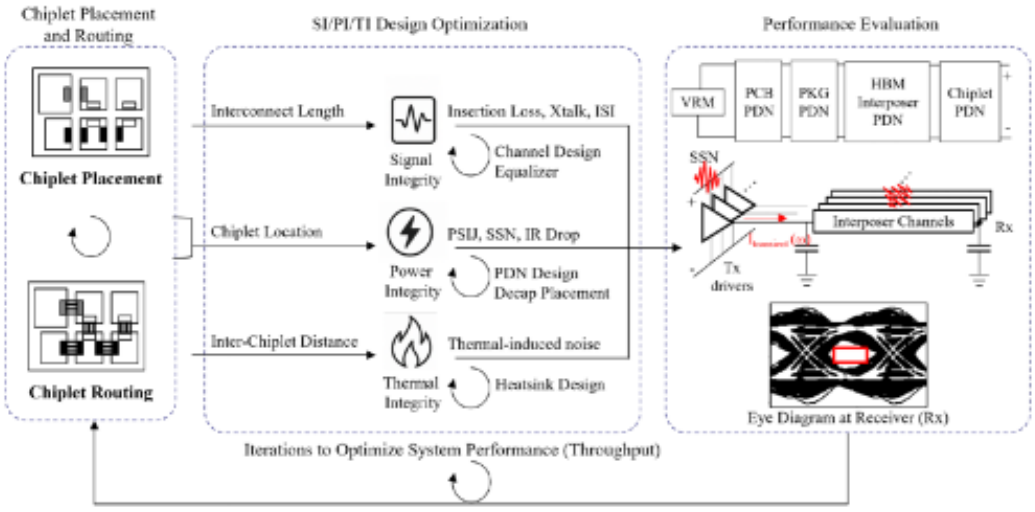


Figure 13.11.9: Each of the multidisciplinary analysis is handled by different AI agents for the performance optimization.

