Power Delivery and Thermal-Aware Arm-Based Multi-Tier 3D Architecture

Lingjun Zhu, Ankit Kaul, Sung Kyu Lim
Georgia Institute of Technology, Atlanta GA, USA
{lingjun, ankit.kaul}@gate.edu, limsk@ece.gatech.edu

Rossana Liu, Rahul Mathur, Xiaoqing Xu, Shidhartha Das, Alejandro Rico, Doug Joseph, Brian Cline
Arm Inc. Austin TX, USA
{rossana.liu, rahul.mathur, xiaoqing.xu, shidhartha.das, alejandro.rico, doug.joseph, brian.cline}@arm.com

Abstract—3D integration is becoming a cost-effective way to incorporate more CPU cores and memory to improve the performance of computing systems. Meanwhile, due to the higher power density, power delivery and thermal issues become more significant in multi-tier 3D ICs. In this paper, we explore and evaluate multiple design options for an Arm Neoverse-based 3D architecture focusing on power and thermals at 7nm process and sub-10µm pitch. Using a rapid voltage-drop and thermal analysis methodology, we model a system with a 32-core CPU layer and up to 4 layers of system-level caches, and quantify the tradeoffs between performance, cost, voltage-drop, and temperature. A 3-layer configuration shows a good balance with 17% IPC gain and 17% lower cost, while incurring 15mV worse voltage drop and 8.5°C higher temperature compared with 2D. Our studies suggest that the co-optimization of system architecture, technology and physical design is key for high-performance 3D systems.

I. INTRODUCTION

As transistor scaling falls behind the trajectory predicted by Moore’s Law, attaining performance, power, area, and cost (PPAC) improvement of digital designs is becoming increasingly challenging. The semiconductor industry’s answer to these challenges is multi-chip heterogeneous systems comprised of functional dies assembled inside a package. However, die-to-die latencies still limit the performance of such systems, motivating the use of 3D-stacked dies. 3D-stacking can enable shorter wires and high-bandwidth die-to-die connections, resulting in overall PPAC gains.

High-density 3D stacking technologies such as face-to-face (F2F) hybrid bonding can enable 3D connections with sub-10µm pitches [1]. Fine-pitch TSV technologies with similar pitches have also been demonstrated, which can allow more than two layers of high-bandwidth connections in a face-to-back (F2B) configuration [2]. Based on these technologies, existing multi-core systems can be envisioned in a multi-tier 3D-stacked configuration with high-capacity, low-latency on-die memory and significant performance improvement compared with 2D systems, as shown by the example in Fig. 1a and Fig. 1b.

However, power delivery and thermal management are two of the primary challenges hindering wide-spread adoption of 3D stacking technologies [3], [4]. Voltage drop and thermal analysis in multi-tier 3D designs are non-trivial, while existing tools mostly target gate-level 2D designs. Although some studies have been done to enable thermal analysis in large-scale 3D designs [5], [6], voltage drop analysis for multi-tier 3D stack has not been studied comprehensively because it requires detailed modeling of the chip power consumption, power grid design, and package parasitics.

In this paper, we address these challenges and present a detailed technology and design space exploration of multi-tier 3D systems based on Arm® Neoverse™ CPU and mesh interconnect with a focus on power delivery and thermal perspectives. Our first contribution is a rapid analysis flow with an order of magnitude shorter runtime than the gate-level flow, facilitating early design-space explorations of multi-tier
TABLE I: Design options for evaluation with the 2D/3D architecture.

<table>
<thead>
<tr>
<th>option</th>
<th>core #</th>
<th>SLC per core</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-tier 3D</td>
<td>32</td>
<td>0.5MB</td>
</tr>
<tr>
<td>3-tier 3D</td>
<td>32</td>
<td>8MB</td>
</tr>
<tr>
<td>5-tier 3D</td>
<td>32</td>
<td>16MB</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>32MB</td>
</tr>
</tbody>
</table>

3D architectures. Second, to the best of our knowledge, this is the first work quantifying trends and trade-offs between performance, cost, voltage-drop and temperature rise for multi-tier 3D designs based on a real-world CPU model and system IPs. The rest of the paper presents the Arm-based design setup, rapid power and thermal analysis flow followed by a detailed 3D design-space exploration section.

