Physical Design Challenges in Modern Heterogeneous Integration

Yao-Wen Chang
Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University
Taipei, Taiwan 10617
ywchang@ntu.edu.tw

ABSTRACT

To achieve the power, performance, and area (PPA) target in modern semiconductor design, the trend to go for More-than-Moore heterogeneous integration by packing various components/dies into a package becomes more obvious as the economic advantages of More-Moore scaling for on-chip integration are getting smaller and smaller. In particular, we have already encountered the high cost of moving to more advanced technology and the high fabrication cost associated with extreme ultraviolet (EUV) lithography , mask, process, design, electronic design automation (EDA), etc. Heterogeneous integration refers to integrating separately manufactured components into a higher-level assembly (in a package or even multiple packages in a PCB) that provides enhanced functionality and improved operating characteristics. Unlike the on-chip designs with relatively regular components and wirings, the physical design problem for heterogeneous integration often needs to handle arbitrary component shapes, diverse metal wire widths, and different spacing requirements between components, wire metals, and pads, with multiple cross-physics domain considerations such as system-level, physical, electrical, mechanical, thermal, and optical effects, which are not well addressed in the traditional chip design flow. In this paper, we first introduce popular heterogeneous integration technologies and options, their layout modeling and physical design challenges, survey key published techniques, and provide future research directions for modern physical design for heterogeneous integration.

CCS CONCEPTS

• **Hardware → Physical design (EDA)**.

KEYWORDS

Heterogeneous Integration, Physical Design, Layout, Partitioning, Floorplanning, Placement, Routing, Silicon Photonics, Optical Routing, Chip-Package-Board Co-Design, Machine Learning

ACM Reference Format:
Yao-Wen Chang. 2024. Physical Design Challenges in Modern Heterogeneous Integration. In *Proceedings of the 2024 International Symposium on Physical Design (ISPD '24), March 12–15, 2024, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626184.3639690

This work was partially supported by AnaGlobe, ASUS, Delta Electronics, Google, Maxeda/Synopsys, TSMC, NSTC of Taiwan under Grant NSTC 110-2221-E-002-177-MY3, NSTC 112-2218-E-002-033, NSTC 112-2223-E-002-020, and NSTC 112-2119-M-002-017.

ISPD '24, March 12–15, 2024, Taipei, Taiwan

ACM ISBN 979-8-4007-0417-8/24/03…$15.00
https://doi.org/10.1145/3626184.3639690

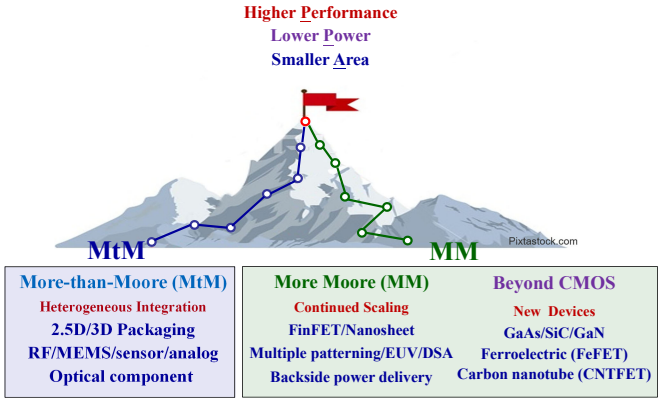

Figure 1: Three main options in semiconductors to achieve the power, performance, and area (PPA) targets, including More than Moore, More Moore, and Beyond CMOS [11].

1 INTRODUCTION

There are quite a few critical drivers for advancing semiconductor technologies, including artificial intelligence (AI), high-performance computing, the Internet of Things (IoTs), and virtually all types of computing. These technology drivers have created a world with massive amounts of information, leading to the era of data explosion. According to DOMO Data Never Sleeps, every minute on the Internet in 2022, 5.9M user queries were conducted on Google, 500 hrs of video were uploaded to YouTube, and 1.7M contents were posted on FaceBook [1]. It is predicted that 147 ZB data will be generated globally in 2024, and the data size is expected to be over 181 ZB next year [7]. The tremendous data explosion drives the urgent need for computing devices with higher performance, better power efficiency, and smaller chip areas (i.e., lower cost).

To achieve the performance, power, and area target, as shown in Figure 1, we have three main directions in semiconductors [11]: (1) More Moore with continued CMOS scaling to extend Moore's Law by using advanced transistor technologies (such as FinFET and nanosheet), patterning technologies (such as multiple patterning, extreme ultraviolet lithography (EUV), and directed self-assembly (DSA)), and interconnect technologies (such as through silicon vias (TSVs), backside power delivery by moving the power delivery network to the backside of a wafer, and optical interconnect); see Figure 2 [11]. (2) Beyond CMOS, with new devices such as compound semiconductors (say silicon carbide (SiC)). (3) More-than-Moore, with 2.5D or 3D heterogeneous integration to pack multiple components into an integrated system, where the components could

be analog/RF devices, sensors, MEMS, biochips, and even optical components. The challenges in these directions have brought many exciting research opportunities.

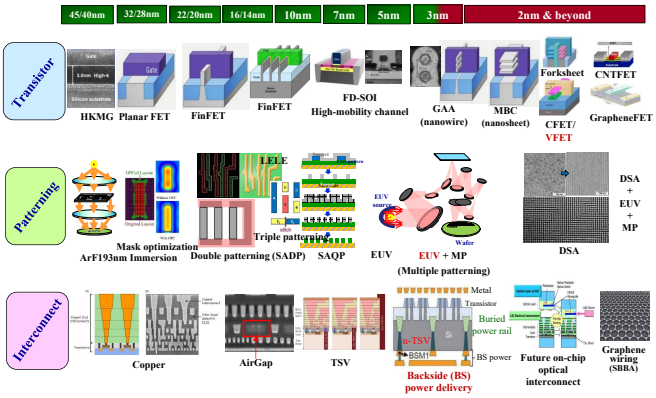

Figure 2: More-Moore technology landscape with transistor, patterning, and interconnect technology evolutions [11].

Are the directions on More Moore and More-than-Moore brand new ideas? Moore already envisioned the two directions and even electronic design automation (EDA) needs in his well-known paper on Moore's Law published in 1965 [54]. He pointed out that the total cost of making a particular system function must be minimized. He envisioned that the integration advantages would bring about a proliferation of electronics, pushing this science and technology into many new areas. On More Moore, he observed that the complexity for minimum component costs would increase at a rate of roughly a factor of two per year, which is the frequently quoted Moore's Law. On More than Moore, very amazingly, he commented that it might prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected [54], often called *heterogeneous integration* nowadays. He even envisioned the EDA necessity for circuit designs that perhaps newly devised design automation procedures could translate from logic diagrams to technological realization without any special engineering [54]. So we are standing right on the shoulder of a giant to develop and envision emerging technologies.

As the economic advantages of More-Moore scaling for chip design are getting smaller and smaller, the trend to go for the More-than-Moore option becomes more obvious. In particular, we already encounter the high cost of moving to more advanced technology nodes (say, the technology development for 2nm and beyond), and the high fabrication cost associated with EUV, mask, process, design, EDA, etc. As a result, only limited foundries such as TSMC, Intel, and Samsung can afford the high manufacturing costs.

In contrast, More-than-Moore heterogeneous integration with 2.5D/3D stacking provides promising solutions to the increasing cost of more-Moore scaling. Heterogeneous integration refers to integrating separately manufactured components into a higher-level assembly that provides enhanced functionality and improved operating characteristics. We could even integrate electrical and optical components in a single package or a chip for better power, transmission bandwidth, and speed. Even better, we have many options for 2.5D/3D stacking, including silicon interposers, integrated fan-out (InFO) packages, chips on wafer on substrate (CoWoS), chiplets, and monolithic 3D integration.