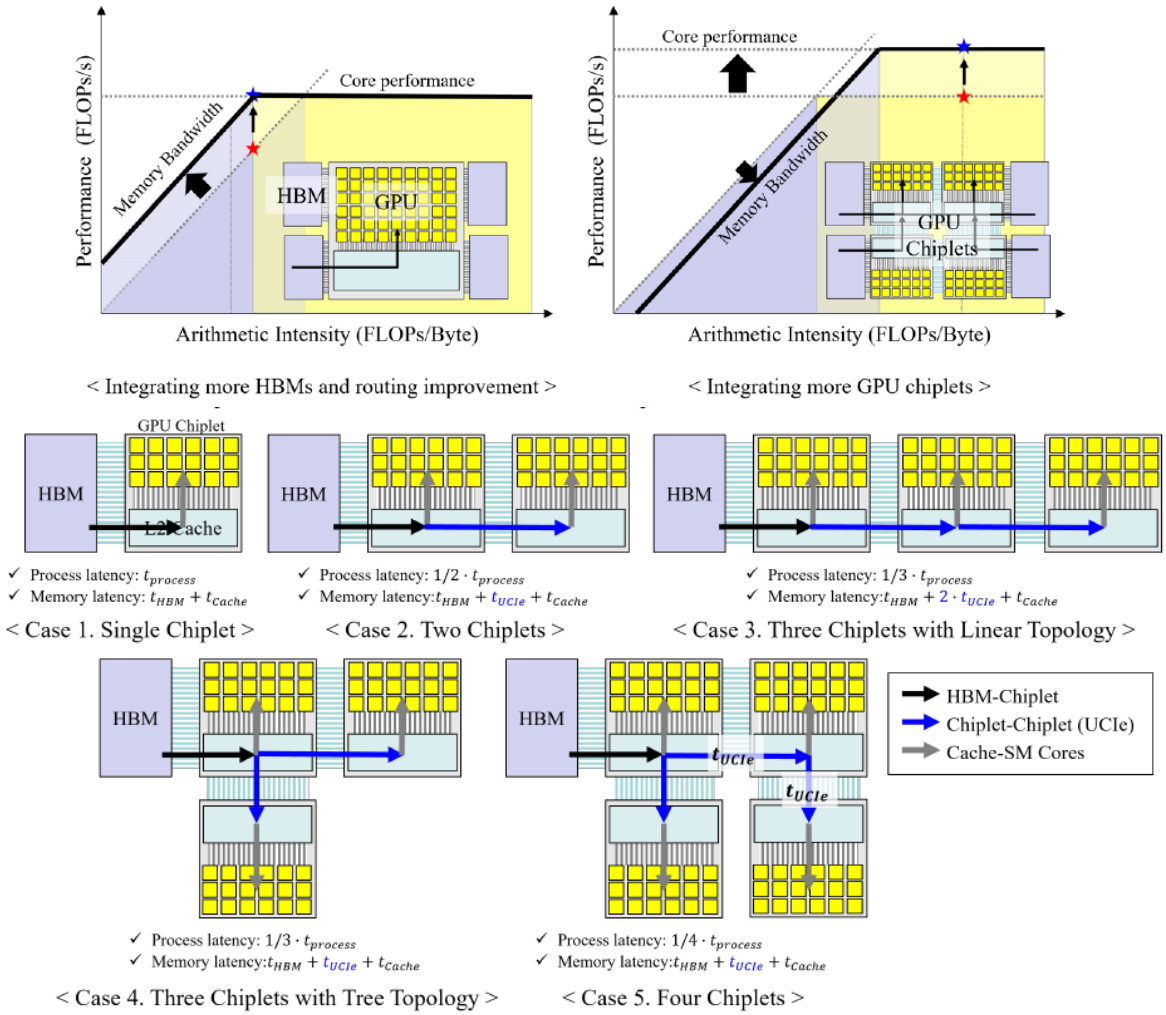


Figure 13.11.10: Architectural trade-offs with varying amounts of HBM, GPU and interconnect for optimal compute performance.

13.11.5 Multi-AI Agents for Hierarchical Reinforcement Learning Optimization of 3DIC

In this example, the 3DIC is partitioned into regions based on the chiplet and substrate locations. Multiple levels of power analysis are performed by agents in a hierarchical manner, with a global reinforcement-learning–based optimization used for overall power-delivery improvement.

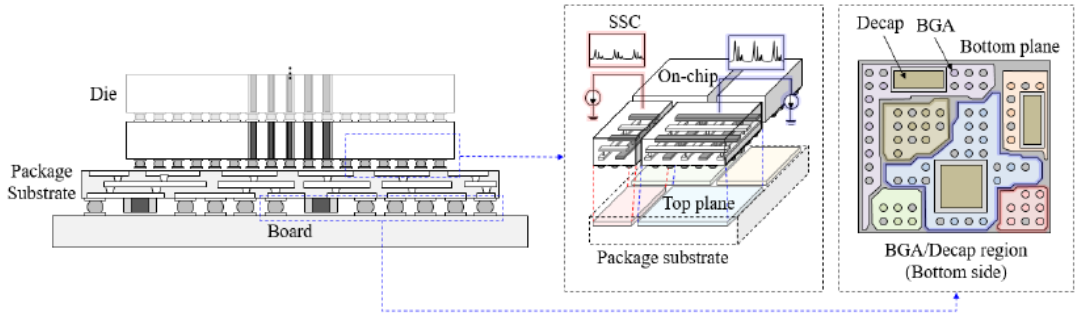


Figure 13.11.11: Different area of chiplets and substrate and partitioned for analysis.