II. DESIGN SETUP

A 32-core high-performance system is configured based on Arm Neoverse V1 CPUs, distributed system-level cache (SLC), and cache-coherent mesh interconnect. The CPU cores are modeled based on Arm Neoverse V1, a 64-bit out-of-order CPU targeting data-center applications in a foundry 7nm process node. The mesh network and snoop filters ensure cache coherence for the multi-core system. Table II and Fig. 2a show the configuration and floorplan of a 2D 32-core system. The system consists of 16 identical tiles, placed in a 4x4 array. Each 2D tile consists of two CPU cores, snoop filters, SLC data banks and associated logic. The SLC capacity in the 2D tile is limited to 0.5MB per core because larger SLCs will increase the tile area and the point-to-point latency in the mesh network, leading to system performance degradation.

The 3D tile has a similar structure to 2D with the CPU cores, mesh interconnect and snoop-filter blocks on one 3D tier, while all the SLC data arrays are moved to separate 3D tiers. As shown in Fig. 2b, on the compute die, the CPU cores, snoop filters, and SLC tags are still connected to the 2D mesh network, and 3D stacking of the SLC data arrays allows us to utilize the entire area under the compute tile for larger capacity. We estimate 8MB SLC capacity per CPU core can be added with one 3D tier.

The number of inter-tier signal nets connecting the compute die to the SLC memory dies is lower than 1500 per tile. This includes a 512-bit tri-state bidirectional data bus and control signals. The 3D signal density requirement is ≤ 10µm pitch, which can be easily satisfied by state-of-the-art hybrid-bonding and TSV-based 3D technologies [1], [2]. 3D stacking of multiple SLC dies with one compute die can increase the memory capacity up to 32MB per core in a 5-tier configuration and provide dramatic performance uplifts as shown in Fig. 1b. It is important to note that the performance speedup is modeled using full-system simulation in gem5 incorporating latencies based on physical design experiments. Additional cycles are not required to access a 2-layer 3D stack as demonstrated in [7]. We estimate an additional cycle latency when stacking more than 3 layers with face-to-back TSVs. This 2D/3D architecture allows us to analyze and simulate the multi-core system with various tier and memory configurations.

In Section V, we will focus on evaluating 4 design options as shown in Table I.

III. RAPID ANALYSIS OF POWER DELIVERY AND THERMAL (RAPIT) FLOW

In a typical IC design flow, the voltage-drop and thermal analysis is usually scheduled at the sign-off stages. The conventional gate-level analysis flow requires a complete implementation database, and thus is not applicable for early design space exploration due to prohibitively high compute requirements and all IP blocks not being ready. Additionally, for 3D design space exploration, it is imperative to prioritize power delivery and thermal aspects due to the higher power density of stacked designs. To overcome these issues, we propose a rapid power and thermal analysis flow for large-scale designs using commercial electronic design automation (EDA) tools.
A. Current Region and Power Grid Generation

The primary goal of early design space exploration is identifying critical challenges and potential solutions for future system designs. In order to enable the quick analysis, our approach is to abstract gate-level details and utilize micro-architectural block-level power to generate representative design floorplans and current regions, as shown in Fig. 3. A system floorplan containing CPU, caches and interconnects is composed, each consisting of multiple current regions (rectangular sub-blocks) representing the average current drawn within the micro-architecture block (e.g., fetch, decode, etc.). The area of the current region matches the footprint area of the micro-architecture block in a post-route reference implementation database of the CPU.

The accuracy of the voltage-drop and thermal analysis is related to the functional block hierarchy and current region size. We find that using the first level of CPU micro-architecture hierarchy to create current regions results in an order of magnitude improvement in runtime with marginal errors.

We extract micro-architectural block-level power through gate-level simulations of three representative benchmarks typically used to evaluate CPU power: dhrystone represents CPU power under typical integer programming; maxpower puts the CPU under an extremely heavy workload and shows the worst-case power consumption; daxpy represents power during vector operations. Based on these power simulations, we calculate the average current for steady-state analysis and generate time-based current waveforms for transient analysis in each current-region block.

The power delivery network (PDN) of the system tile is created with Cadence® Innovus™ based on the metal usage information extracted from a reference implementation in foundry 7nm process. For multi-tier 3D designs, the power grids of multiple dies can be connected together through the die-to-die bumps to create a 3D power grid. The parasitic resistance, inductance, capacitance (RLC) of the full power grid are extracted and connected to the current regions in Cadence Voltus™ Early-Rail Analysis (ERA), which can reflect the on-die power delivery in a real design.

