Full-chip Inter-die Parasitic Extraction in Face-to-Face-Bonded 3D ICs

Yarui Peng[1], Taigon Song[1], Dusan Petranovic[2], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA [2]Mentor Graphics, Fremont, CA, USA
yarui.peng@gatech.edu limsk@ece.gatech.edu

***Abstract*—Face-to-face (F2F) bonded 3D ICs are promising design solutions. However, because of the short die-to-die distance, direct coupling between the metal layers of the top and bottom dies introduces severe signal integrity problems that call for accurate extraction. This study is the first to demonstrate and compare three parasitic extraction methods of F2F-bonded 3D ICs. One is traditional die-by-die extraction, which cannot handle inter-die coupling and E-field sharing. We propose another method, holistic extraction, which treats all layers from both dies simultaneously and captures all inter-die coupling at the cost of high Layout Versus Schematic (LVS) complexity. We also propose an in-context extraction method that accounts for interface layers between dies. Carefully handling double-counting and surface layers issues, in-context extraction is LVS-friendly without losing accuracy. Full-chip analyses show that both of our extraction methods are highly accurate and able to handle various metal layers in several process nodes. It also corrects timing, power, and signal integrity errors introduced by die-by-die extraction. In-context extraction with two interface layers is highly accurate and efficient with an error of 0.9% for total ground capacitance and 0.8% for total coupling capacitance.**

***Index Terms*—Face-to-face, 3D IC, die-by-die, holistic, in-context, extraction**

I. Introduction

To achieve performance improvements at reduced power and smaller footprint than conventional 2D processes, 3D ICs are promising solutions to extend the Moore's Law. Face-to-face (F2F) bonding technology connects top metal layers from both dies using a direct Cu-Cu bonding process with F2F vias [1], [2]. Without using TSVs for inter-die connections, F2F designs achieve a much higher 3D connection density with F2F vias in a few microns [3]. Recent studies have shown that F2F designs have many benefits over F2B designs [4]. However, F2F bonding introduces new parasitics. As shown in Figure 1, inter-die coupling capacitance becomes more significant with a closer die-to-die distance. Many studies [5] have shown that with a direct Cu-Cu bonding, the die-to-die distance is comparable to the thickness of the top inter-layer dielectric (ILD). Therefore, inter-die coupling can no longer be ignored especially with future 3D IC technology where dies are closely bonded or even fabricated in monolithic.

We compare three methods for capacitance extraction of F2F designs. First, die-by-die extraction ignores inter-die coupling and extracts the bottom and top dies individually. Though field solvers are always able to handle any kinds of capacitance extraction with an extremely long runtime, die-by-die extraction is the only option available for full-chip F2F 3D IC extraction. Sign-off parasitic extraction needs to be performed after LVS checking so that parasitics can be netlisted, and die-by-die extraction is considered as "LVS-friendly", since LVS can be done without knowing any geometries from the neighbour die. If different foundries are responsible for fabricating the top and bottom dies, they do not need to share their technology information, since the fabrication process is highly

This work is supported by the Semiconductor Research Corporation (CADTS Task 2239) and Mentor Graphics Corp.

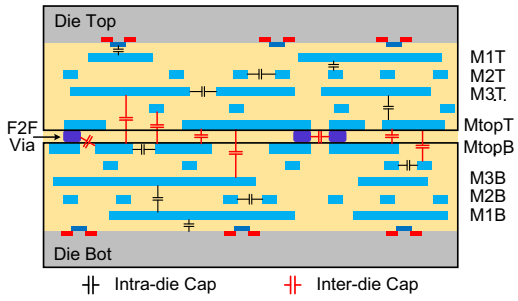

Fig. 1. Cross-sectional view of a F2F-bonded 3D IC structure with interconnect parasitics.

classified, especially for any details regarding device layer fabrication. Therefore, die-by-die extraction is used in commercialized technology, and demonstrated F2F 3D ICs are all based on this extraction technique.

However, new parasitics in F2F designs require to include impacts from the neighbouring die. Our second option is holistic extraction that takes the whole layer stack from both dies into account. This method allows for full extraction between any metal or device layers and is able to capture all E-field interactions between neighbouring dies. However, holistic extraction is extremely challenging computationally. Further, it is especially hard for commercialization and implementation. Because of the complexity of coding an LVS deck that can properly recognize all the devices, and connect two dies stacked on top of each other and may from different processes, holistic extraction requires significant expertise in LVS rule file coding, and it is sometimes impossible to ask foundries to share their top secrets in device fabrication to create a joint LVS rule deck.