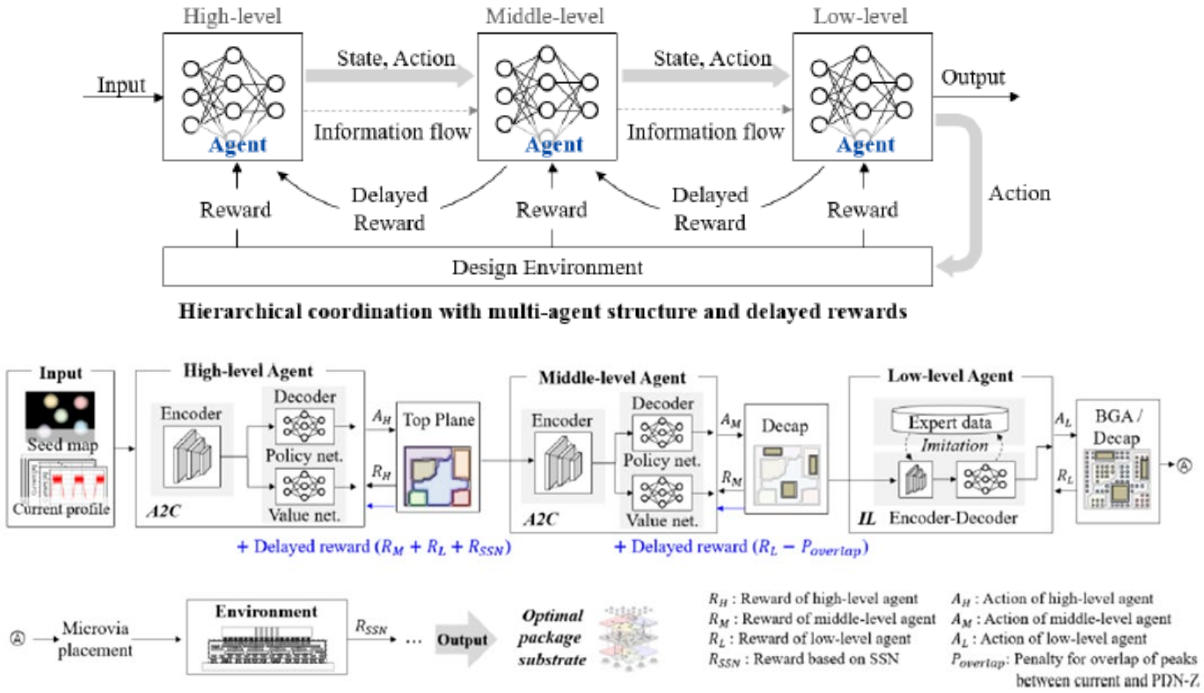


Figure 13.11.12: Hierarchical reinforcement learning for power distribution optimization.

Glossary

2.5D Integration- Packaging approach where multiple dies are placed side-by-side on an interposer.

3D Integration- Stacking of multiple dies vertically with interconnects between layers.

ADK- Assembly Design Kit.

BIST- Built-In Self-Test.

Chiplet- A small semiconductor die designed to be combined with other chiplets to form a complete system.

CoWoS- Chip-on-Wafer-on-Substrate—a 2.5D packaging technology by TSMC using a silicon interposer.

DFT- Design-for-Test.

EMIB- Embedded Multi-die Interconnect Bridge—Intel's technology using a small silicon bridge embedded in an organic package.

HBM- High Bandwidth Memory—a type of stacked memory using TSV inter- connects.

HSUL- High Swing Unterminated Logic—a signaling standard used in die-to-die interfaces.

IVR- Integrated Voltage Regulator.

KGD- Known Good Die.

LVSTL- Low-Voltage Swing Termination Logic—a low-power signaling standard.

PDN - Power Distribution Network—the network of conductors delivering power to integrated circuits.

PHY - Physical Layer—the hardware interface responsible for transmitting raw data bits.

PPACR- Power, Performance, Area, Cost, and Reliability.

RDL - Redistribution Layer—metal layers that reroute I/O pads to different locations on a die or substrate.

SerDes- Serializer/Deserializer—circuitry that converts parallel data to serial for transmission and back.

SiP- System-in-Package—an approach that integrates multiple ICs in a single package.

TSV- Through-Silicon Via—a vertical electrical connection passing through a silicon die.

UCIe- Universal Chiplet Interconnect Express—an open industry standard for die-to-die interconnects.

TWG Volunteers

Vaishnav Srinivas
Jose Schutt-Aine
Christopher Bailey
Atom Watanabe
Daniel Lambalot
Srikanth Rangarajan
Krishnendu Chakrabarty
Puneet Gupta
Archana Cheruliyil
Sung Kyu Lim
David Pan
Ken Willis
John Ferguson
Matthew Grange
Norman Chang
Chris Cheng
Rohit Sharma
Sandeep Goel
Amitabh Das
Pratyush Kamal
Soni Kapoor
Tony Carusone
David Choe
Wendem Beyene
Tahsin Shameem
Yi Zhou
Partho Bhoumik
Dhruv Thapar
Arjun Hati

References

[13.1.1] A. Kabir, Y. Peng, "Chiplet-Package Co-Design For 2.5D Systems Using Standard ASIC CAD Tools," *25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan. 13–16, 2020.

