Parasitic Extraction for Heterogeneous Face-to-Face Bonded 3-D ICs

Yarui Peng, *Student Member, IEEE*, Taigon Song, *Member, IEEE*, Dusan Petranovic, *Member, IEEE*, and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*— Face-to-face (F2F)-bonded 3-D ICs provide higher vertical interconnection densities and cost-effective solutions compared to face-to-back-bonded 3-D ICs. With a bumpless direct-copper-bonding process, the die-to-die distance is significantly reduced to enable a finer F2F via pitch. Unfortunately, this increases interdie parasitic components that require careful extraction. Heterogeneous 3-D ICs are built using dies from different design houses and foundries, potentially using different technology nodes. They again require accurate parasitic extraction across multiple dies and thus call for new computer-aided design methodologies with intellectual property protection. We, for the first time, provide a comprehensive study of three full-chip parasitic extraction methods for homogeneous and heterogeneous F2F 3-D ICs. The traditional die-by-die extraction ignores any interdie coupling and underestimates the total coupling capacitance by 35%. The holistic extraction that takes all dies into account provides the most accurate results at the cost of high layout-versus-schematic (LVS) complexity. The in-context extraction, taking only the interface layers from the neighbor dies, offers tradeoffs between accuracy and complexity. Our study shows that with only two interface layers, in-context extraction offers highly accurate and efficient extraction results with 0.9% error for the total ground capacitance and 0.8% for the total coupling capacitance.**

***Index Terms*— 3-D IC, capacitance extraction, die-by-die, face-to-face (F2F), heterogeneous integration, holistic, in-context.**

I. Introduction

TRADITIONAL technology scaling in sub-20-nm nodes is expensive. To lower cost, reduce power consumption, and increase signal bandwidth on a smaller footprint, 3-D ICs are promising solutions to extend Moore's law. A common 3-D IC technology uses face-to-back (F2B) bonding, which builds through-silicon vias (TSVs) in the silicon substrate as vertical interconnects. With this technology, however, increasing 3-D via density is difficult because TSVs penetrate a thick silicon substrate, and fabricating TSVs with high aspect ratio is prohibitively expensive and complex. Unlike F2B bonding, in which the vertical interconnection density is limited by the TSV size, face-to-face (F2F) bonding technology connects top metal layers from both dies with F2F vias [1]. F2F designs achieve much higher 3-D connection density with F2F vias in a few microns [2].

F2F bonding with copper microbumps can achieve a die-to-die (D2D) distance of 8.4 μm [3]. This distance is comparable to the thickness of a regular redistribution layer [4], which results in large coupling capacitance. Advanced In–Au microbumps can reach a size of 1.6 μm [5], and the gap between tiers can be reduced to 1.5 μm [6], which increases interdie parasitics significantly. Also, larger bonding pressure is required for better connection yields and lower resistance, which results in an even smaller D2D distance [7] and stronger interdie coupling. Moreover, with a direct copper-to-copper bonding process [8], the thickness of the bonding interface layer and copper pads can be reduced to less than 1 μm [9]. This technology is commercialized by various foundries and packaging house [10]. Such a close distance is similar to the thickness of interlayer dielectrics of the top metal layers, which makes parasitic extraction inaccurate without considering the electrical fields (E-fields) from the neighboring die.

Since the top metal layers are often used for some critical nets, such as clocks, power, and other global signals, any inaccurate parasitics translate into inaccurate timing, power, and noise analysis. Also, because F2F bonding provides a close coupling distance between multiple dies, it is extremely suitable for signal [11] or power transmission [12] with contactless transceivers in 3-D ICs with high efficiency. For these designs, any parasitics on the 3-D vias significantly varies the resonance frequency and characteristics of the transmission channel [13]. Therefore, accurate interdie coupling extraction is critical to designs, such as the transceiver implemented with F2F coupling capacitance [14] or inductance coupling [15]. As shown in Fig. 1, interdie coupling capacitance becomes increasingly significant with a closer D2D distance [16]. In future generation with monolithic 3-D ICs, the tier-to-tier distance is reduced smaller than 100 nm, and even devices from the bottom die can be affected by E-field from metal wires of the neighboring die [17]. Therefore, with future technologies, E-fields from multiple dies will heavily interact with each other [18], and interdie coupling parasitics will increase. All these new technologies require accurate extraction of interdie coupling elements aware of layers from all dies.

There are many existing works on parasitic extraction methodologies for interconnects in traditional 2-D ICs. These standard extraction techniques can be divided into

Manuscript received June 17, 2016; revised January 10, 2017; accepted January 29, 2017. Date of publication April 4, 2017; date of current version May 31, 2017. This work was supported by Mentor Graphics Corporation. Recommended for publication by Associate Editor F. H. Al-Hawari upon evaluation of reviewers' comments.

Y. Peng is with the Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR 72704 USA (e-mail: yrpeng@uark.edu).

T. Song and S. K. Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: limsk@ece.gatech.edu).

D. Petranovic is with the Mentor Graphics Corporation, Fremont, CA 94538 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCPMT.2017.2677963

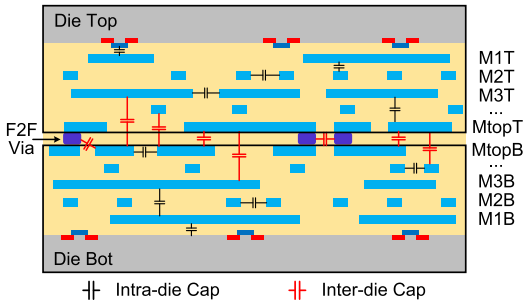


Fig. 1. Cross-sectional view of an F2F-bonded 3-D IC structure with interconnect parasitics.

deterministic methods, such as finite-element method [19] and boundary-element method [20], as well as statistic methods, such as floating random walk [21]. Though techniques based on field solving or random walk can be accelerated further with a hierarchical approach with lookup tables and macromodels [22], they require significant runtime on the full-chip level especially in advanced technology nodes with very fine pitch structures. Therefore, for efficiency reasons, pattern-matching-based extraction is still widely used for large-scale designs, while critical nets are often extracted using field solving.

Some recent works also demonstrate significant parasitic coupling in F2B-bonded packages in both signal [23] and power distribution networks [24]. However, there are few existing works focusing on parasitic impacts on F2F-bonded 3-D IC designs, and all previous works assume a full knowledge of interconnection on both sides. The direct Cu–Cu bonding enables two dies to be tightly connected, thus the close D2D distance requires considering both dies simultaneously for signal and power integrity issues [25]. To enable next generation of heterogeneous F2F integration, it is also critical to define an interconnection interface to ensure that designs from multiple sources can be integrated without violating signal integrity constraints. Though it is always possible to minimize parasitic impacts by inserting large IO drivers with ESD protection circuits, only interdie signal pins can be protected. For intradie signal routing close to the die surface, parasitic components still have large impacts on the delay and noise.

In this paper, we provide a comprehensive study on various extraction methodologies, runtime–accuracy tradeoffs, and full-chip parasitic impacts for heterogeneous F2F integration, and define a practical interconnection interface to enable interdie coupling consideration with intellectual property protection among collaborative companies. We start by introducing various methodologies for F2F interdie coupling extraction and comparing their pros and cons using GDS-level full-chip benchmarks. Then, we analyze full-chip impacts from F2F interdie coupling elements with our path-finding study into future heterogeneous 3-D ICs.

II. Proposed F2F Extraction Methodologies

A. Die-by-Die Extraction

In order to handle various F2F technologies and configurations, we propose and evaluate three extraction methods in this paper. First, die-by-die extraction extracts the bottom and top dies individually similar to current 2-D IC extraction, as shown in Fig. 2(a). It ignores coupling capacitance between dies and can be implemented easily using traditional 2-D extraction engines such as Calibre xACT. Presuming extractions for each die, the only requirement is a method that can stitch together these individual die netlists with parasitics. The die-by-die extraction is accurate as long as the D2D distance is large and the E-fields from both dies do not couple to each other. On the other hand, the die-by-die extraction is also considered as "LVS-friendly," since layout versus scheme can be done without knowing any geometries from the neighbor die. Since any sign-off parasitic extraction needs to be performed after LVS check and all layer patterns are properly netlisted, die-by-die extraction completely decouples designs of each dies, allowing for a faster time-to-market and easier industrial collaboration, which are critical to enable parasitic extraction of heterogeneous 3-D ICs.

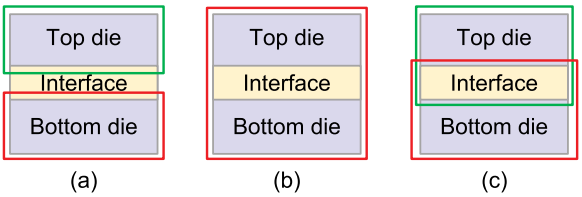


Fig. 2. Comparison of F2F extraction methods. (a) Die-by-die, (b) holistic, and (c) in-context extraction.

