Demonstration of a Power-efficient and Cost-effective Power Delivery Architecture for Heterogeneously Integrated Wafer-scale Systems

Haoxiang Ren
Center for Heterogeneous Integration and Performance Scaling (CHIPS), UCLA
Los Angeles, USA
haoxiang.ren@ucla.edu

Krutikesh Sahoo
Center for Heterogeneous Integration and Performance Scaling (CHIPS), UCLA
Los Angeles, USA
krutikesh@ucla.edu

Tianyu Xiang
Center for Heterogeneous Integration and Performance Scaling (CHIPS), UCLA
Los Angeles, USA
txiang@ucla.edu

Guangqi Ouyang
Center for Heterogeneous Integration and Performance Scaling (CHIPS), UCLA
Los Angeles, USA
guangqiouyang@ucla.edu

Subramanian S. Iyer
Center for Heterogeneous Integration and Performance Scaling (CHIPS), UCLA
Los Angeles, USA
s.s.iyer@ucla.edu

2023 IEEE 73rd Electronic Components and Technology Conference (ECTC) | 979-8-3503-3498-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ECTC51909.2023.00274

***Abstract*— In recent years, wafer-scale engines have emerged as a promising solution for achieving high performance computing (HPC), thanks to their advantages on form factor and scalability. By building a wafer-scale system with fine-pitch die-to-wafer bonding, it is possible to achieve high computing power, large memory capacity, and fast and efficient access to this memory, while ensuring high manufacturing yield and low design complexity. Large wafer-scale systems, however, make enormous demands on power (≥ 50 kW for a 300 mm diameter wafer), and methods to deliver this power efficiently, uniformly, and cost-effectively have not been fully realized yet. Compared to silicon, Gallium Nitride (GaN) switches promise higher conversion efficiency and higher power density, due to GaN's large bandgap, large breakdown electric field, and high electron mobility. In this paper, a dielet-on-GaN interconnect fabric (GaN-IF) vertical structure was demonstrated with a ≤ 10 µm metal bonding pitch for the future three-dimensional integrated voltage regulator (3D-IVR). This allows for intimate integration of the GaN switches with high-performance CMOS logic and passives in the substrate. An average shear force of 160.76 N was achieved on a dielet-to-wafer assembly with a dielet size of 4 mm^2. An effective specific contact resistance of 0.13 Ω-µm^2 was measured for the Cu-Cu bonding interface. A reliability test was performed, showing a resistance change of < 4%. Photoluminance, x-ray diffraction, and Raman spectra were measured to prove that our fabrication and bonding processes are not degrading the quality of the GaN layer. Additionally, a novel architecture is demonstrated, which allows an efficient delivery of power to the wafer-scale system through a dielet-side power delivery network (PDN) that does not require through-silicon vias (TSVs) – which are costly and necessitate the thinning of the substrate. A robust integration process flow for dielet-side power delivery was developed and optimized to obtain desired mechanical and electrical properties of the assembled structure. Power platforms were flip-bonded to the front side of the die-to-wafer assembly to form daisy chains. This is the first work that demonstrates a power-efficient, cost-effective, and heterogeneous power delivery architecture for wafer-scale systems.**

Keywords- power delivery; heterogeneous integration; wafer-scale system; three-dimensional integrated voltage regulator

I. INTRODUCTION

Wafer-scale systems –with many processing cores, large memory capacity, and high bisection bandwidth – satisfy the rapidly rising demand for high-performance computing (HPC). Such large but compact systems substantially facilitate the development of a wide variety of memory-access limited applications such as graph processing, recommendation engines, molecular dynamics simulation, etc. Compared to conventional systems, wafer-level systems exhibit better performance and higher energy efficiency[1][2]. A wafer-scale graphics processing unit (GPU) implementation outperforms the traditional multi-chip module (MCM)-based implementation on PCB by up to 18.9x speedup and 143x energy-delay product (EDP) benefit[3]. This improvement mainly stems from largely improved connection density between the GPU nodes.

There are two methods to create a large system: building a large system-on-chip (SOC) or interconnecting known-good-dielets (KGDs) onto a large substrate with fine I/O pitch. Building a large wafer-scale SOC is impractical due to the following issues: 1. it will suffer from low manufacturing yield, which requires complex redundancy design to promise an acceptable functionality. 2. the nature of homogeneity – substrate materials, technology nodes, etc. – only allows the use of static random-access memory (SRAM) for memory, which is relatively small and may not satisfy some of the memory-hungry applications such as GPUs. The need for terabytes of memory necessitates numerous denser memories

2377-5726/23/$31.00 ©2023 IEEE
DOI 10.1109/ECTC51909.2023.00274

such as dynamic random-access memory (DRAM) or flash and mandates high-bandwidth access to these memories.