[13.1.2] L. Li, P. Chia, P. Ton, M. Nagar, S. Patil, J. Xue, J. DeLaCruz, M. Voicu, J. Hellings, B. Isaacson, M. Coor, and R. Havens, "3D SiP with organic interposer of ASICand memory integration," *Proceedings of IEEE/ECTC*, May 2016, pp. 1445–1450.

[13.1.3] H. Braunisch, A. Aleksov, S. Lotz, and J. Swan, "High-speed performance of Silicon Bridge die-to-die interconnects," *Proceedings of Electrical Performance of Electronic Packaging and Systems (EPEPS)*, San Jose, CA, USA, Oct. 23–26, 2011, pp. 95–98.

[13.1.4] D. Greenhill, R. Ho, D. Lewis, H. Schmit, K. Chan, A. Tong, S. Atsatt, D. How, P. McElheny, K. Duwel, J. Schulz, D. Faulkner, G. Iyer, G. Chen, H. Phoon, H. W. Lim, E. Koay, T. Garibay, "A 14nm 1GHz FPGA with 2.5D transceiver integration," *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2017, pp. 54–55.

[13.2.1] Y.-J. Huang et al., "UCIe-Compliant Chiplet Interconnect Design Lever- aging Cutting-Edge Packaging Technologies," *2025 IEEE Custom Inte- grated Circuits Conference (CICC)*, Boston, MA, USA, 2025, pp. 1–8, doi: 10.1109/CICC63670.2025.10983497.

[13.2.2] S. Y. Hou et al., "Integrated Deep Trench Capacitor in Si Interposer for CoWoS Heterogeneous Integration," *2019 IEEE International Electron De- vices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 19.5.1–19.5.4, doi: 10.1109/IEDM19573.2019.8993498.

[13.3.1] Courtland, Rachel. "Transistors could stop shrinking in 2021." *IEEE Spectrum*, 53, no. 9 (2016): 9-11.

[13.3.2] ITRS, *The International Technology Roadmap for Semiconductors*, 2015, Semiconductor Industry Association,

[13.3.3] Ellsworth Jr, Michael J., and R. E. Simons. "High powered chip cooling—Air and beyond." *Electron. Cooling*, 11, no. 3 (2005): 14-22.

[13.3.4] G. Refai-Ahmed et al., "Establishing Thermal Air-cooled Limit for High Performance Electronics Devices," *2020 IEEE 22nd Electronics Packaging Technology Conference (EPTC)*, Singapore, Singapore, 2020, pp. 347-354, doi: 10.1109/EPTC50525.2020.9315139

[13.3.5] Tuckerman, David B., and Roger Fabian W. Pease. "High-performance heat sinking for VLSI." *IEEE Electron device letters*, 2, no. 5 (1981): 126-129.

[13.3.6] Radmard, Vahideh, Yaser Hadad, Srikanth Rangarajan, Cong H. Hoang, Najmeh Fallahtafti, Charles L. Arvin, Kamal Sikka, Scott N. Schiffres, and Bahgat G. Sammakia. "Multi-objective optimization of a chip-attached micro pin fin liquid cooling system." *Applied Thermal Engineering* (2021): 117187.

[13.3.7] Radmard, Vahideh, Arad Azizi, Srikanth Rangarajan, Najmeh, C. Hiep Hoang, Ghazal, Charles Arvin, Kamal Sikka, Scott N. Schiffres, and Bahgat Sammakia. " Performance Analysis of Impinging Chip-Attached Micro Pin Fin Direct Liquid Cooling Package for Hotspot Targeted Applications" In *2021 36th Semiconductor Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*, pp. 1-9. IEEE, 2020.