B. Flattened vs. Hierarchical Analysis

To extend the single tile to a complete model of the 32-core design, two different approaches are applied: the flattened approach stitches all the current regions from the 16 tiles together and performs the analysis on the entire design as a whole; the hierarchical approach analyzes the worst-case scenario for each tile. When all the CPUs run the same workload, the voltage-drop analysis results of these two flows are very close, as shown in Fig. 4, because of the similar bump configuration in each tile. We use the hierarchical approach for PDN analysis since it reduces the computational runtime and we assume the worst-case power workload for all CPU cores. However, we use the flattened approach for thermal analysis since temperature rise is a function of die area, power hotspots and power density on the entire chip.

C. Runtime Evaluation for Voltage-Drop Analysis

We first apply the current-region-based method to a single-core 2D CPU to evaluate the runtime and accuracy of the flow. We perform transient voltage-drop analysis with the proposed RAPIT flow, and compare its accuracy and runtime with the conventional gate-level approach. Taking the gate-level transient analysis results as a reference, the error of the proposed flow is within 0.8% of the supply voltage (VDD). The proposed flow is able to capture the trend of voltage drop and the distribution of hotspots with various workloads. As shown in Table III, the runtime of the proposed flow is about an order of magnitude better than the gate-level approach, which is important for early design space exploration.
IV. 3D DESIGN MODELING

A. TSV Modeling

We consider via-middle TSVs with three different diameters: 5µm, 2µm, 1µm, as shown in Table II. We estimate parasitic RLC of TSVs based on the empirical formula [8]. The smaller TSVs have a higher resistance than the larger ones, while they can be grouped together as a cluster to reduce the equivalent resistance. Since TSVs penetrate the silicon substrate and require a keep-out zone (KoZ) without standard cell placement, they occupy a non-negligible silicon area in a 3D design. We consider three possible TSV cluster configurations: 1×1, 2×2, and 3×3, and the KoZ area of a cluster is linearly related to the diameter, pitch, and cluster size. TSV configurations have impacts on both voltage drop and area of 3D designs, which needs to be determined carefully.

B. 3D PDN Modeling

We model a robust multi-tier 3D power grid in the 3D designs. We assume that the tiers are bonded in a face-to-back fashion with TSVs and via arrays, and the cross-section view of the PDN is shown in Fig. 5a.

In the 32-core system, the CPU cores belong to a different power and clock domain than the mesh network and SLCs. The CPU power domain current consumption is significantly higher due to higher switching activity versus the system power domain, which consists of the mesh interconnect and SLCs. In order to compensate for this imbalance, the ratio of TSVs assigned to CPU versus system domain is set to 5 : 1 based on the current distribution.

For the 3D design, we consider an organic package substrate with stack-up layers between the 3D-stacked dies and the PCB. The package size is set to 45mm×45mm for the SoC, and the equivalent package resistance and inductance are estimated based on the averaged trace length for each tile. The parasitic RLC of the package, PCB, and other components of the PDN are also estimated based on the design size and empirical values [9] and taken into consideration in the equivalent circuit model in Fig. 5b during voltage-drop analysis.

V. 3D DESIGN SPACE EXPLORATION

A. Impact of TSV Configurations

The first step in our design space exploration is identifying a TSV configuration that balances supply drop and area overhead with minimal impact on design. We sweep the TSV dimensions, number of TSVs in a cluster and the pitch of the TSV clusters for a 2-tier 3D design as shown in Fig. 6. The trend of voltage drop in 2-tier 3D with various TSV configurations can be applied to multi-tier 3D. Worst case voltage drop is measured for each case with maxpower as the CPU workload. With the decreasing TSV pitches, the worst-case voltage drop decreases while the area occupied by the TSV KoZ increases, because lower TSV pitches allow more TSVs to be added into the design.

Compared with the single TSV configuration, the 2×2 TSV cluster provides up to 46% voltage drop reduction, while it also occupies 63% smaller silicon area compared to the 3×3 TSV cluster. Therefore, the 2×2 TSV cluster shows a good balance between the voltage drop and area overhead. Similarly, Fig. 6 shows that the 2µm-diameter TSV can balance the voltage drop and area overhead in a larger range of pitches, while the pitch range for 1µm-diameter TSV and 5µm-diameter TSV is limited by the large TSV resistance and KoZ area, respectively. As a result, we choose the 2µm-diameter TSV with a 2×2 cluster for our multi-core 3D designs, and set the TSV cluster pitch to 44µm to minimize the area overhead and provide a margin for voltage drop in our 3D designs.