Heterogeneous wafer-scale integration, on the other hand, offers the system large compute, large memory, and high bandwidth access to this memory simultaneously, leveraging silicon-based interposer techniques and fine-pitch die-to-wafer bonding. At UCLA CHIPS, previous works have demonstrated the architectural and technological potential of such an integration scheme by developing a silicon-based large-scale substrate called the Silicon Interconnect Fabric (Si-IF)[4]. It borrows the silicon back end of line (BEOL) process materials and techniques and achieves low latency (< 20 ps), and high bandwidth density with an energy per bit of < 0.4 pJ[5]. However, such a large and compact heterogeneous wafer-scale system demands enormous power (≥50 kW for a 300 mm diameter wafer). Assembly of wafer-scale systems for high performance computing will consume several tens of megawatts[6]. However, an approach to efficiently, uniformly, and cost-effectively delivering this power has not been fully realized yet. The current state-of-the-art power delivery schemes for large-scale systems usually have low overall efficiency (<70%). The overall efficiency is the product of the conversion/regulation efficiency and the PDN efficiency. A high conversion efficiency has been achieved thanks to the mature switching regulator circuits and technologies, while most of the power is dissipated by Joule heating ($P=I^2R$) at the PDN. In this work, two efforts are made to reduce the prohibitive PDN power loss for wafer-scale systems while addressing the challenges of heterogeneity and scalability:

The first effort is to reduce the current term (I), which is significant because it is a quadratic term. This requires a high voltage conversion ratio at high efficiency with fast response time and compact size. Utilizing larger bandgap GaN devices allows high voltage conversion thus reducing the current [7]–[13]. Fig. 1 shows a comparison of traditional PDN for wafer-scale systems and the novel PDN architecture in this work. Details of this effort is discussed in section II: a Si-on-GaN three-dimensional (3D) structure for 3D integrated voltage regulators (3D-IVRs) is proposed to increase the power efficiency, and heterogeneous integration of silicon dielet to GaN-interconnect fabric (GaN-IF) is demonstrated.

The second effort is to reduce the resistance term (R) of the PDN. GaN VR splits the system PDN into two parts – PDN_1 and PDN_2, and impedance on both sides is required to be reduced. Placing GaN in a vertically close vicinity to the point of load (discussed in section II) has been explored to decrease the resistance of the PDN_2. To reduce the resistance of the PDN_1, uniformly delivering power to the voltage regulator from the dielet-side/substrate-side of the wafer-scale systems should be considered. In this work, we show that a dielet-side PDN approach enables a cost-effective and scalable solution for wafer-scale systems. Experimental investigations are discussed in section III.

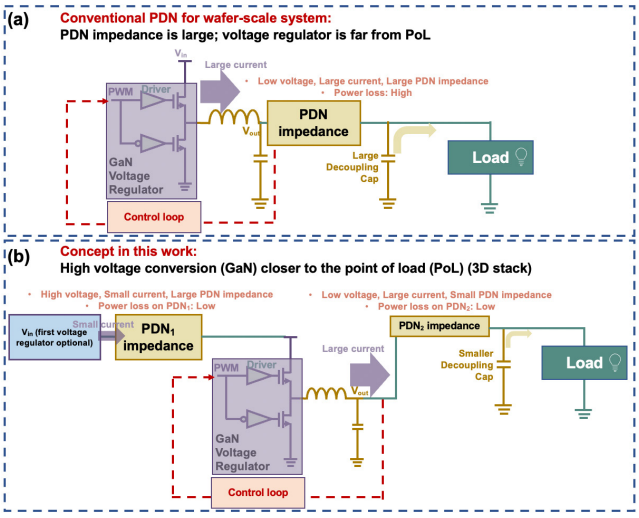

Figure 1. (a) Conventional PDN architecture for wafer-scale systems: large current passes through high-impedance PDN; (b) Concept in this work: high voltage converter (GaN) splits the original PDN into two parts (PDN_1 and PDN_2), and the voltage regulator is vertically close to the point of load.

II. Three-dimential Power Delivery Platform with Fine-pitch Cu-Cu Bonding

In this section, the concept of GaN-IF for efficient power delivery was proposed, and a heterogeneous integration was demonstrated on the GaN-IF platform.

A. GaN-Interconnect Fabric: a compact power delivery and interconnect platform

When power is delivered to a wafer-scale system, the traditional voltage conversion/regulation (usually on the motherboard) leads to unacceptably high voltage drop and prohibitive power loss. Integrated voltage regulators (IVRs) have gained popularity due to their ability to minimize power loss and voltage drop in the PDN, while also providing additional voltage domains. On one hand, the closer IVR is placed to the point of load (PoL), the longer segment of PDN will deliver low current, thus decreasing the PDN power loss. Take a wafer-scale system as an example, if a 48 V-1 V GaN IVR is placed near the dielet rather than on the motherboard, the current will be decreased by 48-fold, yielding a ≥2000-fold power saving on the PDN between the power input and the IVR. This is a substantial power saving for the wafer-size systems. On the other hand, multi-core or multi-dielet can have different voltage domains using a highly granular segmented approach.

The high input/output conversion ratio of IVRs requires III-V devices such as GaN high electron mobility transistors (HEMTs). Compared to silicon laterally diffused metal oxide semiconductor (LDMOS), the GaN HEMT has larger bandgap and higher electron mobility to deliver higher conversion efficiency and power density[14]. GaN switches can work at higher frequencies compared to silicon switches, and this allows miniaturization of the inductors and capacitors to further compact the system and increase the

power density. However, most state-of-the-art GaN IVRs are still integrated into the package side-by-side with the load-chip. These IVRs not only deteriorate the power density by occupying valuable space on the package, but also disrupt the regularity of the load-array. To address these problems, we report a novel three-dimensional integrated voltage regulation (3D-IVR) platform GaN-IF. Fig.2 illustrates an envisioned schematic of a system on the GaN-IF. Silicon utility dielets are bonded on the GaN-IF, serving as the regulation feedback. The inductors and the capacitors can be fabricated on the silicon substrate for a better form factor and lower PDN power loss. The GaN switches are on the Interconnect Fabric, and the load dielets will be heterogeneously bonded onto this GaN-on-Si substrate with a fine bonding pitch. GaN-on-Si is a robust material system compared to other substrates such as sapphire and SiC. Furthermore, all our Si-IF learning is extendible to the GaN-IF system. The proof of concept and a passive demonstration are presented in the rest of this section.