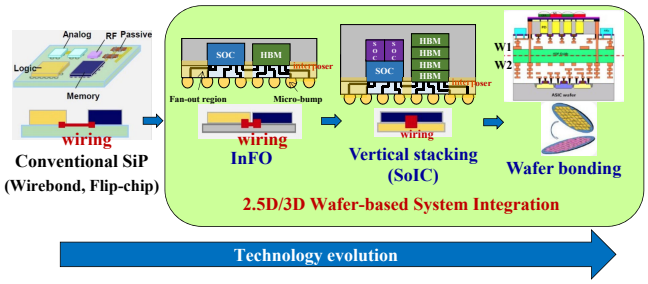

Figure 3: Advanced packaging evolution, with inter-die wiring becoming shorter and shorter.

Packaging technologies are evolving quickly from traditional system-in-package with wire-bond and flip chips to 2.5D InFO packages, 3D vertical stacking, and even wafer bonding, where their inter-die wiring becomes shorter and shorter. With 2.5D/3D heterogeneous integration, we can enjoy better system values on the cost, performance, power, form factor, heterogeneity, and security. However, there are significant issues to handle before we can enjoy the ride, including design complexity, yield, heat, TSV and other component costs, mechanical stress, testing, standardization, and even the whole ecosystem. As the design complexity grows dramatically in modern designs, 2.5D/3D heterogeneous integration becomes effective for system performance, power, and cost optimization. Many heterogeneous integration implementation techniques are proposed [11], as shown in Figure 4, including CoWoS [39], InFO packages [46], interposer-based 2.5D packages [30], silicon bridge dies [42], chiplets, and stacked monolithic 3D ICs [41]. Figure 5 shows TSMC's latest advanced packaging technologies, 3D Fabric, including InFO, CoWoS, and SoIC.

Figure 6 illustrates a generic architecture of a heterogeneous integration system consisting of optical, digital, analog, sensor, MEMS, etc. in a System-in-Package (SiP). Multi-physics domains such as systems, physical, electrical, thermal, mechanical, and optical designs should be addressed when designing a comprehensive physical design flow for such heterogeneous integration. Unlike traditional physical design (say, placement and routing), timing, electrical, thermal, and mechanical issues need to be considered for electrical-optical co-design. With the increasing demands for large bandwidth, high performance, and low-power transmission in modern system designs, electrical wires alone might not meet the design requirements, and optical communication based on nanophotonic devices and interconnections has attracted much attention recently, due to its large bandwidth, low interconnection delay, and low-power consumption. As a result, semiconductor giants like Intel and TSMC are actively exploring the technologies for integrating electrical and optical interconnections in a modern 2.5D/3D heterogeneous integration system, and TSMC has even announced to integrate silicon photonics with optical components in its advanced packages by 2025 [6].

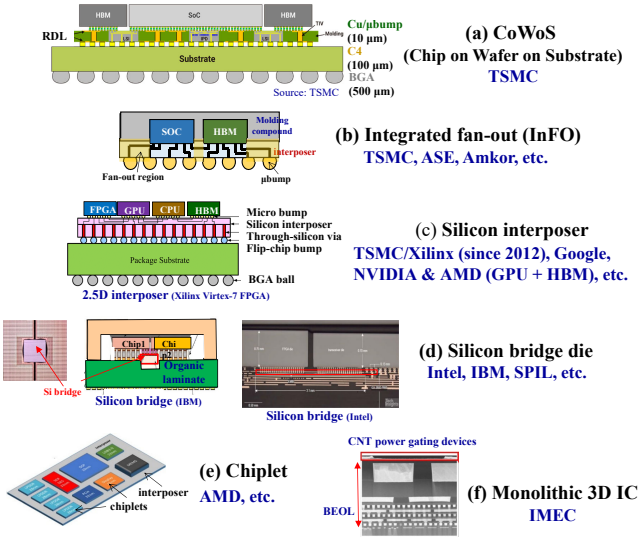

Figure 4: Multiple options for heterogeneous integration.

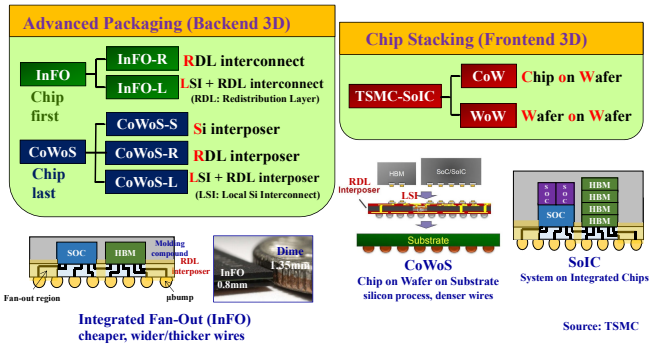

Figure 5: TSMC's 3D Fabric.

To address the heterogeneous integration issues, like the International Technology Roadmap for Semiconductors (ITRS) [4] on devices and the IEEE International Roadmap for Devices and Systems (IRDS) [3] on systems, researchers organized technical Working Groups and published the Heterogeneous Integration Roadmaps [2], first in 2015 and the latest in 2023. The roadmap identifies technology requirements and potential solutions to inspire collaboration between industry and academia to accelerate progress. The 2023 edition of the Heterogeneous Integration Roadmap includes 24 chapters on key applications such as high-performance computing and data centers, IoTs, medical health and wearables, automotive, and technology challenges such as photonics, co-design, modeling and simulation, interconnects, test, cyber security, thermal, wafer-level packaging, reliability, and even the business issue on supply chains.

The rest of this paper is organized as follows. Section 2 presents challenges to achieving heterogeneous integration targets. Section 3 provides some future research directions in physical design for heterogeneous integration, and Section 4 concludes this paper.

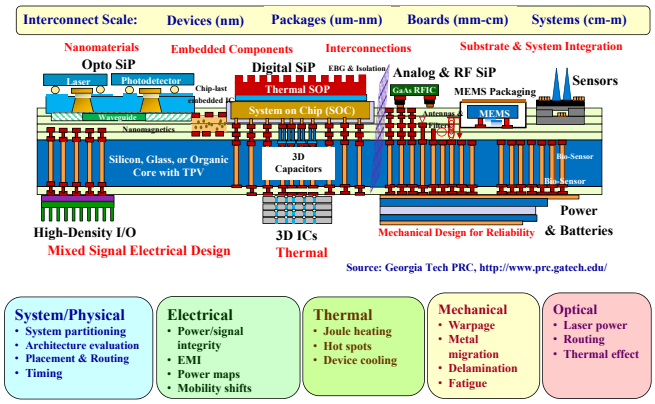

Figure 6: A typical advanced packaging structure with multiple cross-physics domain issues for heterogeneous integration.

2 PHYSICAL DESIGN CHALLENGES

To achieve high-quality 2.5D/3D heterogeneous integration, we encounter significant challenges in physical design (partitioning, floorplanning, placement, and routing) for chip/package/board design and co-design with advanced packages and optical communication, considering multi-physics issues such as physical, timing, electrical, thermal, mechanical, and optical effects. Due to the space limitation, we intend to identify only the most crucial challenges and show examples of handling the challenges in the following, instead of providing comprehensive surveys on these issues and previous work.