B. Holistic Extraction

The second method is the holistic extraction, where all layers from both dies are taken into account during technology calibration and parasitic extraction. As shown in Fig. 2(b), this extraction requires a full knowledge of both dies, from device layer all the way to the top metal layer. By performing a holistic LVS of all dies, the geometry connectivity can be fully netlisted. It can achieve maximum accuracy by capturing all E-fields from both dies; therefore, this paper uses holistic extraction as a reference in our F2F extraction, and compares other extraction methods to it. However, holistic extraction is extremely challenging computationally both during precalibration and runtime. First, considering all layers requires many more library structures to be built, and there are more combinations of possible structures. This can significantly increase calibration time. Ideally, it is upon the system designer to choose vendors for each components.

For heterogeneous integration, it is difficult to consider all possible combinations of different technologies from multiple foundries beforehand, especially with various metal stack configurations. Moreover, it requires coding holistic LVS and extraction rule decks that can properly recognize all devices, connect two dies, and perform extraction for all layers. As dies may be from different technologies, foundries need to share all information of their technologies, including critical layers, such as devices and local interconnects, which are needed for holistic rule decks. For homogeneous 3-D ICs, where both dies are from the same foundry, it is not impossible, but it takes time to regenerate rules for both dies and carefully resolve

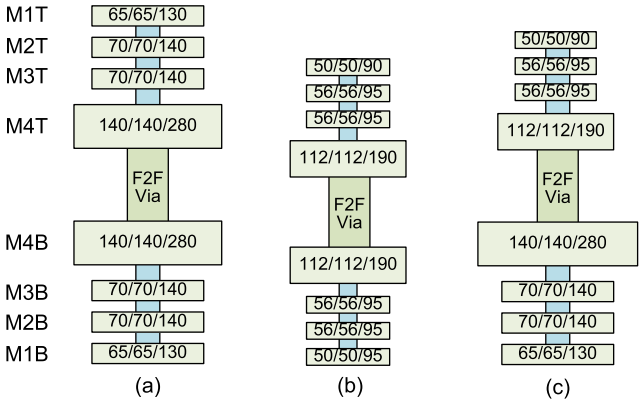


Fig. 3. Technology configurations with 1-μm F2F via thickness. Numbers show the width/spacing/thickness of each metal layer in micrometers. (a) and (b) Homogeneous technology in 45- and 28-nm nodes, respectively. (c) Heterogeneous technology, where bottom die uses 45 nm and top die 28 nm.

any conflicts. However, if multiple foundry technologies are used, it requires foundries to share their critical trade secrets to the public and their competitors. Either one master foundry is responsible to incorporate layers from the neighboring die, and maintain the holistic rule deck, or a third party, likely a packaging house for F2F bonding, is required to combine rule decks, which are generally encrypted by owner foundries. Not only is it time-consuming to resolve conflicts and combine LVS and extraction rules, but also it threatens intellectual property protection of design houses. Designers for both dies need to reveal all of their layouts and netlists, which open doors to backengineering. Therefore, though holistic extraction may be possible with homogeneous 3-D ICs, it may not be efficient and realistic for commercial use, especially with heterogeneous 3-D ICs.

C. In-Context Extraction

To improve parasitic extraction accuracy without imposing the need for detailed information from the neighboring die, in-context extraction is proposed to take advantage of die-by-die extraction without losing much accuracy compared to holistic extraction. Previous study has shown that most of the coupling E-field is formed within limited depth into the other die [26], and we define this as the "coupling depth." Therefore, to efficiently perform extraction without sacrificing accuracy, in-context extraction only takes a few layers, called "interface layers," from the neighboring die into account during both technology calibration and parasitic extraction. As shown in Fig. 2(c), similar to die-by-die extraction, top and bottom dies are extracted separately, but both are extracted with the knowledge of interface layers. Dies with interface layers from the neighbor are called "in-context dies." Though extra layers are still needed, in-context extraction significantly reduces the number of precalibrated structures required since structures with small dimensions, such as devices and local metals, are not included.

Also, for advanced technologies nodes, thousands of design rules strictly applying to those small structures can also be avoided, and it is much easier to handle top metals with

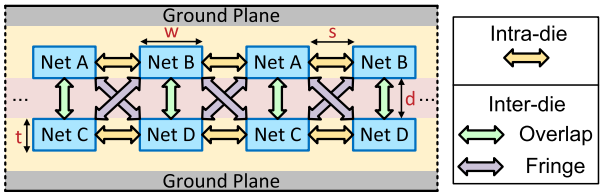


Fig. 4. Raphael structure for capacitance extraction. Both the top and bottom dies contain repeated layout patterns. D denotes the D2D distance, while w, s, and t denote wire width, spacing, and thickness, respectively.

larger dimensions. On the other hand, in-context extraction still remains LVS-friendly because only a few top metal layers are needed to code an LVS deck for in-context dies. New rule decks can be calibrated incrementally by reusing existing rule decks. Note that by revealing the noncritical properties, such metal dimensions and dielectric properties are not critical issues, and calibration of heterogeneous 3-D IC technologies need only extend the existing 2-D rule files. Therefore, this approach reduces the complexity of handling all layers simultaneously and can be carried out independent of device fabrication process. Previous work [27] implemented the in-context extraction with homogeneous technology, and in this paper, we are focused on heterogeneous 3-D IC integration.

III. Field Sharing Analysis

To find out how two E-fields from both dies interact with each other, we build a test structure shown in Fig. 4. The ground planes are located 3 μm away from wires, and wire width (w) and thickness (t) are fixed as 0.8 and 1.2 μm, which are the same as top metal layer dimensions in a 45-nm technology. We duplicate patterns of Nets A and B on the top die, and Nets C and D on the bottom die. In this structure, the coupling capacitance can be divided into three groups: intradie coupling capacitance, interdie overlapping capacitance, and interdie fringe capacitance. The repeated patterns ensure that any capacitors of the same kind have the same value. Therefore, intradie coupling, interdie overlapping, and interdie fringe capacitance can be represented by Cap AB (or Cap CD), Cap AC, and Cap AD (or Cap BC), respectively. Note that because of the symmetric structure, total intradie capacitance can be represented by 2× Cap AB, while total interdie capacitance can be represented by 1× Cap AC plus 2× Cap AD. Capacitance is extracted assuming an infinite wire length with a 2-D extraction with a unit of fF/μm.

First, we vary the D2D distance (d) from 0.5 to 8 μm and find out its impact on the coupling capacitance. Field solver extraction results are shown in Fig. 5, where capacitance values are measured by the average of ten wires on each die. The wire spacing (s) is fixed as 0.9 μm, which is the minimum spacing of M4–M6 in the target technology. With a closer D2D distance, interdie coupling capacitance (represented by Cap AC) increases significantly, while interdie fringe capacitance (represented by Cap AD) increases slightly. Also, because of the E-field sharing from the neighbor die, with a closer D2D distance, intradie coupling capacitance decreases. It only changes slightly when dies are far from

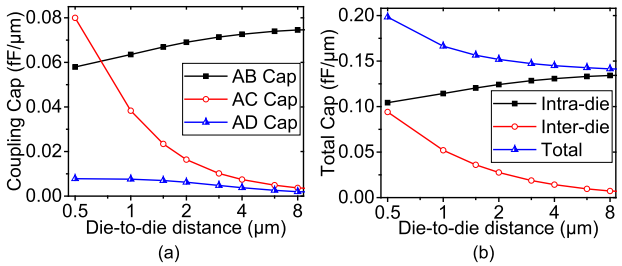


Fig. 5. D2D distance (=d in Fig. 4) impact. (a) Single capacitor extraction, A–D are nets in Fig. 4. (b) Total capacitance extraction.

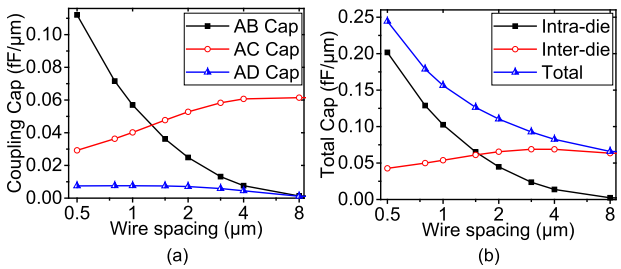


Fig. 6. Wire-to-wire spacing (=s in Fig. 4) impact. (a) Single capacitor extraction, A–D are nets in Fig. 4. (b) Total capacitance extraction.

each other, but it decreases significantly when D2D distance is less than 5 μm since the E-field sharing from the other die is much stronger. With a D2D distance smaller than 1 μm, interdie coupling capacitance becomes comparable to intradie coupling capacitance even with minimum wire spacing. Therefore, interdie coupling can no longer be ignored with a close D2D distance. Shown in [16], if the D2D distance is similar to the top metal dimensions, interdie coupling becomes comparable to the intradie coupling of the top wires. Note that the total capacitance always increases with a closer D2D distance, and the portion of interdie coupling keeps increasing as well. Therefore, die-by-die extraction, which is unaware of the neighboring die and ignores interdie E-field sharing, cannot extract the interdie coupling capacitance accurately when D2D distance is smaller than 5 μm in this technology.

