Low Power Monolithic 3D IC Design of Asynchronous AES Core

Neela Lohith Penmetsa[1], Christos Sotiriou[2], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[2]University of Thessaly, Greece
neelalohith@gatech.edu, chsotiriou@inf.uth.gr, limsk@ece.gatech.edu

***Abstract*—In this paper, we demonstrate, for the first time, that a monolithic 3D implementation of an asynchronous AES encryption core can achieve up to 50.3% footprint reduction, 25.7% improvement in power, 34.3% shorter wirelength and 6.06% reduced cell area compared to its 2D counterpart, at identical (ISO) performance. We also demonstrate that combining asynchronous circuits with 3D integration can yield a peak power reduction of 63.9% compared to the equivalent synchronous realisation. We also verified that the asynchronous implementation of the encryption core is more tolerant to monolithic 3D tier-tier variation compared to its synchronous counterpart. To the best of our knowledge, this is the first paper to discuss the mutual benefits of asynchronous and monolithic 3D IC integration.**

I. Introduction

One approach to tackling variability issues in modern VLSI circuits is to exploit asynchronous design techniques. Instead of using a rigid external clock reference calibrated at worst-case conditions, we generate internal clocks based on actual, typical-case conditions. Such circuits automatically tune their internal clocks to optimal timing conditions at any given process and operating conditions. Furthermore, this adaptivity can be exploited even in subthreshold conditions [1], [2], where synchronous operation is very difficult for external control. Asynchronous circuits don't come without drawbacks, and these include a more complex design methodology along with power, performance and area (PPA) overheads due to the clock generation and handshaking circuitry, which must be very carefully managed and minimized.

3D ICs have emerged as one of the most promising solutions for sustaining Moores law. 3D ICs enable high density integration through die-stacking, which reduces power dissipation and increases performance compared to 2D ICs. The most prominent 3D ICs are Through Silicon Via (TSV)-based, but their integration density is limited by the significant area overhead and large pitch of TSVs. Monolithic 3D is an emerging solution that enables much higher integration density than TSV-based 3D, because of the extremely small size of monolithic inter-tier vias (MIV) [3]. Figure 1 compares a typical TSV-based and monolithic 3D structure.

In this paper, we present the design and implementation of both synchronous and asynchronous versions of the AES encryption core using monolithic 3D IC technology. We demonstrate significant PPA savings compared to a traditional 2D IC implementation. To the best of our knowledge, this is the first comprehensive analysis which combines 3D IC design with asynchronous circuits. We show that it is mutually beneficial to combine the domains of asynchronous and 3D integration as their respective strengths and weaknesses complement each other. Asynchronous circuits supplement 3D ICs with better thermal control, power supply integrity and variation tolerance. In return, 3D ICs help manage the PPA overheads of asynchronous circuits. Our study is based on GDSII layouts and industry standard sign-off analysis flows.

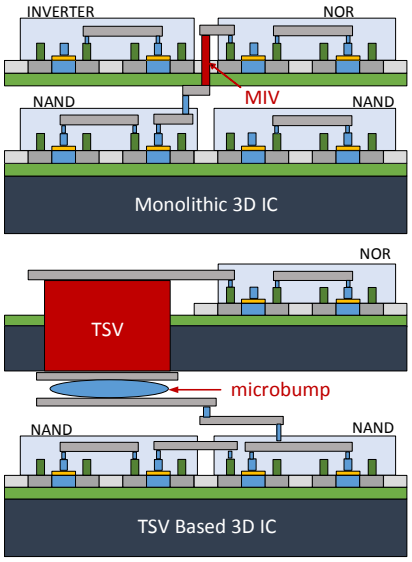

Fig. 1. Monolithic and TSV-based methodologies for 3D integration. Typical TSV diameter is 5um compared to 100nm diameter of an MIV.

II. Design Methodology and Implementation

This section presents the design and implementation of both synchronous and asynchronous versions of the AES encryption core using monolithic 3D IC technology. This experiment is done to study the PPA savings compared to a traditional 2D IC implementation.

A. Benchmark Design

In this work a custom, high performance pipelined Advanced Encryption Standard (AES) RTL is implemented. The ubiquity and the importance of an AES core is the main motivation behind its selection. AES encryption cores are present in thousands of real products, with a diversity of form