To improve parasitic extraction accuracy without imposing the need for holistic extraction, we propose in-context extraction. Instead of a single run with every layers, in-context extraction only takes a few neighbouring layers, called interface layers, into consideration during one extraction stage. The top and bottom dies are extracted separately but both are extracted with the knowledge of interface layers. To differentiate this with die-by-die extraction, we call dies with interface layers from the neighbouring die as in-context dies. By providing enough geometry and material data, in-context extraction is able to capture most inter-die coupling elements since major capacitive coupling fields are limited between a few neighbouring layers. This approach requires more efforts than die-by-die extraction, but it still remains LVS-friendly. This is because only a few top metal layers are needed to code an LVS deck for in-context dies, which is much easier than capturing both complex device layer geometries. And foundries only need to share dimensions of their top metal layers to enable a close-to-optimum solution. Therefore, this approach reduces the complexity of handling all layers simultaneously and can be carried out independent of device fabrication process.

978-1-4673-8388-2/15/$31.00 ©2015 IEEE

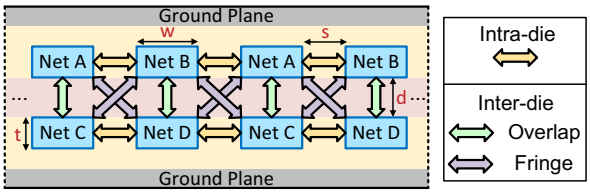

Fig. 2. Raphael structure for capacitance extraction. Both the top and bottom dies contain repeated layout patterns. D denotes the die-to-die distance while w, s, and t denote wire width, spacing, and thickness, respectively.

In this paper, we make the following contributions: (1) We develop algorithms and CAD flows for both holistic and in-context extraction by combining commercial and in-house tools. (2) We provide a comprehensive study and comparison among all three F2F extraction methods (i.e., die-by-die, holistic, and in-context extraction) using full-chip 3D ICs. (3) We study impacts of E-field sharing and surface layer handling on extraction accuracy in detail. (4) We investigate inter-die coupling, timing, power and signal integrity issues with full-chip analyses.

II. Field Sharing Analysis

In this section, we validate our motivation using a field solver and analyze impacts that affect inter-die coupling. To find out how large is inter-die coupling compared with intra-die coupling, we build a test structure shown in Figure 2, where the top die is a repeated pattern of Net A and B, and the bottom die is a repeated pattern of Net C and D. The ground planes are located 3µm away from wires, and wire width (w) and thickness (t) are fixed as 0.8µm and 1.2µm, which are the same as top metal layer dimensions in a 45nm technology. Coupling capacitance in this structure can be divided into three groups: intra-die coupling capacitance, inter-die overlapping capacitance, and inter-die fringe capacitance. This special layout is used so that every capacitor in one group has a same value. Therefore, intra-die coupling, inter-die overlapping, and inter-die fringe capacitance can be represented by capacitance between Net A and B (Cap AB), Net A and C (Cap AC), and Net A and D (Cap AD), respectively. Capacitance is extracted assuming an infinite wire length thus it has a unit of fF/µm.

The raphael extraction results with various die-to-die distances are shown in Figure 3, and total capacitance results are measured with ten wires on each die. Wire spacing is fixed as 0.9µm, which is required by design rules. With a closer die-to-die distance, inter-die coupling capacitance increases significantly, while inter-die fringe capacitance increases slightly. And with a die-to-die distance smaller than 1µm, inter-die coupling capacitance becomes comparable to intra-die coupling capacitance even with minimum wire spacing. This clearly demonstrates that inter-die coupling cannot be ignored with a close die-to-die distance. Another interesting observation comes from E-field sharing of the neighbouring die. With a closer die-to-die distance, the neighbour die shares more E-fields between wires. This results in a reduction in intra-die coupling capacitance as dies get closer. Overall, total capacitance always increases with a closer die-to-die distance, and the portion of inter-die coupling keeps increasing as well. Therefore, die-by-die extraction, which is unaware of the neighbouring die and ignores the E-field sharing, cannot extract the inter-die coupling capacitance accurately.