Then, we vary the wire spacing while keeping the D2D distance to be 1 μm. Raphael extraction results are shown in Fig. 6. With a large wire spacing, both intradie coupling and total coupling capacitance decrease. However, the interdie coupling capacitance percentage increases with a larger wire pitch. Also, E-field sharing from neighboring wires within the same die is weaker, thus stronger coupling is formed between overlapped wires across dies, which is the major portion in the interdie capacitance. As a result, total interdie capacitance increases with a wire spacing up to 3 μm. However, if wire spacing increases further, intradie E-field sharing is very weak, and the increase of overlap capacitance (Cap AC) saturates. Therefore, total interdie capacitance slightly decreases with smaller fringe capacitance (Cap AD). Overall, interdie capacitance becomes comparable to intradie capacitance with wire spacing larger than 1 μm. From these results, interdie coupling cannot be ignored in designs with sparsely routed top metal layers, while intradie E-field sharing cannot be ignored with densely routed wires during parasitic extraction.

IV. Die-by-Die and Holistic Extraction Flows

In this section, we demonstrate our computer-aided design (CAD) flows of all three extraction methods discussed in Section II.

A. Die-by-Die Extraction

The CAD flow of homogeneous die-by-die extraction is shown in Fig. 7. If a heterogeneous technology is used, two sets of extraction rules can be calibrated independently.

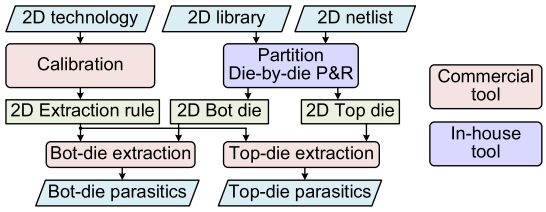


Fig. 7. CAD flowchart of our die-by-die extraction.

Since currently no commercial design tool is able to handle timing and power optimization of 3-D designs, commercial 3-D ICs have their dies designed separately only with predefined 3-D vias as interface to the neighboring dies, which reduces CAD complexity and accelerates the design process. LVS can be done similarly as a 2-D design to match the layouts or extract a netlist for parasitic extraction. After parasitics are obtained from both top and bottom dies, designers need to include a top-level netlist, which describes 3-D connections and I/O interfaces between dies, as well as a top-level parasitic file, which includes capacitance of 3-D vias. The full-chip analysis can be easily performed by merging of all the parasitics with a connected netlist for the whole system. Ignoring interdie capacitance, this flow is widely adopted for both F2F and F2B designs, and it is the fastest approach and the only feasible way nowadays.

B. Holistic Extraction

Compared to the die-by-die approach, holistic extraction requires considering all layers simultaneously. The metal layers located in the bottom die are denoted with a postfix of "B," while the metal layers in the top die are with "T." With F2F bonding, top metal layers from both dies are heavily coupled. Especially when only a few metal layers are used, the interdie coupling capacitance consumes a large portion of total coupling capacitance. However, there is currently no commercial full-chip extraction engine, which is able to handle two device layers simultaneously. Therefore, we implement the holistic extraction flow shown in Fig. 8 by considering the top die device layer as a conducting layer. This will introduce some errors mostly on the M1T layer in holistic extraction. However, it still gives reasonable results since parasitics inside standard cells should be extracted separately and included in the post-layout cell netlist, and its timing and power impacts should be considered by cell characterization. Since most M1

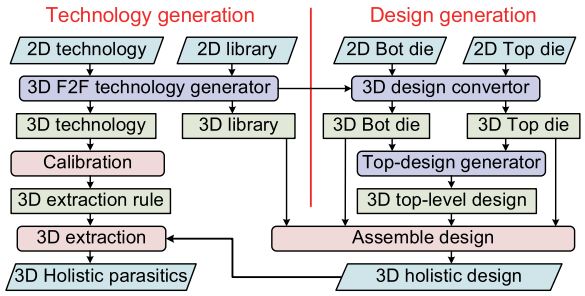


Fig. 8. CAD flowchart of our holistic extraction.

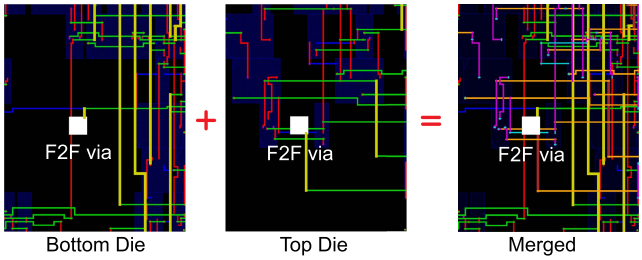


Fig. 9. 3-D holistic design generation.

areas are used for intracell connections, and we perform full-chip top-level extraction with very few M1 wires for intercell connections, only a small portion of coupling capacitance is extracted on M1 layers of both dies.

First, to generate a holistic technology, a technology generator reads the 2-D technology and library and duplicates metal layers and cells in the F2F fashion. The generated 3-D technology and library contain all metal layers as well as the bottom die substrate and device layer. With all layers calibrated, holistic extraction is able to fully cover all E-field interactions inside the F2F bonding layer as well as any E-field sharing impacts from metal layers. Since current physical design flow implements each die in 3-D ICs separately, we implement a CAD flow to generate the 3-D holistic design from die-by-die designs. A 3-D design converter takes in both designs and converts each design according to the output of the 3-D technology generator. Then, by taking the LEFs of both dies, our top-design generator creates a top-level layout, which has the same footprint as the 3-D chip, but only contains dies and F2F via connections. Note that cells from both dies overlap on the floorplan, thus it cannot pass the geometry check. Luckily, all major extraction tools are able to handle this routed layout. Fig. 9 shows our design merging process.

V. In-Context Extraction Flow

A. Technology and Design Generation

Unlike holistic extraction, which handles multiple substrate and device layers simultaneously, in-context extraction does not require creating new extraction engines for multiple dies. For naming convenience, we use "In-C:N" to denote in-context extraction with N interface layers per die. Note that holistic extraction can be considered as a special case of in-context extraction, where all metal layers become interface layers. Also, our flow is able to handle heterogeneous 3-D ICs even

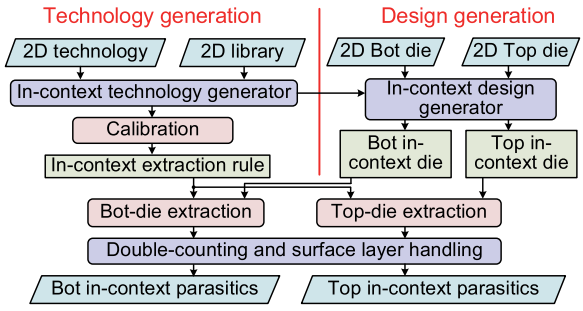


Fig. 10. CAD flowchart of our in-context extraction.

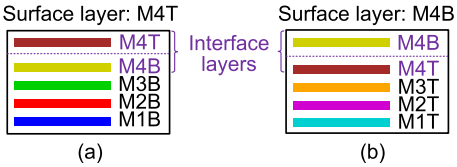


Fig. 11. Sample in-context interconnect technology with four metal layers. (a) Bottom in-context die. (b) Top in-context die.

with mismatched die footprint or unsymmetrical F2F-bonded designs in which the number of metal layers or interface layers from top and bottom dies is not the same.

To enable such extraction, for each in-context die, we must include enough data about connectivity and geometries from its neighboring die. Our in-context flow for homogeneous 3-D ICs is shown in Fig. 10. If a heterogeneous design is used, in-context extraction rules for bottom and top in-context dies require being calibrated separately. For technology generation, we simply extend the basic 2-D technology and library to create in-context technology files so that there are minimum changes to the technology description files. Also, incremental calibration can be applied to reuse existing rule decks and ensure the silicon-validated 2-D extraction rules are unchanged.

An example with four metal layers and one interface layer per die is shown in Fig. 11. We call the outmost metal layer in our in-context technology "surface layer," though no metal layer is physically located at the surface of the chip in real F2F technology. For example, with the metal stack (In-C:1) shown in Fig. 11, M4B and M4T layers are surface layers of top and bottom in-context dies, respectively. Similarly for In-C:2 extraction, M3B and M3T become surface layers of top and bottom dies. The surface layer is special since it has one missing neighbor layer in the in-context technology. Since each in-context technology only includes one substrate and device layer, it can be calibrated similarly as a traditional 2-D technology.

Similar to holistic extraction, our generator takes in design files from both dies and renames the cells and layers. However, only interface layers are included during design layout merging, and other layers as well as cells from the neighboring die are discarded. Fig. 12 shows an in-context design generation process for technology shown in Fig. 11. After in-context designs are generated, they are extracted similarly as the die-by-die flow. Since most of the interdie E-fields are

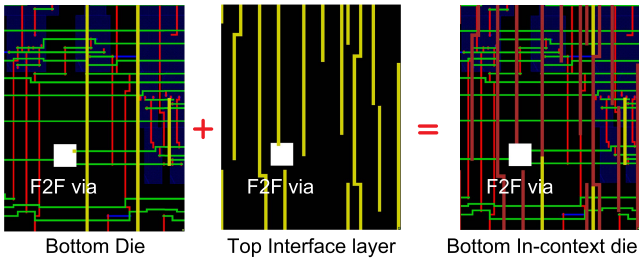


Fig. 12. 3-D in-context design generation.

formed within neighbor layers, in-context extraction provides a close-to-optimum solution with easy implementations. Also, it is much easier to avoid intellectual property issues with heterogeneous designs.

