# Next-Generation Design and Technology Co-optimization (DTCO) of System on Integrated Chip (SoIC) for Mobile and HPC Applications

Y.-K. Cheng, F. Lee, M.-F. Chen, J. Yuan, T.-C. Huang, K.-J. Chen, C.-T. Wang, C.-L. Chen, C.-H. Tsai, and Douglas Yu
R&D, Taiwan Semiconductor Manufacturing Company, Ltd., HsinChu, Taiwan, email: yk_cheng@tsmc.com

2020 IEEE International Electron Devices Meeting (IEDM) | 978-1-7281-8888-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEDM13553.2020.9372005

***Abstract*****—**This paper demonstrates the next-generation design and technology co-optimization (DTCO) of system on integrated chip (SoIC) for mobile and HPC applications, where the SoIC technology was proposed to integrate multi-chips with different functionality and technology into a single SoC chip. The new DTCO includes overall die partitioning, die integration, and interconnect. These methodologies can be used for improving time-to-market and trade-off between performance and cost. In this paper, two prototypes of stacking CPU and memory dies are demonstrated with 15% performance gain and 30% average point-to-point distance reduction.

## I. Introduction

Future high-performance computing (HPC) systems require very high cache capacity and large memory bandwidth to reduce latency and increase bitrate. Improving cache performance essentially helps to reduce the miss rate, miss penalty, and hit time. However, existing cache memory with low density obstructs the increase of cache capacity, the limited block size, and the pin terminal count, called memory bandwidth wall. To address these challenges, 2.5D/3D die-stacking technologies using through-silicon via (TSV) are becoming a promising candidate for meeting the demand of future HPC systems. They can increase the on-die cache memory capacity to several gigabytes along with orders of magnitude higher bandwidth. System-on-Integrated-Chip (SoIC) technology is one of the best solutions for logic and large-capacitor memory integration.

## II. SoIC Technology

The SoIC enables the system scaling with better merits of cost and performances by implementing SoC partition and re-integrating dies. SoIC technology features Face-to-Face (F2F), Face-to-Back (F2B) [1], sub-μm bond pitch, 3D memory integration, and low temperature (LT) bonding and stacking [2]. The F2F bonding introduces very short interconnect to allow close proximity between chip-to-chip and chip-to-package communication by TSV, through-dielectric-via (TDV), and back-side redistribution layer (BSRDL) as in Fig. 1 (a). The F2B bonding provides the flexibility of die stacking, which could benefit the system with optimized electrical, mechanical, and thermal performance as in Fig. 1 (b). The sub-μm bond pitch enables cell/circuit level interconnection to achieve intra-chip interaction among dies. The 3D memory integration applies multiple die stacking to realize high capacity and bandwidth cache under the optimization of system performance and die yield as in Fig. 1 (c). The low temperature bonding and stacking technology facilitates the system integration with thermal-sensitive dies, e.g., DRAM, as shown in Fig. 2. Finally, SoIC can be integrated into an advanced wafer level system integration (WLSI) platform to achieve greater functionality and enhanced system performance at increasingly competitive cost.

## III. Design and Technology Co-optimization

The SoIC technology affects the traditional SoC chip design. A three-dimension multi-die integration raises design complexity such as system architecture, floorplan, sign-off conditions for timing, IR drop, and verification. This DTCO section addresses these design challenges from 2D to 3D chip designs. Several methods focused on die partitioning, die integration, and I/O interconnect designs will be shown in the remaining of the section.

### *A. Die partitioning*

In SoIC technology, the SoIC bond can be used to implement new system architectures due to a small form factor (sub-10 μm thickness) and a small bump pitch (sub-10 μm pitch). The above benefit is useful for logic-to-memory and logic-to-logic integration. As an example of CPU and cache, the latency and capacity of cache usually dominates CPU throughput and cache miss. To increase the throughput, a good solution is to enlarge the cache capacity, which also increases the latency. Based on SoIC technology, however, many small-size SRAM dies can be integrated into a single SRAM chip with large capacity and low latency. Furthermore, the SRAM or uncritical functions (e.g., I/O, mix-signal, and analog circuits) can be partitioned and implemented using a sub-node technology for cost saving. After die partitioning, long horizontal paths are changed to short vertical ones, so that the performance, power and area can have significant improvement.