The raphael extraction results with various wire spacings are shown in Figure 4, where the die-to-die distance is fixed as 1µm. With a large wire spacing, both intra-die coupling and total coupling capacitance decrease. However, the inter-die coupling capacitance percentage increases with a wider wire pitch. This is because with a larger wire

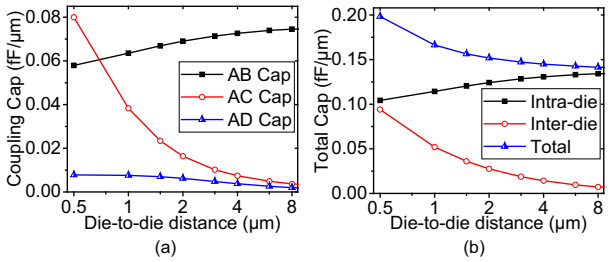

Fig. 3. Die-to-die distance (= d in Figure 2) impact. (a) Single capacitor extraction, A to D are nets in Figure 2; (b) total capacitance extraction.

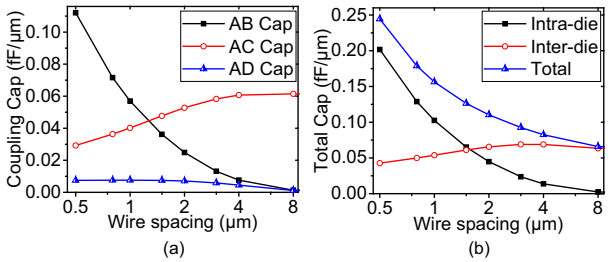

Fig. 4. Wire-to-wire spacing (= s in Figure 2) impact. (a) Single capacitor extraction, A to D are nets in Figure 2, (b) total capacitance extraction.

spacing, E-field sharing from neighbouring wires in the same die is weaker, thus more coupling is formed between overlapped wires across dies. As a result, total inter-die capacitance increases with a wire spacing up to 3µm. However, because of a longer distance, fringe capacitance between Net A and D decreases, this results in a slight reduction in total inter-die capacitance with a wire spacing larger than 3µm. Overall, inter-die capacitance becomes comparable to intra-die capacitance with a wire spacing larger than 1µm. Therefore, inter-die coupling cannot be ignored in designs with sparsely-routed top metal layers, and E-field sharing between wires also significantly affects inter-die coupling capacitance.

III. Die-by-die And Holistic Extraction

A. Die-by-die Extraction

Die-by-die extraction uses the same technology and libraries as the traditional 2D process. A sample technology with four metal layers is shown in Figure 5 (a) for die-by-die extraction, and it is the same as a 2D technology. Since no commercial design tool is able to handle timing and power optimization of 3D designs, current 3D ICs have dies designed separately. Both the top and bottom dies can share the same 2D technology and extraction rule files, thus a certain technology only needs to be calibrated once. After parasitics are obtained from both top and bottom dies separately, designers need to include a top-level netlist, which describes 3D connections between dies, as well as a top-level parasitic file which includes capacitance of all F2F vias, since both dies are extracted unaware of 3D interconnects. Ignoring the inter-die capacitance, this flow is widely adopted for both F2F and F2B designs, and it is the fastest approach by excluding any interface layers.

B. Holistic Extraction

Compared with the die-by-die approach, holistic extraction needs to consider all layers simultaneously as shown in Figure 5 (b). 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.

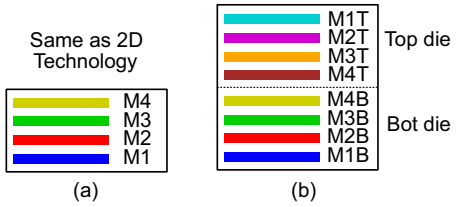

Fig. 5. Sample interconnect technologies with four metal layers. (a) Die-by-die extraction, and (b) holistic extraction.

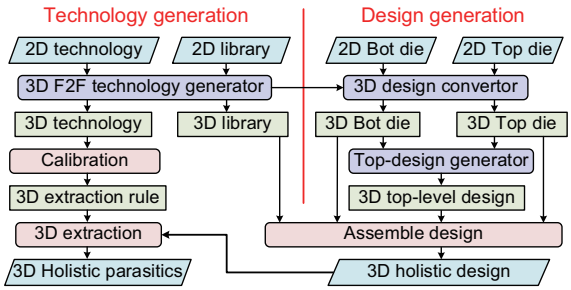

Fig. 6. CAD flow chart of our holistic extraction.

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 Figure 6 without considering the top die device layer. This will introduce some errors mostly on the M1T layer in holistic extraction. However, this is still reasonable since parasitics inside standard cells should be extracted separately, and we are performing a full-chip cell-level extraction with very few M1 wires for inter-cell connections.