B. Double Counting Correction

Though in-context extraction provides a way to stay compatible with current CAD tools and extract dies separately, the interface layers need to be handled with extra care to avoid any inaccuracy. First is to avoid double-counting the capacitance formed between the interface layers. If we directly add parasitics from both dies together, the capacitance will be significantly overestimated since interface layers are extracted both in the top and bottom in-context designs.

To solve the double-counting problem, we extract capacitance with their geometry information annotated into the SPEF file. Then, we implement a parasitic analyzer, which reads the extended SPEF file and looks up the capacitance layer connection one-by-one. An intuitive way to solve the double-counting is to divide every double-counted capacitance by half. It is effectively calculating the average between top and bottom in-context parasitics. We call this method "In-C halved," and the method by simply merging both in-context parasitics as "In-C original." With an In-C halved extraction, overestimation of interdie coupling can be corrected. However, this is still not fully accurate since the overestimated capacitance is not exactly twice as large as their correct value. Neither bottom die nor top die has the full information of the whole design, and even for the same capacitor, its value is different in two dies because the extraction environment is not the same in both dies.

C. Surface Layer Correction

Another issue, which also affects the in-context extraction accuracy, is the surface layer handling. Shown in Fig. 11, surface layers of both in-context dies are outmost metal layers missing one neighbor layer in the metal stack. However, with in-context designs, E-field sharing impacts are not fully considered since a few metal layers are missing during the technology calibration. Most E-field interactions are between neighboring metal layers, and surface layers are mostly affected by inaccurate extraction. Unlike other metal layers where E-field sharing from both sides is taken care of, the capacitance extracted on the surface layer only considers the E-field sharing from one of its neighboring layers. The In-C halved method is able to correct the double-counting but unable to fix the inaccurate surface layer capacitance.

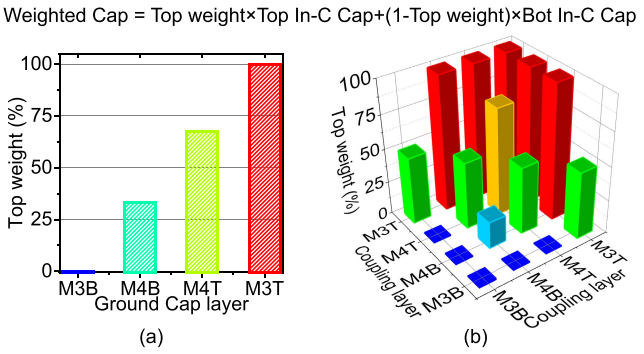


Fig. 13. Correction weight for top in-context die in a two-tier 3-D IC with two interface layers per die. (a) Ground cap scaling factor. (b) Coupling cap scaling factor.

To solve this issue, we propose an "In-C weighted" method. The motivation is simple, as we observe that a surface layer in one in-context die is not the surface layer in the other in-context die. For example, as in Fig. 11, ground capacitance on M4T cannot be extracted accurately with bottom in-context die because layers from M1T to M3T are missing. However, it is accurate in the top die, where M4T is not the surface layer and has both its neighboring layers. Therefore, when stitching together capacitance of both dies, imbalanced weights should be used depending on how close a layer is to the surface.

To implement this, we use a parameter D for each metal layer as the distance to surface. In any in-context technology, the surface layer has a D value of zero, while D increments by one for each metal layer below the surface layer. For example, in Fig. 11, D value of M2B is 3 in the bottom in-context technology, while D value of M3T is 2 in the top in-context die. Generally, with a larger distance to surface, more E-field sharing can be considered for that layer. We define an R ratio for each interface layer as the ratio between its D values in the bottom in-context die and the top in-context die. It is used as a weight to merge capacitance extracted from both dies. To combine calculation of both ground capacitance and coupling capacitance, we define the R ratio of the ground layer as 1:1.

Then, we can calculate the capacitance from interface layers based on a weighted average from both dies. Note that we do not need to handle capacitance, which is not double-counted. As long as the total weight of both dies is equal to 1, there is no overestimation in interdie coupling. Therefore, for a double-counted capacitor connecting two layers, we normalize the product of R ratios of these layers to 1, and use it as the weight between the bottom in-context die and the top in-context die. Fig. 13 shows an example with four metal layers and two interface layers. Our in-context extraction gives larger weights to layers far from the surface so that the inaccuracy from E-field sharing impact is mitigated. As in the example, larger weight is given to ground capacitance in M3T in the top die, but M3B in the bottom die. Also, we use half from bottom die and half from top die for coupling between M4T and M4B.

VI. Full-Chip Extraction Results

In this section, we build a 64-point FFT (FFT64) circuit in a 45-nm technology shown in Fig. 3(a) and apply all three

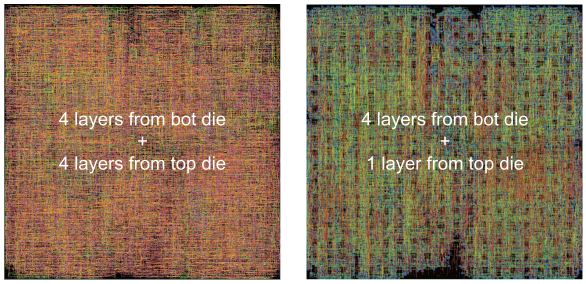


Fig. 14. Layouts of FFT64 benchmark using four metal layers. (a) Holistic and (b) in-context with one metal layer from the other die for the interface.

TABLE I
HOLISTIC EXTRACTION OF F2F COUPLING CAPACITANCE. CAPACITANCE VALUE IS IN fF

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	5.76							
M2B	3.03	381						
M3B	17.1	147	1261					
M4B	0.13	396	231	1826				
M4T	0.03	18.6	9.9	1184	1311			
M3T	0.14	0.69	140	18.6	196	1226		
M2T	0.00	2.58	0.72	46.7	369	148	442	
M1T	0.00	0.01	0.28	0.12	0.28	25.3	4.63	7.54

extraction methods on it for comparison. The F2F via has a size of 1 $\times$ 1 μm ^{2}, and the F2F bonding layer is 1 μm in thickness and filled by SiO_2 with a relative permittivity of 3.9. We implement the flows described in Section IV and generate design layouts in all three styles: die-by-die, holistic, and in-context. Fig. 14 shows FFT64 design shots. This design is routed up to M4 and has a footprint of 380$\times$ 380 μm ^{2} with 38 K gates, which is similar to a digital block in a modern system. The F2F via resistance is assumed as 1 Ω connecting between M4B and M4T.

A. Interdie Versus Intradie Breakdown

First, we analyze how much coupling in an F2F design is contributed by interdie coupling using holistic extraction shown in Table I. In our extraction, both ground capacitance and coupling capacitance are extracted. Table I is symmetric, thus only the lower triangle is shown. It can be divided into three parts: intrabottom-die coupling, intratop-die coupling, and interdie coupling. As results show, intradie coupling is still the most dominant portion in total capacitance, and most interdie coupling is between top metal layers of both dies. The M4B-to-M4T coupling contributes to 83% of all interdie coupling. The interdie coupling contributes to 34% of the total coupling capacitance on M4B and 39% of the total coupling capacitance on M4T layer. We also observe a noticeable contribution from interdie coupling on total coupling capacitance of second-topmost layers (8.4% and 9.1% for M3B and M3T, respectively). For lower metal layers, the contribution from interdie coupling is negligible. Overall, interdie coupling contributes to 23% in total coupling capacitance in the F2F-bonded FFT64 design.

The results validate two of our motivations: 1) interdie coupling is not negligible especially for the top metal layers; therefore, die-by-die extraction is not sufficient for accurate extraction of F2F designs and 2) interdie coupling E-fields are mostly limited between a few metal layers because of E-field shielding from metal wires. In this configuration, the coupling depth is around two metal layers. Therefore, it is safe to ignore a few metal layers in our in-context extraction, which still captures most interdie coupling E-fields. From the results, we conclude that our holistic extraction is highly accurate to capture all E-field interactions inside F2F designs.

TABLE II
DIE-BY-DIE EXTRACTION ERROR ANALYSIS AGAINST HOLISTIC EXTRACTION. CAPACITANCE IS IN fF

	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	26.2	949	1,808	3,703	3,089	1,755	1,013	38.2	12,381
D-D	20.1	856	1,620	1,955	1,413	1,399	747	21.2	8,032
Err	-6.06	-93.4	-187	-1,747	-1,676	-356	-266	-17.0	-4,349
Err %	-23%	-9.8%	-10%	-47%	-54%	-20%	-26%	-45%	-35%

B. Die-by-Die Versus Holistic Extraction

Then, we analyze how much error is introduced with die-by-die extraction. The total extracted ground capacitance is very similar between die-by-die extraction (39476fF) and holistic extraction (39247fF) with only a 0.58% difference. This is because the substrate, which serves as the ground plane, is far from interface layers. Most differences between these two methods come from coupling capacitance. As shown in Table II, die-by-die extraction underestimates total coupling capacitance by 35% compared to holistic extraction. Though with more metal layers in each die, percentage difference between die-by-die and holistic extraction will be smaller, but accurate extraction is still essential for critical nets on the top metal layer. Therefore, we conclude that die-by-die extraction cannot accurately capture all coupling capacitance and E-field interactions inside the F2F designs, especially for technologies with a close D2D distance.