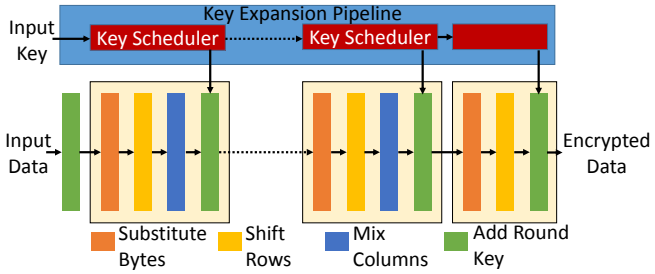

Fig. 2. AES architecture

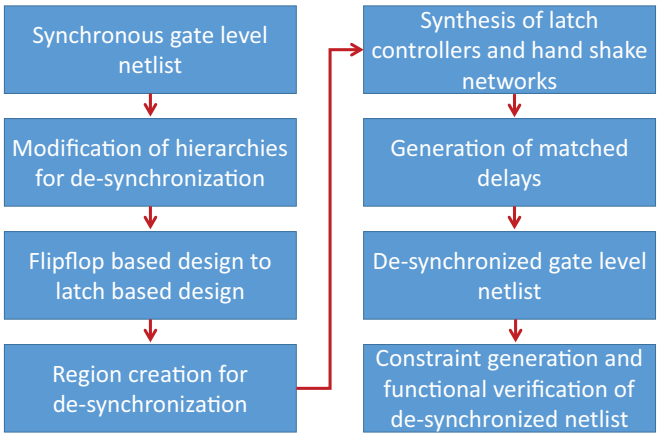

Fig. 3. De-synchronization flow overview.

factor, ranging from ultra-low power sensor networks to high-performance server processors. Typically, depending on the end product's target encryption rate, AES cores are designed for various throughput speeds. Figure 2 shows the top level architecture of the AES encryption standard. It takes a plain input text and an AES key and performs 10 rounds of data transformations on it to generate the encrypted output. The current AES implementation used for this work is optimized for encrypting 128-bit data packets into a 128-bit cipher text, using an AES key of the same size. The design is a deep pipelined architecture, which dumps out encrypted data packets at every clock cycle, with an input to output latency of 41 clock cycles. Standard data packets and their pre-encrypted ciphers are used to functionally validate the design.

B. Logic Synthesis and De-synchronization Flow

As discussed in the previous section, this work uses a de-synchronization methodology [4], which presents a fairly simple framework for converting a synchronous gate-level netlist into an asynchronous equivalent. The high-level flow diagram of the conversion process is shown in Figure 3. First, the AES RTL is synthesized using Design Compiler and conventional synchronous constraints. Next, the post-synthesis netlist is de-synchronized, according to the following steps:

1) Modification of the design's hierarchy to facilitate de-synchronization.
2) Conversion of the synchronous design's Flip-Flops to Latches. Each Flip-Flop is split to its corresponding Master and Slave latches.
3) Automated region creation for de-synchronization. In this step, we assign each standard cell of the netlist to a de-synchronization region which is controlled by its corresponding handshake controller.
4) Synthesis of 2-phase (or 4-phase) latch controller templates for implementing the handshake protocols between the de-synchronized regions. In this work we use only 2-phase latch controllers. C-elements are used to synchronize the hand-shake protocols across regions.
5) Automated Combinational Logic (C.L.) matched delay generation using delay chains. Delay chains are inserted into the netlist, so as to match, the corresponding C.L. cloud path delay (must be greater than the C.L. delay), as shown in Figure 4. Delay chains act as bundled-data completion detection signals. We actually implemented delay chains using higher V_t cells, as this ensures that the delay chain is always slower than the combinational path, even at lower V_{DD}.
6) Constraint generation: Data setup timing check points are extracted from the synchronous netlist and are used to generate the timing constraints for the de-synchronization flow which can aid optimization during the place and route stages.

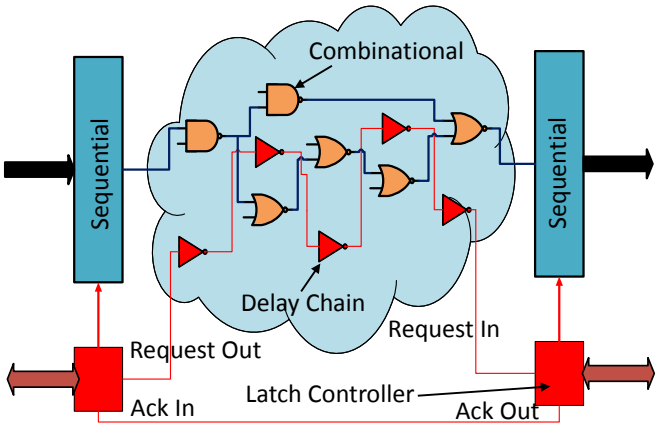

Fig. 4. Synthesis of handshake controller and insertion of matched delays.

C. 2D Physical Design Flow

Current study is based on a 28-nm PDK. We take the design through typical physical design stages like floorplanning, placement, clock tree synthesis, routing and physical verification. The post-routed databases are used to perform parasitic extraction. The GDSII-level design data is then analyzed using industry standard tools like PrimeTime.