Our holistic extraction flow contains both in-house tools and commercial tools for design automation. For holistic technology generation, a technology generator reads the 2D technology and library, and duplicates metal layers and cells in the F2F fashion as shown in Figure 5(b). The cells located in the top and bottom dies are given with different cell names and their pin layers are changed accordingly. The generated 3D technology and library contain all metal layers as well as the bottom die substrate and device layer. Also, with holistic technology files, it is possible to have devices fabricated in different technologies for the top and bottom dies. Thus our method can easily be extended to handle mixed-technology designs. For F2F bonding layer, we adopt the method used in [4], in which F2F connections are modeled as F2F vias between top metal layers of both dies. Unlike die-by-die extraction, where the extraction engine cannot analyze the F2F bonding layer and ignores inter-die coupling capacitance, 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. After the 3D technology is calibrated, extraction rules are generated to handle any 3D holistic designs with the same vertical structure. Therefore, even though more runtime is needed for holistic field solving and technology calibration, once these extraction rules are generated, the runtime of full-chip holistic extraction is still comparable with die-by-die extraction, since all layers are extracted in a single run.

To merge the top and bottom dies, both of which are designed in 2D fashion, we implement a CAD flow for generating the 3D holistic design based on die-by-die designs. First, a 3D design convertor takes in both designs and converts all layers and cells as well as design netlists according to the output of the 3D technology generator. Then,

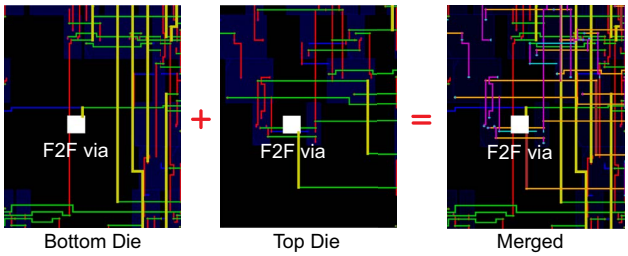

Fig. 7. 3D holistic design generation.

by taking the LEFs of both dies, our top-design generator creates a top-level design which has the same footprint as the 3D chip but only contains 3D connections between top and bottom dies. In this design, both top dies and bottom dies are design blocks, and they are overlapped in the floorplan. Only F2F connections and I/O pins of the 3D design are presented in this design, and the top-design generator automatically inserts F2F vias to connect block pins from both dies. With all three designs ready (*i.e.*, the top-level design as well as two converted 3D dies), we use the assemble design command to read die info into the top level. By tricking the tool and treating both dies as design blocks, we successfully generate the 3D holistic design, which bears all design information of the 3D chip and passes all connectivity and geometry verification. With the 3D design and extraction rules, all parasitics can be extracted holistically.

IV. In-Context Extraction

In this section, we present our first-of-its-kind in-context extraction flow. Our goal is to use a similar flow as traditional die-by-die extraction but with inter-die extraction accuracy similar to that of the holistic design.

A. Technology and Design Generation

Unlike holistic extraction, where current CAD tools cannot handle multiple substrate and device layers simultaneously, in-context extraction eliminates the need to create new extraction engines for 3D structure, and our flow is fully compatible with all major CAD tool vendors. 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 unsymmetrical F2F bonded designs in which number of metal layers or interface layers from top and bottom dies are not the same.

To enable such extraction, for each in-context die, we must include enough data about geometries and materials from the neighboring die's routing results. Our in-context flow is shown in Figure 8, assuming both dies are fabricated with the same technology. For technology generation, we start with the 2D technology and library to create in-context technology files. An example with four metal layers and one interface layer per die is shown in Figure 9. For the bottom die, we need to add the top die interface layer, which is recognized as M5 by CAD tools. Similarly for the top die, the M4B layer is recognized as the new M5 layer. Note that if both in-context dies have exact the same layer stack, only one technology calibration step is needed. We call the top metal layer in our in-context technology as the surface layer, though no metal layer is physically located at the surface in F2F bonding. For example, with a metal stack (In-C:1) shown in Figure 9, M4B and M4T layers are surface layers of top and bottom in-context technologies, respectively. For In-C:2 extraction, M3B and M3T become surfaces layers of top and bottom

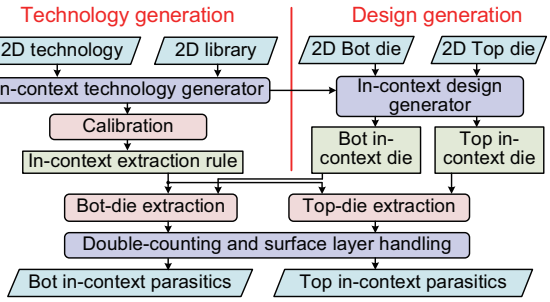

Fig. 8. CAD flow chart of our in-context extraction.