- **3D system-level partitioning:** Mixed-logic and -memory 3D partitioning into multiple dies with different technologies for PPA optimization incurs significant challenges. Process technologies for multiple dies and corresponding interconnect/via/bump parameters need to be considered for the overall cost optimization.
- **Cross-physics domain floorplanning:** Integrating multiple dies into an advanced package may suffer from severe electrical, thermal, and mechanical issues such as substrate noises, Joule heating, stress, warpage, and delamination problems. For example, warpage caused by the mismatch in coefficients of thermal expansion between different manufacturing materials, leading to deformation and malfunction in the integrated package, is a significant challenge in advanced packaging. As a result, the industry is eager to find a solution for warpage optimization [33]. The work by Hsu et al. [35] proposes the first warpage-aware floorplanning algorithm for heterogeneous integration, where a fast qualitative warpage model for a multi-die package based on Suhir's modeling (other than the time-consuming finite element analysis) is adopted to model warpage effects, and simulated annealing based on the transitive closure graph representation is used to optimize the warpage-aware floorplan design.
- **Multi-die, multi-technology placement:** 3D/2.5D placement problems have been studied extensively [34]. Most previous work considers multiple dies/tiers with the same process technology; so only "static" physical constraints (e.g.,

same dimensions for a circuit component even in different dies/tiers) are considered. However, the dies integrated in an advanced package could be fabricated with heterogeneous technology nodes and connected by various technologies, so the dimensions of a circuit component (a standard cell or a macro) are "dynamically" changed with different die assignments. This dynamic physical constraint creates a challenging chicken-and-egg problem seldom seen before.

In the 2022 and 2023 CAD Contest at ICCAD, Synopsys provided two versions of the face-to-face (F2F) two-die placement problem, where signals are connected by hybrid bonding terminals [36, 37]. The most significant challenge is that each die can be fabricated by different technologies, so the dimensions of a block (a standard cell or a macro) change with die assignments. See Figure 7 for an example of F2F two-die placement.

To address this emerging challenge, researchers considered the dynamic physical constraint to optimize the 3D placement effectively [8, 13]. The work [13] presents a multi-technology weighted-average wirelength model [34], a weighted inter-die-connection cost to control the net-degree distribution, and a via-cell co-optimization technique to further improve placement quality.

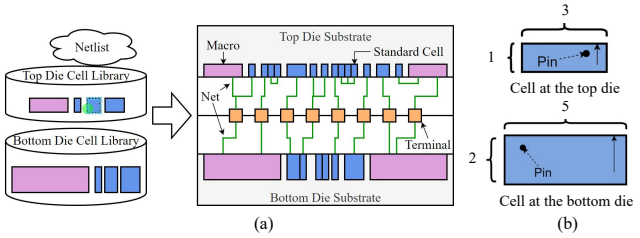

Figure 7: An example face-to-face (F2F) two-die placement. (a) Hybrid bonding terminals connect signals between the top and bottom dies. (b) Each die can be fabricated by different technologies, so block shapes/dimensions change with assignments to the top or bottom die.

- **Chip-package routing:** Due to the fast-growing design complexity, it is desirable to have more I/O counts for packaging technologies. The flip-chip package is introduced for larger I/O counts and higher integration density. The flip-chip package connects a chip to an external circuitry such as package carriers or printed circuit boards (PCBs). It can provide larger areas for I/Os, achieve higher performance with shorter interconnections, and keep signals isolated from environmental hazards or influences [44]. Figure 8 shows a typical flip-chip package structure. A chip is flipped over and mounted on a package carrier. To facilitate design changes, an extra metal layer, called a Redistribution Layer (RDL), is added to establish the interconnection between I/O pads and bump pads, i.e., RDL routing.

  The flip-chip package can be categorized into two major types [44]: (1) the *peripheral-I/O* flip chip [27, 28] and (2) the *area-I/O* flip-chip [22, 23], as illustrated in Figure 9. The area-I/O flip-chip routing is generally harder than the peripheral-I/O one because the routing region in an area-I/O flip-chip is more congested and irregular, as illustrated in Figure 9(b). As a result, Voronoi Diagram and Delaunay Triangulation are often applied to handle their irregularity [22, 23]. Nevertheless, the area-I/O flip-chip can often achieve shorter wirelength and better flexibility.

  The RDL routing can be classified into three major types [44]: (1) *free-assignment* (FA) routing [22, 23, 27, 28], (2) *pre-assignment* (PA) routing [25, 26, 45], and (3) *unified-assignment* (UA) routing [29]. For FA routing, a router can assign each I/O pad to an arbitrary bump pad, where the generic maximum-flow formulation can often be employed, as illustrated in Figure 11. In contrast, connections between I/O and bump pads are predefined for pre-assignment routing, and the assignments cannot be changed. Pre-assignment routing is harder than free-assignment one because the mapping between I/O and bump pads impose more restricted constraints [44]. Consequently, previous work often resorts to integer linear programming (ILP) or dynamic programming. ILP is a generic method for optimization with a brute-force nature that enumerates all possible solutions blindly, thus typically time-consuming. Dynamic programming leverages the regularity of the bump and I/O structures to decompose the problem into well-structured, ordered subproblems to minimize routing detours. For UA routing, some net assignments between I/O and bump pads are predefined, and some are not. Consequently, FA and PA routing shall be considered simultaneously to achieve better design quality [9, 15, 29, 46, 62]. As commented in [44], the network-flow-based formulation well suits the free-assignment routing problem because it can determine the global routes concurrently, and the flow running freely in a network well matches the intrinsic nature of free assignment of nets in a flip chip (see Figure 11). The key to a network-flow-based formulation lies in an accurate tile model where the capacities of nodes and edges can precisely capture the routing resource. There are two popular tile models: (1) the rectangular-tile model is naturally formed with the regular structure of a bump-pad region [27, 28], while the triangular-tile model is formed with Delaunay triangulation [9, 10, 22, 23]. The triangular-tile model is typically more flexible for various applications (area-I/O flip chips, irregular structures, any-obtuse-angle routing, etc.), but it is less accurate in capacity modeling for a single triangular region and the interaction between two adjacent triangular regions needs to be considered for accurate modeling. Figure 10 classifies flip-chip routing problems with corresponding publications based on the I/O pad structures and pad assignment styles.

  More advanced flip-chip structures incur various bump-pad structures such as hexagonal array (Y-architecture) patterns that can achieve higher integration density and larger I/O counts [55] or multiple routing layers with irregular bump-pad structures due to different technologies used to manufacture packaged chips [9, 14, 62]. Traditional RDL routers use only 90- and 135-degree turns for routing. In contrast,

routing for advanced RDLs can be any obtuse angle, leading to larger routing solution spaces and shorter total wirelength [18]. Further, differential pairs and impedance matching among signals and obstacles often need to be addressed in advanced packaging [10, 24].

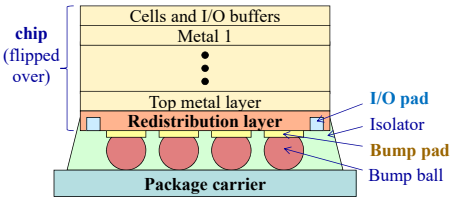

Figure 8: A typical flip-chip package structure.

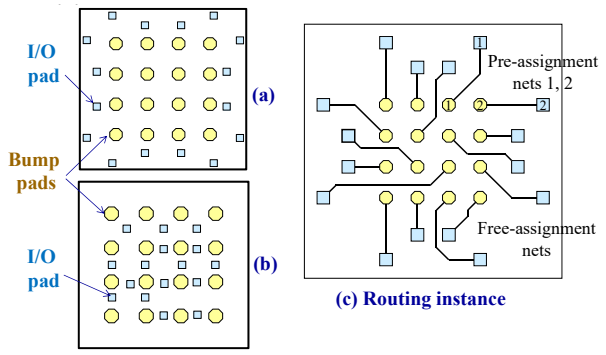

Figure 9: RDL routing structures. (a) Peripheral I/Os. (b) Area I/Os. (c) A routing instance with pre-assignment nets 1 and 2 and other free-assignment ones.

