2025 Design, Automation & Test in Europe Conference (DATE 2025)

A 3D Design Methodology for Integrated Wearable SoCs: Enabling Energy Efficiency and Enhanced Performance at Iso-Area Footprint

H. Ekin Sumbul, Arne Symons, Lita Yang, Huichu Liu, Tony F. Wu, Matheus Trevisan Moreira,
Debabrata Mohapatra, Abhinav Agarwal, Kaushik Ravindran, Chris Thompson, Yuecheng Li, Edith Beigne
Reality Labs, Meta, Menlo Park, CA, USA
Contact Author: *ekinsumbul@meta.com*

***Abstract*—Augmented Reality (AR) System-on-Chips (SoCs) have strict power budgets and form-factor limitations for wearable, all-day use AR glasses running high-performance applications. Limited compute and memory resources that can fit within the strict industrial design area footprint of an AR SoC, however, create performance bottlenecks for demanding workloads such as Pixel Codec Avatars (PiCA) group-calling which connects multiple users with their photorealistic representations. To alleviate this unique wearables challenge, 3D integration with hybrid-bonding technology offers energy-efficient 3D stacking of more silicon resources within the same SoC footprint. Implementing such 3D architectures, however, is another challenge as current EDA tools and flows offer limited 3D design control. In this work, we present a 3D design methodology for robust 3D clock network and datapath design using current EDA tools. To validate the proposed methodology, we implemented a 3D integrated prototype AR SoC housing a 3D-stacked Machine Learning (ML) accelerator utilizing TSMC SoIC™bonding technology. Silicon measurements demonstrate that the 3D ML accelerator enables running PiCA AR group call at 30 frames-per-second (fps) by 3D-expanding its memory resources by 4× to achieve 2× better energy-efficiency when compared to a 2D baseline accelerator at iso-footprint.**

***Index Terms*—IC Design Methodology, 3D-stacked SoC, 3D-stacked ML Accelerator, Augmented/Virtual Reality**

I. Introduction

Augmented Reality (AR) System-on-Chips (SoC) have stringent power, performance, and area challenges, whereby the specialized SoC is required to fit within the industrial design (ID) form-factor of wearable AR glasses while running high-performance applications on a tight power budget for an all-day battery use. As the area occupied by circuits becomes high demand within the AR SoC, enabling complex AR workloads with the limited amount of compute and memory resources that can fit in a given footprint becomes challenging.

One such complex AR workload is Pixel Codec Avatars (PiCA) [1], which connects people on AR devices with their photorealistic avatars. PiCAs are rendered on the AR displays by running Machine Learning (ML) algorithms and graphics pipelines at target frame-rates. When multiple users connect in a group call, running multiple ML inferences simultaneously at 30 frames-per-second (fps) becomes a challenging task for the AR SoC with its limited area and power budgets.

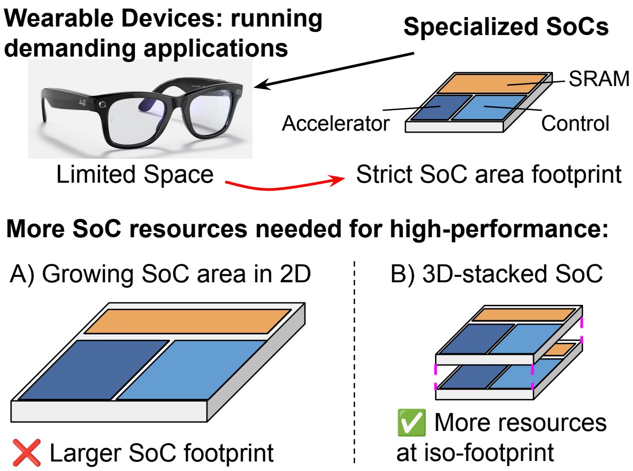

Wearable Devices: running demanding applications
Specialized SoCs
SRAM
Accelerator
Control
Limited Space
Strict SoC area footprint
More SoC resources needed for high-performance:
A) Growing SoC area in 2D
B) 3D-stacked SoC
❌ Larger SoC footprint
✅ More resources at iso-footprint

Fig. 1. 3D-integration of compute and memory resources within the same area footprint for form-factor limited wearable SoCs.

An enabling technology for this unique AR SoC challenge is 3D integration to add more hardware resources in the Z-direction within the same area footprint and ID form-factors (Fig.1). Compared to Through-Silicon-Via (TSV) based 3D integration, hybrid-bonding based stacking offers energy-efficient 3D interconnection at higher density and bandwidth [2] [3], with recent demonstrations of sub-micron bonding pitches [4] [5]. Hybrid bonding further offers non-restrictive 3D signal routing without keepout regions, which allows new opportunities to design 3D architectures as implementing 2D and 3D signal wires becomes a similar circuit routing problem.