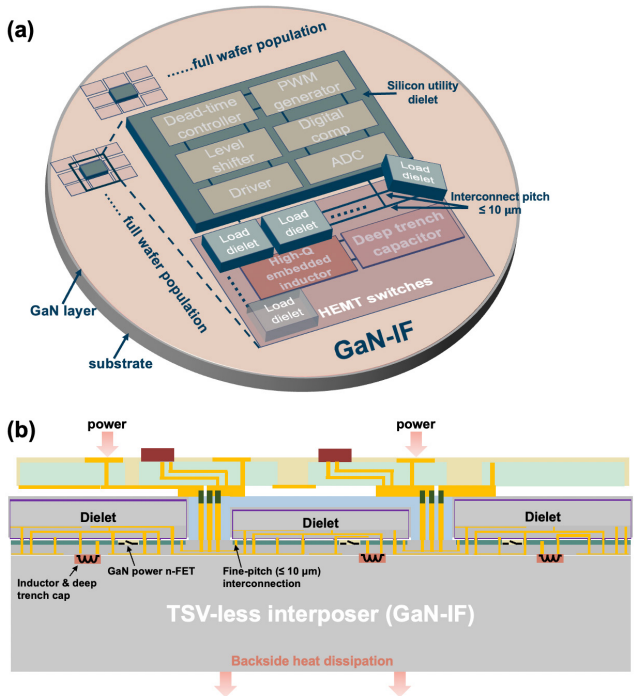

Figure 2. (a) 3D schematic of the GaN-IF: a wafer-scale heterogeneous die-to-wafer integration with embedded power switches, inductors, and capacitors. Dielets, GaN devices, L and C are 3D stacked; (b) Cross-section schematic of the PDN architecture in this work.

B. Experiment

To evaluate the mechanical robustness, specific contact resistance, electrical continuity, and reliability of the fine-pitch Cu-Cu direct bonds in a Si-to-GaN heterogeneous assembly, a specific daisy-chain structure was chosen and fabricated. Each assembly of dielets-to-wafer comprised a total of 180 daisy chains. Out of the 180 chains, 15 were chosen for testing. Each chain features a length of 3600 µm and contains 180 bonding pillars, providing an adequate test bed for characterizing the electrical performance of the Cu-Cu direct bonds. GaN-on-(111)Si substrate was chosen for its excellent thermal conductivity, mechanical strength, and affordability. 400 nm of GaN was grown using hydride vapor phase epitaxy (HVPE) by Kyma Technologies, and a thin (200 nm) layer of AlN was sandwiched by the GaN layer and the Si substrate to be used as a stress buffer. The quality and uniformity of the GaN layer were evaluated by measuring photoluminescence (PL) spectra, x-ray diffraction (XRD) spectra, and Raman spectra at five different locations (including the center and the edge). The dielet-on-GaN-IF fabrication and bonding process flow is shown in Fig. 3.

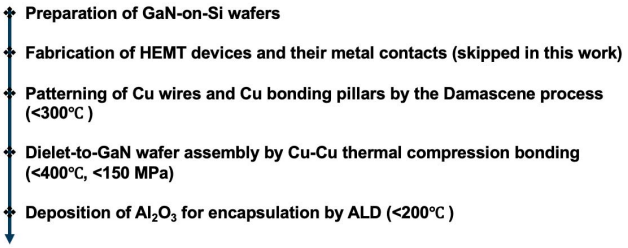

Figure 3. The integration process flow of the dielet-on-GaN-IF assembly.

The Damascene process was performed on both the GaN-on-Si wafer (substrate, Fig. 4(c)) and another silicon wafer (from which the dielets were sourced, Fig. 4(d)). The dicing street (Fig. 4(d)) was specifically designed and recessed to avoid dielet edge chipping and cracking during the dicing process[15]. Then the silicon wafer was protected using a spin-on organic film and diced into 2×2 mm^2 pieces (Fig. 4(d)). The top surface of the GaN-IF wafer and the diced dielets were terminated with Cu pillars and Cu pads, respectively. The top dielectrics on the substrate were recessed for 200 nm to ensure a successful tacking of Cu to Cu during the bonding process[16]. Both the GaN-IF substrate and the Si dielets were cleaned, followed by an additional acetic acid cleaning to remove the surface cupric oxide. After a 30-second Ar/H_2 reducing plasma surface treatment, the known-good-quality silicon dielets were picked, aligned, and tacked to the GaN-IF substrate with a 100 µm inter-dielet spacing (Fig. 4(b)) under a relatively low temperature of 120°C, a bonding force of 200 N, and a bonding time of 10 seconds. Subsequently, the assembly was subjected to a one-hour annealing process at 400°C under a top-down pressure of 137.5 MPa in a high vacuum environment to enhance the bonding strength[17]. The bonded samples (Fig. 4(a)) were then encapsulated with a 15 nm of Al_2O_3 layer using atomic layer deposition (ALD) process.