### *B. Die integration*

There are three kinds of die integration schemes, i.e., F2F, F2B, and Side-by-Side, as shown in Fig. 1. For F2F stacking, the SoIC bond interconnects can reduce the latency between two chips. Meanwhile, the TSV and TDV can provide good I/O and power delivery network (PDN) paths. In the SoIC structure, the TSV is connected to higher metal layers rather than lower ones, because lower metal layers have a large RC delay. It can directly provide PDN paths to top metals for both dies. In addition, the TSV can achieve low IR drop compared to current via pillars from $M_1$ to $M_{top}$ in a traditional PDN design. However, the stacking-die PDN can still incur extra voltage drop through TSV. To achieve better IR results, a TSV array design can reduce the worst static IR drop from 2.3% to 1.2% at FFG/125°C corners, as shown in Fig. 3.

For F2B stacking, both dies are integrated with the face down, where the bottom die takes the benefit of direct bumping in terms of short-path PDN. So, a high power die is better to be the bottom one in order to eliminate IR drop impact caused by TSV. However, interconnects of bottom die are also blocked by TSV insertion. As a result, the interconnect density is limited. To increase interconnect density, partial I/Os and power domains can be shared for both dies, as shown in Fig. 4. For a Side-by-Side integration, the BSRDL can provide a high-density horizontal interconnection. That is useful to integrate a CPU die with a high-bandwidth memory (HBM) or wide-IO memory. A system architecture that uses HBM or WIO as the L3 or LLC cache can increase CPU throughput with a small form factor.

### C. *I/O Interconnect*

In order to achieve better signal communication between different dies, a Lite-IO is proposed to meet the requirement of high speed, small cell area, and 10V CDM protection, as shown in Fig. 5. For implementation, the Lite-IO also can be compatible with the small-pitch TSV, TDV, and SoIC bond for the vertical and horizontal integration. Notably, the Lite-IO achieves lower power consumption and lower cell area compared to a conventional GPIO design.

## IV. SoIC Implementation

### A. *SoIC-PTV1*

To simulate die partition for high bandwidth and short latency applications, SoIC-PTV1 is developed by cutting L2 cache out of non-CPU portion of a quad-core ARM A72 design to establish a F2F 3D stacking, as shown in Fig. 6. The SRAMSS and CPUSS dies are implemented using 5nm FinFET technology, and the PKG die is solely implemented by the BSRDL. Figure 7 shows the design partition diagram of CA72 CPU. There are 32 SRAM macros (2MB) used for L2 data cache in SRAMSS die. The interconnect speed is 1GHz when CA72 $F_{max}$ is operating at 2 GHz and 0.75V.

To take advantage of the reduced RC load of SoIC bonds, simple Lite-IO scheme is employed for the vertical interconnection between CPUSS and SRAMSS. The Lite-IO consists of a regular buffer and ESD protection devices. Its electrical characteristics are shown in Fig. 5. Compared to regular buffers, ESD devices in Lite-IO introduce additional capacitance which increases delay and power consumption. In spite of that, Lite-IO and SoIC bonding still achieves the best power and area efficiency, when compared to other I/O and package interconnect schemes [3]-[6] owing to its simplified circuitry and short connection, as shown in Table I.

For IR drop prediction, IR drop is TSV current multiplied by TSV resistance, where TSV current is calculated by total power, VDD, and TSV count. Table II shows that TSV IR drop is 0.85% of VDD at worst $R_{TSV}$. Compared to the simulation results from a commercial IR drop simulation tool, the worst TSV IR is 0.992%. The predicted results correlate well with simulation and can be used at the floorplan stage to optimize TSV placement to mitigate IR drop impact.

Figure 8 shows the steady state thermal results for the three dies stacking at FFGNP/0.825V corner. The most power consuming die, CPUSS, has 88.4°C of junction temperature, along with 87.9°C and 87.3°C for SRAMSS and PKG dies, respectively. Based on this thermal model, we can experiment on several density parameters to verify different kinds of stacking system scenarios. For example, with SoIC bond density reducing from 10% to 1%, the temperature impact is only 0.3°C.

