Abstract
A novel chip I/O technology, which enables high-bandwidth signaling, embedded microfluidic cooling and power delivery for high-performance 2.5D (silicon interposer) and 3D integrated circuits is presented. It features annular-shaped fluidic microbumps with 150 µm inner diameter and 210 µm outer diameter and fine-pitch electrical microbumps with 25 µm diameter and 50 µm pitch. Silicon dice with the novel electrical and fluidic I/O interconnections and die-level embedded microfluidic cooling were successfully fabricated and assembled. Following assembly, the measured resistance of a single electrical microbump is 13.50 mΩ ± 1.82 mΩ. Fluidic testing was conducted by pumping DI water into the bonded dice at a flow rate of up to 50 mL. No leakage or pressure drop change occurred during testing, and thus, demonstrating the feasibility of the novel fluidic I/O interconnections. In addition, the impact of the number of power delivery microbumps and die thickness (as a result of embedded microfluidic cooling) on power supply noise is analyzed using a compact physical model.

I. Introduction
Current and future high-performance computing systems are constrained by off-chip signaling bandwidth, cooling, and power delivery. Silicon interposer (2.5D) and 3D IC technologies have attracted significant interest because they enable high-bandwidth chip-to-chip communication, which helps address the memory wall problem and improve system performance [1][2]. There are also other compelling advantages that include lower power dissipation due to shorter interconnects, smaller form factor and heterogeneous integration [3]. However, cooling and power delivery are challenging for these densely integrated high-performance 2.5D and 3D systems.

Silicon interposer technology has been widely studied recently due to its ultra-fine pitch wiring density. Xilinx has reported a silicon interposer with fine-pitch wires using conventional back-end of the line (BEOL) processes and ultra-fine pitch microbumps ranging from 30 µm and 60 µm [4]. Multiple FPGA dice were assembled side-by-side on a silicon interposer, which provides high interconnect connectivity (more than ten thousand interconnections between two dice), high-bandwidth density and low power [5]. IBM has reported 50 µm pitch microbumps for 3D stacking and silicon interposer application [6][7] and fine-pitch wires for high-speed signaling [8]. Thus, compared to conventional interconnections, silicon interposer technology significantly improves integration density and system performance. TSV based 3D integration is another example of a novel emerging interconnect technology that is currently being extensively explored and deemed as a promising solution for high-performance computing systems by significantly shortening interconnect length [2].

Thermal management is a critical issue for high-performance systems, especially for 2.5D and 3D ICs. Power density and thermal resistance increase as the number of stacked dice increases. According to the International Technology Roadmap for Semiconductors (ITRS) 2012 [9], power density of one processor die is approximately 100 W/cm². Stacking multiple high-power dice leads to a large power density. Moreover, thermal resistance increases for the bottom most die in the stack since 3D stacking makes it harder to dissipate heat through the heat sink, which is attached to the top of the stack. Thus, conventional forced air cooling is no longer feasible. Previous work has demonstrated that embedded microfluidic cooling is a promising solution for high power density cooling. For a given power density, the temperature of a die with embedded microfluidic cooling can be 30 °C lower than that of a die with forced air cooling [10]. Moreover, embedded microfluidic cooling can be integrated into each die in a 3D stack, which makes it a feasible cooling solution for high-performance 3D stacks [11].

Power delivery is another major challenge facing 3D ICs. Power supply noise (PSN) increases due to increasing power density and TSV parasitics. Moreover, the ever decreasing power supply voltage for future systems leaves smaller noise margins. Thus, the difficulty in PSN suppression is increasing. Adding larger on-die decoupling capacitors and power delivery microbumps are two effective ways of suppressing PSN [12]. However, on-die decoupling capacitors consume precious silicon die area.

In this paper, a novel chip I/O technology is presented to simultaneously address the cooling, signaling and power delivery challenges of future high-performance computing systems. It features fluidic microbumps and fine-pitch electrical microbumps, which are fabricated using conventional processes. This paper is organized as follows: Section II describes 2.5D and 3D systems with embedded microfluidic cooling and the novel chip I/O interconnections. Section III presents design considerations of the fluidic microbumps. In Section IV, the fabrication process of the novel interconnects is presented. Section V presents assembly and test results. Finally, Section VI investigates the impact of the number of power delivery microbumps and die thickness due to embedded microfluidic cooling on PSN followed by the conclusion in Section VII.
II. Novel Electrical and Fluidic I/Os for 2.5D and 3D ICs

Fig. 1 illustrates 2.5D and 3D integrated ICs using the novel I/O technology under consideration to enable high-bandwidth signaling, embedded microfluidic cooling, and power delivery. As an example, two high-performance processors with embedded microfluidic cooling can be assembled onto a silicon interposer, as shown in Fig. 1(a), using the proposed interconnects. Moreover, memory dice can be stacked on top of the processors (which is not possible with an air-cooled heat sink approach).