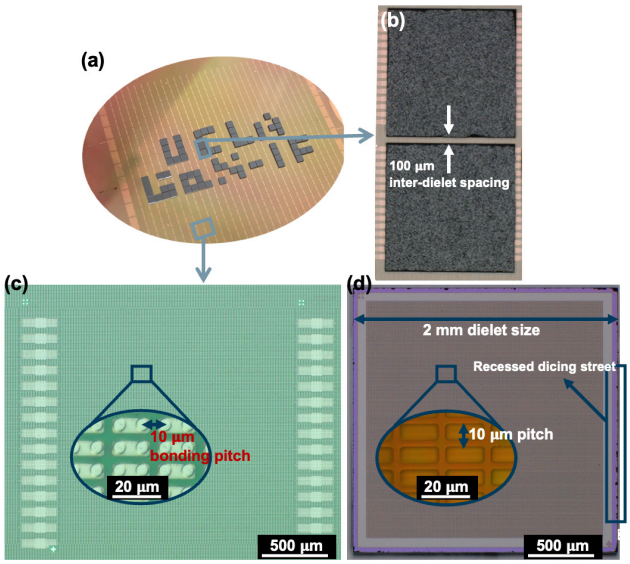

Figure 4. (a) Demonstration of fine-pitch heterogeneous integration: 2 mm by 2 mm silicon dielets are flip-chip bonded to a GaN-on-Si wafer with a pattern of "UCLA GaN-IF"; (b) Zoomed-in view of two adjacent dielets with an inter-dielet spacing of 100 µm; (c) Top view of GaN-IF substrate before dielet assembly (inset: zoomed-in view of the bonding pillars with 10 µm pitch); (d) Front-side view of the silicon dielet before assembly (inset: zoom-in view of the bonding pads with 10 µm pitch). The dicing street is recessed to eliminate the effect of edge chipping and cracking during the saw dicing process.

To check the alignment accuracy and the bonding interface quality, a focused ion beam (FIB) cross section was fabricated, and a cross-sectional view was obtained using scanning electron microscopy (SEM). Five different dielets were selected from the entire GaN-IF for conducting shear tests. A four-probe test was conducted on 45 different daisy chain links on the bonded samples. To prove that our fabrication and bonding processes are compatible with GaN devices, GaN layer quality was evaluated again by PL, XRD, and Raman on a control sample that went through the same processes described above. To examine the performance of the passivated samples under extreme temperature and humidity conditions, the unbiased highly accelerated stress test (UHAST) at 130°C and 85% relative humidity for 96 hours (JEDEC JESD22-A118 test condition A) was performed. After subjecting the samples to the UHAST, the shear tests and four-probe measurements were repeated.

C. Results and Discussion

To scale the interconnect pitch down to ≤10 µm, it is crucial to achieve a die-to-substrate alignment accuracy of ≤1 µm with an exceptional bonding quality. The bonding interface was meticulously analyzed using a Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) to appraise the alignment accuracy, as illustrated in Fig. 5(a). The three-dimensional cross-sectional view of the bonded die-to-wafer stack revealed the heterogeneous integration of silicon (top dielet) on top of the GaN-on-Si (bottom substrate) with a fine-pitch of 10 µm and an alignment accuracy of ±1 µm in both the x and y directions. Fig.5(b) shows a zoomed-in view of the Cu-Cu bonding interface, suggesting enough Cu grain growth and an excellent void-less interconnection.

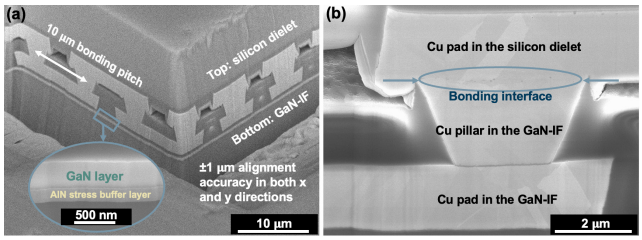

Figure 5. (a) 3D cross-sectional view of the bonded die-to-wafer stack featuring 3D heterogeneous integration of silicon (top dielet) on top of the GaN-on-Si (bottom substrate); (b) Zoomed in view of the Cu-Cu bonding interface featuring void-less interconnection.

Shear tests were conducted on five different die-to-substrate assemblies to benchmark the mechanical properties of the bonds. Excellent bonding strength was shown in Fig. 6(a) with an average shear force of 160.76 N, which is three times higher than the military specification (50 N)[18]. The shear strength variation can be attributed to the different Cu pillars' flatness/uniformity from site to site on the substrate, misalignment from assembly to assembly, and/or measurement errors.

The resistance of each daisy chain was characterized by a four-probe voltage-current measurement. We observed complete connectivity across 45 measured daisy chains on three different dielets. In Fig. 6(b), the current-voltage characteristics of the daisy chain are depicted by red boxes, with the red dotted line indicating a linear regression of the average resistance of 4.65 Ω. The variation in the measurement can be attributed to different translational/rotational misalignments from dielet to dielet, rotational misalignment on a single dielet-to-substrate assembly, and/or measurement errors. To determine the contact resistance of an individual pillar, we de-embedded the effect of the Cu pad resistances from the measured daisy chain resistance. We found that the average specific contact resistance per Cu-Cu bond is 0.13 Ω-µm^2.

Following the reliability test (UHAST), the average shear force was quantified as 138.47 N, exhibiting a 13.8% variation in the bonding strength, as demonstrated in Fig. 6(a). Moreover, the average resistance of the 45 links increased to 4.83 Ω, indicating only a 3.87% change in the electrical property, as shown in Fig. 6(b). These reliability results demonstrate the effectiveness of the ALD encapsulation process.