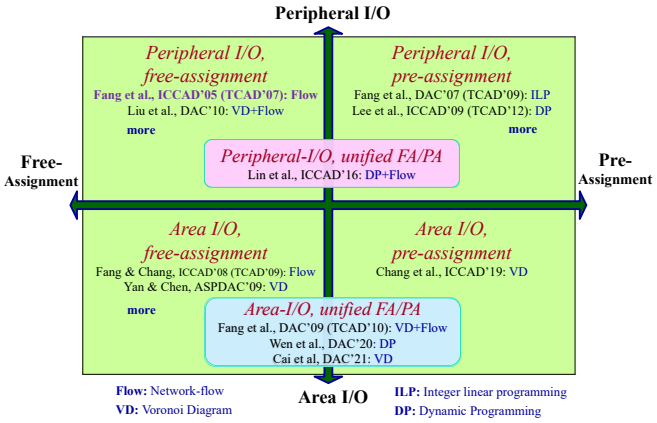

Figure 10: Existing RDL routing works regarding the I/O pad structures and pad assignment styles.

- **Package-board routing:** For package-board routing, the typical objective is to develop routability-driven solutions under complex constraints. Dense inline and staggered bump structures [31, 32], stringent electrical/physical constraints (skew/impedance matching for a matched group, differential pairs, etc.) [10, 24], and complex design rules (obstacles, obtuse angle routes, irregular component shapes, non-uniform wire widths/spacing, irregular stacked vias, etc.) [14, 18, 48, 49] impose substantial challenges for advanced package-board routing.

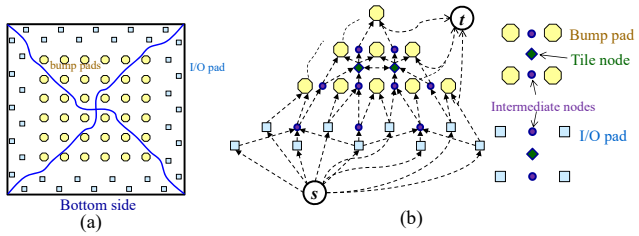

Figure 11: Network-flow-based formulation for free-assignment routing.

- **Chip-package-board co-design:** Heterogeneous integration technologies have complicated the designs of chips, packages, and PCBs. Chip-package-board co-design is strongly desired to improve the deign convergence and quality of the whole system. Figure 12 illustrates chip-package-board co-design, where a simple combination of traditional tools for each level (chip, package, or board) is insufficient to achieve high-quality designs. Take Figures 12(b) and (c) as examples. If the layout design between different levels is not considered simultaneously, the final layout quality could be only suboptimal, incurring longer and more congested wirings [56], as shown in Figure 12(b). The work [24] presented an ILP-based placement and routing formulation for chip-package-board co-design considering differential pairs. Although the algorithm guarantees to find an optimal solution, this algorithm is time-consuming.
  Lee et al. presented an efficient and effective board-driven Λ-shaped chip-package-board layout co-design flow [43], where the layout information among chip, package, and board is well conveyed bidirectionally. The co-design flow comprises two passes, first from boards to chips and then from chips to boards. The first pass sends the PCB input information to the package and then the chip, where the chip keeps the design information from the PCB and the package for reference but defers its design decision because the information might lack the global view. With the information from chips at the second pass, the package and board design can have a more global view to fix the potentially wrong decisions made during the first pass, while some partial results from the first pass can still be reused. By exchanging the bi-directional information among different levels, the decisions made in earlier stages can be refined in later stages to achieve the desired solutions. Specifically, the co-design flow consists of four major stages: (1) inverse escape routing, (2) package-aware I/O placement, (3) flip-chip bump reassignment, and (4) RDL and escape rerouting. Here, escape routing refers to the algorithm used to route the I/O or bump

pads to the lines that can escape to the area surrounding the die.

For modern heterogeneous integration, a flexible, robust, and vertically integrated physical design flow considering multiple objectives and complex constraints is urgently needed.

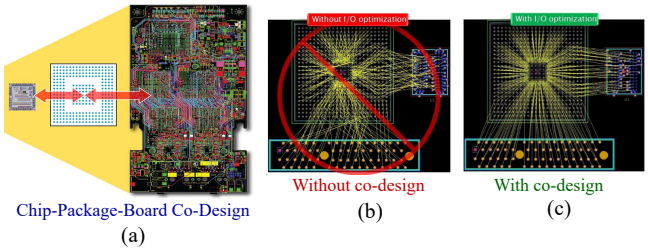

Figure 12: Chip-package-board co-design [56]. (a) Chip-package-board integration of a system. (b) If such co-design is not considered simultaneously, the final layout quality could be only suboptimal, incurring longer and more congested wirings. (c) Co-design is considered to achieve the desired solution quality.

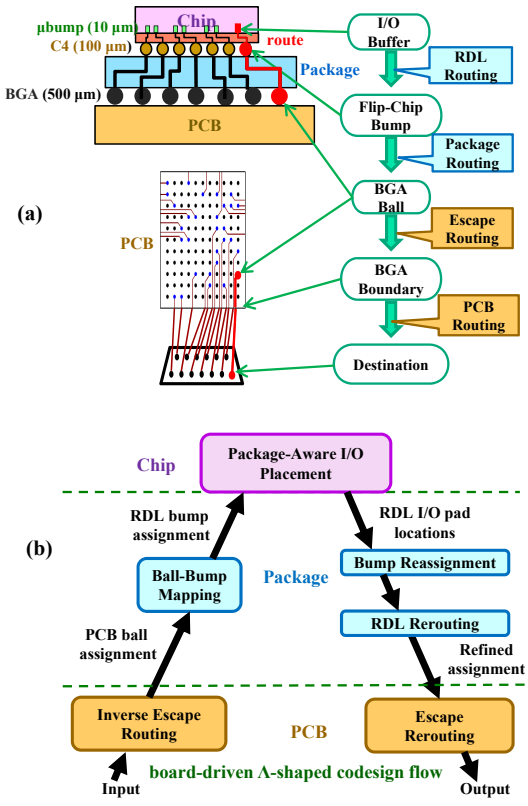

Figure 13: Chip-package-board co-design flow [43]. (a) Upper part: Cross view of a single chip with flip-chip packaging mounted on a PCB; Lower part: Aerial view of the BGA on the PCB; Right-hand side: Corresponding routing problems on the signal path. (b) The board-driven Λ-shaped co-design flow.

- **Cross-physics-domain consideration:** Modern heterogeneous integration incurs challenges across multiple physics domains, such as system, physical, electrical, thermal, mechanical, and optical designs. Figure 14 lists some critical challenges for each domain. Due to space limitations, only some most critical challenges are highlighted. For system design, for example, we encounter significant challenges on mixed-logic and -memory 3D system-level partitioning into multiple dies with different technologies for power, performance, thermal, and area optimization, system-level heterogeneous integration modeling and simulation, heterogeneous device applications and computing, architecture with heterogeneous integration components, cost evaluation and decision, hardware security and reliability, etc. Tools for 3D partitioning, floorplanning, placement, routing, and post-layout optimization with different technologies for multiple dies must be reinvented for the physical domain. We also need to handle issues with chip-package-board co-design and cross-domain timing analysis/optimization. For electrical effects, power/signal integrity and EMI prevention, backside power delivery, stacking P/G network optimization, stacking cross-die static timing analysis (STA) and electrostatic discharge (ESD), inter-die coupling, and substrate RLC modeling and optimization become more challenging than ever. For thermal effects, stacking interconnect, full-system thermal analysis and electromigration reliability, Joule heating, hotspot detection/handling become significant issues. For the mechanical effects, we first need to model the warpage and stress due to the mismatch in thermal expansion coefficients between different materials and develop warpage- and stress-aware optimization techniques [35]. Delamination, metal migration simulation and optimization, and fatigue reliability incur emerging challenges. For the optical domain, optical routing [53], thermal-aware power device placement [40, 60], electrical and optical co-design [51], system-based optical device analysis and optimization, and laser power network synthesis have become critical challenges that require decent solutions.