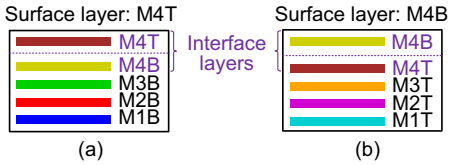

Fig. 9. A sample in-context interconnect technology with four metal layers.

in-context technologies, respectively. After the in-context technology is generated, they can be calibrated similarly as the traditional 2D technology.

For in-context design generation, similar to holistic extraction, our generator takes in both dies' design files. But when routing wires are merged into a in-context design, we only include layers of interest in each in-context design, and other layers as well as devices are discarded. The layers are renamed accordingly to technology generator outputs. Figure 10 illustrates this in-context design generation process. However, unlike the holistic design, where cells from both dies are sharing the same design, in-context extraction does not need to modify any geometry or timing library files for cells, thus the same 2D library as the die-by-die flow is used, which is an advantage for LVS and connectivity check. Then both in-context dies are extracted similarly as the die-by-die flow. Note that our flow is intended for validation of in-context extraction, thus we assume both dies are LVS-clean and do not perform a real LVS check before our extraction. However, when the extraction flow is implemented for sign-off verification, the extraction tool can no longer assume netlists are LVS-clean, therefore the layouts must first be netlisted then extracted. With in-context extraction, most of inter-die E-fields are captured, thus it provides a close-to-optimum solution with easy implementation.

B. Double Counting Correction

In-context extraction is more LVS-friendly and compatible with current CAD tools than holistic extraction. However, special issues exist for in-context extraction, especially for handling the interface layers. If we directly read parasitics from both dies incrementally, inter-die capacitance will be significantly overestimated because of double-counted parasitics in interface layers. Since interface layers are extracted both in the top and bottom in-context designs, any ground capacitance of the interface layers are calculated twice. Also, any coupling capacitors, both of whose nodes are in interface layers, are also double-counted. For example, coupling between M3T and M4B is double-counted, but the coupling between M4T and M2T is not.

To solve the double counting problem, we implement an SPEF analyzer in C++, which reads an extended SPEF file with geometry

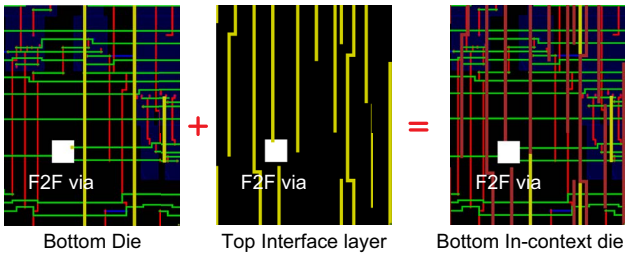

Fig. 10. 3D in-context design generation.

information and look 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 value between top and bottom in-context parasitics. We call this method as "In-C halved" and the method simply merging both in-context parasitics as "In-C original." With an In-C halved extraction, overestimation of inter-die coupling is resolved.

C. Surface Layer Correction

Another issue which also affects the in-context extraction accuracy is the surface layer handling. Shown in Figure 9, surface layers of both in-context dies are metal layers without any top neighbour layer in the metal stack. As discussed in Section II, E-field sharing in the F2F design significantly affects coupling capacitance. However, with in-context designs, E-field sharing impacts are not fullly taken into consideration since a few metal layers are missing during the technology calibration. Since most E-field sharing happens between neighbour metal layers, surface layers are mostly affected by inaccurate extraction. Unlike other metal layers where E-field sharing from both sides are taken care of, the capacitance extracted on the surface layer only considers the E-field sharing from its bottom neighbour layer. The In-C halved method is able to correct the double-counting but unable to fix the inaccurate surface layer capacitance.