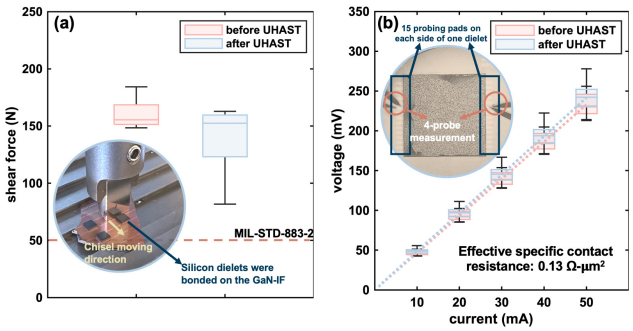

Figure 6. (a) Shear forces of the dielet-to-GaN-IF assembly before (red) and after (blue) the UHAST reliability test with the inset showing the testing methodology; (b) Voltage-current characteristics of the dielet-to-GaN-IF assembly before (red) and after (blue) the UHAST reliability test with the inset illustrating the measurement setup.

Fig.7(a), (b), and (c) illustrate that the GaN layer quality was intact after the metal layers fabrication process and the bonding process by presenting the comparison of PL, XRD, and Raman results. In the PL measurement, excitation wavelengths of 280 nm, 290 nm, and 300 nm were chosen. Five sites across the wafer from center to the edge were measured. Tool setup, aperture size, sample position, and measured sites were consistent before and after the processes. The solid lines represent the average of the measured data, and the colored shadow regions represent the error bar. In all the PL measurements, the peak position, the full width at half maximum (FWHM), and the peak intensity displayed a variation less than 5%, indicating the stable GaN crystal quality during the entire process. In the XRD measurement, the GaN peak was offset to zero, and the FWHM was 1653 arcsecs and 1675 arcsecs before and after processes, further implying that the quality of the film did not change. In the Raman measurement, a laser beam with a wavelength of 488 nm was used, and a Raman shift peak at 566 cm^{-1} was obtained for the sample both before and after the processes, revealing no strain in the GaN layer after the processes.

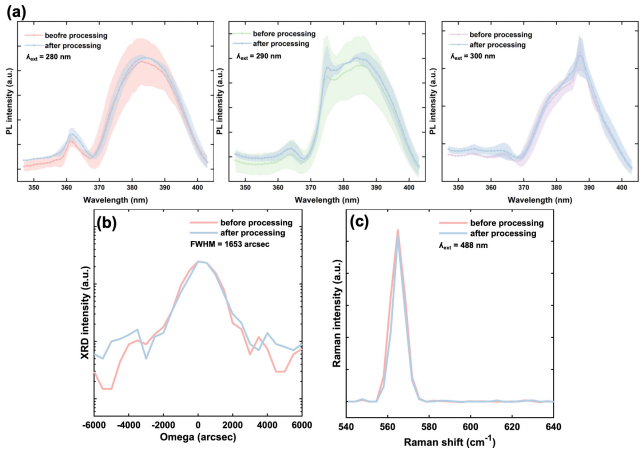

Figure 7. (a) Photoluminance spectra of the GaN layer before and after processing with excitation wavelengths of 280 nm, 290 nm, and 300 nm; (b) XRD GaN peak spectrum of the GaN layer before and after processing; (c) Raman spectrum of the GaN layer before and after processing with an excitation wavelength of 488 nm.

III. Demonstration of a TSV-less Dielet-side Power Delivery Architecture

In our previous work, the dielet-side power delivery architecture was proposed, and a surrogate structure was built[19]. This uniform power delivery scheme manifests itself with much lower power loss compared to delivering power from the peripheral of the wafer, and with much lower cost compared to fabricating through-wafer vias (TWVs) and delivering power from the substrate-side. In this section, a full demonstration of a dielet-side power delivery architecture with thin-dielet-to-wafer assembly and PCB-to-assembly bonding is presented.

A. Experiment

To demonstrate the robustness of the process and the electrical continuity, and to measure the specific contact resistance, a daisy chain test structure was designed. The process flow is illustrated in Fig. 8. 500-μm-thick 4-inch silicon wafers were used as the substrate, and 100-μm-thick and 200-μm-thick 4-inch silicon wafers were used for fabricating the dielets. Thin wafers patterns were designed and fabricated using the Ti/Cu (20/200 nm) lift-off process, and the wafers were saw-diced into 5 mm by 5 mm dielets. The substrate wafers patterns were designed and fabricated using the single Damascene process. A 2-layer FR-4-based PCB was designed and fabricated for testing. The PCB size is 9.85 mm by 9.85 mm, and the thickness is 600 μm. Electroless nickel immersion gold (ENIG) was used for the pads surface termination. The photographs in Fig. 10(a) and (b) depict the meandering path of the daisy chain within the PCB. The complete chain was routed between the substrate wafer and the test PCB using through-polymer vias (TPVs) as depicted in the schematic diagrams of Fig. 11(a) and (b). Each chain is 96.42 mm and 98.42 mm long for assemblies with 100-μm-thick dielets and 200-μm-thick dielets respectively, and each chain contains 18 TPVs and 36 solder-to-Cu interconnections. Four of the thin silicon dielets were bonded on the silicon substrate by Cu-Cu thermal compression bonding (Fig.9(a)). The bonding parameters were optimized to obtain the maximum bonding strength and the minimum damage on the dielets.