For the physical design of de-synchronized designs, it is ensured that the delay chains are placed near the respective combinational logic to track variations as accurately as possible. We also break any timing loops caused by the handshake controllers manually, as the synchronous 2D tool is not capable of recognizing them. In addition, a pseudo clock tree synthesis is performed to distribute the low-skew,

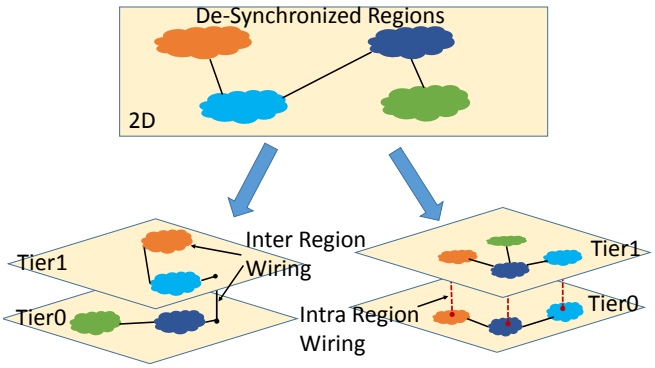

Fig. 5. Partitioning of De-Synchronized regions

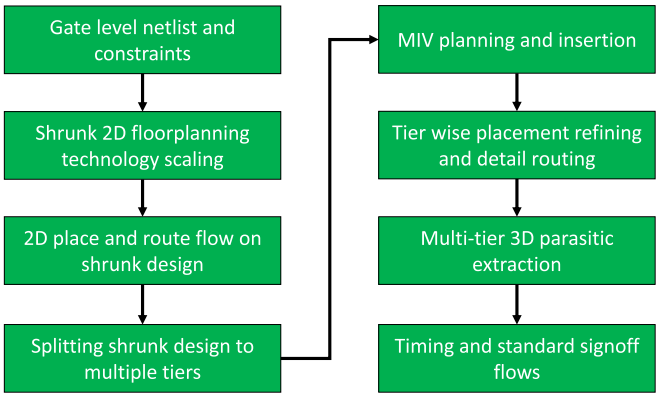

Fig. 6. Monolithic 3D flow

local latch triggering pulses of each handshake controller to its corresponding set of latches. Finally, after the routing stage of de-synchronized design, a recalibration step is done to fine tune the delay chains, to account for any delay mismatch between the latter and their corresponding C.L. clouds, caused by timing perturbations by the placement and routing steps.

D. 3D Integration Choice: TSV vs Monolithic

Initial preference for this work was to use a TSV based integration. In a TSV based integration the netlist is partitioned in to two tiers using a simple min-cut algorithm. The min-cut strategy strives for an area balance between both the tiers while minimizing the cut-size which is equivalent to the TSV count. As shown in the Figure 5 the De-synchronized netlist has several regions where each region only communicates with a limited number of adjacent regions. Partitioning regions on to multiple tiers would lead to half the regions split on to one die while other half on to another die. In this strategy TSVs are used for inter-region wiring. Since each region is still effectively a 2D design in itself, not much benefit is obtained from this strategy. Min-cut experiments with this style of partitioning lead a 2 tier design with 230 TSVs to achieve the required area balance on both the tiers. A second type of folding strategy is shown in the same picture. In this style of partitioning each region is folded on to multiple tiers with TSVs used for intra-region wiring. The advantage of this strategy is each region is now split on to multiple-tiers there by enabling effective optimization of intra-region interconnects. However using TSVs for this folding scheme has its own drawbacks. Since a typical TSV size is about 5um x 7um, their count has to be limited to about 15-20% of the total die area. This would put a limitation on the number of intra-region wires crossing the tiers thus limiting the optimal solution. Secondly the area overhead due to the TSVs as they take up considerable silicon area adds to the burden from de-synchronization. Hence what ever area and performance benefit achieved by 3D integration would be offset by this over head. Finally, TSV parasitics play a significant role in timing as they add a considerable amount of capacitance depending on various factors. This will indirectly impact the timing and performance of the de-synchronized design. Taking all the above factors into consideration a decision was taken to use monolithic integration approach for this work. Monolithic Inter-tier VIAs (MIV) have very small sizes compared to TSVs (in the order of 100nm) and present significantly low parasitics. This allows us to use a large number of MIVs with minimal impact to area or performance. Such an approach would be suitable for intra-region folding. The MIV model used in this work has a capacitance of 0.1fF and resistance of 16Ω.