Using state-of-the-art EDA tools and flows to implement such 3D stacked architectures, however, creates a design challenge as these tools currently offer a limited 3D control for 3D routing at the intra-circuit-block levels while multi-die, multi-corner On-Chip Variations (OCV) compound the timing closure problem. Although current 3D IC design tools offer in-package 2.5D/3D solutions for multi-die or chiplet co-design [6] [7], finer grained 3D routing within the micro-architecture or the circuit block remains a design challenge. To 3D-extend various subsystems within the SoC, a scalable 3D circuit design approach is needed.

Several recent studies present 3D architectures demonstrating energy efficiency improvements, however there are still limited details on a 3D design methodology that can scale up to SoC-level in a robust fashion. [8] proposed a closely coupled 3D compute and memory architecture for DNNs while only providing a very high level design overview. [9] demonstrated a 3D DNN accelerator for edge devices, exploiting spatial and

978-3-9826741-0-0/DATE25/© 2025 EDAA

temporal localities, but only shared simulation-based results. Our previous work [10] showcased an AR SoC design with silicon results demonstrating energy-efficiency gains for AR workloads, however did not expand on 3D design considerations. Lastly, [11] discussed a 3D design approach for face-to-face (F2F) stacked heterogeneous 3D ICs, but only focused on physical design aspects of cell placement and metal stack choices targeting a relatively small-scale SoC subsystem.

In this work, we present a robust 3D-aware design methodology that enables using existing EDA tools and flows to implement 3D stacked architectures. As a demonstration vehicle, a prototype 3D integrated AR SoC is implemented by utilizing TSMC SoIC™Wafer-on-Wafer bonding technology to 3D-stack two 7nm dies face-to-face (F2F) at $<2\mu$m bond pitch. Using our methodology, we successfully 3D-extended two main SoC subsystems to implement a 3D-stacked ML accelerator and a 3D stacked global SRAM memory.

As its main contribution, this work shares 3D design considerations for implementing 3D architectures targeting wearable SoCs utilizing hybrid-bonding by providing details on 3D cross-die circuit boundary analysis, clock network, datapath design, and timing closure, power delivery, testing, and thermal considerations. To validate the efficacy of the presented 3D design methodology on silicon, we further analyze and deploy a representative model of photorealistic face avatars 5-user group-call application on the 3D stacked test-chip. Silicon results demonstrate that the novel way of architecting the ML accelerator in 3D at negligible 3D access cost allows for overcoming memory limitations at iso-footprint to run inference on the representative 5-user AR group call model at 30fps, which is not feasible to meet with the 2D baseline ML accelerator with its limited resources at the same footprint. The 3D-stacked ML accelerator with a 3D expanded 4MB SRAM achieves this at $\sim 2\times$ better energy-efficiency for PiCA inference when compared to the 2D baseline configuration with a 1MB SRAM.

II. Workload Analysis & Architecture Study

Photorealistic avatar telepresence has two main building blocks [1]: the Encoder which captures and encodes the user's facial expressions to a latent code [12], and the Decoder which decodes the latent code to render the user's face. The Decoder is based on Deep Neural Networks (DNN) to compute texture and geometry of the face and graphics pipelines for rendering [1], with its DNN trained for a unique user for the best performance. For a face group call of five users (i.e. 1-to-4 call) each pair of AR glasses decode four other users' face models for every frame at 30fps. Since the decoder DNN is unique to each user, however, running inference can not be batched over multiple users, which creates a performance bottleneck for the future group calling workloads.

A. Workload Profiling on a 2D Baseline Design

To analyze this bottleneck, we use a representative PiCA Decoder DNN model with its architecture details taken from [1], with a reported model size of 5.5M parameters. We define a target SoC configuration as a realistic baseline to estimate the performance of running the representative DNN. To be consistent with existing works targeting AR SoCs, the baseline 2D system is designed with an ML accelerator of 1K MACs [12] with 1MB local buffer and a control core (Fig.2(a)). To allocate the same scratchpad capacity of 5.5 MBs to each of two possible additional subsystems in a real SoC, e.g. a CPU and another accelerator, the baseline includes a shared system SRAM of 16MB (consistent with [13]) built by tiling 1MB partitions connected over a Network-on-Chip (NoC) (Fig.2(a)).