To solve this issue, we propose an "In-C weighted" method. The motivation is simple, as we observe that the surface layer in one in-context die is not the surface layer in the other in-context die. For example, as in Figure 9, capacitance on M4T can not be extracted accurately with bottom in-context die, but they can be extracted accurately in the top in-context die, where M4T is not the surface layer. To implement this, we use a parameter D for each metal layer as the distance to the surface. In an 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 Figure 9, the M2B layer has a D of three in the bottom in-context technology, while the M3T layer has a D of two in the top in-context die. 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. 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 are not double-counted. As long as the total weight of both dies is equal to one, there is no overestimation in inter-die coupling. Therefore, for a double-counted capacitor, the weight ratio between the bottom in-context die and the top in-context die is proportional to the product of R ratios of both layers the capacitor connects to. Figure 11 illustrates a sample calculation of the weigh ratio in a technology with four metal layers and two interface layers. Using this ratio, our in-context extraction algorithm gives more

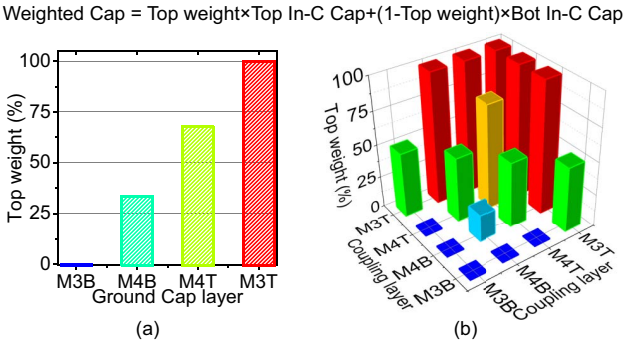

Fig. 11. Correction weight for top in-context die in a 2-tier 3D IC with two interface layers per die.

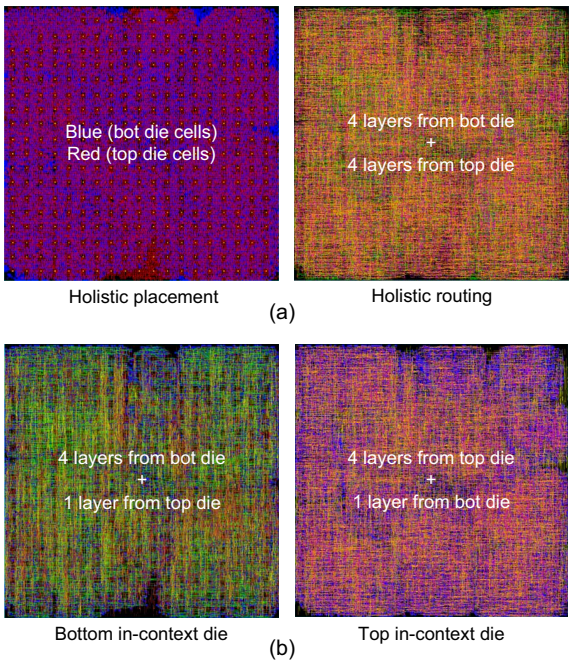

Fig. 12. GDSII layouts of FFT64 benchmark using four metal layers. (a) holistic, (b) in-context with 1 metal layer from the other die for the interface.

weights to layers far from the surface so that the inaccuracy from E-field sharing impact is mitigated.

V. Full-chip Extraction Results

We build a 64 point FFT (FFT64) circuit routed up to M4 using a 45nm technology for most studies in this paper. The F2F via has a size of 1μm × 1μm, and the F2F bonding layer is 1μm in thickness and filled by SiO_2 with a relative permittivity of 3.9. The F2F via resistance is assumed as 1Ω. The FFT64 design has a footprint of 380μm × 380μm with 38K gates. Figure 12 shows FFT64 design shots.

A. Inter-die vs. Intra-die Breakdown

We analyze how much coupling in a F2F design is contributed by inter-die coupling using holistic extraction results as shown in Table I. Table II summarizes total intra-die coupling capacitance vs. total inter-die coupling capacitance for each metal layer. As results shown, most inter-die coupling is between top metal layers of both dies. The inter-die coupling contributes to 34% of the total coupling

TABLE I
Holistic extraction of F2F coupling capacitance. Capacitance valus is in fF.

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

TABLE II
Breakdown of coupling capacitance shown in Table I into intra-die vs. inter-die.

	M1B	M2B	M3B	M4B	M4T	M3T	M2T	M1T	Total
Intra	26.0	927	1,656	2,454	1,876	1,595	963	37.8	9,536
Inter	0.18	21.9	151	1,249	1,212	160	50.0	0.42	2,845
Inter %	0.7%	2.3%	**8.4%**	**34%**	**39%**	**9.1%**	4.9%	1.1%	23%

capacitance on M4B and 39% of the total coupling capacitance on M4T layer. We also observed a noticeable contribution from inter-die 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 inter-die coupling is negligible. Overall, inter-die coupling contributes to 23% in total coupling capacitance in the F2F-bonded FFT64 design.