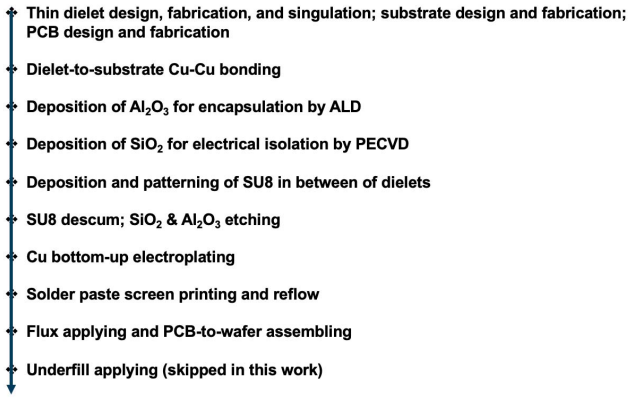

Figure 8. The integration process flow of the dielet-side PDN structure.

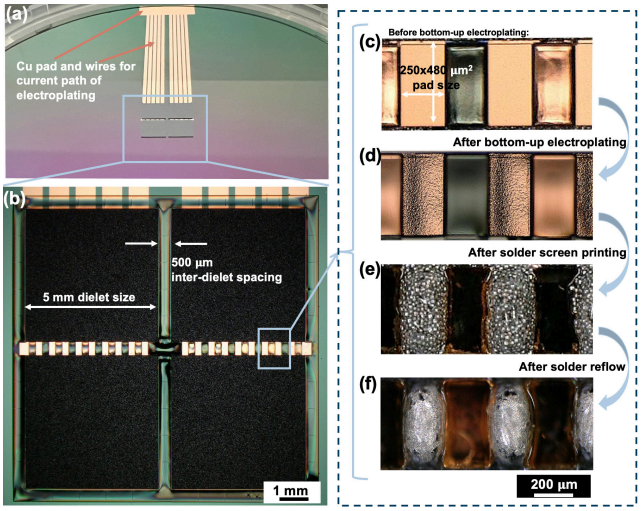

Figure 9. (a) A photograph shows that four 100 µm thin dielets were bonded on a silicon wafer by Cu-Cu thermal compression bonding, and the Cu patterns were for the bottom-up electroplating; (b) Top view of the assembly with the patterned SU8; (c) Zoomed in view of the Cu pads and SU8 before electrochemical deposition; (d) Zoomed in view of the Cu pads and SU8 after electrochemical deposition; the wrinkle patterns and shadows imply short and long wavelength roughness/flatness on the plated Cu surface; (e) Zoomed in view of the Cu pads and SU8 after solder screen printing; (f) Zoomed in view of the Cu pads and SU8 after solder reflow.

The bonded assembly was then coated with a thin layer of alumina by ALD for encapsulation, followed by a thin layer of silicon oxide deposition by plasma-enhanced chemical vapor deposition (PECVD) for electrical isolation and SU8-to-substrate adhesion. Then a SU8 film was coated with a thickness equal to the dielets' height. TPVs were photo-patterned (Fig. 9(b)), and the SU8 was hard-baked as it constitutes a permanent structure on the wafer-scale system. $SF_6/CF_4/O_2$ plasma descum process was performed to remove the residue of SU8 in the TPVs. Then the silicon oxide was dry etched by CHF_3/C_4F_8/Ar inductively coupled plasma reactive ion etching (ICP-RIE) and the alumina was wet etched by a tetramethylammonium hydroxide (TMAH) solution. After completing the Cu bottom-up electroplating process as depicted in Fig. 9(c) and (d), the subsequent steps of SAC305 solder screen printing and reflow were performed as shown in Fig. 9(e) and (f), respectively. Flux paste was then applied to the reflowed solder balls and the PCB pads were aligned to the solder balls and bonded together by another reflow process (Fig. 10(b) and (c)). The assembly was finally cut into a smaller piece for testing. The electrical resistances of different lengths of segments on two individual daisy chains on the assemblies with 100-µm-thick dielets and 200-µm-thick dielets were characterized using the four-probe measurement methodology.

B. Results and Discussion

The cross-sectional views of the assembled structures were observed. Cut lines were marked in Fig. 10(b). Cross-sectional view of the A-A' is presented in Fig. 10(d), indicating the complete connectivity of the daisy chain. Zoomed in view of the interconnection is shown in Fig. 10(f). The presence of total thickness variation (TTV) and non-uniformity of the electroplated Cu underscored the need for a thickness-compliant material, such as screen-printed solder. The total thickness was controlled by the thickness of the stencil, which compensated for the TTV and resulted in successful bonding. The cross-sectional view of B-B' is shown in Fig. 10(e), demonstrating a full three-dimensional (3D) stacking structure of the top-side power platform (PCB) on a die-to-wafer bonded large-scale system.