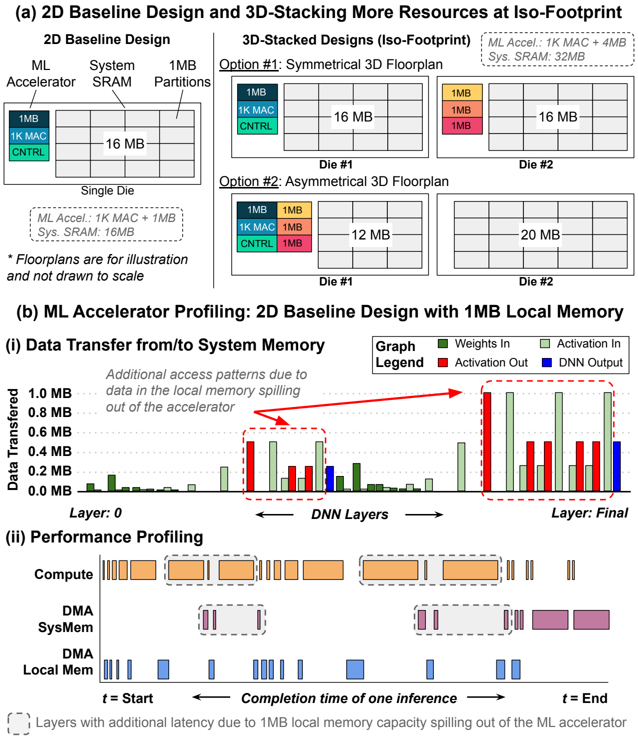

(a) 2D Baseline Design and 3D-Stacking More Resources at Iso-Footprint
2D Baseline Design
ML Accelerator
System SRAM
1MB Partitions
1MB
1K MAC
CNTRL
16 MB
Single Die
ML Accel.: 1K MAC + 1MB
Sys. SRAM: 16MB
* Floorplans are for illustration and not drawn to scale
3D-Stacked Designs (Iso-Footprint)
ML Accel.: 1K MAC + 4MB
Sys. SRAM: 32MB
Option #1: Symmetrical 3D Floorplan
1MB
1K MAC
CNTRL
16 MB
Die #1
1MB
1MB
1MB
16 MB
Die #2
Option #2: Asymmetrical 3D Floorplan
1MB
1MB
1K MAC
1MB
CNTRL
1MB
12 MB
Die #1
20 MB
Die #2
(b) ML Accelerator Profiling: 2D Baseline Design with 1MB Local Memory
(i) Data Transfer from/to System Memory
Graph Legend
Weights In
Activation In
Activation Out
DNN Output
Additional access patterns due to data in the local memory spilling out of the accelerator
Data Transfered
1.0 MB
0.8 MB
0.6 MB
0.4 MB
0.2 MB
0.0 MB
Layer: 0
DNN Layers
Layer: Final
(ii) Performance Profiling
Compute
DMA SysMem
DMA Local Mem
t = Start
Completion time of one inference
t = End
Layers with additional latency due to 1MB local memory capacity spilling out of the ML accelerator

Fig. 2. (a) 2D baseline design and 3D floorplan alternatives to expand its hardware resources at iso-footprint. (b) Pixel Codec Avatars (PiCA) DNN model analysis shows limited size of the 2D baseline ML accelerator's 1MB local memory creating performance bottlenecks.

A cycle-accurate analysis on data-movement and latency of deploying the model on this 2D baseline design is provided in Fig.2(b). First, SoC-level data-movement in between the ML accelerator and the global system SRAM is highlighted in Fig.2(b)(i), showing the amount of data transfer for weights and activations in MBs across DNN layers. The profiling shows that the necessity of tiling the workload to fit the layers within a 1MB buffer size requires the local memory to spill outside of the accelerator and access the system SRAM as an additional memory resource, which in turn creates more data-movement at the SoC-level. As a result, the baseline accelerator with the 1MB buffer configuration requires additional data movement of $\sim$4 MBs in and $\sim$4 MBs out from/to the system SRAM. In parallel, this increased data-movement creates additional latency cost for the baseline design with the 1MB buffer as highlighted in the performance profile graph in Fig.2(b)(ii), which provides accelerator operation latencies across DNN inference time. DMA based memory access requests are shown for both the local and system memories and data transfers are overlapped with compute. Fig.2(b)(ii) shows the associated latency cost of spilling for the spill-layers, estimated to be 39.1% of the overall latency for the baseline design with 1MB buffer per inference.

B. Improving the 2D Baseline with 3D-Stacking

For a single DNN inference, the profiling indicates that the realistic baseline design can meet the 30fps target. However extending the profiling to 1-to-4 group call scenario, the ML accelerator has to run inference on four user DNNs in a time-multiplexed way per frame, and misses the target fps due to the spilling behavior of its 1MB buffer configuration. To solve this, our analysis indicates that the 1K compute resources is adequate to achieve the target 30fps only if the 1MB buffer can be expanded by more than 2MBs to eliminate the data spilling. As a result, the ML accelerator local buffer can be extended to a total of 4MBs to improve the performance of the baseline design. For the global memory, the SoC needs 22MBs of system SRAM to store four models persistently to eliminate the high energy cost of fetching model data from an off-chip memory at every frame. To enable model persistence for energy-efficiency and also to provide the same 10MB memory space to other feasible subsystems as the baseline, the system SRAM can be doubled to 32MBs.