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

B. Die-by-die vs. Holistic Extraction

We analyze how much error is introduced by die-by-die coupling. Total extracted ground capacitance is very similar between die-by-die extraction (39476fF) and holistic extraction (39247fF) with only a 0.58% difference. Most differences come from coupling capacitance. Die-by-die extraction results are shown in Figure III. Note that all inter-die coupling is ignored by this extraction method. This results in significant underestimation in total coupling capacitance. We also observe large underestimation with die-by-die extraction on coupling capacitance of top metal layers. This is because die-by-die extraction ignores the F2F bonding layer, thus top metal layer capacitance is extracted inaccurately. Overall, die-by-die extraction underestimates total coupling capacitance by 35% compared with holistic extraction, as shown in Table IV. And most errors come from the top metal layers of both dies. Therefore, we conclude that die-by-die extraction cannot accurately capture all coupling capacitance and E-field interactions inside the F2F designs.

C. In-Context vs. Holistic Extraction

We compare extraction results of in-context extraction with holistic extraction, which is assumed as our golden model. Note that since holistic extraction cannot handle the top die substrate and device layer, M1T layer parasitics extracted with holistic extraction are less reliable. Table V shows extraction results with in-context extraction

TABLE III
DIE-BY-DIE EXTRACTION OF F2F COUPLING CAPACITANCE. CAPACITANCE IS IN fF.

Layer	M1B	M2B	M3B	M4B	Layer	M4T	M3T	M2T	M1T
M1B	5.33	2.36	12.3	0.09	M4T	905	203	305	0.16
M2B	2.36	337	139	377	M3T	203	1055	127	13.6
M3B	12.3	139	1216	253	M2T	305	127	313	2.46
M4B	0.09	377	253	1325	M1T	0.16	13.6	2.46	4.97

TABLE IV
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%

with two interface layers per die (In-C:2). Since M1 and M2 are not interface layers, any inter-die coupling capacitance on those layers is ignored by in-context extraction, but since the inter-die coupling contributions are small, 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 layer is much easier than analyzing multiple device layers.

Table VI summarizes the extraction comparison between in-context and holistic extraction. As results shown, our in-context extraction is highly accurate in both ground capacitance and coupling capacitance layer by layer. Since our in-context extraction ignores a few inter-die coupling elements, 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 0.3% and 0.7%, respectively, indicating that almost all inter-die 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

We provide our in-context extraction results with various interface layer handling methods discussed in Section IV-C. Table VII summarizes full-chip extraction results. When merging parasitics of interface layers for in-context extraction, interface layer handling significantly affects extraction accuracy. As results indicate, 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 neighbour layers differ.

By dividing every capacitance in half, extraction errors are 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. With our proposed method using a weighted average, the accuracy of the in-context extraction improves significantly. Total coupling capacitance errors for M3B and M3T are reduced by -0.3% and -1.4%, respectively,

TABLE V
IN-CONTEXT EXTRACTION OF F2F COUPLING CAPACITANCE. WE USE TOP 2 METAL LAYERS FOR THE INTERFACE. CAPACITANCE IS IN fF.

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

TABLE VI
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%	-1%	-0.8%

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 the capacitance value affects the 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. However, we also study the in-context extraction accuracy with various numbers of interface layers. Table VIII summarizes these results. Interestingly, even with only one interface layer per die, in-context extraction is quite accurate. Total coupling capacitance only has a 2.9% error compared with holistic extraction, which can actually be regarded as In-C:4 for a technology with four metal layers. 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 with holistic extraction. Note that since in-context extraction still ignores some inter-die coupling, thus it generally extracts less coupling capacitance than holistic extraction. From these results, we conclude that most of inter-die coupling capacitance can be extracted even with one interface layer from each die. If higher accuracy is needed, more interface layers can be included into the in-context extraction to provide detailed consideration of the neighbouring die and metal layers.

VI. POWER AND NOISE ANALYSIS

A. Impact of Inter-die Coupling on 3D Nets

Since inter-die coupling are mostly between top metal layers of both dies, we measure important metrics on 3D nets one by one in Primetime. Except for the clock pin, which is assumed to be an ideal network, all other 329 F2F vias are measured in detail. The results are shown in Figure 13, where each dot represents one 3D net, and the X axis value is the result with holistic extraction. As results show, using die-by-die extraction, number of aggressors is significantly underestimated for each 3D net, because aggressors

TABLE VII
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 VIII
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

from the neighbour die are ignored. However, with our in-context extraction, most aggressors are correctly captured even with one interface layer per die. With more interface layers included, more inter-die aggressors are captured. Similarly, wire capacitance of each 3D net is underestimated with die-by-die extraction as well.