C. In-Context Versus Holistic Extraction

To validate our in-context extraction, we compare extraction results to holistic extraction, which is assumed as our golden model. However, there is no extraction tool available that handles two device layers and substrates simultaneously. Therefore, we build the holistic technology with the bottom die as the device substrate, while we use a ground plane as boundary conditions to replace top die substrate. On the other hand, in-context extraction does not have this limitation with a single device substrate for each in-context die. Therefore, the in-context extraction can be more accurate on layers close to the top die substrate. Note that since only the top die grounded substrate is replaced, only ground capacitance of layers close to the top die substrate is affected. Targeting a coupling depth of two layers, Table III shows extraction results of in-context extraction with two interface layers per die (In-C:2). Since M1 and M2 are not interface layers, interdie coupling capacitance on those layers is zero with in-context extraction. However, the in-context extraction still

TABLE III

IN-CONTEXT EXTRACTION OF F2F COUPLING CAPACITANCE. WE USE TOP TWO METAL LAYERS FOR THE INTERFACE. CAPACITANCE IS IN fF

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	5.76							
M2B	3.02	380						
M3B	17.2	148	1265					
M4B	0.13	399	235	1818				
M4T	0.03	18.9	9.88	1165	1303			
M3T	0.14	0.54	127	17.8	195	1218		
M2T	0	0	0.48	43.6	365	149	438	
M1T	0	0	0.19	0.09	0.25	25.6	4.63	7.27

TABLE IV

IN-CONTEXT EXTRACTION ERROR ANALYSIS AGAINST HOLISTIC EXTRACTION. CAPACITANCE IS IN fF

	Ground capacitance								
	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	1,136	6,588	9,240	3,878	2,664	8,320	6,306	1,117	39,247
In-C	1,137	6,583	9,249	4,159	2,639	8,183	5,986	949	38,886
Err	1.10	-4.20	9.00	281	-24.9	-136	-319	-168	-361
Err%	0.1%	-0.1%	0.1%	7.2%	-0.9%	-1.6%	-5.1%	-15%	-0.9%
	Coupling capacitance								
	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	26.2	949	1,808	3,703	3,089	1,755	1,013	38.2	12,381
In-C	26.3	950	1,803	3,679	3,058	1,734	1,001	38.0	12,287
Err	0.15	0.81	-5.15	-24	-31.0	-21.3	-12.3	-0.22	-93.3
Err%	0.6%	0.1%	-0.3%	-0.7%	-1.0%	-1.2%	-1.2%	-0.6%	-0.8%

remains as highly accurate since the interdie coupling contributions from M1 and M2 are small, and negligible errors are introduced. If higher accuracy is desired, more interface layers can be added into in-context extraction, and LVS complexity is still much lower than holistic extraction since adding a few interconnect layers with large dimensions is still much easier than analyzing multiple device layers or local interconnection layers.

Table IV summarizes the extraction comparison between in-context and holistic extraction. As results show, for all layers, our in-context extraction is highly accurate in both ground capacitance and coupling capacitance. Since our in-context extraction ignores a few interdie coupling elements on M1 and M2, total capacitance extracted with our in-context flow is underestimated slightly. As results show, total ground capacitance is underestimated only by 0.9%, and total coupling capacitance is underestimated only by 0.8%. Note that coupling capacitance errors on M4B and M4T are only 0.7% and 1.0%, respectively. These two interdie coupling elements are largest in absolute value, indicating that almost all interdie coupling capacitors are captured with our in-context extraction. Therefore, we can conclude that our in-context extraction is highly accurate and efficient to capture most E-field interactions inside the F2F designs without adding too much CAD complexity.

D. Impact of Interface-Layer Handling

Previous results are extracted based on the In-C weighted method, which corrects both double-counting and surface layer errors. We compare various interface-layer handling methods discussed in Section V-C. Table V summarizes full-chip extraction results with three handling methods on M3B and M3T. As results show, interface-layer handling significantly affects extraction accuracy. If the coupling capacitance is simply added up from both dies, the In-C original method overestimates coupling capacitance in the interface layer significantly. The total coupling capacitance errors for M3B and M3T are 77% and 112%, respectively. Total coupling capacitance is also overestimated for M4B and M4T as well. Note that even for the same capacitor, its capacitance value is different when extracted with bottom and top in-context dies because its context and the E-shield sharing from neighbor layers differ.

TABLE V

COMPARISON OF INTERFACE-LAYER HANDLING METHODS. UNIT OF TOTAL COUPLING CAPACITANCE OF EACH LAYER IS fF

Layer	Method	M3B	M4B	M4T	M3T	Total	Err	Err%
M3B	Holistic	1261	231	9.9	140	1642	-	-
	original	2220	413	16.4	255	2904	1262	77%
	halved	1110	206	8.2	127	1452	-190	-12%
	weighted	1265	235	9.9	127	1637	-5.27	**-0.3%**
M3T	Holistic	140	18.6	196	1226	1581	-	-
	original	255	32.9	377	2682	3347	1766	112%
	halved	127	16.4	188	1341	1673	92.3	5.8%
	weighted	127	17.8	195	1218	1559	-22.4	**-1.4%**

TABLE VI

IMPACT OF THE INTERFACE-LAYER COUNT ON EXTRACTION ACCURACY. "IN-C:N" DENOTES IN-CONTEXT EXTRACTION WITH N INTERFACE LAYERS PER DIE. CAPACITANCE IS IN fF

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Holi	26.2	949	1808	3703	3089	1755	1013	38.2	12,381
In-C:1	26.1	953	1701	3708	2994	1604	994	37.8	12,018
In-C:2	26.3	950	1803	3679	3058	1734	1001	38.0	12,287
In-C:3	26.2	949	1794	3671	3057	1745	1012	38.2	12,292

By dividing every capacitance in half, extraction errors are significantly reduced to -12% and -5.8% for M3B and M3T, respectively. However, the extraction accuracy is still not high enough because E-field sharing impacts are not handled well for surface layers as discussed in Section V-C. With our proposed method using a weighted average, our in-context extraction is highly accurate compared to holistic extraction. Total coupling capacitance errors for M3B and M3T are reduced to -0.3% and -1.4%, respectively, which is almost negligible for full-chip analyses. Our interface-layer capacitance handling does not affect the number of coupling capacitance, thus the number of aggressors is the same, but it affects the coupling strengths of the aggressors. Overall, we can conclude that our in-context extraction algorithm using weighted average to handle interface layers is highly effective and accurate.

Previous in-context extraction results are based on two interface layers per die. Table VI summarizes results with various numbers of interface layers. Interestingly, even with only one interface layer per die, in-context extraction is quite accurate. Total coupling capacitance has only a 2.9% error compared to holistic extraction, which can actually be regarded as In-C:4 for a technology with four metal layers.

TABLE VII

F2F EXTRACTION FOR 28-nm LDPC BENCHMARK WITH TWO DIES, EACH ROUTED UP TO M6. CAPACITANCE IS IN pF

Method	Holi	D-D	In-C:1	In-C:2	In-C:3
Characterization time (min)	21	10	12	14	17
Total Coupling Cap	226	206	222	224	225
Err	-	-19.9	-3.68	-1.61	-1.37
Err%	-	-8.8%	-1.8%	-0.7%	-0.6%

With more interface layers, accuracy increases. Total coupling capacitance errors of In-C:2 and In-C:3 are -0.76% and -0.68%, respectively, compared to holistic extraction. Note that since in-context extraction still ignores some interdie coupling, it generally extracts less coupling capacitance than holistic extraction. From these results, we conclude that most interdie coupling capacitance can be extracted even with one interface layer from each die. If higher accuracy is required, more interface layers can be included in the in-context extraction to provide detailed consideration of the neighboring die and metal layers.

E. Extraction With 28-nm LDPC

To extend our study into advanced technology nodes, we also design an low-density parity-check (LDPC) benchmark in a 28-nm technology with routing up to M6. Table VII summarizes the extraction results of all three methods. Technology characterization is performed on a Linux server using 12 cores in parallel. As expected, the interdie coupling portion reduces in 28-nm technology. This is because M4–M6 wire dimensions (0.112 μm in width and 0.19 μm in thickness) decrease compared to those in 45-nm technology (0.14 μm in width and 0.28 μm in thickness). Also, the wire pitch in the 28-nm technology is smaller than that in the 45-nm technology, which makes the routing denser and increases intradie coupling. However, we still observe noticeable underestimation of -8.8% in total coupling capacitance from die-by-die extraction, as a result of ignoring interdie coupling. As results show, our in-context extraction methods using one to three interface layers per die are highly accurate compared to holistic extraction. Therefore, we conclude that our extraction methods are highly flexible and can handle various metal structures in different technology nodes.

VII. FULL-CHIP ANALYSIS

In this section, we present our full-chip timing, power, and signal integrity analysis results of our FFT64 benchmark using Primetime.