To achieve these system improvements, we perform floor-planning based architectural studies. The total area footprint of 1K MACs, 1MB SRAM, and the control core is found to be same as three 1MB SRAM partitions. Therefore a straightforward approach would be doubling the accelerator and global memory areas in 2D to achieve the system improvements, however this breaks the strict ID form-factor requirements of an AR SoC. Alternatively, 3MBs additional memory can be opportunistically 3D stacked on top of the 2D baseline ML accelerator to expand the local buffer size to a total of 4MBs at iso-footprint (Fig.2(a)). Similarly, 3D-stacking the same number of memory partitions on top of the 16MB SRAM would double the available capacity to 32MBs within the same footprint (Fig.2(a)). Another approach is doubling the ML accelerator area in 2D while only 3D-extending the global SRAM to achieve iso-footprint at the SoC-level (Fig.2(a)). Such an asymmetrical 3D floorplan, however, brings added energy cost on data movement as the total interconnect length increases [14]. We estimate that extending the ML accelerator local buffer to 4MB in 2D costs up to 14% more data access energy when compared to the symmetrical 3D-stacked floorplan. Moreover the asymmetrical 3D floorplan of the global SRAM costs up to 8% more data access energy as opposed to the symmetrical 3D floorplan.

Based on these design considerations, we expand the hardware resources of the realistic 2D baseline with the symmetrical 3D architecture floorplan at iso-footprint to enable running the PiCA 1-to-4 group call at 30fps in an energy-efficient manner. To demonstrate this solution, we implement a prototype 3D AR SoC which further includes a 2D CPU to orchestrate model deployment. Next, we present our 3D design methodology to enable this implementation.

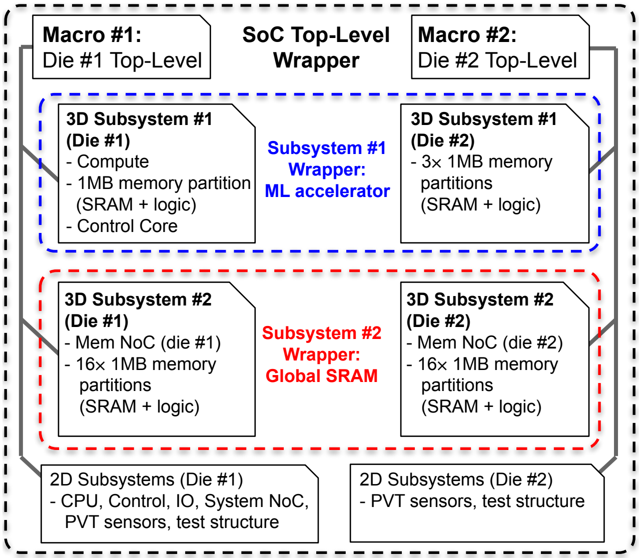


Fig. 3. RTL Hierarchy for 3D: Two stacked dies implemented as two macros instantiated within a top-level RTL wrapper.

III. 3D Circuit Design Methodology

3D circuit design for a stacked SoC requires careful timing considerations, as the OCV sources from the two dies combine to create hundreds of Process-Voltage-Temperature (PVT) corners for timing closure on the 3D datapaths [10]. As such, implementing a 3D clock-tree and closing setup and hold times on over thousands of 3D signal paths at the SoC-level using automated clock-tree-synthesis (CTS) and Static-Timing-Analysis (STA) flows of conventional EDA tools create hard to solve design challenges. To guarantee correct 3D-IC functionality using the conventional EDA tools, a systematic and robust approach is needed.

A. 3D cross-die circuit boundary analysis

To implement our 3D IC design methodology, we start by establishing the 3D cross-die circuit boundary. For this, we take advantage of the minimal RC loads of the SoIC bonds [2] [3]. Since the RC load of the bond is small, driving a 3D interconnect load becomes a similar circuit design problem to driving a medium to long distance 2D interconnect, as they exhibit similar wire-loads. To substantiate this approach, we performed circuit simulations to compare RC of 2D wires and 3D bonds. The average distance from middle of the MAC block to 1MB SRAM banks is estimated as $\sim 500\mu m$ during floorplanning. Considering half of this distance as a typical medium-distance wire length driven on intermediate-level metal layers, we performed parasitic-extracted post-layout simulations to estimate the associated RC of a repeater chain switching full-rail. Comparing this with the additional load that the 3D bond introduces, our estimations show that the 3D bonding constitutes only a 2% additional energy for driving a signal (at nominal Vdd, TT, 25°C). This analysis informed us to implement the 3D boundary at the circuit block level where typically medium to long distance signals are driven such as an SRAM bank or a NoC router. With this approach the added energy and latency cost of a 3D interconnect could be minimized to become negligible when compared to 2D routing, enabling using conventional EDA routing methods.