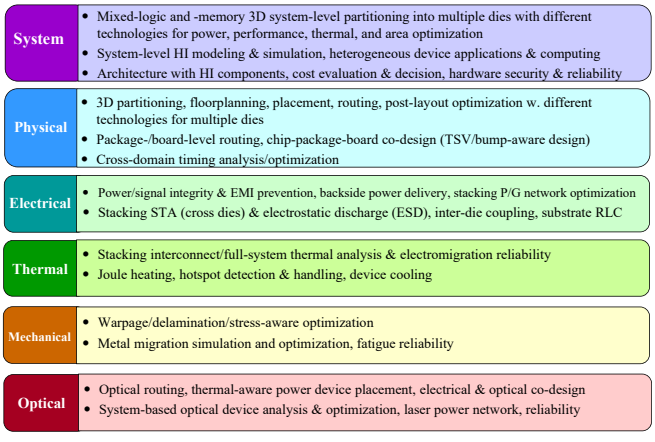

Figure 14: Multiple physics-domain challenges for heterogeneous integration.

- **Silicon photonics and optical routing:** Silicon photonics uses photons to transmit data and signals at higher speed, bandwidth, and power efficiency than copper interconnects that suffer from limited bandwidth, current leakage, and crosstalk between adjacent wires. As a result, silicon photonics has shown great promise for data centers and sensing applications. In 2015, researchers demonstrated the first photonic-electronic hybrid processor that uses light for ultrafast communications by packing two processor cores with more than 70 million transistors and 850 photonic components onto a 3-by-6-millimeter chip [59]. An optical network-on-chip can use silicon-on-insulator (SOI) photonics chips as an interposer, so photonics has become a valuable resource for many-core systems. TSMC has announced the inclusion of silicon photonics components in its advanced packages by 2025 [5]. Silicon photonics brings many research challenges and opportunities, including optical placement and routing [17, 20, 52, 53], electrical and optical co-design and co-packaging [47, 51]. The optical devices are very thermal-sensitive, so it is critical to explore thermal-aware optical design (device placement), power device placement, laser power reduction and network synthesis, and laser integration [40, 51, 60].
  As the design complexity increases dramatically, the optical network-on-chip (ONoC) with wavelength division multiplexing (WDM) becomes an attractive alternative for on-chip signal transmission for large bandwidth and low power consumption. Figure 15 shows the WDM waveguide structure, where electrical signals are converted to optical ones, transmitted through optical waveguides, and then converted back to electrical signals when received. Lasers are used for optical transmission in the optical waveguides, and microring resonators (MRRs) are used to control the signal transmission directions. MRRs are critical components in a WDM system for controlling signal transmission direction, which have small sizes and low power consumption but are sensitive to temperature fluctuations and process variation. By using multiplexers and demultiplexers, signals with different wavelengths can be clustered into a single WDM waveguide for transmission, saving more routing resources and reducing potential crosses [52, 53].
  The operating power reached at all receiving ports should be larger than a threshold value to maintain the signal integrity. The maximum transmission loss in an ONoC determines the minimum required laser power, so maximum transmission loss minimization is crucial in ONoC designs [17, 20, 51–53]. The transmission loss in optical routing comprises several major sources, including crossing loss (incurred by two intersecting wires), bending loss (occurs when a signal is transmitted through a bending wire), through loss (induced when a signal passes an MRR without activating it), path loss (induced when a signal is transmitted through a long wire), and drop loss (induced when a signal is switched from one waveguide to another). Besides, using WDM waveguides could also induce wavelength power when multiple signals are clustered into a WDM waveguide, and different wavelengths are needed for signal transmission, requiring more laser power. The WDM-aware optical routing problem to minimize total wirelength, transmission loss, and wavelength power under some constraints, such as the WDM waveguide capacity, is critical to the advanced packaging with silicon photonics components [20, 52, 53].

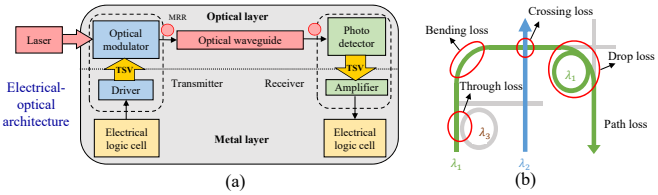

Figure 15: Optical routing. (a) Waveguide architecture: A signal from an electrical component is converted into an optical signal, transmitted on an optical layer, and then converted back to an electrical signal when received. (b) An optical routing instance with loss illustrations.

- **Design culture issues:** There are significantly different practices/cultures in chip and PCB designs (see Figure 16). Chip design instances are often large-scale, regular, homogeneous and arranged with uniform design rules for wire widths, spacings, and component dimensions; in contrast, PCB designs often encounter relatively small-scale, irregular, heterogenous components and must consider complex design rules with non-uniform wire widths and clearance rules and even 3D irregular and highly heterogeneous components. Besides, power loops, heat sink insertion, and electromagnetic interference are also influential considerations in PCB layout designs, making the design more complex.
  To achieve higher design quality and generalize the design automation capability for PCB and package designs, we shall learn from the chip-level design methodology to make design instances more regular and uniform and follow some reference design flow to simply the design rules. In this way, design tools can be generalized, and their capability can be maximized to achieve high-quality PCB/package designs. To achieve this ultimate goal, we would need an ecosystem to bring experts from various domains to work together, where all areas of EDA, circuit design, manufacturing, and fabrication equipment across geopolitical borders are essential to the advancement. Together, we can go very far!
- **Machine-learning-assisted optimization:** Emerging artificial intelligence (AI) technologies have reshaped the EDA landscape. EDA and AI are complementary, like yin and yang in oriental philosophy. For example, the tremendous AI needs have driven us to redefine the architecture and the way of computing, like neuromorphic and in-memory computing, which have empowered AI applications. In an AI chip that implements a deep neural network, its very dense interconnections between two layers of neurons cause severe routability problems never seen before. As a result, even leading layout design tools cannot handle the routing congestion problems for AI chips well [16]. The interactions between AI and EDA (AI for EDA and EDA for AI) have

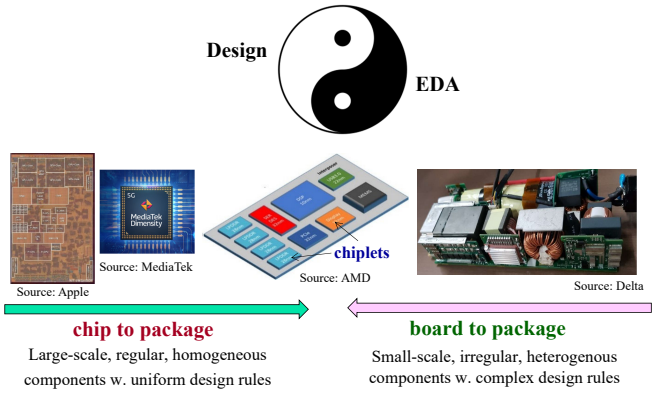

Figure 16: Cultural differences between on-chip and PCB designs for advanced packaging designs.