B. 2D vs. 3D PDN Analysis

With the TSV configuration fixed, we compare the 32-core 2D design and a 2-tier 3D design to show the power delivery challenges and voltage-drop limits in 3D. To ensure a fair
Y μm

TSV clusters, the area of the 3D dies also increases by 6.8%. Due to the KoZ of the 2×2 2-µm-diameter TSVs is non-negligible, introducing a considerable voltage increase in each cluster. Moreover, the parasitic RLC of the TSVs can be inserted.

The thermal properties of the 3D designs with 2 to 5 tiers are evaluated by the steady-state thermal simulation, and the longest resistive path from the power source. The worst-case voltage drop occurs on the topmost compute die, because it has the highest current density. The worst-case voltage drop increases by 3.0mV per tier on average, due to the additional current and TSV parasitics, while the worst-case voltage drop in the 5-tier 3D design is still less than 6.3%. This increase is not dramatic, since the memory dies have relatively low current consumption and the TSV configuration enables a low-resistance power delivery path to the compute die. The worst-case voltage drop occurs on the topmost compute die, because it has the highest current and the longest resistive path from the power source.

In 3D designs with multiple SLC tiers stacked with compute die on top, the power TSV cluster pitch impacts the memory die floorplan. If the SRAM instances used in the SLC banks are larger than the TSV cluster pitch, we have to replace them with multiple smaller SRAM instances to allow power TSV clusters at the required pitch to minimize voltage drop. Generally, smaller instances have a lower array efficiency, i.e., a larger portion of the area is occupied by peripheral circuits versus bitcell arrays. Based on a 7-nm SRAM technology, we estimate that when the TSV cluster pitch is 50 µm, the SRAM instances introduce a significant area overhead (60%), compared with the case when the pitch is 150 µm.

To mitigate this issue, it is feasible to place power TSV clusters in the SRAM instance periphery as indicated in Fig. 9. This approach requires careful co-optimization of the system floorplan, 3D technology, and SRAM physical design, which also highlights the value of early-stage design space exploration to build realistic multi-core 3D design.

D. Voltage-Drop Analysis in Multi-Tier 3D

We extend the 3D design up to 5 tiers by stacking more memory dies using the balanced TSV configuration (2-µm-diameter TSV with a 2×2 cluster and a 44µm pitch). Then we perform the steady-state voltage-drop analysis on the design with the maxpower benchmark. Fig. 10 shows the trends of worst-case voltage drop increase with increasing tier number. The 1-tier and 2-tier designs are corresponding to the ones discussed in section V-B. From the 2-tier design to the 5-tier design, the worst voltage drop increases by 3.0mV per tier on average, due to the additional current and TSV parasitics, while the worst-case voltage drop in the 5-tier 3D design is still less than 6.3%. This increase is not dramatic, since the memory die has relatively low current consumption and the TSV configuration enables a low-resistance power delivery path to the compute die. The worst-case voltage drop occurs on the topmost compute die, because it has the highest current and the longest resistive path from the power source.

E. Thermal Simulation in Multi-Tier 3D

The thermal properties of the 3D designs with 2 to 5 tiers are evaluated by the steady-state thermal simulation, and the system floorplan, 3D technology, and SRAM physical design.
trough the cooling techniques. A 3-tier configuration provides a good balance for the design metrics with less than 5.7% static power consumption, the temperature rise is up to 19°C in the 5-tier 3D stack, which may lead to a reduced thermal headroom.

VI. CONCLUSIONS

A comprehensive design-space exploration of an Arm Neoverse based multi-core multi-tier 3D design is presented in this paper. We demonstrate that up-to-5-tier 3D stacked designs with expandable SLC can be designed and analyzed with a rapid analysis flow. A 3-tier configuration provides a good balance for the design metrics with less than 5.7% static voltage drop, at the cost of 6.8% TSV KoZ area overhead, which also requires a significant redesign of SRAM instance to enable feed-through TSVs. The thermal simulation shows the temperature rise is up to 19°C in the 5-tier 3D stack, which may lead to a reduced thermal headroom.

REFERENCES