B. 3D design overview using conventional EDA flows

With the 3D boundaries decided, we then design the RTL to implement a 3D hierarchy as illustrated in Fig.3. Two dies are treated as two large macros that are connected at the top-level with an RTL wrapper where 3D bonds are treated as pins to integrate the two stacked dies, such that the 3D bonds can be constrained as conventional pins using SDC later in the design flow. Inline with the 3D cross-die boundary analysis, the two 3D extended subsystems of ML accelerator and global memory are sub-divided to two sub-blocks and dissagragated at the memory bank and memory NoC boundaries, respectively, as shown in Fig.3. Similar to the top-level, these sub-blocks are again connected with RTL wrappers at the top-level with 3D bonds as pin definitions in the stubs. Any 2D subsystem, such as the CPU or test structures are designed as conventional sub-blocks under the die macros. With this hierarchical approach, we were then able to use state-of-the-art EDA tools, following conventional design steps and files of RTL design, synthesis, and place-and-route (PnR). As such, the presented 3D flow is compatible to the current industry practices and standards, where multiple teams or third-party vendors may interact without blockers. As a result, the presented flow uses design constraints at the macro level such as SDC, CDC, and UPF, and conventional signoff tools for timing and power to implement and validate all the timing and power paths at the combined PVT corners. At the end of the flow, two tapeout ready GDS files are generated.

C. 3D clock network design considerations

The next step in the flow is implementing the 3D clock network that spans to two dies. The main challenge is using the 2D CTS flows for multiple dies with a single clock source (such as a PLL) which can create long divergent clock paths due to the combined OCV effects [10]. However, since the 2D EDA tool works with two separate modules at the top-level in our approach, the endpoints of the divergent clock paths may connect to govern any of the 3D signal pins. As a result, closing timing at the SoC-level quickly becomes an endless non-converging cycle of STA, where one 3D path is fixed only to violate another 3D path due to the long chain of divergent clock paths. To overcome this issue, we implemented the 3D clock-tree based on a 3D clock-forwarding approach. This method mitigates the risks of OCV effects by minimizing insertion delay of clock signals feeding into capturing flops of a 3D path, thus shortening the divergent clock paths on the 3D signal routes. For correct 3D-clock forwarding, we set constraints to enforce implementing the 3D signal routes as flop-to-flop paths only, as illustrated in Fig.4(a).

D. 3D datapath design considerations

Based on the 3D-clock forwarding method and flop-to-flop signal path constraint, two main connection styles are implemented for the 3D datapaths as illustrated in Fig.4(a): 3D links for 3D-extending pipeline stages with deterministic clock cycle requirements and 3D-extended NoC for paths with dynamic data traffic. As such, 3D links are used for extending the ML accelerator local memory, and the 3D-extended NoC extends the global SRAM memory. The 3D links are inserted as pipeline stages in the existing pipeline of the ML accelerator so that the throughput of the accelerator remains the same. This also enables maintaining the same access latency to 2D vs. 3D memory banks, and therefore eliminates any requirements to change the compiler's view of the accelerator. The 3D NoC routers communicate with multiple 1MB SRAM banks on both dies and implement the interconnect fabric of the overall global SRAM memory. Synchronizers were not placed on the 3D datapaths to avoid additional and unnecessary pipe-stage delays in both cross-die ways, which would incur latency costs to memory accesses. The final 3D design decision is enforcing all the communication to the 1MB SRAM banks on the top-die to be initiated at the bottom die to avoid creating "sneak paths" on the top-die that could recreate STA issues.

E. 3D timing closure considerations

Final step of 3D clock network and datapath design is timing closure. Inline with the hierarchical RTL design approach, we first performed STA iterations and timing closure fixes at the single die level for the two dies separately by constraining the 3D pins using SDC. Then 3D die-crossing setup & hold time fixes are performed by integrating the two dies as macros at the top-level. Any timing violations caught at this level need iterations to update all SDCs simultaneously to fix the violating paths at the SoC-level. Especially for shared and divided clocks paths spanning to two dies that may have long bi-directional routes, timing closure may need clock insertion delay fixes and rebalancing iterations at CTS, as illustrated in Fig.4(b). By employing these 3D-aware design rules and steps, we successfully designed an SoC-level 3D clock-network using a state-of-the-art 2D EDA tool to correctly implement more than 33,000 3D signals by closing timing on more than hundreds of combined multi-die multi-PVT-corners.

F. 3D testing and validation considerations

Based on the presented design approach, conventional behavioral and back-annotated gate-level simulations were used to verify the design at RTL, synthesis, and PnR stages. To validate the signal bonds post-tapeout, a boundary scan that stitches all the cross-die SoIC bonds is implemented as shown in Fig.4(c) (only the scan mode is shown). As the cross-die paths are unidirectional, the driver and receiver sides of the 3D paths are grouped as "set" and "detect" pairs on both dies, respectively. Receiving side flip-flops are implemented as scannable flops such that they perform as regular flops during functional mode and as scan-flops during scan mode. Additional scan flops are inserted to implement the "set" flops which only work in scan mode by being muxed in. In scan mode, a scan pattern is shifted in to the scan-chain of one die. Then the other scan-chain residing on the other die is clocked once to capture the pattern, and is subsequently shifted out