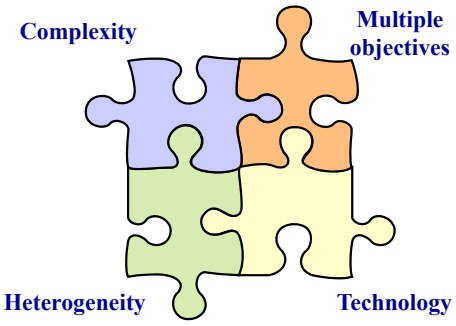

Figure 17: Four dimensions in future design/EDA research directions.

brought significant challenges and great opportunities for emerging AI (especially machine/deep learning) and EDA research and business, and have drawn much attention in the literature [12, 19, 21, 38, 50, 57].

3 FUTURE RESEARCH DIRECTIONS

Heterogeneous integration has brought many design challenges, thus creating substantial research opportunities. Below, we provide some of the most needed research directions in physical design for heterogeneous integration.

- **General directions:** Modern design/EDA challenges share some common grounds and can be explored from four major directions: complexity, multiple objectives, heterogeneity, and technology. Take the above-mentioned 2.5D/3D die-to-die placement problem as example. Modern design requirements have reshaped this problem substantially, which becomes very challenging because we need to handle ultra-large-scale designs with tens/hundreds of millions of objects (i.e., the complexity challenge). Besides wirelength, we must also consider other placement constraints such as blockages, routability, timing, manufacturability, reliability, etc. (i.e., the multi-objectives requirements). Further, the placed objects could be very different in their sizes, and their sizes also depend on their die assignments. In a modern design, we could have a mixed-size die-to-die placement with thousands of large macros and tens/hundreds of millions of small standard cells, and the size changes with different die assignments further complicate the problem (i.e., the heterogeneity challenge). And emerging technologies (advanced packaging, lithography, reliability, machine learning, etc.) often create new challenges. As a result, advanced design methodology and leading commercial tools cannot even handle modern die-to-die placement problems well.
- **Multi-die partitioning and placement with different technologies:** The dies fabricated with different technologies and integrated into an advanced package lead to dimension changes with different die assignments, creating the above-mentioned challenging chicken-and-egg problem. As a result, multi-die partitioning and placement with different technologies and interconnect/via/bump parameters for PPA optimization is desired.
- **Obstacle-avoiding routing with irregular pad/via structures:** Modern advanced packaging often integrates dies with different manufacturing technologies. As a result, the placement of bump and I/O pads could be irregular or mixed with regular and irregular structures. Further, multiple RDL layers with flexible vias can be placed everywhere in the package, and some regions could be blocked for prerouted nets or signal integrity protection to achieve the desired performance. The irregularity, flexibility, obstacles, and multi-layer structures substantially complicate the RDL/substrate routing process. It is desirable first to derive an accurate obstacle-aware routing tile model to capture the routing resource well and then develop an effective routing algorithm to handle the above-mentioned complex issues.
- **Any-obtuse-angle routing:** Traditional RDL routers use only 90- and 135-degree turns for routing. With technological advances, routing in RDLs can be any obtuse angle, incurring larger solution spaces and shorter interconnection [18]. To develop an effective and efficient any-obtuse-angle RDL router, we shall first derive an accurate model to capture routing capacity, even with obstacle considerations, so that a router can optimize routing cost metrics under various constraints.
- **Warpage-aware modeling and optimization:** Warpage is a significant challenge that needs urgent solutions. It is desirable to develop an efficient yet sufficiently accurate warpage model [58, 61] that can be incorporated into a warpage-aware physical design process to optimize the effects.
- **Optical placement and routing:** For optical placement and routing, wirelength and various transmission losses (e.g., crossing, bending, path, and drop losses) need to be minimized considering the thermal effects on optical devices. Further, signal nets with different wavelengths can be clustered into a single WDM waveguide for signal transmission to save routing resources and reduce transmission loss.
- **Cross-die optical-electrical co-design/-analysis/-modeling/-simulation:** The use of optical layer resources can reduce

signal transmission delay and routing congestion in the electrical layer effectively. Nevertheless, optical devices are very sensitive to thermal effects, degrading functionality and performance in the optical layer significantly, and cross-die timing/thermal analysis further complicate the optical-electrical co-optimization. Optical-electrical co-design with the awareness of resource competition, obstacle avoidance, and thermal hot spots is required to achieve the desired solutions with minimized power consumption, timing, thermal impact, and wirelength.

- **Chip-package-board co-design/-modeling/-simulation:** A flexible, robust, and vertically integrated chip-package-board co-design flow considering multiple objectives and various constraints in cross dies and cross-physics domains is urgently needed to facilitate advanced heterogeneous integration. It is desirable to derive efficient yet sufficiently accurate chip-package-board co-analysis, co-modeling, and co-simulation techniques first and incorporate these techniques to enable the co-design flow and optimization.
- **PCB placement and routing with multiple routing layers and complex constraints:** As illustrated in Figure 16, modern PCB designs often comprise 3D irregular, highly heterogenous components, even with concave geometries. Further, there are typically multiple routing layers with highly non-uniform routes of different widths, where the width of a high current-density wire could be tens of the width of a regular signal wire. As a result, wire area (not wire length), component shapes, pad/wire clearance, power loops, heat sinks, and electromagnetic interference with complex design rules all need to be considered for the modern PCB placement and routing.
- **Machine-learning-assisted physical design for heterogeneous integration:** Researchers have adopted machine-learning-based techniques to accelerate and improve the traditional physical design process [38]. For the physical design (partitioning, floorplanning, placement, routing, etc.) problems for heterogeneous integration, we could model netlists into graphs. Machine learning (ML) on graphs, such as the graph neural network (GNN), can transform all graph attributes (including nodes, edges, global context, hierarchy, and timing) for optimization. Design objectives and constraints can be modeled as features for ML, and thus various ML-based techniques can then be applied to assist the physical design with complex objectives and constraints. Exciting research outcomes for machine-learning-assisted physical design in heterogeneous integration are expected to be available soon.

4 CONCLUSIONS

In conclusion, we have many technology options to achieve the holy grail of PPA goals for future computing. We also have multiple complementary options to achieve the continued scaling and heterogeneous integration goals. To achieve the ultimate goal, we would need an ecosystem to bring experts from various domains to work together, where all areas of EDA, circuit design, manufacturing, and fabrication equipment across geopolitical borders are essential to the advancement. Together, we can go very far! Technological challenges will not limit our progress, but human imagination can bring us to the endless frontier! To conclude this paper, we want to share Einstein's most inspiring quote that the human spirit must prevail over technology.