E. Inter-Tier Variation

Handling variation in 3D IC is extremely important as this might offset the performance benefit arising due to 3D integration. This is one of the main motivating factors to explore asynchronous circuits for 3D integration as they do not have a global clock and are proven to operate reliably when subjected to process variations. In a TSV based 3D IC, both within die and die to die variations contribute to the overall variations [5]. Moreover, variations in RC properties of through-silicon vias (TSVs) also add to total delay variations in 3D ICs. Hence, methodologies are required to reduce the effect of within-chip and chip-to-chip variations in 3D ICs. Monolithic 3D ICs differ from TSV-based 3D ICs in that tiers are fabricated sequentially. The devices and interconnects of the top tier are fabricated on top of an already existing front end-of-line (FEOL) and back end-of-line (BEOL). During the processing of the top tier, care must be taken to prevent damage to the devices and interconnects of the bottom tier. If we wish to use copper on the bottom tier, laserscan anneal has been proposed for the dopant activation on the top tier. This method only results in localized heating, thereby preventing any damage to the devices and interconnects on the bottom tier. However, this process results in considerably degraded transistors, and the PMOS and NMOS performance degrade by 27.8% and 16.2% respectively [6]. We model these degraded transistors in our analysis by assuming assume up to 15% performance degradation on an average in all the devices on the upper tier. Considering these factors we do a functional verification which accounts for these variations to see how the synchronous and its de-synchronized counterparts fare.

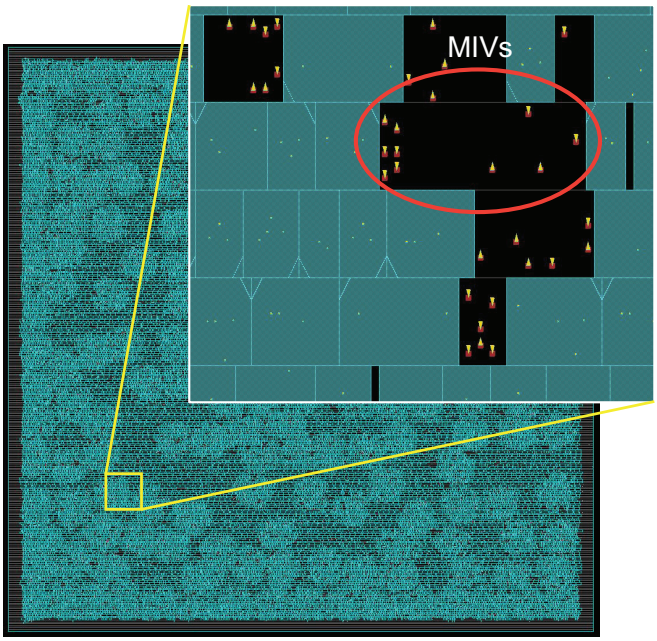

Fig. 7. MIVs are inserted into the whitespace between the standard cells.

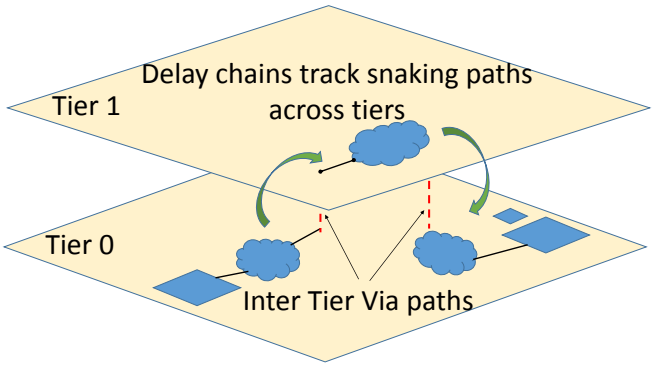

Fig. 8. Snaking timing paths in 3D designs

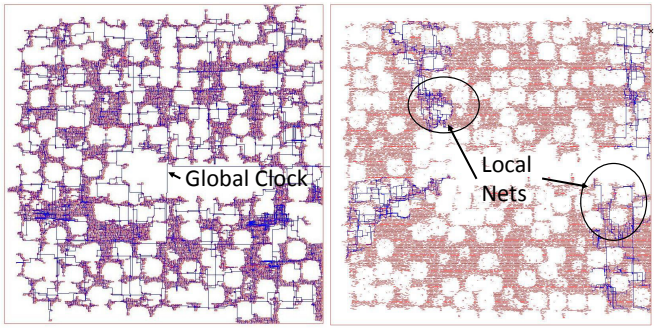

Fig. 10. Localized interconnects in De-Synchronized designs

F. 3D Physical Design Flow

This section presents the description of RTL-GDSII CAD flow for monolithic 3D ICs [7]. A mix of industry standard tools and custom tools are used in this approach. In this work, we focus only on two tier designs. A block diagram of the flow steps is shown in Figure 6. Once we obtain a gate-level synthesized netlist, we make use of an industry standard tool (SoC Encounter) to place all the standard cells on to a shrunk footprint corresponding to that of a monolithic 3D IC. In order to do this, first the chip width and height are shrunk, as well as the width and height of all the standard cells by a factor of 0.707. Then the traditional 2D flow is run as described in the 2D physical design flow section to obtain a shrunk 2D design.