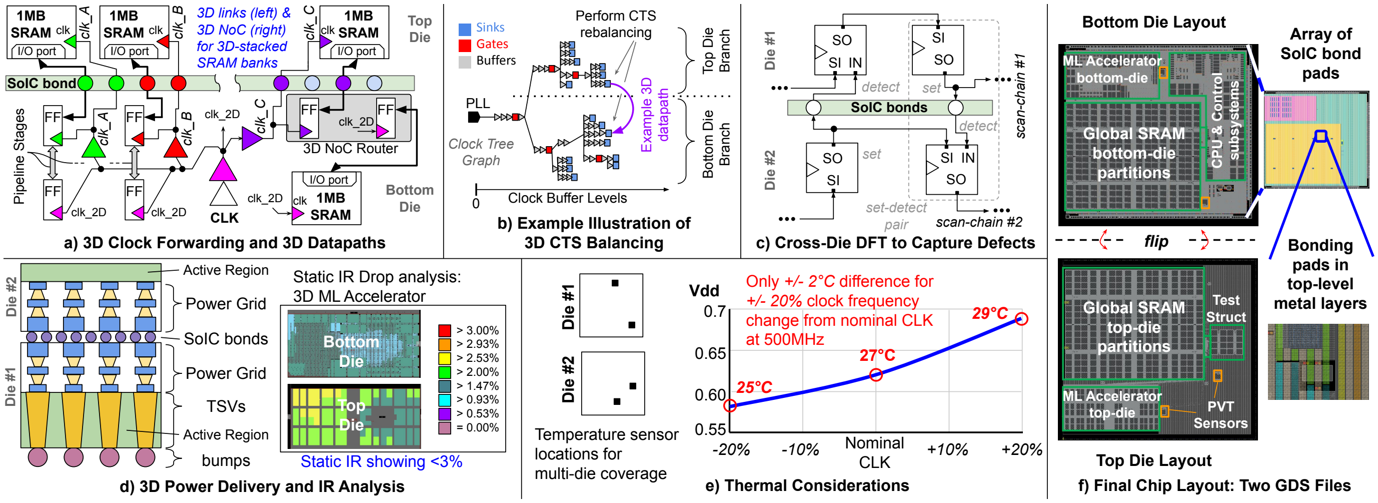


Fig. 4. Overview of the methodology: (a)-(e) 3D design considerations. (f) Final GDS for the two dies, implemented with the proposed methodology.

to detect any 3D bond defects. Since this covers only one direction of the 3D paths (e.g. bottom to top die direction) scanning order is switched and repeated to the detect defects on the reverse direction (e.g. top to bottom die). This way, conventional DFT insertion flows are used and existing buffers driving the 3D bonds are reused during scan mode. As a result, each 3D bond is validated for defects on silicon.

G. 3D power delivery considerations

Power is delivered to the bottom die first from the backplane at the bottom via TSVs and then distributed to the top die through SoIC bonds (Fig.4(d)). For this, TSV islands are designed specifically for power delivery with a power grid via-ladder all the way to the top layers. Performing static and dynamic IR drop analysis, we verified that IR drop was dominated by the regular interconnect via stack and not the TSVs or SoIC bonds. For instance for the 3D-extended ML accelerator, Fig.4(d) shows the IR drop in the 3D subsystem to be less than the foundry suggested drop-rates. As a result, conventional Signal/Power Integrity measures and via-insertion methods were employed. With this approach, more than 6M SoIC bonds were implemented for power delivery.

H. 3D thermal and yield considerations

One of the main concerns for 3D stacked chips is thermal considerations. To analyze and report our post-silicon findings, we placed four PVT sensors on the two stacked dies (with two PVT sensors on each) as shown in Fig.4(e). We then ran extensive benchmarks on silicon on a range of Voltage-Frequency points to analyze the average on-chip temperature collected from the PVT sensors. As shown in Fig.4(e), our silicon measurements show a maximum difference of 2°C in between top and bottom dies when ranging the frequency from nominal to +/-20%. Typically for high-performance and large area SoCs such as dataserver chips and GPUs at hundreds of Watts to kilo-Watts TDP ranges, thermal issues pose huge challenges which are exacerbated with 3D stacking. However our findings indicate that for the domain-specific AR SoCs at the edge with smaller areas at orders of magnitude lower TDP targets at mili-Watts to Watts ranges [13] the temperature variation does not pose a major risk for 3D stacking. In parallel, as yield is exponentially related to the inverse of the die area, we argue that comparatively smaller AR SoC areas present yield advantages for using Wafer-on-Wafer (WoW) stacking process. As such the presented prototype 3D-stacked SoC showed yield metrics within expectations during tapeout.