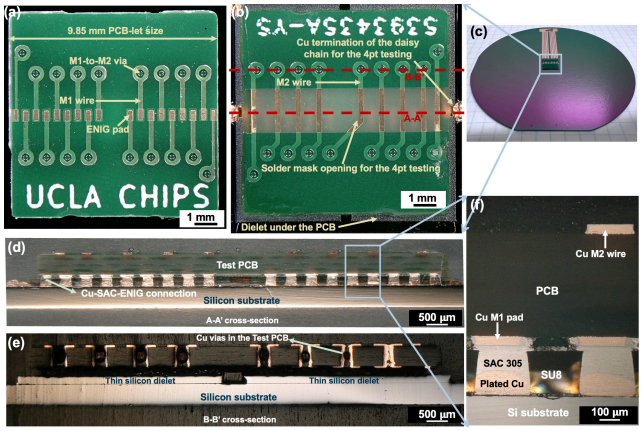

Figure 10. (a) Front-side photograph of the designed PCB; (b) Top view of the PCB-on-dielets-on-wafer stack; (c) Photograph of the full wafer after all the dielet-side PDN integration processes; (d) A cross-sectional view of the A-A' cut demonstrates a complete connectivity of the daisy chain; (e) A cross-sectional view of the B-B' cut present 3D stacking of the PCB, dielets, and the substrate; (f) A zoomed in cross-sectional view of the TPVs and solder interconnections shows the non-uniformity of Cu plating and the thickness compensation of the solder.

Daisy chains measured on two different samples were found to have no opens and shorts. The resistance of the chain was calculated based on the current-voltage characteristics measured on the sample with 100-µm-thick dielets (Fig. 11(c)). As depicted in Fig. 10(b), nine M2 wires were exposed on the backside of the PCB, and segments of different lengths were selected to evaluate the specific contact resistance of SAC305/Cu (Fig. 11(a)). For each length below 7 segments, three measurements were taken on different segments of the same length. For the length of 7 segments, two measurements were taken on different segments of the same length. For the length of 8 segments and the full chain, only one measurement was taken. A de-embedding technique was applied to eliminate the effect of the resistance of Cu vias and wires from the measured daisy chain resistance. Our analysis demonstrated an average specific contact resistance of 36 mΩ-cm^2 per SAC305-Cu bond. This relatively high value may be due to the formation of Cu_3Sn and Cu_6Sn_5 intermetallic compounds (IMCs) during the double reflow process.

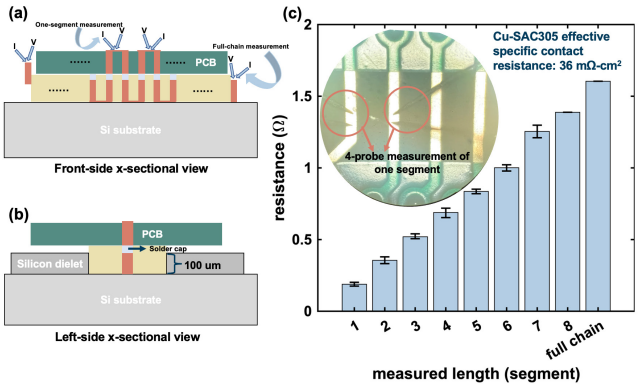

Figure 11. (a) Front-side cross-sectional schematic of the daisy chain and an illustration of the 4-pt test; (b) Left-side cross-sectional schematic of the daisy chain; (c) Resistance vs. measured length of the daisy chain with the inset showing the 4-pt measurement setup for one segment.

IV. Conclusion

In this paper, we propose a novel PDN architecture for wafer-scale systems and discuss two efforts of implementing this structure. Firstly, we constructed a dielet-on-GaN-on-silicon vertical structure with fine-pitch (≤10 μm) interconnects for the future compact 3D-integrated voltage regulator (3DIVR). The dielet-to-wafer assembly achieved an average shear force of 160.76 N, and a specific contact resistance of 0.13 Ω-μm^2 was measured for the Cu-Cu bonding interface. The reliability test revealed a resistance change of <4%. Measured PL, XRD, and Raman spectra verified that the quality of the GaN layer was not compromised during our fabrication and bonding processes. Secondly, a dielet-side PDN architecture without TSVs was demonstrated. A robust integration process flow of the dielet-side PDN was developed and optimized to obtain desired mechanical and electrical properties of the assembled structure. The total thickness variation (TTV) of through-polymer vias (TPVs) was minimized by using the solder paste screen printing technique. The power platform test PCB was flip bonded to the front side of the die-to-wafer assembly. Daisy chains were designed and built to benchmark the dielet-side PDN. Complete connectivity of the full chain was presented, and the specific contact resistance was extracted. This successful demonstration paves the way for constructing a power-efficient, cost-effective, and heterogeneous power delivery architecture for wafer-scale systems.

Acknowledgment

This work was supported in part by the Semiconductor Research Corporation (SRC) JUMP ASCENT, SRC JUMP CHIMES, National Science Foundation (NSF), UCOP MRPI MRP-17-454999, and the UCLA CHIPS consortium. We thank Nano Research Facility (NRF) and California Nano Systems Institute (CNSI) for their support during the fabrication process development. The authors gratefully acknowledge Prof. Mark S. Goorsky, Prof. Qibing Pei, Noah Bodzin, Ziqing Han, Michael Liao, Kaicheng Pan, and Randall Irwin for the measurements and valuable discussions.

References

[1] K. Rocki et al., "Fast Stencil-Code Computation on a Wafer-Scale Processor," no. arXiv:2010.03660. arXiv, Oct. 07, 2020. doi: 10.48550/arXiv.2010.03660.

[2] S. Pal et al., "Designing a 2048-Chiplet, 14336-Core Waferscale Processor," in 2021 58th ACM/IEEE Design Automation Conference (DAC), 2021, pp. 1183–1188. doi: 10.1109/DAC18074.2021.9586194.