[13.3.8] Radmard, Vahideh, Arad Azizi, Srikanth Rangarajan, Najmeh, C. Hiep Hoang, Ghazal, Charles Arvin, Kamal Sikka, Scott N. Schiffres, and Bahgat Sammakia. "Direct Micro-Pin Jet Impingement Cooling for High Heat Flux Applications." In *2020 The Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (IEEE ITherm)* (To appear)

[13.3.9] Hoang, Cong Hiep, Srikanth Rangarajan, Sadegh Khalili, Bharath Ramakrisnan, Vahideh Radmard, Yaser Hadad, Scott Schiffres, and Bahgat Sammakia. "Hybrid microchannel/multi-jet two-phase heat sink: A benchmark and geometry optimization study of commercial product." *International Journal of Heat and Mass Transfer* 169 (2021): 120920.

[13.3.10] Khalili, Sadegh, Srikanth Rangarajan, Vadim Gektin, Husam Alissa, and Bahgat Sammakia. "An Experimental Investigation on the Fluid Distribution in a Two-Phase Cooled Rack Under Steady and Transient Information Technology Loads." *Journal of Electronic Packaging*, 142, no. 4 (2020).

[13.3.11] Drummond, Kevin P., Doosan Back, Michael D. Sinanis, David B. Janes, Dimitrios Peroulis, Justin A. Weibel, and Suresh V. Garimella. "A hierarchical manifold microchannel heat sink array for high-heat-flux two-phase cooling of electronics." *International Journal of Heat and Mass Transfer*, 117 (2018): 319-330.

[13.3.12] Liang, Gangtao, and Issam Mudawar. "Review of pool boiling enhancement by surface modification." *International Journal of Heat and Mass Transfer*, 128 (2019): 892-933.

[13.3.13] Patil, Chinmay M., and Satish G. Kandlikar. "Pool boiling enhancement through microporous coatings selectively electrodeposited on fin tops of open microchannels." *International Journal of Heat and Mass Transfer*, 79 (2014): 816-828.

[13.3.14] Wong, Kin Keong, and Kai Choong Leong. "Saturated pool boiling enhancement using porous lattice structures produced by Selective Laser Melting." *International Journal of Heat and Mass Transfer,* 121 (2018): 46-63.

[13.4.1] Praful P and Bailey C (2025) "Warpage in wafer-level packaging: a review of causes, modelling, and mitigation strategies". Front. Electron. 5:1515860. doi: 10.3389/felec.2024.1515860

[13.4.2] Lau, J. H. (2019b). Recent advances and trends in fan-out wafer/panel-level packaging. ASM International, 618 141. doi:10.1115/1.4043341

[13.4.3] Jithdrriyan P. Bailey, C. Dasgupta A. Physics-informed machine learning framework for predictive warpage control in fan-out wafer level packaging, ITHERM 2026.

[13.4.4] Huber et-al (2024). "Analyzing the influence of rdl stack-up on wafer warpage in FOWLP through experimental and numerical investigations," in Electronic components and technology conference (ECTC). 10.1109/ectc51529.2024.00163

[13.4.5] Shangguan, D. (Ed.). (2024). Direct Copper Interconnection for Advanced Semiconductor Technology (1st ed.). CRC Press. https://doi.org/10.1201/9781003408673

[13.4.6] Benedict San Jose; Cliff Sandstrom; Jan Kellar; Craig Bishop; Tim Olson, "Adaptive Patterning and M-Series for High Density Integration," in Embedded and Fan-Out Wafer and Panel Level Packaging Technologies for Advanced Application Spaces: High Performance Compute and System-in-Package , IEEE, 2022, pp.125-154, doi: 10.1002/9781119793908.ch5.

[13.4.7] E. Hudson, D. Baldwin, T. Olson, C. Bishop, J. Kellar and R. Gabriel, "Deca & Cadence Breakthrough Heterogeneous Integration Barriers with Adaptive Patterning™," 2021 IEEE 71st Electronic Components and Technology Conference (ECTC), San Diego, CA, USA, 2021, pp. 45-49, doi: 10.1109/ECTC32696.2021.00019.

[13.4.8] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics, V 378, 2019, pp 686-707, doi.org/10.1016/j.jcp.2018.10.045.

[13.5.1] J. H. Lau, "Current Advances and Outlooks in Hybrid Bonding," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 15, no. 4, pp. 651–681, 2025.