The next step is to split the shrunk 2D design into multiple tiers to obtain a DRC clean design with MIVs inserted into the whitespace between the standard cells (Figure 7). There are various sub-steps involved here. First, all the standard cells are expanded back to their original sizes, which will cause a lot of overlaps in their placement. Next, placement bins are created in a traditional fashion. A partitioner is then used to split the cells from each bin onto top and bottom tiers such that area balance is maintained within each placement bin. Once this step is completed, each tier is routed separately and a tier-level parasitic extraction is done. Then custom tools are used to create a 3D parasitics database by stitching all the individual tiers and MIV parasitics together. In the final stage, this information is used along with 3D netlists to perform timing and functional sign-off flows.

G. Partitioning of Delay Chains

In 3D de-synchronized designs, delay chains must also be partitioned, and further on, in the same way as their corresponding C.L. segments, so as to track delay variations as tightly as possible.

During C.L. folding across multiple tiers, several timing paths snake across tiers, as shown in Figure 8. Hence, the C.L. partitioning must dictate the delay chain partitioning. Each snaking path across tiers is tracked with a corresponding delay segment. Thus, not only do delay chains respond to tier-tier variations, but further on, this feature renders the 3D de-synchronized design more variation tolerant than its synchronous counterpart. Delay chain folding is performed after the netlist partitioning and MIV placement step. A custom script analyzes the number of tier-tier transitions in the combinational paths and corrects the corresponding delay chain points, so as to create similar tier-tier connections.

III. Results and Analysis

A. Functional Verification and Power Simulations

PrimeTime-based timing analysis is performed on all the designs using the extracted parasitics and the post-routed gate-level netlists. From this timing analysis, timing delays are extracted for each cell of the design into a standard delay format (SDF) file. This file is used to back-annotate timing delays in gate-level functional simulations. Both synchronous and de-synchronous designs are functionally verified with real time encryption work loads. Basic system level verification of the de-synchronized design is shown in Figure 9 The advantage of de-synchronized design is its ease of interfacing with other synchronous designs. Input request and output

TABLE I
ISO-PERFORMANCE (0.25NS) COMPARISON FOR VARIOUS IMPLEMENTATION FLAVORS. WL IS WIRELENGTH

Parameter	Sync 2D	Sync 3D (Δ% wrt Sync 2D)	DeSync 2D	DeSync 3D (Δ% wrt DeSync 2D)	Sync 2D vs DeSync 2D (Δ% wrt Sync 2D)	Sync 2D vs DeSync 3D (Δ% wrt Sync 2D)
footprint (mm^2)	0.504	0.25 (-50.3%)	0.504	0.25 (-50.3%)	0.0%	-50.3%
cell area (mm^2)	0.400	0.373 (-6.75%)	0.425	0.399 (-6.11%)	6.25%	-0.25%
buffer count	31757	26440 (-16.7%)	34292	29834 (-13.0%)	7.98%	-6.05%
total Wirelength (m)	3.03	2.09 (-31.0%)	3.06	2.01 (-34.3%)	0.99%	-33.66%
avg Wirelength (um)	20.27	14.582 (-28.1%)	18.20	13.18 (-27.5%)	-10.21%	-34.97%
MIV Count	-	91520	-	83832	-	-

TABLE II
POWER COMPARISON OF 2D AND 3D DESIGNS IN WATTS

Parameter	Sync 2D	Sync 3D (Δ% wrt Sync 2D)	DeSync 2D	DeSync 3D (Δ% wrt DeSync 2D)	Sync 2D vs DeSync 2D (Δ% wrt Sync 2D)	Sync 2D vs DeSync 3D (Δ% wrt Sync 2D)
Switching power (W)	0.1171	0.0824 (-29.6%)	0.1361	0.0981 (-27.9%)	16.2%	-16.2%
Cell power (W)	0.0529	0.0423 (-20.0%)	0.0513	0.0372 (-27.4%)	-3.02%	-29.6%
Leakage power (W)	0.0221	0.0198 (-10.4%)	0.0225	0.0205 (-8.88%)	1.80%	-7.23%
Total Power (W)	0.1921	0.1444 (-24.8%)	0.2098	0.1557 (-25.7%)	9.21%	-18.9%
Peak Power (W)	1.39	1.302 (-6.33%)	0.602	0.47 (-21.9%)	-56.6%	-66.18%