### B. *SoIC-PTV2*

To demonstrate SoIC for HPC applications, a SoIC-PTV2 is implemented to simulate heterogeneous F2B stacking, as shown in Fig. 9. One 7nm SRAM chip with 4 x 8MB stacks is designed to support multiple SRAM dies stacking for large SRAM capacity requirement. The logic die is 5nm CPU chip with 4 quad-cores A72 and two L3 caches with 4MB to build up logic-to-logic 3D connections for logic dies stacking. The new system can have 15% performance gain, with the assumption that the L2 hit rate is 60%, average L3 access time is 5 times of L2, and the latency from cluster to L3 is 3 times of L2. Figures 10 and 11 show the CPU die logic diagram and SRAM die floorplan. They can be used for F2F or F2B integration with this symmetric floorplan. The average point-to-point distance can be reduced by 30% with 50% topology area shrink from 2D side-by-side to 3D stacking.

## V. Conclusion

The DTCO methodology for SoIC technology is established with system architecture, floorplan, TSV/TDV assignment, and timing/IR/thermal verifications. The proposed Lite-IO achieves low power and small cell area. Critical components are verified by the test vehicles of SoIC-PTV1 and SoIC-PTV2. Finally, the CPU and Memory integration using SoIC technology is demonstrated to achieve 15% performance gain and 30% average point-to-point distance reduction with 50% topology area shrink from original 2D integration to 3D integration.

## References

[1] M.-F. Chen *et al.*, “System on Integrated Chips SoIC$^{TM}$ for 3D Heterogeneous Integration,” in *Proc. IEEE ECTC*, May 2019, pp. 594-599.

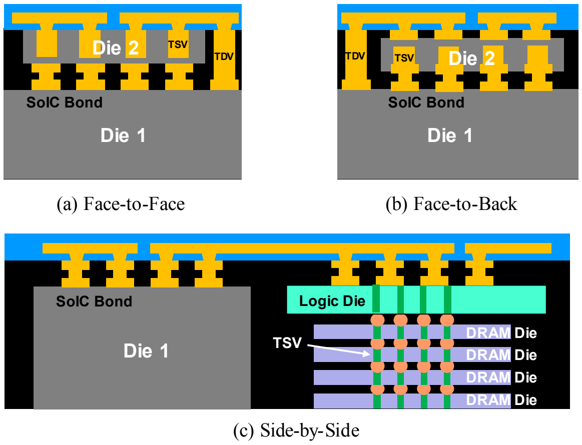
[2] C.-H. Tsai *et al.*, “Low Temperature SoIC$^{TM}$ Bonding and Stacking Technology for 12/16-Hi High Bandwidth Memory (HBM),” in *Proc. IEEE Symposium on VLSI Technology*, Jun. 2020.

[3] P. Vivet *et al.*, “A 220GOPS 96-Core Processor with 6 Chiplets 3D-Stacked on an Active Interposer Offering 0.6ns/mm Latency, 3Tb/s/mm$^2$ Inter-Chiplet Interconnects and 156mW/mm2@ 82%-Peak-Efficiency DC-DC Converters,” in *Proc. IEEE ISSCC*, Feb. 2020, pp. 46-48.

[4] M.-S. Lin *et al.*, “A 7nm 4GHz Arm®-Core-Based CoWoS® Chiplet Design for High Performance Computing,” in *Proc. IEEE Symposium on VLSI Circuits*, Jun. 2019, pp. C28-C29.

[5] N. Beck *et al.*, “Zeppelin: An SoC for Multichip Architectures,” in *Proc. IEEE ISSCC*, Feb. 2018, pp. 40-42.

[6] David Greenhill *et al.*, “A 14nm 1GHz FPGA with 2.5D Transceiver Integration,” in *Proc. IEEE ISSCC*, Feb. 2017, pp. 54-55.