[3] S. Pal, D. Petrisko, M. Tomei, P. Gupta, S. S. Iyer, and R. Kumar, "Architecting Waferscale Processors - A GPU Case Study," in 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2019, pp. 250–263. doi: 10.1109/HPCA.2019.00042.

[4] S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for AI systems," IBM Journal of Research and Development, vol. 63, no. 6, Art. no. 6, 2019, doi: 10.1147/JRD.2019.2940427.

[5] K. Sahoo, U. Rathore, S. Chandra Jangam, T. Nguyen, D. Markovic, and S. S. Iyer, "Functional Demonstration of < 0.4-pJ/bit, 9.8 μm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric," in 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), May 2022, pp. 2104–2110. doi: 10.1109/ECTC51906.2022.00332.

[6] S. S. Vazhkudai et al., "The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems," in SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, Nov. 2018, pp. 661–672. doi: 10.1109/SC.2018.00055.

[7] M. Gong, H. Chen, X. Zhang, R. Jain, and A. Raychowdhury, "A 94% Peak Efficiency 48-to-1-V GaN/Si Hybrid Converter With Three-Level Hybrid Dickson Topology and Gradient Descent Run-Time Optimizer," IEEE Journal of Solid-State Circuits, pp. 1–13, 2022, doi: 10.1109/JSSC.2022.3228233.

[8] Y. Guan, C. Cecati, J. M. Alonso, and Z. Zhang, "Review of High-Frequency High-Voltage-Conversion-Ratio DC–DC Converters," IEEE Journal of Emerging and Selected Topics in Industrial Electronics, vol. 2, no. 4, pp. 374–389, Oct. 2021, doi: 10.1109/JESTIE.2021.3051554.

[9] N. Desai et al., "A 32-A, 5-V-Input, 94.2% Peak Efficiency High-Frequency Power Converter Module Featuring Package-Integrated Low-Voltage GaN nMOS Power Transistors," IEEE Journal of Solid-State Circuits, vol. 57, no. 4, pp. 1090–1099, Apr. 2022, doi: 10.1109/JSSC.2022.3141779.

[10] A. Barner, J. Wittmann, T. Rosahl, and B. Wicht, "A 10 MHz, 48-to-5V synchronous converter with dead time enabled 125 ps resolution zero-voltage switching," in 2016 IEEE Applied Power Electronics Conference and Exposition (APEC), Mar. 2016, pp. 106–110. doi: 10.1109/APEC.2016.7467859.

[11] E. Aklimi, D. Piedra, K. Tien, T. Palacios, and K. L. Shepard, "Hybrid CMOS/GaN 40-MHz Maximum 20-V Input DC–DC Multiphase Buck Converter," IEEE J. Solid-State Circuits, vol. 52, no. 6, pp. 1618–1627, Jun. 2017, doi: 10.1109/JSSC.2017.2672986.

[12] H.-Y. Chen et al., "33.1 A Fully Integrated GaN-on-Silicon Gate Driver and GaN Switch with Temperature-compensated Fast Turn-on Technique for Improving Reliability," in 2021 IEEE International Solid- State Circuits Conference (ISSCC), Feb. 2021, vol. 64, pp. 460–462. doi: 10.1109/ISSCC42613.2021.9365828.

[13] X. Yang et al., "33.4 An 8A 998A/inch3 90.2% Peak Efficiency 48V-to-1V DC-DC Converter Adopting On-Chip Switch and GaN Hybrid Power Conversion," in 2021 IEEE International Solid- State Circuits Conference (ISSCC), Feb. 2021, vol. 64, pp. 466–468. doi: 10.1109/ISSCC42613.2021.9366005.

[14] M. Parvez, A. T. Pereira, N. Ertugrul, N. H. E. Weste, D. Abbott, and S. F. Al-Sarawi, "Wide Bandgap DC–DC Converter Topologies for Power Applications," Proceedings of the IEEE, vol. 109, no. 7, Art. no. 7, Jul. 2021, doi: 10.1109/JPROC.2021.3072170.

[15] H. Ren, Y.-T. Yang, and S. S. Iyer, “Recess Effect Study and Process Optimization of Sub-10 μm Pitch Die-to-wafer Hybrid Bonding,” in 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), May 2022, pp. 149–156. doi: 10.1109/ECTC51906.2022.00034.

[16] S. Jangam and S. S. Iyer, “Silicon-Interconnect Fabric for Fine-Pitch (≤10 μm) Heterogeneous Integration,” IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 11, no. 5, Art. no. 5, 2021, doi: 10.1109/TCPMT.2021.3075219.

[17] K. Sahoo, H. Ren, and S. S. Iyer, “A High Throughput Two-Stage Die-to-Wafer Thermal Compression Bonding Scheme for Heterogeneous Integration,” in 2023 IEEE 73rd Electronic Components and Technology Conference (ECTC), May 2023.

[18] “MIL-STD-883 - 2 -2 MECHANICAL TEST METHODS MICROCIRCUITS 2.” http://everyspec.com/MIL-STD/MIL-STD-0800-0899/MIL-STD-883-2_56325/ (accessed Feb. 24, 2023).

[19] H. Ren, S. Pal, G. Ouyang, R. Irwin, Y.-T. Yang, and S. S. Iyer, “TSV-less Power Delivery for Wafer-scale Assemblies and Interposers,” in 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), May 2022, pp. 1934–1939. doi: 10.1109/ECTC51906.2022.00303.