A. Impact of Interdie Coupling on 3-D Nets

Since interdie coupling is mostly between top metal layers of both dies, we focus on 3-D nets, which connect between bottom and top dies. Except for the clock net, which is assumed to be an ideal network, the rest of the 329 F2F vias are measured. Other 2-D nets have fewer routing wires on the top metal layers and are less affected by interdie coupling. The results are shown in Fig. 15, where each dot represents one

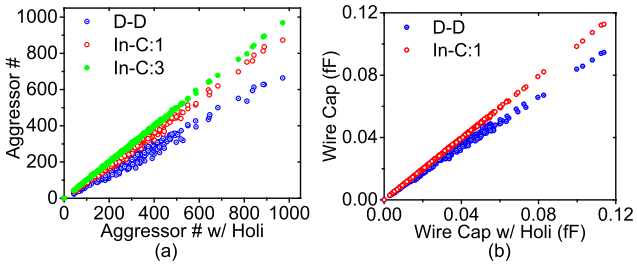


Fig. 15. Full-chip comparison of die-by-die (D-D) and in-context (In-C) against holistic extraction (Holi) on 3-D nets, each of which is represented by one dot. (a) Aggressor count. (b) Wire capacitance.

TABLE VIII

FULL-CHIP COMPARISON OF DIE-BY-DIE (D-D), HOLISTIC (HOLI), AND IN-CONTEXT (IN-C) EXTRACTION WITH ONE INTERFACE LAYER PER DIE

metric	Holi	D-D	Err%	In-C	Err%
Longest path delay (ns)	3.90	3.66	-6.2%	3.83	-1.8%
3D nets switching power (mW)	1.05	1.01	-3.5%	1.04	-0.4%
Total switching power (mW)	12.1	11.9	-1.7%	12.0	-0.8%
Total coupling cap on 3D nets (fF)	4.37	2.96	-32%	4.21	-3.7%
Total wire cap on 3D nets (fF)	10.8	9.35	-13%	10.7	-1.1%
Average aggressor # on 3D nets	285	200	-30%	253	-11%
Max noise on 3D nets (mV)	41.3	30.40	-26%	38.8	-6.1%

3-D net, and its X value is the result with holistic extraction. As results show, using die-by-die extraction, the number of aggressors is significantly underestimated for 3-D nets because aggressors from the neighbor die are ignored. However, with our in-context extraction, most aggressors are correctly captured even with one interface layer per die.

B. Full-Chip Power, Performance, and Noise

To find out how large interdie coupling impacts are on the full-chip metrics, we compare three extraction methods with full-chip analyses as shown in Table VIII. The longest path reported by Primetime is a 3-D path, which starts from a register in the top die, goes to the bottom die through an F2F via, and ends on another register in the top die. Since parasitics of interdie coupling mainly affect wires on the top metal layer, 3-D paths are heavily affected by interdie coupling. As results show, without interdie coupling, die-by-die extraction underestimates the longest path delay by 6.2%. Also, total wire capacitance on 3-D nets is underestimated by 13%. Therefore, die-by-die extraction is not enough for accurate full-chip analysis. Note that although interdie coupling capacitance is a large portion of total coupling capacitance, ground capacitance and pin capacitance are major contributors to the capacitive load of a net. Therefore, interdie coupling only slightly affects the switching power consumption of F2F designs. From our results, ignoring interdie coupling and the F2F bonding interface layers, die-by-die extraction underestimates 3.5% of total switching power on 3-D nets, while we only observe 1.7% underestimation on the switching power.

However, in terms of signal integrity, interdie coupling shows much larger impacts, especially on top metal layer wires. Total coupling capacitance reported on 3-D nets is underestimated significantly by 32%. Similarly, average

number of aggressors for 3-D nets is also underestimated by 30%. Because of fewer aggressors and a weaker coupling, the maximum noise on 3-D nets is underestimated by 26% with die-by-die extraction as well. Therefore, for sign-off verification and post-silicon analysis, where highly accurate parasitic extraction is required, die-by-die extraction introduces significant errors, and interdie coupling needs to be handled carefully.

With our in-context extraction, most of the interdie coupling and E-field interaction are captured accurately. As results show, the timing error is only 1.8% even using our in-context extraction with one interface layer per die, and total switching power is underestimated by only 0.8%. For signal integrity analyses, in-context extraction is also able to capture most coupling aggressors. For 3-D nets, only 3.7% and 1.1% underestimation is observed on total coupling capacitance and total wire cap, respectively. The max noise underestimation is only 6.1% with in-context extraction. Note that only one interface layer per die is included, and more coupling aggressors will be captured using in-context extraction with more interface layers. However, their coupling strengths are relatively weak, thus their impacts are much smaller.

VIII. Extraction for Heterogeneous 3-D ICs

Previous designs are still based on the homogeneous technology where the fabrication processes of both bottom and top dies are the same and designers have a full knowledge of the connectivity and geometry of the system. As discussed in Section II, though in-context extraction provides a fast and accurate approximation and is easier for implementation, holistic extraction is still the most accurate solution and can be implemented without problem. Once CAD tools are completely migrated to handle multiple dies, holistic extraction provides a straightforward solution. However, when multiple vendors are responsible for design and fabricating different dies, in-context extraction is preferred to protect intellectual property and decouple the design process with multiple companies. In this section, we discuss several issues in heterogeneous integration and the tradeoffs with in-context extraction. We also implement a heterogeneous design and perform full-chip extraction to validate our in-context flow.

A. Methodology

For accurate parasitic extraction, the connectivity (or netlists) information of both dies is required. However, with heterogeneous integration, it may not be possible because of intellectual property protection. This results in tradeoffs between extraction accuracy and CAD complexity. An example is shown in Fig. 16 with two nets. Net A is in the top die, and net B is in the bottom die. Both nets span across two layers with multiple wire segments. For an in-context extraction with one interface layer, various handling methods can be applied for heterogeneous integration. If the extraction engine has a full knowledge of the connectivity, as shown in Fig. 16(a), the extraction can be performed with correct E-field distribution, and all extracted capacitance can be netlisted correctly. In this case, capacitances C1 and C2 can be further reduced into one. However, if only the layout geometry is known, as shown in Fig. 16(b) and (c), there are two ways of handling the interface layer. Note that current analysis engine generally ignores floating nets, so either it can assume that all wires on the neighboring die are independent signal nets or they are grounded wires.

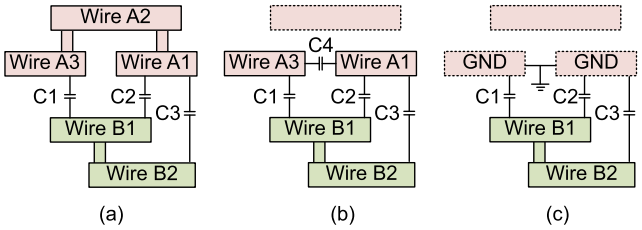


Fig. 16. Three cases of in-context extraction with one interface layer: (a) with connectivity information of the interface layer, (b) assumes signal nets, and (c) assumes ground nets.

However, both methods have to sacrifice the extraction accuracy. In Fig. 16(b), since wires A1 and A3 belong to different nets, it introduces an extra coupling capacitor C4 between them. Because of the E-field sharing represented by C4, some E-fields are redistributed to coupling between wires A1 and A3. This results in smaller capacitance from C1 to C3. On the other hand, wires A1 and A3 become two independent signal nets, which also differ from Fig. 16(a). As for Fig. 16(c), all the capacitance can be extracted as ground capacitance, but parasitics between two dies are completly decoupled. This results in some errors in noise and delay analyses as well. If net B is a victim, since both wires A1 and A3 are aggressors in Fig. 16(a), they generate noises through capacitors C1–C3 when switching. However, these capacitors become grounded in Fig. 16(c). Not only interdie aggressors are missing, but also the total ground capacitance on net B increases, which makes net B become more difficult to switch. Therefore, the coupling noise on net B is underestimated. On the other hand, the impact on the timing comes from Miller effects. In Fig. 16(a), the worse case delay is when net A and net B are switching to the opposite directions. Because of the Miller capacitor C1–C3, the delay of both nets is larger. However, Fig. 16(c) can only provide an average estimation for the delay after interdie capacitance is decoupled.

Since Primetime does not consider Miller effects on timing and power, we rebuild the environment of each 3-D net and perform Hspice simulation one-by-one for worst-case timing and noise analysis. All aggressors are assumed to have the same waveform switching in the opposite direction to victim nets, and we measure the delay and noise on each victim net between coupled capacitance as in Fig. 16(a) and decoupled capacitance as in Fig. 16(c). The results are shown in Fig. 17. As results show, with decoupled capacitance, the worst-case delay and noise are underestimated by 4.7% and 17.3% on average. Note that for a full timing path, the difference is small since most 2-D nets are not affected much. However, if the signal integrity is critical, designers need to provide both layout geometries as well as netlist connectivity for the interface layer to allow maximum accuracy with

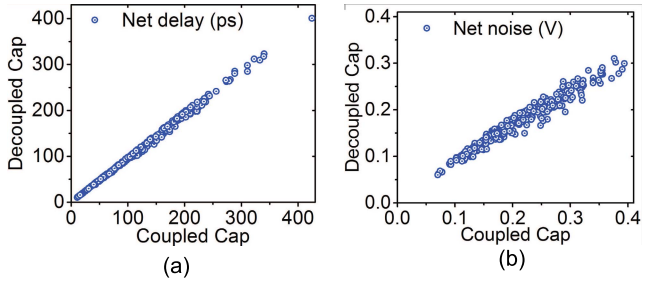


Fig. 17. Interdie decoupling impact on 3-D nets. (a) Worst-case delay. (b) Worst-case noise.

in-context extraction. This can be done by providing an annotated GDS file for the interface layers, where wire geometries are labeled with their connectivity information.