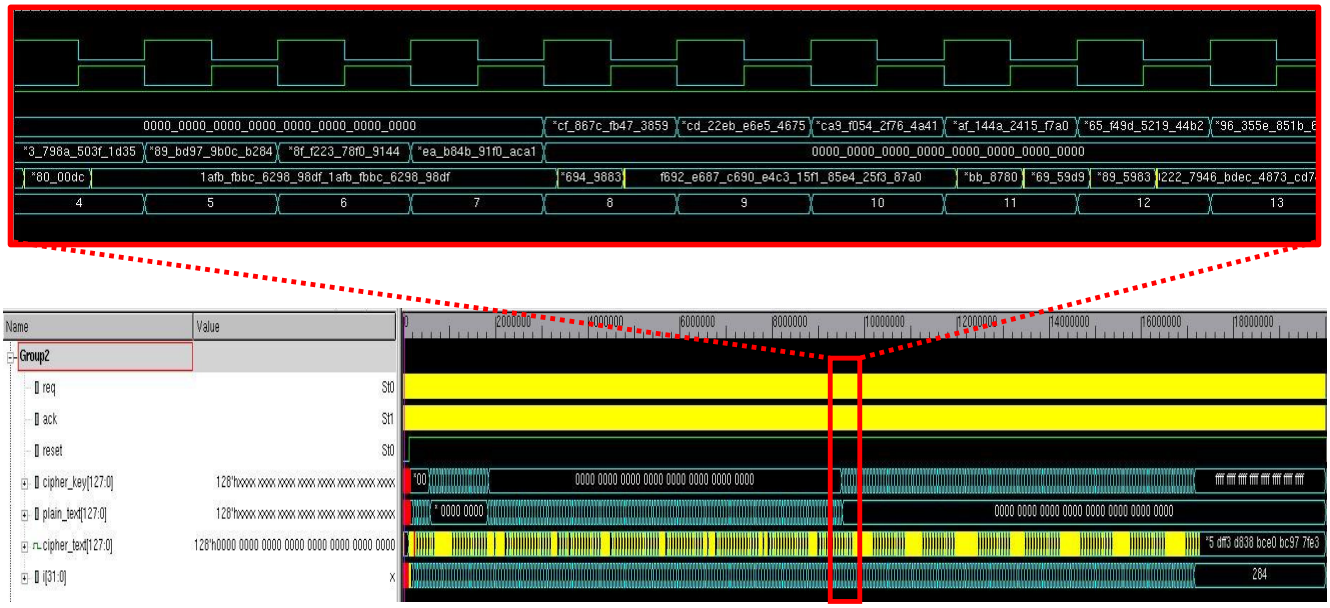

Fig. 9. Functional Verification of the De-Synchrnonized design

acknowledge of the de-synchronized blocks can be driven by an external interface clock while ignoring their corresponding acknowledge and request signals respectively. Several pre-calculated encryption work loads are used to verify correctness of operation and generate a value change dump (VCD) file containing the switching activities of all the gates. We use this file for accurate real time power simulations.

B. Footprint and Wirelength Reduction

Both synchronous and de-synchronized designs are implemented in 2D and monolithic 3D. Various key metrics such as wirelength, footprint area, cell area and buffer count are presented in Table I. This work primarily focuses on ISO-performance comparisons, and hence the critical path delays of all implementations have been optimized to be 0.25ns. This bound is decided because of the speed limitation from the 2D de-synchronous design.

From Table I, we first observe that while the 2D footprint is forced to be the same between synchronous and de-synchronized designs, the cell area in the latter goes up. This is because de-synchronized designs can reach a slightly higher utilization than synchronous counterparts due to the absence of global interconnects. Each de-synchronized region only interacts with its neighboring region which facilitates a tighter packing. However, we observe that de-synchronized design has higher buffer count and total wirelength. This is due to the area and interconnect overhead from various hand-shaking controllers. This matches existing literature, where

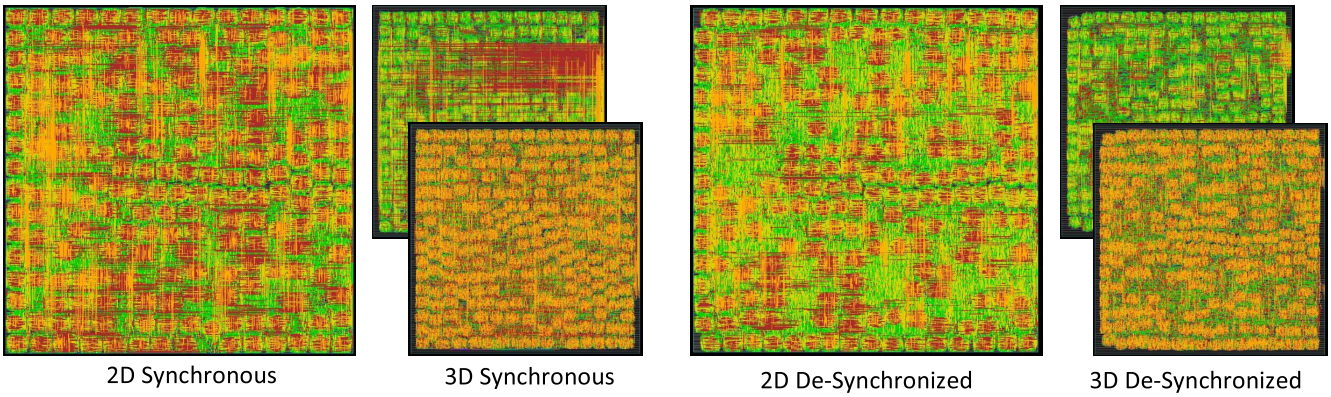

Fig. 11. GDSII Layouts of 2D and 2-tier 3D synchronous and de-synchronized AES designs. 2D footprint is 710x710um, and 3D is 500x500um. We observe that de-synchronous has fewer global interconnects.