REFERENCES

[1] 2022. *DOMO data never sleeps 10.0.* https://www.domo.com/data-never-sleeps
[2] 2023. *Heterogeneous Integration Roadmap 2023 Edition.* https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2023-edition.html
[3] 2023. *IEEE International Roadmap for Devices and Systems (IRDS).* https://irds.ieee.org
[4] 2023. *International Technology Roadmap for Semiconductors (ITRS).* http://www.itrs2.net/
[5] 2023. *Silicon photonics reshapes the future; TSMC, ASE ready to seize market share.* https://www.trendforce.com/news/2023/12/04/news-silicon-photonics-reshapes-the-future-tsmc-ase-ready-to-seize-market-share/
[6] 2023. *TSMC bets on silicon photonics to enable more powerful ChatGPT.* https://asia.nikkei.com/Business/Tech/Semiconductors/TSMC-bets-on-silicon-photonics-to-enable-more-powerful-ChatGPT
[7] 2024. *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025.* https://www.statista.com/statistics/871513/worldwide-data-created
[8] P. Liao abnd Y. Zhao, D. Guo, Y. Lin, and B. Yu. 2023. Analytical die-to-die 3D placement with bistratal wirelength model and GPU acceleration. In *arXiv:2310.07424.*
[9] Y.-J. Cai, Y. Hsu, and Y.-W. Chang. 2021. Simultaneous pre- and free-assignment routing for multiple redistribution layers with irregular vias. In *Proc. of DAC.* San Francisco, CA.
[10] Y.-H. Chang, H.-T. Wen, and Y.-W. Chang. 2019. Obstacle-aware group-based length-matching routing for pre-assignment area-I/O flip-chip design. In *Proc. of ICCAD.* Westminster, CO.
[11] Y.-W. Chang. 2020. EDA for more-Moore and more-than-Moore designs: challenges and opportunities. In *Proc. of ICCAD.* Virtual Event.
[12] T. Chen, S. Xiong, H. He, and B. Yu. 2023. TRouter: thermal-driven PCB routing via nonlocal crisscross attention networks. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* 42, 10 (October 2023), 3388–3401.
[13] Y.-J. Chen, Y.-S. Chen, W.-C. Tseng, C.-Y. Chiang, Y.-H. Lo, and Y.-W. Chang. 2023. Late breaking results: Analytical placement for 3D ICs with multiple manufacturing technologies. In *Proc. of DAC.* San Francisco, CA.
[14] Y.-T. Chen and Y.-W. Chang. 2022. Obstacle-avoiding multiple redistribution layer routing with irregular structures. In *Proc. of ICCAD.* San Diego, CA.
[15] C.-H. Chiang, F.-Y. Chuang, and Y.-W. Chang. 2020. Unified redistribution layer routing for 2.5D IC packages. In *Proc. of ASPDAC.* Bejing, China.
[16] Y. Chou, J.-W. Hsu, Y.-W. Chang, and T.-C. Chen. 2021. VLSI structure-aware placement for convolutional neural network accelerator units. In *Proc. of DAC.* San Francisco, CA.
[17] Y.-K. Chuang, K.-J. Chen, K.-L. Lin, S.-Y. Fang, B. Li, and U. Schlichtmann. 2018. PlanarONoC: concurrent placement and routing considering crossing minimization for optical networks-on-chip. In *Proc. of DAC.* San Francisco, CA.
[18] M.-H. Chung, J.-W. Chuang, and Y.-W. Chang. 2023. Any-angle routing for redistribution layers in 2.5D IC packages. In *Proc. of DAC.* San Francisco, CA.
[19] S. Das, J. R. Doppa, D. Kim, P. P. Pande, and K. Chakrabarty. 2015. Optimizing 3D NoC design for energy efficiency: A machine learning approach. In *Proc. of ICCAD.* Austin, TX.
[20] D. Ding, B. Yu, and D. Z. Pan. 2012. GLOW: A global router for low-power thermal-reliable interconnect synthesis using photonic wavelength multiplexing. In *Proc. of ASPDAC.* Sydney, Australia.
[21] Y. Duan, X. Liu, Z. Yu, H. Wu, L. Shao, and X. Zhu. 2023. RLPlanner: Reinforcement learning based floorplanning for chiplets with fast thermal analysis. In *arXiv:2312.16895.*
[22] J.-W. Fang and Y.-W. Chang. 2008. Area-I/O flip-chip routing for chip-package co-design. In *Proc. of ICCAD.* San Jose, CA, 518—-522.
[23] J.-W. Fang and Y.-W. Chang. 2010. Area-I/O flip-chip routing for chip-package co-design considering signal skews. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* 29, 5 (May 2010), 711—-721.
[24] J.-W. Fang, K.-H. Ho, and Y.-W. Chang. 2008. Placement and routing for chip-package-board co-design considering differential pairs. In *Proc. of ICCAD.* San Jose, CA.
[25] J.-W. Fang, C.-H. Hsu, and Y.-W. Chang. 2007. An integer linear programming based routing algorithm for flip-chip design. In *Proc. of DAC.* San Diego, CA, 606—-611.
[26] J.-W. Fang, C.-H. Hsu, and Y.-W. Chang. 2009. An integer linear programming based routing algorithm for flip-chip designs. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* 28, 1 (January 2009), 98—-110.