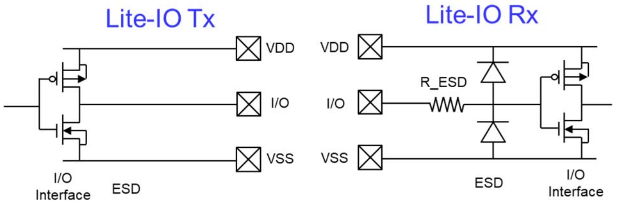


(a) Face-to-Face

(b) Face-to-Back

(c) Side-by-Side

Fig. 1. SoIC structures (a) Face-to-Face (b) Face-to-Back (c) Side-by-Side.



| Hspice simulation | D8/D4 | Lite-IO D8/D4 |
|---|---|---|
| Propagation Delay | 1x | 1.57x |
| Avg. power per channel from D8 to D4 | 1x | 2.18x |
| ESD cap | - | 1 fF |
| CDM | - | 10 V |
| Cell area | 0.19278 μm$^2$ | 1.071 μm$^2$ |
| Energy | 0.000799 pJ/bit | 0.001745 pJ/bit |

Fig. 5. Lite-IO schematic and performance comparison.

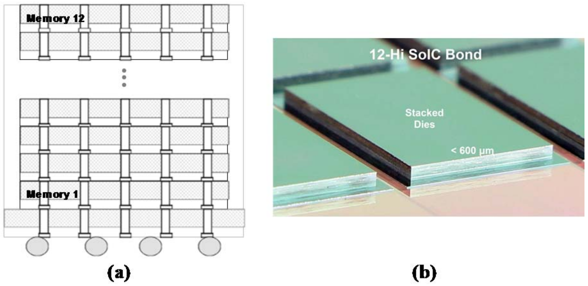


Fig. 2 (a) 3D memory structure (b) Sample photo of LT bonding and stacking.

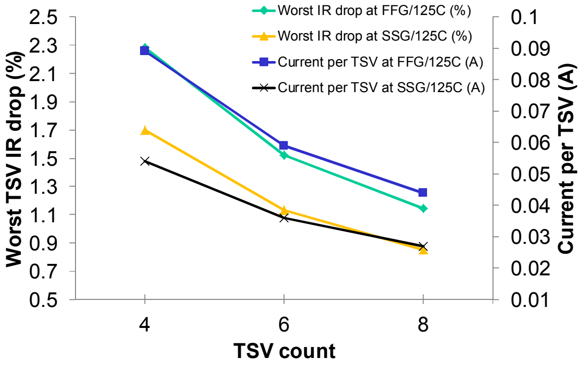


Fig. 3. The worst IR drop result for different TSV counts and corners.

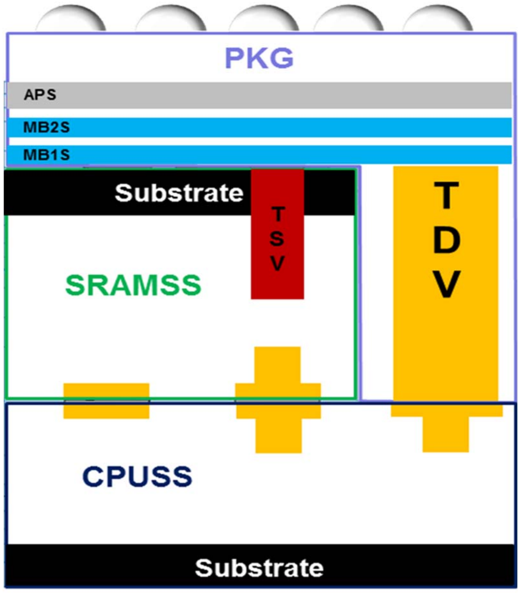


Fig. 6. SoIC-PTV1 structure (Face-to-Face).

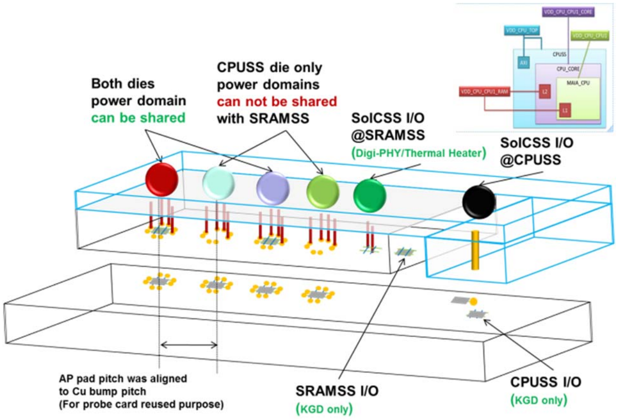

Both dies power domain can be shared
CPUSS die only power domains can not be shared with SRAMSS
SoICSS I/O @SRAMSS (Digi-PHY/Thermal Heater)
SoICSS I/O @CPUSS
AP pad pitch was aligned to Cu bump pitch (For probe card reused purpose)
SRAMSS I/O (KGD only)
CPUSS I/O (KGD only)

Fig. 4. IO/bumping/testing/power domains consideration.