Conventionally, each processor is packaged individually and assembled onto a motherboard. The packaged processors communicate using the board-level low-bandwidth density and high-power signaling wires. Moreover, the use of conventional bulky air-cooled heat sinks can keep two packaged processors far apart from each other, up to several inches away [13], which degrades interconnect performance and energy dissipation. The combination of silicon interposer high-density wiring and die-level embedded microfluidic cooling enables high-power processors (memory) to be assembled very close to each other. This drastic reduction in interconnect length, along with a larger number of electrical microbumps, enhances bandwidth density and reduces power dissipation.

The fluidic microbumps under consideration are used to create continuous fluidic paths between the silicon interposer and the die as well as between stacked dice, as shown in Fig. 1. A top view of fluidic microbumps and electrical microbumps is shown on the right side of Fig. 1(a).

The fluidic I/Os can be shared or dedicated to each die in the stack. In Fig. 1(b), two processors share one set of inlet/outlet, which implies that the flow rate and flow direction cannot be adjusted for each of the stacked dice. Counter flows can help minimize temperature gradients, as shown by the arrows. In Fig. 1(c), each processor has its own fluidic inlet/outlet. Flow rate and flow direction can be adjusted independently for each die based on power demands [11].

Fluidic microbumps are the most critical part of the fluidic network. Two columns of fluidic microbumps, one for inlet and one for outlet, are placed on two opposite edges of the dice, as shown in Fig. 1(a). Coolant is pumped into the silicon interposer through an inlet port, and guided into the assembled dice through one column of fluidic microbumps. Once the coolant flows into the die, it flows across the on-die embedded micropin-fin heat sink (to reject the heat), as shown in Fig. 2 [11]. The coolant exits the die through another column of fluidic microbumps (outlet) and enters the silicon interposer. Finally, the coolant exits the interposer through an outlet port to a heat exchanger. We have previously shown that the silicon micropin-fin heat sink (Fig. 2) used in this work can dissipate 103.4 W/cm² with a junction temperature of 47.9 °C [10].

III. Fluidic Microbump Design

The annular-shaped fluidic microbump is defined by its inner diameter and outer diameter. The inner diameter must be no smaller than the diameter of the die-level fluidic vias. To find the appropriate diameter for the fluidic vias, the pressure drop within the vias and the die area consumption of the fluidic vias must be considered simultaneously. Pressure drop within a fluidic via is calculated using the following formulas [14].
Pressure drop in Pascals, \( v \) is velocity in m/s, \( L \) is length of fluidic via in m, \( \rho \) is density of coolant in kg/m\(^3\), \( D \) is diameter of fluidic via in m, and \( f \) is friction factor. If the fluidic flow is laminar, the friction factor is calculated by

\[
f = \frac{64}{Re}
\]

where \( Re \) is Reynolds Number, which can be calculated by

\[
Re = \frac{1000 \cdot v \cdot D}{\nu}
\]

where \( \nu \) is kinematic viscosity in centistokes.

Die area consumption of fluidic vias can be simply calculated by

\[
p = \frac{A_{\text{via}} \cdot N}{A_{\text{chip}}} = \frac{\pi \cdot (D/2)^2 \cdot N}{A_{\text{chip}}} \times 100\%
\]

where \( p \) is percentage of fluidic via area, \( A_{\text{via}} \) is area of a fluidic via, \( N \) is number of fluidic vias, and \( A_{\text{chip}} \) is total die area.

Assuming that the fluidic via is 200 µm in length and a flow rate of 70 mL/min is used, the pressure drop within a fluidic via and percentage area occupied by the fluidic vias are plotted in Fig. 3 as functions of fluidic via diameter. The pressure drop and percentage area consumed should be as small as possible. Smaller pressure drop improves reliability and reduces pumping power, while smaller percentage area saves precious die area. By setting the thresholds to 15 kPa and 0.5% for pressure drop and percentage area, respectively, we find the feasible diameter range to be from ~70 µm to ~125 µm, as shown by the shaded region of Fig. 2. In this work, 100 µm was selected as the diameter of our fluidic vias. Based on this value, we selected the inner diameter of fluidic microbumps to be 150 µm, which affords a 25 µm gap between fluidic microbump and via. This gap is required to avoid possible via solder clogging and to compensate for flip-chip assembly alignment error.

The final dimensions of the electrical microbumps, the fluidic microbumps, and the fluidic vias are listed in Table 1.

### IV. Fabrication of Silicon Die and Interposer with the Novel Electrical and Microfluidic Interconnects

The fabrication processes of silicon die and interposer are described in this section, as shown in Fig. 4. With respect to the die fabrication, we start with a 4 inch silicon wafer. A layer of SiO\(_2\) is deposited on one side of the wafer. Micropin-fin heat sink and fluidic vias are etched on the back side of the wafer using two BOSCH etching steps \([10]\). Next, a seed layer (Ti/Cu) is deposited on the side of the wafer opposite to the micropin fin heat sink to enable the electroplating of fine-pitch wires (8 µm width/2 µm thick) and copper pads for the fluidic microbumps and electrical microbumps. Ni and solder are electroplated on the copper pads to form fluidic microbumps and electrical microbumps. Following this process step, the seed layer is stripped and the microbumps are reflowed. With respect to the silicon interposers, they are fabricated using processes similar to the die fabrication, as shown in Fig. 4.