B. Routing Direction Impact

Another issue with heterogeneous integration lies in routing directions of metal layers. If wires on the neighboring layers are routed in the same direction, it is more likely that several wires are routed along in a long range. This will significantly increase the coupling between wires on the neighboring layers. Therefore, in common modern designs, wires on the neighboring layers are routed in orthogonal directions to avoid large coupling capacitance, except for M1, which may be routed in the same direction of its neighboring layer for manufacturing alignment issues. Previous design assumes a homogeneous technology such that metal stack configurations of both dies are the same. Therefore, the coupling capacitance is mainly formed between two top metal layers, which are routed in the same direction. This helps the in-context extraction to achieve high accuracy when only one interface layer is included.

However, in a heterogeneous design, where the designer and manufacturer of both dies are different and dies are designed separately, routing directions of top metal layers are likely to be orthogonal. This significantly changes the interdie coupling E-field distribution in the interface layers. Intuitively, interdie coupling may reduce because smaller coupling capacitance is formed between top layers of both dies. However, nonneighboring interface layers are routed with the same direction, which significantly increases the interdie coupling between them. For example, if M4B and M4T are routed in the orthogonal direction, the coupling between them will reduce. However, the coupling between M4B and M3T as well as the coupling between M3B and M4T increases since they are routed in the parallel direction. Therefore, if top metal layers are changed from parallel routing direction to orthogonal routing direction, its impact on interdie coupling depends on the technology configuration, such as metal dimensions and dielectric properties, as well as design layouts, which determine the wirelength distribution of each layers. Interdie coupling may increase or decrease depending on E-field distribution.

To illustrate this, we design our FFT circuit with top metal layers routed in orthogonal directions for comparison. To avoid changing the wirelength distribution, we redesign the top die by keeping its cell placement and F2F via locations the same, while rotating the routing directions of all its layers 90°. Then, we perform an incremental routing on the top die to fix any design violations. After the designs are generated, we perform holistic and in-context extraction on the new design and compare it to the original one. However, since we focus on heterogeneous designs, which are unaware of its neighboring die before bonding, in-context extraction results are divided into two parts, one for bottom die and one for top.

TABLE IX
HOLISTIC EXTRACTION OF FFT WITH ORTHOGONAL TOP METAL LAYERS. CAPACITANCE IS IN fF

Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
M1B	5.76							
M2B	3.02	380						
M3B	17.0	146	1268					
M4B	0.13	396	234	1824				
M4T	0.24	1.36	343	51.3	1278			
M3T	0.02	12.9	4.69	492	214	1681		
M2T	0.02	0.15	9.63	1.76	243	128	377	
M1T	0.00	0.02	0.02	0.33	0.14	5.97	5.02	7.10

TABLE X
IN-CONTEXT EXTRACTION ERRORS. NUMBER OF INTERFACE LAYERS IS ATTACHED AFTER THE DIE. CAPACITANCE IS IN fF

Die	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
bot:1	-0.03	-13.11	-12.30	-487	9.04			
top:1				-76.11	-348	-89.92	-10.09	-0.13
bot:2	-0.01	0.56	-8.93	24.17	1.46	26.26		
top:2			-62.67	-35.7	-62.53	-47.79	-2.69	-0.10
bot:3	0.00	0.30	-2.57	8.95	-3.07	11.50	-1.37	
top:3		-3.24	-29.65	-17.25	-30.8	-18.03	-1.51	-0.07

Table IX shows the holistic extraction of the redesigned FFT. As results show, unlike the original design where the maximum interdie capacitance is between M4B and M4T, in this design with orthogonal top metal layers, the maximum interdie coupling is between nonneighboring layers. The interdie coupling between M4 layers significantly decreases to 214 fF because of the routing direction change. Therefore, the coupling depth of this design increases to around two metal layers. This also changes the in-context extraction accuracy, as shown in Table X. As results indicate, because the interdie coupling increases significantly, in-context extraction on each individual die with only one interface layer is no longer accurate enough. The coupling depth is not fully covered by one interface layer, so adding more interface layers is necessary. By including two interface layers, it is guaranteed that at least one layer with horizontal routing direction and one layer with vertical routing direction will be included. The extraction error is significantly decreased. Furthermore, benefits of including three interface layers are small since it is out of the coupling depth. Therefore, we conclude that in-context extraction of heterogeneous 3-D IC needs to include at least enough interface layers covering the coupling depth: most likely, one interface layer if top layers of both die are routed in with parallel direction, and two layers if routed in orthogonal directions. Note that orthogonal top-layer routing is not a problem if designers have full knowledge to both dies, including layouts and connectivity as in homogeneous designs. This is because weighted interface-layer handling methods are able to correct the extraction error by combining both dies.

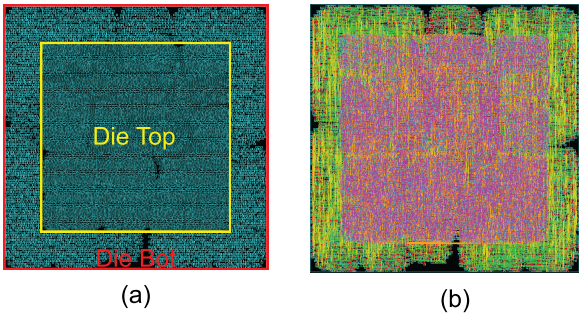


Fig. 18. Layout shots of the FFT design, whose top die is in 28 nm and bottom die in 45 nm. (a) Placement. (b) Routing.

TABLE XI
HOLISTIC EXTRACTION AND IN-CONTEXT EXTRACTION OF FFT SHOWN IN FIG. 18. CAPACITANCE IS IN fF

	Layer	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T
Holi	intra-die	32.63	1056	1865	2602	1768	2161	1651	56.84
	inter-die	0.20	15.46	134	781	677	146	105	1.40
INC	intra-die	32.98	1081	1876	2626	1752	2145	1623	56.29
	inter-die	0.21	11.03	118	764	669	130	93.39	1.13

C. Full-Chip Extraction of Heterogeneous Technologies

With heterogeneous integration, it is possible that top and bottom dies are designed and fabricated in different technology nodes. To illustrate this, we redesign our FFT circuits with heterogeneous integration shown in Fig. 3(c). The top die is designed in 28 nm, and the die footprint size is measured at 300 μm square. As shown in Fig. 18, the bottom die is still in a 45-nm node, and the cell placement is the same as previous designs with a footprint size of 380 μm. However, in order to fit F2F vias into the top die footprint, F2F via placement is scaled while the F2F via dimensions are the same. Bottom and top dies are still bonded with a 1-μm dielectric layer in between. We perform holistic extraction and in-context extraction with two interface layers on this design, and results are shown in Table XI. For the bottom die, the coupling capacitance is smaller for its top layers since the top die is shrunk, which leaves an empty region to its boundary. This also results in a reduction in total interdie coupling since the D2D distance is unchanged. However, if bonding technology improvement is considered, which requires a thinner interdie dielectric layer, the interdie coupling is still comparable to previous designs. As results show, our in-context extraction is still accurate for designs with heterogeneous integration and remains LVS-friendly. However, if extraction of each die is conducted independently, including layers at least covering interdie coupling depth is required for high accuracy.

As a summary, die-by-die extraction is cost-efficient and does not require new CAD tools. It is accurate on designs with thick die interface layers and small interdie coupling capacitance. Holistic extraction, in contrast, is the most complex and time-consuming procedure, but provides the highest accuracy across various technologies. It is more suitable for homogeneous integration or designs in which information about both designs is provided beforehand. However, it requires updating current CAD infrastructures with multiple-die handling, which will take some time before it is widely adopted. In-context extraction entails fewer layers, and does not require the simultaneous extraction of two device layers, which introduces significant difficulties for LVS checking. Foundries need only to reveal their interface layers, but they do not need to share important device fabrication details with in-context extraction. This can further accelerate the commercial adoption of in-context extraction on heterogeneous F2F 3-D ICs.

IX. CONCLUSION

In this paper, we compared three extraction methods in F2F 3-D ICs. We demonstrated the impact of E-field sharing and determined that interdie coupling cannot be ignored in F2F-bonded 3-D ICs. We implemented a holistic extraction method for homogeneous integration and found that it is the most accurate at capturing all interdie coupling. We also proposed an in-context extraction method for heterogeneous integration that is compatible with traditional CAD tools, but includes interface layers from a neighboring die during extraction. While die-by-die extraction underestimates total coupling capacitance, holistic extraction more accurately estimates coupling capacitance by capturing all interdie coupling but with higher complexity. Our in-context extraction is highly accurate and captures most E-field interactions across dies. In addition, being LVS-friendly, it can be easily implemented to simplify collaboration across multiple companies.

REFERENCES

[1] D. H. Kim *et al.*, "Design and analysis of 3D-MAPS (3D massively parallel processor with stacked memory)," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 112–125, Jan. 2015.