B. Full-chip Power and Noise Results

To find out how large inter-die coupling impacts have on the full-chip metrics, we compare the Primetime analysis results with die-by-die extraction to results with holistic extraction as shown in Table IX. The longest path reported by Primetime is a 3D path which starts from a register in the top die, goes to the bottom die through a F2F via, and ends on another register in the top die. As results show, ignoring inter-die coupling, die-by-die extraction underestimates the longest path delay by 6.2%. This indicates that inter-die coupling cannot be ignored for F2F designs. Also, total wire capacitance on 3D nets is underestimated by 13% with die-by-die extraction. Note that though inter-die coupling capacitance is a large portion of total coupling capacitance, ground capacitance and pin capacitance are major contributors to the load of a net. Therefore, inter-die coupling only affects slightly on the switching power consumption of F2F designs. From our results, ignoring inter-die coupling and the F2F bonding interface layers, die-by-die extraction underestimates 3.5% of total switching power on 3D nets, while we only observe 1.7% underestimation on the switching power.

However, in terms of signal integrity, inter-die coupling plays an important role. Total coupling capacitance reported on 3D nets is underestimated significantly by 32%. Similarly, average number of aggressors for 3D nets is also underestimated by 30%. Because of reduced aggressor count and strength, the maximum noise on 3D nets is underestimated by 26% with die-by-die extraction as well. These results indicate that die-by-die extraction introduces significant errors into full-chip signal integrity analyses, and the inter-die coupling needs to be handled carefully for sign-off verifications.

With our in-context extraction, errors introduced by die-by-die extraction are almost corrected completely. As results show, the timing error is only 1.8% even using our in-context extraction with one interface layer per die, and negligible errors are observed on power metrics as well. For signal integrity analyses, in-context

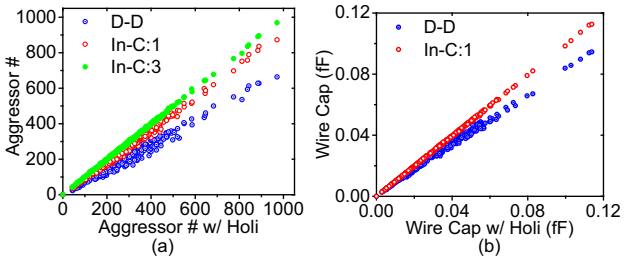

Fig. 13. Full-chip comparison of die-by-die (D-D) and in-context (In-C) against holistic extraction (Holi) on 3D nets, each of which is represented by one dot. (a) aggressor count, (b) wire capacitance.

TABLE IX
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%

extraction is also able to capture most of coupling aggressors. For 3D nets, only 3.7% and 1.1% underestimation is observed on total coupling capacitance and total wire cap, respectively. And the max noise error decreases to 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 smaller.

VII. CONCLUSION

In this paper, we compared three extraction methods in F2F 3D ICs. We proposed an in-context extraction method that was compatible with traditional CAD tools but included interface layers from neighboring dies during extraction. We demonstrated the impacts of E-field sharing and inter-die coupling cannot be ignored in F2F-bonded 3D ICs. Die-by-die extraction underestimates total coupling capacitance, while holistic extraction is able to capture all inter-die coupling at the cost of high complexity. Our in-context extraction is highly accurate, and captures most E-field interactions across dies. It also remains LVS-friendly and can easily be implemented for mixed-process extraction.

REFERENCES

[1] 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 *IEEE Electronic Components and Technology Conf.*, May 2014, pp. 338–341.

[2] D. H. Kim *et al.*, "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," *IEEE Trans. on Computers*, vol. 64, no. 1, pp. 112–125, Jan 2015.

[3] C. S. Tan *et al.*, "Three-Dimensional Wafer Stacking Using Cu-Cu Bonding for Simultaneous Formation of Electrical, Mechanical, and Hermetic Bonds," *IEEE Transactions on Device and Materials Reliability*, vol. 12, no. 2, pp. 194–200, June 2012.

[4] M. Jung *et al.*, "On enhancing power benefits in 3D ICs: Block folding and bonding styles perspective," in *Proc. ACM Design Automation Conf.*, June 2014, pp. 1–6.

[5] L. Peng *et al.*, "Ultrafine Pitch (6 μm) of Recessed and Bonded Cu-Cu Interconnects by Three-Dimensional Wafer Stacking," *IEEE Trans. on Electron Devices*, vol. 33, no. 12, pp. 1747–1749, Dec 2012.