[13.5.2] K. Maet al., "Process Technology Enablers for Aggressive Scaling of Hybrid Bonding Temperature and Pitch," in IEEE 75th Electronic Components and Technology Conference (ECTC), 2025, pp. 1901–1907.

[13.5.3] "International Technology Roadmap for Semiconductors," https://www.semiconductors.org/wp-content/uploads/2018/08/2011Test.pdf, 2011, accessed: Dec. 2025.

[13.5.4] M. Slessoret al., "Strategies for Wafer Probe in a Chiplet World," https://nccavs-usergroups.avs.org/wp-content/uploads/2022/07/Joint-UG622-6-LeongA.pdf, 2022, accessed: Dec. 2025.

[13.5.5] "Case Study: Challenges when Probing High Pad Count ICs and How FormFactor Overcomes These Challenges," https://www.formfactor.com/blog/2018/case-study challenges-when-probing-high-pad-count-ics/, 2018, accessed: Dec. 2025.

[13.5.6] mbelectronique.fr/Web PDF/1260 20191209091515 54 InfinityXT Specifications.pdf, 2019, accessed: Dec. 2025.

[13.5.7] "3D microprinted probes for testing at sub 20 µm pitch," swtestasia.org/2023proc/pdf/F04_01_KOELMANS_SWTestAsia2023.pdf, 2023, accessed: Dec. 2025.

[13.5.8] http://psirep.com/system/files/2025-08/FormFactor Probe Station Accessories Catalog.pdf, 2025, accessed: Dec. 2025.

[13.5.9] T. Fukushimaet al., "Integration of Damage-less Probe Cards Using Nano-TSV Technology for Microbumped Wafer Testing," in IEEE International 3D Systems Integration Conference (3DIC), 2021, pp. 1–4.

[13.5.10] "An Inside Look at 3D-DfT Standard IEEE Std 1838 2019," https://www.3dincites.com/2020/03/an-inside-look-at-3d-dft-standard-ieee-std-1838-2019, 2020, accessed: Dec. 2025.
[13.5.11] S. K. Goel et al., "Physical-Aware Interconnect Test for Multi-Die Systems Using 3Dblox Open Standard," in IEEE International Test Conference (ITC), 2024, pp. 451–459.
[13.5.12] P.-Y. Chuang et al., "Generating Test Patterns for Chiplet Interconnects with Optimized Effectiveness and Efficiency," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2024.
[13.5.13] P. Bhoumik et al., "Fault Modeling and Testing of Chiplet-to-Chiplet Interconnects in Fan-out Wafer-Level Packaging*," in IEEE International Test Conference (ITC), 2025, pp. 357–366.
[13.5.14] ——, "SMART: Scalable and Modular Architecture for Routing-Aware Testing of Fan-out Wafer-Level Packages*," in IEEE International Test Conference (ITC), 2025, pp. 310–319.
[13.5.15] Y.-C. Huang et al., "Efficient Built-In Self-Test Scheme for Inter-Die Interconnects of Chiplet-Based Chips," in IEEE International Test Conference (ITC). IEEE, 2024, pp. 149–156.
[13.5.16] "Testing For Thermal Issues Becomes More Difficult," https://semiengineering.com/testing-for-thermal-issues-becomes-more-difficult/, 2024, accessed: Dec. 2025.
[13.5.17] Synopsys, "Multi-Die System White Paper," https://www.synopsys.com/content/dam/synopsys/solutions/multi-die/white-papers/multi-die-wp. pdf, accessed: 2025-12-10.
[13.5.18] AMD, "XGMI Configuration — AMD Instinct Virtualization Driver," accessed: 2025-12-10.
[13.5.19] Intel, "Intel CSME Security White Paper," https://www.intel.com/content/dam/www/public/us/en/security-advisory/documents/intel-csme-security-white-paper.pdf, accessed: 2025-12-10.
[13.5.20] AMD, "AMD Pro Technologies Security White Paper," https://www.amd.com/content/dam/amd/en/documents/products/processors/technologies/amd-pro-technologies-security-white-paper.pdf, accessed: 2025-12-10.
[13.5.21] Caspia Technologies, "CODAx One-Pager (FINAL)," https://caspiatechnologies.com/wp-content/uploads/2025/10/CODAx-One-Pager-FINAL.pdf, accessed: 2025-12-10.
[13.5.22] MITRE, "Common Weakness Enumeration (CWE)," https://cwe.mitre.org/, accessed: 2025-12-10.
[13.5.23] MITRE, "Common Vulnerabilities and Exposures (CVE)," https://cve.mitre.org/, accessed: 2025-12-10.
[13.5.24] D. D. Sharma, G. Pasdast, Z. Qian, and K. Aygun, "Universal chiplet interconnect express (UCIe): An open industry standard for innovations with chiplets at package level," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 9, pp. 1423–1431, 2022.
[13.5.25] PCI-SIG, "PCI Express Base Specification, Revision 6.0," https://pcisig.com/specifications/pciexpress, accessed: 2025-12-10.
[13.5.26] CXL Consortium, "CXL 2.0 Specification," https://computeexpresslink.org/wp-content/uploads/2024/02/CXL-2.0-Specification.pdf, accessed: 2025-12-10.
[13.5.27] UALink Consortium, "UALink 200 Specification, v1.0 (Evaluation Copy)," accessed: 2025-12-10.