IV. Silicon Demonstration and Model Deployment

Enabled by the presented design methodology, 3D clock network and datapath timing is successfully closed at a target 500MHz nominal clock frequency (at 0.75V, TT, 25°C) over 33K 3D cross-die paths using conventional EDA tools. Testing, functional verification, and power delivery considerations further mitigated any major implementation blockers. DRC and LVS clean final layouts of the two dies are shown in Fig.4(f) with a zoomed in view of the 3D bonding pads at top-level metal layers. As a result, a 3D-stacked prototype AR test-chip in TSMC 7nm F2F WoW SoIC bonding technology was successfully taped out, where two subsystems are re-architected in 3D at iso-footprint to double the memory capacity of the system SRAM to 32MBs and quadruple the local memory of the baseline ML accelerator to 4MBs. Die shots and chip specifications are provided in Fig.5. 3D paths on silicon are first tested against defects in the lab using the cross-die boundary scans. Then we fully deployed the representative telepresence DNN model on the test-chip for realistic silicon demonstration. The practicality of the presented design methodology is validated with two sets of measurement results, first enabling negligible 3D data access cost compared to 2D, and subsequently accelerating AR subsystems by overcoming memory limitations at iso-footprint at better energy-efficiency utilizing hybrid-bonding technology.

A. Firmware and Model Deployment

To provide a complete picture of our methodology, details of model deployment on a 3D architecture are shared first. The application processor (AP) allows deploying models to the ML accelerator by providing a layer of Application Programming

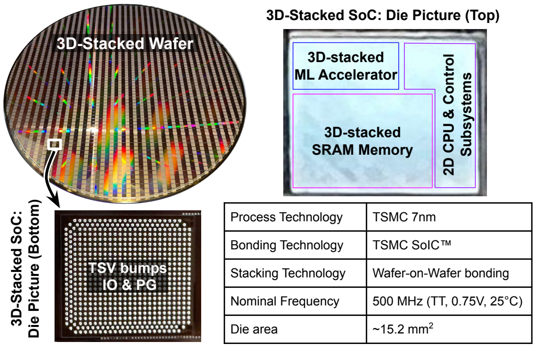
Interfaces (APIs) that are facing the user space. First an initialization API brings up the power, clock, and reset domains of the ML accelerator while loading a firmware layer executing on the control core inside the accelerator. Selection of whether the ML accelerator operates in 2D or 3D mode is done in this step. Next, an API loads the model into the accelerator where AP registers the model. Finally the user uses another API to run inference where AP requests the accelerator to receive inputs and execute. With this process, the user has freedom to map models, inputs, and outputs to any memory available in the user space. Therefore from the user's perspective, there is no difference in how the model is deployed, as the only difference between 2D and 3D modes is the availability of more internal memory for the accelerator as the memories of both dies are in the same address map. This is a crucial feature, since such a strategy allows the integration of the 3D memories to be transparent from a firmware perspective, enabling leveraging the benefits of 3D integration without imposing penalties on design complexity or additional design costs on the firmware.

B. Memory Access Energy Cost of 3D vs. 2D

Next, we substantiate the efficacy of our methodology to achieve minimal 3D access energy cost. For this, the 3D vs. 2D memory access energy cost is measured on the 4x 1MB SRAM banks of the ML accelerator while running inference on the PiCA model. Our silicon measurements show that the 2D baseline ML accelerator with 1MB buffer achieves 0.91 mJ energy per inference (500MHz, 0.75V, 30fps). We then measured the difference between 2D and 3D memory bank access energy by switching the activated 1MB bank from the bottom die (i.e. 2D) to one of the three 3D 1MB banks on the top die. Rerunning inference with each 1MB bank activated one at a time, our measurements show less than 1% difference over active-running mode, and no measurable difference over active-standby mode across all four banks. As a result, our silicon measurements show that 3D access energy cost for the local SRAM buffers implemented with our 3D circuit design method for extending the ML accelerator is negligible when compared to the total inference energy.

C. PiCA 1-to-4 Group Call Enablement

Finally, SoC-level energy and performance benefits of the prototype 3D-stacked AR SoC for PiCA 1-to-4 group-call decoding are demonstrated in Fig.6. Running PiCA 1-to-4 group call is compared on two ML accelerator configurations on silicon: the 2D baseline design with 1MB buffer and the 3D-extended design with 4MB buffer, both with 1K MACs. To capture the SoC-level benefits, silicon measurements include the energy of the ML accelerator, the 32MB global SRAM memory, and the in-between interconnect fabric. By eliminating the additional SoC-level data-movement combined with negligible 3D access energy cost, the 3D-stacked ML accelerator with 4MB buffer achieves 48% lower Energy x Delay product, or ∼2× better energy-efficiency, per PiCA inference, when compared to the 2D baseline design with 1MB buffer configuration, at iso-footprint (Fig.6, left). We estimate that this gain comes from eliminating a total of 8 MBs of data movement at the SoC-level to save 14.5% and 39% energy and latency, respectively. Results of running four decoders consecutively in a time-multiplexed way to render a single frame of the PiCA 5-user group call at 500MHz is provided in Fig.6, right. The 2D baseline design with 1MB configuration does not meet the strict 30fps (or 33.3ms) performance target due to the latency overhead of memory spilling as discussed previously. In comparison, by minimizing the additional data-movement to save latency, the 3D-stacked accelerator with the 4MB buffer successfully enables PiCA 5-user group call at 30fps while sparing 27.5% of the timing budget. These silicon results demonstrate that the presented work enables 3D-extending the AR subsystems in a scalable manner to achieve an application feature that was previously not feasible.