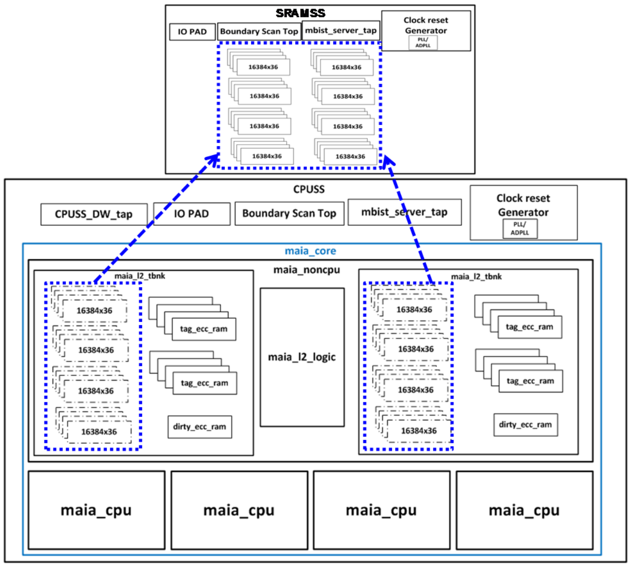


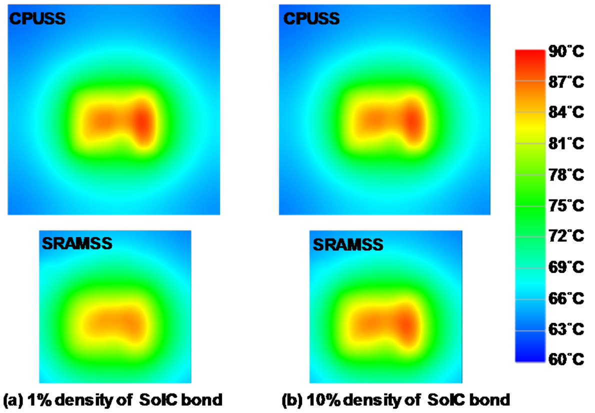
Fig. 7. CA72 design partition diagram, where there are more than 2500 interconnections between L2 and non-CPU portion.

|  | This work | ISSCC'20 [3] | VLSI'19 [4] | ISSCC'18 [5] | ISSCC'17 [6] |
|---|---|---|---|---|---|
| Technology | 5nm FinFET | 28 nm FDSOI | 7nm FinFET | 14nm FinFET | 14nm FinFET |
| Vsw (V) | 0.5 | 1.2 | 0.9 | low-swing | - |
| Bus Width | 2586 | 156 | 320 | 256 | - |
| Channel | SoIC<br>< 10 µm | Active interposer<br>50 µm | CoWoS<br>500 µm | MCM | EMIB<br>1000 µm |
| Die-to-Die Bump Pitch (µm) | < 10 | 20 | 40 | >100 | 55 |
| Data Rate (Gb/s/pin) | 2 | 1.21 | 8 | 5.3 | 2 |
| IO Power Eff. (pJ/bit) | 0.001745 | - | 0.073 | - | - |
| PHY Power Eff. (pJ/bit) | 0.001745 | 0.59 | 0.56 | 2 | 1.2 |
| Bandwidth Density (Tb/s/mm$^2$) | 2.45 | 3 | 1.6 | - | 1.5 |

Table I. The comparison table of Lite-IO and prior works.

| Method | Corner | Power (W) | Bump count | TSV count | SoIC bond count | $IR_{TSV}$ (%) | $IR_{SoIC\ bond}$ (%) | $IR_{via\ of\ SoIC\ bond}$ (%) |
|---|---|---|---|---|---|---|---|---|
| Prediction | Best R | 4.090 | 28 | 224 | 224 | 0.283 | 0.018 | 0.121 |
|  | Typical R | 4.090 | 28 | 224 | 224 | 0.567 | 0.021 | 0.243 |
|  | Worst R | 4.090 | 28 | 224 | 224 | 0.850 | 0.026 | 0.364 |
| Simulation by commercial tool | Worst R | 4.090 | 28 | 224 | 224 | 0.992 | 0.002 | 0.443 |

Table II. The IR drop prediction and simulation results of TSV and SoIC bond at SSGNP/0.675V/125°C corner, where the total power is 4.09 W at SSGNP/0.675V/125°C corner, the ratio of bump and TSV is 1:8, and the $F_{max}$ of CPU is 2 GHz.