### V. Assembly and Testing

Flip-chip bonding was utilized to assemble the fabricated die on the silicon interposer. Electrical and fluidic tests were conducted to experimentally verify the feasibility of the novel electrical and fluidic interconnections.
A. Flip-chip Bonding

As shown in Fig. 6, the test die described in the previous section was flip-chip bonded onto the silicon interposer. Table 2 lists relevant details on the test die and its flip-chip assembly process.

![Fig. 6. Flip-chip bonding of the test die on the interposer](image)

Alignment accuracy is critical to the success of the electrical and fluidic I/O bonding. X-ray images of the bonded die were captured to verify alignment accuracy. Fig. 7 illustrates the micropin-fins and electrical micro-solder joints after bonding. From this image, we conclude that the microbumps are well aligned (electrical measurements are reported in next section).

A. Resistance Measurements

To check the electrical interconnections, resistance of a single micro-solder joint was measured using the 4-point measurement technique.

![Fig. 5. (a) Fluidic microbumps, electrical microbumps and fine-pitch wires (top); (b) silicon interposer (lower left); (c) Micropin-fins and fluidic vias (lower right)](image)

Current is injected into the target micro-solder joint, while the voltage drop across the micro-solder joint is measured. Fig. 8 illustrates the 4-point measurement setup for micro-solder joints.

Resistances of 8 solder joints on 3 different bonded dice were measured using the 4-point technique. Fig. 9 illustrates the measurement results. The average resistance is 13.50 mΩ ± 1.82 mΩ. These values are consistent with the results reported in [6] and demonstrate that the electrical microbumps are well bonded.

![Fig. 7. X-ray image of Electrical microbumps and micropin-fins](image)

B. Fluidic Testing

Following electrical resistance measurement, a glass slide was used to cap the micropin-fin heat sink and inlet/outlet ports/tubes were attached to the bottom of the interposer, as shown in Fig. 10. At this stage, the sample was ready for fluidic testing. Coolant (DI water) was pumped into the
sample while the pressure drop between the inlet and the outlet was measured.

To test fluidic sealing, DI water was pumped into the sample for 4 hours at flow rates of 30 mL/min and 50 mL/min. No leakage was observed during the test. The pressure drop between the inlet and the outlet was stable, as shown in Fig. 11, which indicates no leakage occurred during the test and verifies the sealing of the fluidic microbumps.

Fig. 11. Fluidic sealing test (continuous pumping of DI water for 4 hours, at flow rates of 30 mL/min and 50 mL/min)

VI. Power Supply Noise (PSN)

Besides cooling and signaling, power delivery is another major challenge for future high-performance computing systems. In this section, the impact of the number of power delivery microbumps and die thickness due to embedded microfluidic cooling on PSN will be investigated using a compact physical PSN model [12]. No temperature dependence is taken into account in the model.

The embedding of microfluidic cooling increases die thickness. As reported in [10], the micropin-fin used in our work is 200 µm in height. Assuming die thickness is 50 µm without micropin-fin heat sink, die thickness increases to 250 µm by including the micropin-fin heat sink. This results in TSV length also being increased to the same value, which increases resistive and inductive parasitics to the power delivery network, and thus yielding larger PSN.

Assuming each stacked die dissipating 100 W/cm², 10% on-die decoupling capacitance, and die silicon area of 184 mm², PSN is simulated for two 4-die stacks. One consists of four 50 µm dice, and the other consists of four 250 µm dice. Fig. 13 shows the simulation results. As expected, the thicker stack has larger PSN (red solid line in Fig. 12) relative to the thinner stack (blue dashed line in Fig. 12). In both cases, PSN decreases drastically when the number of power delivery microbumps increases. About 80% of PSN reduction is achieved when the number of microbumps and TSVs increases from 2,048 to 32,768. Thus, increasing the number of power delivery microbumps is effective in PSN suppression and critical to future high-performance 3D ICs.

VII. Conclusions

Silicon dice featuring novel I/O interconnections and embedded micropin-fin heat sink were fabricated and flip-chip bonded to a silicon interposer. The electrical and fluidic interconnections were verified experimentally. The average resistance of a micro-solder joint is 13.50 mΩ ± 1.82 mΩ. No leakage occurred during the fluidic testing. Increasing the number of power delivery microbumps suppresses PSN effectively. Thus, our proposed novel I/O technology represents a promising solution to the high-bandwidth signaling, cooling and power delivery needs of future 2.5D and 3D ICs.
Acknowledgments

The authors would like to thank DARPA and Sandia National Laboratories for supporting this project, and Georgia Tech cleanroom staff Chris White for helping with the bonding process.

References