[2] C. S. Tan *et al.*, "Three-dimensional wafer stacking using Cu-Cu bonding for simultaneous formation of electrical, mechanical, and hermetic bonds," *IEEE Trans. Device Mater. Rel.*, vol. 12, no. 2, pp. 194–200, Jun. 2012.

[3] Y. J. Chang *et al.*, "Electrical characterization and reliability investigations of Cu TSVs with wafer-level Cu/Sn-BCB hybrid bonding," in *Proc. Tech. Program VLSI Technol. Syst. Appl.*, Apr. 2012, pp. 1–2.

[4] T. Lacrevaz *et al.*, "Electrical broadband characterization method of dielectric molding in 3-D IC and results," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 4, no. 9, pp. 1515–1522, Sep. 2014.

[5] M. Murugesan *et al.*, "High density 3D LSI technology using W/Cu hybrid TSVs," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2011, pp. 6.6.1–6.6.4.

[6] M. Motoyoshi *et al.*, "Stacked SOI pixel detector using versatile fine pitch-bump technology," in *Proc. IEEE Int. 3D Syst. Integr. Conf.*, Jan. 2012, pp. 1–4.

[7] H.-G. Lee *et al.*, "Wafer-level packages using B-stage nonconductive films for Cu pillar/Sn-Ag microbump interconnection," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 5, no. 11, pp. 1567–1572, Nov. 2015.

[8] L. Peng *et al.*, "Ultrafine Pitch (6 μm) of recessed and bonded Cu-Cu interconnects by three-dimensional wafer stacking," *IEEE Trans. Electron Devices*, vol. 33, no. 12, pp. 1747–1749, Dec. 2012.

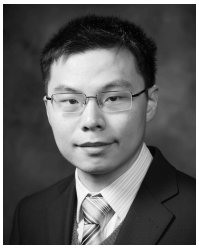
[9] L. Benaissa *et al.*, "A vertical power device conductive assembly at wafer level using direct bonding technology," in *Proc. IEEE 24th Int. Symp. Power Semiconductor Devices ICs (ISPSD)*, Jun. 2012, pp. 77–80.

[10] S. Gupta *et al.*, "Techniques for Producing 3D ICs with high-density interconnect," in *Proc. Int. VLSI Multilevel Interconnection Conf.*, Sep. 2004, pp. 1–5.

[11] J. Wilson *et al.*, "Fully integrated AC coupled interconnect using buried bumps," *IEEE Trans. Adv. Packag.*, vol. 30, no. 2, pp. 191–199, May 2007.

[12] S. Han and D. D. Wentzloff, "0.61W/mm^2 resonant inductively coupled power transfer for 3D-ICs," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2012, pp. 1–4.

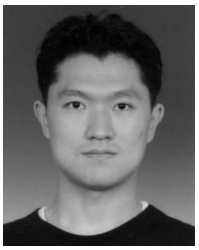
[13] H. Kim *et al.*, "A wideband on-interposer passive equalizer design for chip-to-chip 30-Gb/s serial data transmission," *IEEE Trans. Compon., Packag. Manuf. Technol.*, vol. 5, no. 1, pp. 28–39, Jan. 2015.
[14] M. T. L. Aung, E. Lim, T. Yoshikawa, and T. T. H. Kim, "Design of simultaneous bi-directional transceivers utilizing capacitive coupling for 3DICs in face-to-face configuration," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2, pp. 257–265, Jun. 2012.
[15] J. Xu *et al.*, "2.8 Gb/s inductively coupled interconnect for 3D ICs," in *Proc. Symp. VLSI Circuits*, Jun. 2005, pp. 352–355.
[16] T. Song *et al.*, "Coupling capacitance in face-to-face (F2F) bonded 3D ICs: Trends and implications," in *Proc. IEEE Electron. Compon. Technol. Conf.*, May 2015, pp. 529–536.
[17] P. Batude *et al.*, "3DVLSI with coolcube process: An alternative path to scaling," in *Proc. Symp. VLSI Technol.*, Jun. 2015, pp. T48–T49.
[18] M. L. Fan *et al.*, "Investigation and optimization of monolithic 3D logic circuits and SRAM cells considering interlayer coupling," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jun. 2014, pp. 1130–1133.
[19] G. Chen, H. Zhu, T. Cui, Z. Chen, X. Zeng, and W. Cai, "ParAFEMCap: A parallel adaptive finite-element method for 3-D VLSI interconnect capacitance extraction," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 2, pp. 218–231, Feb. 2012.
[20] T. El-Moselhy, I. M. Elfadel, and L. Daniel, "A Markov chain based hierarchical algorithm for fabric-aware capacitance extraction," *IEEE Trans. Adv. Packag.*, vol. 33, no. 4, pp. 818–827, Nov. 2010.
[21] W. Yu *et al.*, "Utilizing macromodels in floating random walk based capacitance extraction," in *Proc. Design, Autom. Test Eur.*, Mar. 2016, pp. 1225–1230.
[22] Y. Zhou, Y. Zhang, V. Sarin, Q. Qiu, and W. Shi, "Macro model of advanced devices for parasitic extraction," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 35, no. 10, pp. 1721–1729, Oct. 2016
[23] Y. Araga *et al.*, "Measurements and analysis of substrate noise coupling in TSV-based 3-D integrated circuits," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 4, no. 6, pp. 1026–1037, Jun. 2014.
[24] G. Kumar *et al.,* "Design and demonstration of power delivery networks with effective resonance suppression in double-sided 3-D glass interposer packages," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 6, no. 1, pp. 87–99, Jan. 2016.
[25] Z. Li, Y. Li, and J. Xie, "Design and package technology development of face-to-face die stacking as a low cost alternative for 3D IC integration," in *Proc. IEEE Electron. Compon. Technol. Conf.*, May 2014, pp. 338–341.
[26] T. Song and S. K. Lim, "Die-to-die parasitic extraction targeting face-to-face bonded 3D ICs," *J. Inf. Commun. Converg. Eng.*, vol. 13, no. 3, Sep. 2015, pp. 172–179.
[27] Y. Peng *et al.*, "Full-chip inter-die parasitic extraction in face-to-face-bonded 3D ICs," in *Proc. IEEE Int. Conf. Comput.-Aided Design*, Nov. 2015, pp. 649–655.



Yarui Peng (S'12) received the B.S. degree from Tsinghua University, Beijing, China, in 2012, and the M.S. and Ph.D. degrees from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2014 and 2016, respectively.

He joined the Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR, USA, as an Assistant Professor, in 2017. His current research interests include computer-aided design, analysis, and optimization for emerging technologies and systems, such as 2.5-D wafer-level-packaging and 3-D ICs, and high-efficiency digital designs and memory systems. He is also interested in design automation of high-bandgap power electronics and mobile electrified systems.

Dr. Peng was a recipient of the best-in-session award in SRC TECHCON'14 and the best student paper award in ICPT'16.



Taigon Song (M'16) received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2007, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2009, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2015.

He is currently a Senior Research and Development Engineer with Synopsys Inc., Mountain View, CA, USA. He has authored over 40 publications in international conferences and journals. His current research interests include modeling, design, extraction, and analysis in the advanced technologies, including 3-D integrated circuits.

Dusan Petranovic (M'92) received the B.S. degree from the University of Belgrade, Belgrade, Serbia, the M.S. degree from the Worcester Polytechnic Institute, Worcester, MA, USA, and the Ph.D. degree from the University of Montenegro, Podgorica, Montenegro.

He was a Professor and the Chairman of the EE Department, University of Montenegro. He spent six years teaching with the Harvey Mudd College, Claremont, CA, USA. He joined the LSI Logic Advanced Development Laboratory as a member of Technical Staff, where he is involved in interconnect modeling. He was also a Consultant for NASA and NOVA Management Inc. He is currently an Interconnect Modeling Technologist with the Design to Silicon Group, Mentor Graphics Corporation, Fremont, CA, USA, where he is involved in all aspects of parasitic extraction. He holds 15 U.S. patents and has authored numerous journal and conference papers.

Sung Kyu Lim (M'01–SM'05) received the B.S., M.S., and Ph.D. degrees from the University of California at Los Angeles, Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently a Dan Fielder Endowed Chair Professor. His research on 3-D IC reliability is featured as research highlight in the communication of the ACM in 2014. His 3-D IC test chip published in the IEEE International Solid-State Circuits Conference (2012) is generally considered the first multicore 3-D processor ever developed in academia. He has authored the book *Practical Problems in VLSI Physical Design Automation* (Springer, 2008). His current research interests include modeling, architecture, and electronic design automation (EDA) for 3-D ICs.

Dr. Lim has served on the Technical Program Committee of several premier conferences in EDA. He was on the Advisory Board of the ACM Special Interest Group on Design Automation from 2003 to 2008. He was a recipient of the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006. He received the Distinguished Service Award in 2008. He received the Best Paper Awards from the IEEE Asian Test Symposium (2012) and the IEEE International Interconnect Technology Conference (2014). He received the Class of 1940 Course Survey Teaching Effectiveness Award from the Georgia Institute of Technology (2016). He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS from 2007 to 2009. He has been an Associate Editor of the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS since 2013.