(a) 1% density of SoIC bond (b) 10% density of SoIC bond

Fig. 8. SoIC-PTV1 thermal simulation results (a) 1% density of SoIC bond (b) 10% density of SoIC bond, where the total power of CPUSS and SRAMSS is 10.2946 W and 0.3608 W. Simulation assumptions include heatsink with 300 W/m$^2$-C heat transfer coefficient, dummy package/PCB with 70% metal density, and ambient temperature condition at 25°C.

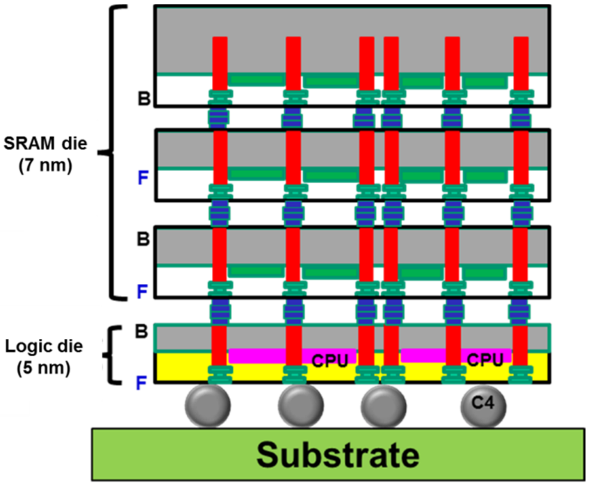


Fig. 9. SoIC-PTV2 stacking view

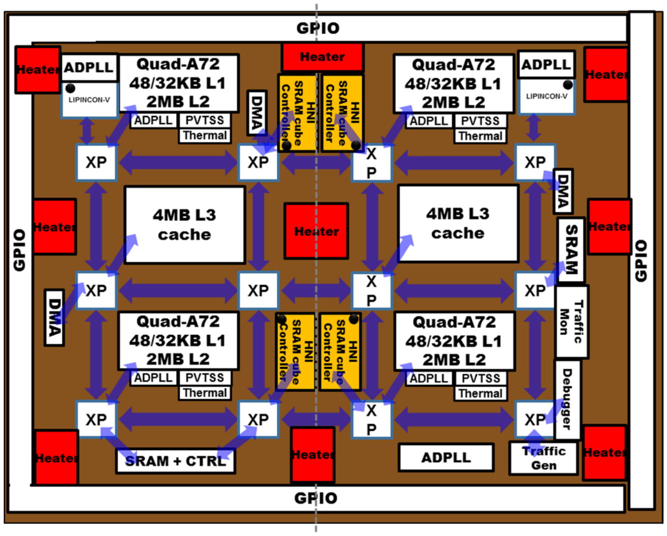


Fig. 10. CPU die logic diagram.

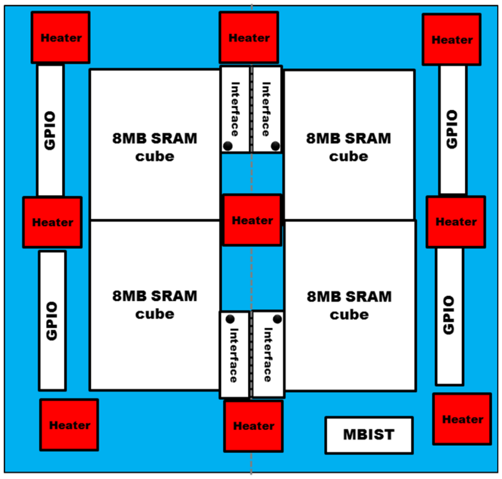


Fig. 11. SRAM die floorplan view