[27] J.-W. Fang, I.-J. Lin, Y.-W. Chang, and J.-H. Wang. 2007. A network-flow based RDL routing algorithm for flip-chip design. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* 26, 8 (August 2007), 1417−−1429.
[28] J.-W. Fang, I-J. Linm, P.-H. Yuh, Y.-W. Chang, and J.-H. Wang. 2005. A routing algorithm for flip-chip design. In *Proc. of ICCAD*. San Jose, CA, 753—-758.
[29] J.-W. Fang, M.D.F. Wong, and Y.-W. Chang. 2009. Flip-chip routing with unified area-I/O pad assignments for package-board co-design. In *Proc. of DAC*. San Francisco, CA, 336—-339.
[30] Y.-K. Ho and Y.-W. Chang. 2013. Multiple chip planning for chip-interposer codesign. In *Proc. of DAC*. Austin, TX.
[31] Y.-K. Ho, H.-C. Lee, and Y.-W. Chang. 2011. Escape routing for staggered-pin-array PCBs. In *Proc. of ICCAD*. San Jose, CA.
[32] Y.-K. Ho, X.-W. Shih, Y.-W. Chang, and C.-K. Cheng. 2013. Layer minimization in escape routing for staggered-pin-array PCBs. In *Proc. of ASPDAC*. Yokohama, Japan.
[33] C. Hou. 2020. TSM ecosystem for innovation. In *TSMC Symposium*.
[34] M.-K. Hsu, V. Balabanov, and Y.-W. Chang. 2013. TSV-aware analytical placement for 3-D IC designs based on a novel weighted-average wirelength model. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* 32, 4 (April 2013), 497—-509.
[35] Y. Hsu, M.-H. Chung, Y.-W. Chang, and C.-H. Lin. 2022. Transitive closure graph-based warpage-aware floorplanning for package designs. In *Proc. of ICCAD*. San Diego, CA.
[36] K.-S. Hu, H.-Y. Chi, I.-J. Lin, Y.-H. Wu, W.-H. Chen, and Y.-T. Hsieh. 2023. 2023 ICCAD CAD Contest Problem B: 3D placement with macros. In *Proc. of ICCAD*. San Francisco, CA.
[37] K.-S. Hu, I.-J. Lin, Y.-H. Huang, H.-Y. Chi, Y.-H. Wu, and C.-F. C. Shen. 2022. ICCAD CAD Contest Problem B: 3D placement with D2D vertical connections. In *Proc. of ICCAD*. San Diego, CA.
[38] G. Huang, J. Hu, Y. He, J. Liu, M. Ma, Z. Shen, J. Wu, Y. Xu, H. Zhang, K. Zhong, X. Ning, Y. Ma, H. Yang, B. Yu, H. Yang, and Y. Wang. 2021. Machine learning for electronic design automation: a survey. *ACM Trans. Design Automation of Electronic Systems* 26, 5 (May 2021).
[39] P. K. Huang, C. Y. Lu, W. H. Weiand Christine Chiu, K. C. Ting, Clark Hu, C.H. Tsai, S. Y. Hou, W. C. Chiou, C. T. Wang, and Douglas Yu. 2021. Wafer level system integration of the fifth generation CoWoS®-S with high performance Si interposer at 2500 mm^2. In *Proc. of ECTC*. Virtual Event.
[40] F. Jiao, S. Dong, B. Yu, B. Li, and U. Schlichtmann. 2018. Thermal-aware placement and routing for 3D optical networks-on-chips. In *Proc. of ISCAS*. Florence, Italy.
[41] D. H. Kim, K. Athikulwongse, and S. K. Lim. 2009. A study of through-silicon-via impact on the 3D stacked IC layout. *IEEE Trans. Very Large Scale Integration Systems* 21, 5 (August 2009), 862–874.
[42] C.-C. Lee and Y.-W. Chang. 2023. Floorplanning for embedded multi-die interconnect bridge packages. In *Proc. of ICCAD*. San Francisco, CA.
[43] H.-C. Lee and Y.-W. Chang. 2012. A chip-package-board co-design methodology. In *Proc. of DAC*. San Francisco, CA.
[44] H.-C. Lee, Y.-W. Chang, and P.-W. Lee. 2010. Recent research development in flip-chip routing. In *Proc. of ICCAD*. San Jose, CA, 404–410.
[45] P.-W. Lee, C.-W. Lin, and Y.-W. Chang. 2009. An efficient pre-assignment routing algorithm for flip-chip designs. In *Proc. of ICCAD*. San Jose, CA.
[46] B.-Q. Lin, T.-C. Lin, and Y.-W. Chang. 2016. Redistribution layer routing for integrated fan-out wafer-level chip-scale package. In *Proc. of ICCAD*. Austin, TX.
[47] D. Liu, Z. Zhao, Z. Wang, Z. Ying, R. T. Chen, and D. Z. Pan. 2018. OPERON: optical-electrical power-efficient route synthesis for on-chip signals. In *Proc. of DAC*. San Francisco, CA.
[48] Q. Liu, D. Lin, C. Chen, H. He, J. Chen, and Y.-W. Chang. 2023. A matching based escape routing algorithm with variable design rules and constraints. In *Proc. of DAC*. San Francisco, CA.
[49] Q. Liu, Q. Tang, J. Chen, C. Chen, Z. Zhu, H. He, J. Chen, and Y.-W. Chang. 2023. Disjoint-path and golden-pin based irregular PCB routing with complex constraints. In *Proc. of DAC*. San Francisco, CA.
[50] Y.-C. Lu, S. S. K. Pentapati, L. Zhu, K. Samadi, and S.-K. Lim. 2020. TP-GNN: A graph neural network framework for tier partitioning in monolithic 3D ICs. In *Proc. of DAC*. Virtual Event.
[51] Y.-S. Lu, K.-C. Chen, Y.-L. Hsu, and Y.-W. Chang. 2022. Thermal-aware optical-electrical routing codesign for on-chip signal communications. In *Proc. of DAC*. San Francisco, CA.
[52] Y.-S. Lu, S.-J. Yu, and Y.-W. Chang. 2020. A provably good wavelength-division-multiplexing-aware clustering algorithm for on-chip optical routing. In *Proc. of DAC*. San Francisco, CA.
[53] Y.-S. Lu, S.-J. Yu, and Y.-W. Chang. 2022. On-chip optical routing with provably good algorithms for path clustering and assignment. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* 41, 11 (November 2022).
[54] G. Moore. 1965. Cramming more components onto integrated circuits. *Electronics* 38, 8 (April 1965), 114–117.
[55] S.-R. Nie, Y.-T. Chen, and Y.-W. Chang. 2022. Y-architecture-based flip-chip routing with dynamic programming-based bend minimization. In *Proc. of DAC*. San Francisco, CA.
[56] J. F. Park. 2010. Board driven I/O planning and optimization. In *Proc. of ICCAD*. San Jose, CA, 395–397.
[57] Sung Joo Park, Bumhee Bae, Joungho Kim, and Madhavan Swaminathan. 2017. Application of machine learning for optimization of 3-D integrated circuits and systems. *IEEE Trans. Very Large Scale Integration Systems* 25, 6 (June 2017), 1856–1865.
[58] E Suhir. 1989. Interfacial stresses in bimetal thermostats. *J. Appl. Mech.* (1989).
[59] C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, L. Alloatti, Mi. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin, B. R. Moss, R. Kumar, F. Pavanello, A. H. Atabaki, H. M. Cook, A. J. Ou, J. C. Leu, Y.-H. Chen, K. Asanović, R. J. Ram, M. A. Popović, and V. M. Stojanović. 2015. Single-chip microprocessor that communicates directly using light. *Nature* 528 (December 2015), 534—-538.
[60] A. Truppel, T.-M. Tseng, and U. Schlichtman. 2020. PSION2:optimizing physical layout of wavelength-routed ONoCs for laser power reduction. In *Proc. of ICCAD*. Virtual Event.
[61] M.-Y. Tsai and Y.-W. Wang. 2019. A Theoretical Solution for Thermal Warpage of Flip-Chip Packages. *IEEE Trans. Component Package Manufacturing Technology* 10, 1 (2019), 72–78.
[62] H.-T. Wen, Y.-J. Cai, Y. Hsu, and Y.-W. Chang. 2020. Via-based redistribution layer routing for InFO packages with irregular pad structures. In *Proc. of DAC*.

A BIOGRAPHY

Yao-Wen Chang received the B.S. degree from National Taiwan University (NTU) in 1988, and M.S. and Ph.D. degrees from the University of Texas at Austin in 1993 and 1996, respectively, all in computer science. He is a Fellow of the ACM and the IEEE and was the President of the IEEE Council on Electronic Design Automation (CEDA) 2020/2021. He is currently the Dean of the College of Electrical Engineering and Computer Science at NTU. His current research interests lie in EDA.

He has co-authored one textbook on EDA (Morgan Kaufmann, 2009) and one book on routing (Springer, 2007), 16 U.S. patents, and 360+ ACM/IEEE conference/journal papers in these areas, including 91 papers in DAC (#1 worldwide), 76 in ICCAD (#2), and 82 in TCAD (#4) with highly cited works on floorplanning, placement, routing, manufacturability, and FPGA. He published the world's most DAC+ICCAD+TCAD papers. His NTUplace3 placer was the core engine of the Digital Custom Placer of SpringSoft, acquired by Synopsys in 2012 for USD 400M+, and his NTUplace4 received three champions at top EDA contests and is the core engine of the leading placer, MaxPlace, by Maxeda which he co-founded in 2015, acquired by Synopsys in 2023.

Dr. Chang received four awards at the 50th DAC in 2013 for the DAC Prolific Author, the Longest Publication Streak (now 25 years from 1999 to now, #1 worldwide), etc. Dr. Chang has received 23 top-3 ACM/IEEE EDA contest awards (#1 worldwide) with seven champions. He has received eleven Best Paper Awards (including one at DAC'17) and the 2020 ASP-DAC Prolific Author Award. He has received two NTU distinguished teaching awards (the highest honor for the top 1% of teachers for ten years), nine NTU excellent teaching awards, and many research awards.

Dr. Chang has been an associate editor of IEEE TCAD, IEEE TVLSI, IEEE Design & Test of Computers, etc. He has served as the general/program chair of ICCAD and ISPD and the program chair of ASP-DAC and FPT, and on the IEEE CEDA, DAC, and ICCAD Executive Committees, and the ASP-DAC Steering Committee. He is a recipient of the 2015 IEEE CEDA Outstanding Service Award and the 2012 ACM Recognition of Service Award. He has served as the chair of the EDA Consortium of Taiwan's Ministry of Education. He was the founding chair of Taiwan's CAD Contest 1999–2001, becoming the most popular CAD Contest at ICCAD since 2012. He has been a technical consultant for MediaTek, Realtek, and Faraday Technology and an independent board director of Genesys Logic (and also in process for MediaTek).