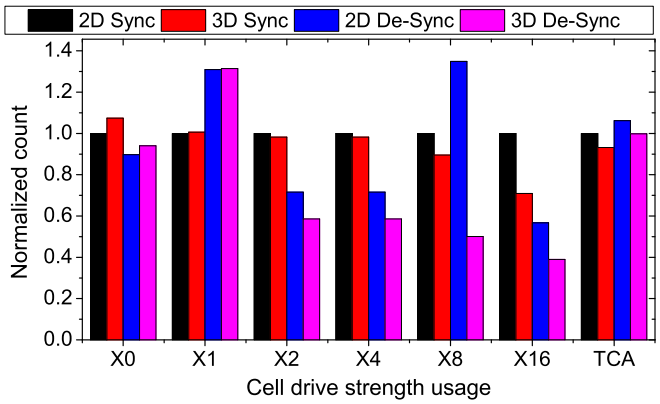

Fig. 12. Comparison of cell usage of various drive strengths normalized to 2D-Sync (X0 being the smallest). TCA is Total Cell Area.

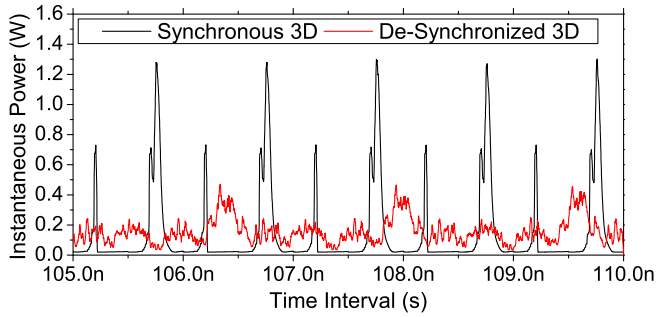

Fig. 13. Transient power analysis of 3D Sync and 3D De-sync

asynchronous designs have an area and wirelength penalty compared to their synchronous counterparts. This is one of the reasons asynchronous designs are are not widely used today. Note that the average wirelength is lower in de-synchronized designs due to the absence of long global connections.

To overcome these limitations in de-synchronized 2D, it is implemented in a monolithic 3D fashion. The footprints and routed die-level screenshots of all implementations are shown in Figure 11. From this figure and Table I, we see that 3D offers a 50.3% footprint reduction over 2D. 3D ICs can operate faster than our target timing constraints, but since we are performing ISO-performance comparisons, we can trade performance for power saving. Optimizing 3D ICs for a frequency less than what they are capable of will lead to significant buffer count and power savings.

As a result of the footprint reduction and close proximity of cells in 3D designs compared to 2D, we see significant reduction of wirelength in 3D designs. From de-synchronized 2D to de-synchronized 3D, we see about 34.3% reduction in total wirelength and 27.5% reduction in average wirelength. This leads to de-synchronized 3D having lower wirelength and using fewer gates overall than the 2D synchronous design. Therefore, monolithic 3D IC technology can overcome all the shortcomings of this asynchronous design style. We discuss how asynchronous operation helps monolithic 3D in the next sections.

C. Power Reduction

The power results obtained from vector based power simulations are presented in Table II. We observe that de-synchronized 2D consumes about 9.2% more power than its synchronous counterpart. This power overhead is due to the handshake controllers and splitting of flip-flops into master-slave latch pairs, and is in line with the results in the previous section. After analyzing final 3D and 2D designs with standard real time test vectors, we observed significant power savings in de-synchronized 3D of up to 25.7% total power reduction compared to de-synchronized 2D and 18.9% percent reduction compared to 2D synchronous.

As mentioned in the last section, 3D can meet the timing target more easily, and hence uses fewer gates overall. This effect is quantified in Figure 12, where we plot the cell usage in each design grouped by size. We observe both fewer cells overall, as well as fewer larger cells. This also leads to a reduction in the total cell area as shown in this figure.