[13.8.1] R. S. Patti, "Three-Dimensional Integrated Circuits and the Future of System- on-Chip Designs," in *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214–1224, 2006.
[13.8.2] R. Tummala, *Fundamentals of Device and Systems Packaging: Technologies and Applications*, McGraw-Hill, 2019.
[13.8.3] D. Milojevic et al., "Ultrafine Pitch 3D Stacked Integrated Circuits: Technology, Design Enablement, and Application," in *Handbook of 3D Integration, Volume 4: Design, Test, and Thermal Management*, Wiley, 2019.
[13.8.4] S. K. Lim et al., "EDA Needs, Gaps, and Open Problems for Wafer-Bonded Digital Logic 3D ICs," *IEEE Trans. Components, Packaging, and Manufacturing Technology*, 2026.
[13.8.5] L. Zhu and S. K. Lim, "INVITED: Design Automation Needs for Monolithic 3D ICs: Accomplishments and Gaps," *ACM/IEEE Design Automation Conference*, 2023.
[13.8.6] Q. Hao et al., "Survey of Chiplet Technology: SoC Architecture, Interconnect, EDA, and Advanced Packaging," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2025.
[13.8.7] S. K. Lim et al., "EDA Needs, Gaps, and Open Problems for Wafer-Scale Glass Interposers," *IEEE Trans. Components, Packaging, and Manufacturing Technology*, 2026.
[13.8.8] J. Xie et al., "Architectural benefits and design challenges for three-dimensional integrated circuits," *IEEE Asia Pacific Conference on Circuits and Systems*, 2010.