3D-Stacked Wafer
3D-Stacked SoC: Die Picture (Top)
3D-stacked ML Accelerator
3D-stacked SRAM Memory
2D CPU & Control Subsystems
3D-Stacked SoC: Die Picture (Bottom)
TSV bumps IO & PG

Process Technology	TSMC 7nm
Bonding Technology	TSMC SoIC™
Stacking Technology	Wafer-on-Wafer bonding
Nominal Frequency	500 MHz (TT, 0.75V, 25°C)
Die area	~15.2 mm^2

Fig. 5. Wafer and die shots during lab testing with technology details.

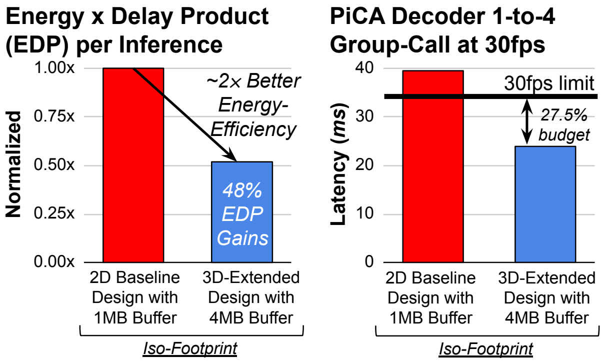


Fig. 6. Silicon Measurements: Energy-efficiency and performance results for two ML accelerator configurations targeting PiCA 1-to-4 group-call.

V. Conclusions

In this work, a robust 3D circuit design methodology is presented to implement 3D architectures targeting power and area constrained wearable SoCs. The method is demonstrated on silicon with a prototype 3D integrated AR SoC using currently available EDA tools. Enabling the advantage of negligible 3D access cost of SoIC bonding, the 3D-stacked SoC is showcased to enhance AR circuits by achieving new AR application features at higher energy-efficiencies while keeping the same footprint in the valuable AR SoC area.

Acknowledgment

The authors would like to thank Prof. Dr. Marian Verhelst for their valuable discussions.

REFERENCES

[1] S. Ma et al., “Pixel Codec Avatars,” IEEE/CVF CVPR 2021.

[2] C. C. Hu et al., “3D Multi-chip Integration with System on Integrated Chips (SoIC™),” IEEE Symp. on VLSI Tech., 2019

[3] D. C. H. Yu, C. -T. Wang and H. Hsia, “Foundry Perspectives on 2.5D/3D Integration and Roadmap,” IEEE IEDM 2021.

[4] B. Zhang et al., “Scaling Cu/SiCN Wafer-to-Wafer Hybrid Bonding down to 400 nm interconnect pitch,” IEEE 74th Electronic Components and Technology Conference (ECTC), 2024

[5] S.A. Chew et al., “700nm pitch Cu/SiCN wafer-to-wafer hybrid bonding,” IEEE 24th Electronics Packaging Technology Conference (EPTC), 2022

[6] K. Larsen and M. Swinnen, “Successful 3DIC Multi-Die Silicon System Design Using Synopsys 3DIC and Ansys Multiphysics Analysis,” Microelectronics Packaging and Test Engineering Council (MEPTEC), Road to Chiplets: Design Integration, May 10-22 2022.

[7] J. Park, “3D-IC (3DHI) Design Challenges,” Microelectronics Packaging and Test Engineering Council (MEPTEC), Road to Chiplets: Design Integration, May 10-22 2022.

[8] G. Murali et al., “On Continuing DNN Accelerator Architecture Scaling Using Tightly Coupled Compute-on-Memory 3-D ICs,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 31, no. 10, pp. 1603-1613, Oct. 2023.

[9] Y. Wang et al., “An Edge 3D CNN Accelerator for Low-Power Activity Recognition,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 40, no. 5, pp. 918-930, May 2021

[10] T. F. Wu et al., “A 3D integrated Prototype System-on-Chip for Augmented Reality Applications Using Face-to-Face Wafer Bonded 7nm Logic at <2um Pitch with up to 40% Energy Reduction at Iso-Area Footprint,” IEEE ISSCC 2024.

[11] L. Bamberg et al., “Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs,” DATE, 2020.

[12] H. E. Sumbul et al., “System-Level Design and Integration of a Prototype AR/VR Hardware Featuring a Custom Low-Power DNN Accelerator Chip in 7nm Technology for Codec Avatars,” IEEE CICC 2022.

[13] K. Kaiser, D. Patil and E. Beigne, “A prototype 5nm custom sensor SoC for Augmented Reality/Virtual Reality targeting Smartglasses with embedded computer vision, audio, security and ML,” IEEE Symp. on VLSI Tech., 2023

[14] L. Yang et al., “Three-Dimensional Stacked Neural Network Accelerator Architectures for AR/VR Applications,” in IEEE Micro, vol. 42, no. 6, pp. 116-124, 1 Nov.-Dec. 2022.