So far, we have discussed the benefits monolithic 3D brings to asynchronous. However, asynchronous operation also mitigates many potential issues in monolithic 3D ICs. Although there is a slight increase in average power from 3D

synchronous to 3D de-synchronous, we see a huge reduction of 63.9% in terms of peak power (Table II). Peak current is a primary concern in the design of power distribution networks especially for 3D ICs. Such peaks determine the maximum voltage drop and probability of failure due to electro-migration. This may lead to performance gaurdbands in 3D ICs, which asynchronous operation helps gets rid off. Since 3D ICs have double the thermal density of 2D designs, it is critical to reduce thermal fluctuations. These fluctuations make the heat removal process more difficult and may penalize design metrics. We have characterized the power spectrum of 3D synchronous and 3D de-synchronous designs based on standard real time encryption workloads. As shown in Figure 13, 3D de-synchronous has the best power profile with almost negligible fluctuations compared to its synchronous counterpart.

D. Performance Benefit

All the previous results have assumed that asynchronous and synchronous have an identical worst case stage delay of 0.25ns. Our AES core has 41 such stages as it is pipelined for maximum throughput. In a synchronous system, the operating frequency is limited by slowest stage which naturally slows down the faster stages. However, in the de-synchronized design, since every stage is locally timed, the latency of the circuit is equal to the sum of delays in each pipeline stage. When a single packet of data is sent for encryption, we observe that the synchronous design has a total input to output latency of 10.25ns. In contrast, the de-synchronized design has a total latency of 6.33ns, which is a significant improvement.

We have also designed for the best performance that each implementation flavor can achieve. 2D synchronous can achieve a critical path delay of 0.24ns while 3D synchronous is 20% faster with a critical path of 0.20ns. Similarly, 2D de-synchronous can achieve a critical path delay of 0.25ns while 3D de-synchronous is 16% faster with a critical path of 0.21ns. We still observe that 3D de-synchronous can operate 12.5% faster than 2D synchronous.

E. Variation aware functional analysis

As explained earlier, we model performance degradation of up to 15% for each cell on the top tier. All the designs are done with typical corner libraries and have a timing guardband of 10ps on all the required margins. The variations are introduced as timing derates in primetime analysis and the SDF files used for functional simulations are altered accordingly. We noticed that synchronous designs face timing violations in the presence of variations and lead to functional errors during verification when operated at the target frequency. Hence for correct functional operation a frequency hit is necessary. Some alternative methods have been proposed [8] where variation aware floor-planning and placement has been proposed to deal with this problem. However de-synchronized 3D AES version is more tolerant to this effect. We noticed correct functional operation even with 15% performance degradation in the upper tier. As the delay chains span across tiers, they get equally impacted by performance degradation and thus replicating the variation in the combinational path delays they are tracking.

IV. Conclusions

In this paper, for the first time, we studied the synergistic benefits of 3D IC and asynchronous circuits. We demonstrated that the power, performance and area overhead in asynchronous designs can be reduced significantly by using monolithic 3D IC integration. At the same time, asynchronous circuits can help monolithic 3D IC designs with better variation tolerance, power supply integrity and thermal characteristics. By switching to monolithic 3D, we obtain significant footprint reduction of the AES core, which facilitates encryption capabilities into products of various form factors. At the same, time de-synchronization gives the 3D IC-based AES design modular capabilities and mitigates some of its negative effects. As a future work we plan to do a full-scale variation analysis of 3D ICs with asynchronous circuits and also compare the 3D integration benefits of different asynchronous schemes.

References

[1] O. C. Akgun, J. Rodrigues, and J. Sparso, "Minimum-Energy Subthreshold Self-Timed Circuits: Design Methodology and a Case Study," in *Proceedings of the International Symposium on Asynchronous Circuits and Systems*, 2010.

[2] M. Lotse, M. Ortmanns, and Y. Manoli, "A Study on self-timed asynchronous subthreshold logic," in *Proc. IEEE Int. Conf. on Computer Design*, 2007.

[3] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2014.

[4] J. Cortadella, A. Kondratyev, L. Lavagno, and C. P. Sotiriou, "De-Synchronisation: Synthesis of Asynchronous Circuits from Synchronous Specifications," in *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2006.

[5] C. O. F. Akopyan, D. Fang, S. J. Jackson, and R. Manohar, "Variability in 3-D integrated circuits," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2008.

[6] B. Rajendran, R. S. Shenoy, D. J. Witte, N. S. Chokshi, R. L. DeLeon, G. S. Tompa, and R. F. W. Pease, "Low Thermal Budget Processing for Sequential 3-D IC Fabrication," in *IEEE Trans. on Electron Devices*, 2007.

[7] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," in *Proc. Int. Symp. on Physical Design*, 2014.

[8] S.Panth, K. Samadi, Y. Du, and S. K. Lim, "Power-Performance Study of Block-Level Monolithic 3D-ICs Considering Inter-Tier Performance Variations," in *Proc. ACM Design Automation Conf.*, 2014.