[13.8.9] K. Banerjee et al., "3-D ICs: a novel chip design for improving deep- submicrometer interconnect performance and systems-on-chip integration," in *Proceedings of the IEEE*, vol. 89, no. 5, pp. 602–633, 2001.
[13.8.10] S. Kim et al., "Signal Integrity and Computing Performance Analysis of a Processing-In-Memory of High Bandwidth Memory (PIM-HBM) Scheme," in *IEEE Trans. on Components, Packaging and Manufacturing Technology*, vol. 11, no. 11, pp. 1955–1970, Nov. 2021.
[13.8.11] M. Choi et al., "Substrate-Embedded Microfluidic Cooling of Distributed Vertical Power Delivery Architectures for High-Performance Computing Processors," *IEEE Trans. on Components, Packaging and Manufacturing Technology*, vol. 15, no. 9, pp. 1912–1920, 2025.
[13.8.12] S.-Y. Lee et al., "Signal-Integrity-Aware Interposer Bus Routing in 2.5D Heteroge- neous Integration," *IEEE Asia and South Pacific Design Automation Conference*, 2022.
[13.8.13] E. Haque et al., "Invited: EDA for Heterogeneous Integration," *ACM/IEEE Design Automation Conference*, 2025.
[13.8.14] H.-H. S. Lee and K. Chakrabarty, "Test Challenges for 3D Integrated Circuits," in *IEEE Design & Test of Computers*, vol. 26, no. 5, pp. 26–35, 2009.
[13.8.15] S. Hwang et al., "Real-Time 3-D Thermal Simulation of Advanced Packages via Generative Adversarial Networks," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 44, no. 7, pp. 2439–2450, 2025.
[13.8.16] A. Graening et al., "CATCH: A Cost Analysis Tool for Co-Optimization of Chiplet-Based Heterogeneous systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2025.
[13.8.17] UCIe Consortium, "Universal Chiplet Interconnect Express (UCIe) Specification," 2022.
[13.8.18] SEMI, "Heterogeneous Integration Roadmap (HIR)," SEMI, 2023.
[13.8.19] B. W. Ku et al., "Compact-2D: A Physical Design Methodology to Build Two- Tier Gate-level 3D ICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 6, pp. 1151–1164, 2020.
[13.8.20] V. Hsu, "2.5D & 3DIC Advanced Packaging: An EDA Perspective," *International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, 2022.
[13.8.21] J. Kim et al., "Architecture, Chip, and Package Co-design Flow for Interposer- based 2.5D Chiplet Integration Enabling Heterogeneous IP Reuse," *IEEE Trans- actions on Very Large Scale Integration Systems*, vol. 28, no. 11, pp. 2424–2437, 2020.
[13.8.22] Moongon Jung et al., "TSV Stress-Aware Full-Chip Mechanical Reliability Analysis and Optimization for 3D IC," *Communications of the ACM*, vol. 57, no. 1, pp. 107–115, 2014.
[13.11.1] P. Kashyap, N. Rouf, Y. Choi, C. Cheng, P. Franzon, "RAG-EM: Retrieval- Augmented Generation for Electromagnetic System Design," *IEEE EMC+SIPI 2025*.
[13.11.2] P. Kashyap, Y. Wen, N. Rouf, Y. Choi, C. Cheng, P. Franzon, "Agentic AI for SI/PI: From concept to complete design execution," *DesignCon*, 2025.
[13.11.3] H. He, N. Chang, A. Kumar, J. Yang, W. Xia, L. Lin, J. Yen, H. Jiang, R. Ranade, "Parallel Per-tile Activation with Linear Superposition of Thermal Response for Solving Arbitrary Power Pattern in 3DIC Thermal Simulation," *IEEE MLCAD*, 2024.
[13.11.4] Z. Lu, A. Kumar, N. Chang, H. He, L. Lin, W. Xia, J. Yang, G. Qu, "Multimodal Learning-Based Thermal Solver for 3DICs with Arbitrary Chip Dimensions and Power Distributions," *IEEE MLCAD*, 2025.
[13.11.5] P. Kashyap, C. Cheng, et al., "Foundational Model Approach for SI/PI Analysis Using Large Language Model Techniques," *DesignCon*, 2025.
[13.11.6] Y. Choi, C. Cheng, R. Kashyap, P. Franzon, W. Sun, M. Chughtai, "Data-driven PAM4 SerDes Modeling w/ Generative Adversarial Network," *DesignCon 2023*.
[13.11.7] P. Kashyap, P. Ravichandiran, L. Wang, D. Baron, C. Wong, T. Wu, P. Franzon, "Thermal Estimation for 3D-ICs through Generative Networks," *IEEE 3DIC*, 2023.
[13.11.8] M. Adams, X. Li, L. Boucinha, S. Kher, P. Banerjee, J. Gonzalez, "Hybrid Digital Twins: A Primer on Combining Physics-Based and Data Analytics Approaches," *IEEE Software*, 2022.
[13.11.9] H. Kim, F. Berto, S. Choi, K. Kim, T. Shin, K. Son, Y. Kim, J. Kim, "Towards A Fully Customizable Chiplet Ecosystem: Signal Integrity-Constrained Throughput Maximization via Symmetricity Guided Imitation Learning," *DesignCon 2026*.

[13.11.10] S. Ryu, S. Lee, D. Park, S. Lee, H. Kim, J. Lee, S. Yong, S. Song, H. Kim, J. Kim, S. Ahn, “Holistic Design Optimization of 3D-IC Package Substrate Interconnections in Multiple Power Domain Environments based on Hierarchical Reinforcement Learning,” *DesignCon 2026*.