3D Silicon Integration: A Novel Approach for Immediate Implementation

Peter C. Salmon
Vice President
Salmon Technologies, LLC
200 E. Dana St. #8
Mountain View, CA 94041
peter@salmontech.com
650-814-1076 (mob)

Introduction

Many new approaches are being considered for 3D integration of semiconductor systems. Several of them propose through-silicon-vias (TSVs) as part of the solution. However, the TSV solution requires redesign of the integrated circuit (IC) chips in a system, a substantial undertaking. For multi-core servers it is estimated that following this approach will take around 5 years to implement, to say nothing of the huge design expense. Even when such a design is fully implemented, it will be difficult and expensive to make incremental improvements.

Alternative Approach

An alternative approach is described in FIG. 1, one that builds on available materials and processes and could be fielded in a much shorter time.

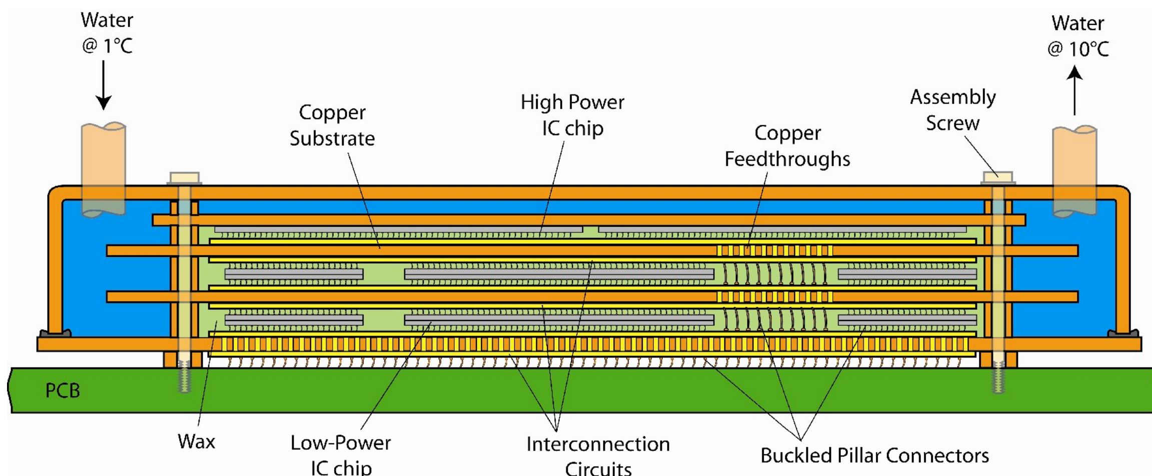

FIG. 1. Proposed 3D electronic system

In FIG. 1 chilled water enters at around 1°C and exits at around 10°C. It passes through a cooling channel having a typical width of 1 mm, formed between two copper sheets; the width is sufficient to achieve a good flow of water. The highest powered chips in the

© Salmon Technologies, LLC, 2008

system such as multi-cored processors are placed in the top row, where they can be cooled at a peak power density of over 1kW per square centimeter; this will be discussed further in reference to the thermal model depicted in FIG. 2. Lower powered chips such as memory chips are placed in the lower rows. They are cooled through their host copper substrate which is chilled at each end by immersion in cool water.

Vertical connections are provided by copper feedthroughs in the copper sheets, combined with large size buckled pillar connectors as shown. The buckled pillar connectors are described in detail in a separate paper[1]. They are formed as an adaptation of a stud bump, using copper or gold wire. The wire is terminated at one end using a ball bond, and then extended to form a flying lead. The flying lead is terminated using electronic flame off (EFO) or a spark discharge. Following planarization the extended wire tip is captured in a cup-shaped receiving feature provided in the surface of the interconnection circuit. Finally, all of the pillars are compressed a few percent in length by the assembly screws, thereby forming buckled pillars. The amount of compression is precisely controlled using spacers between stacking layers. The buckled pillars have a common flexure direction as shown. The stud bumping equipment allows for considerable flexibility. Using programmable automation, connectors of any required length are created, and different wire diameters may be used. Since the stud bumps can be fabricated at fine pad pitches, and the ball bonds can be formed on pads of almost any size and metallurgy, nearly every chip currently in use can be used in the proposed system integration.

A proprietary wax-like material may optionally be used to fill the assembled layers and provide structural support during assembly. The chosen wax is hydrophobic and impervious to water. An independent water seal may also be provided to isolate the internal volume from the water jacket. For convenience, the stacked assembly may be carried on a printed circuit board, PCB, although electronic systems of this type may also be constructed without using any conventional PCBs.

The assembly method is solder-free, and thus avoids the reliability and environmental issues associated with solder. It is conveniently re-workable at each level of assembly, and no good parts are wasted while finding a totally yielded system; this is in contrast to most stacked circuit assemblies of today. The test method will be further described below.

The copper panels are typically 0.5mm thick. A typical size is 18 x 24 inches, from which approximately 24 module sheets of the type depicted in FIG. 1 are fabricated. The interconnection circuits are fabricated using direct laser imaging on the full panel size. The space for dielectric material provided at each copper feedthrough is created using a conventional milling tool that accesses both sides of the panel.

[1] http://ap.pennnet.com/display_article/339446/36/ARTCL/none/none/1/Solder-free-Connectors-Using-Buckled-Pillars/

© Salmon Technologies, LLC, 2008

Primary Benefits
The primary benefits of the method are:

1. Volume reduction which can be as high as 50:1 for servers.
2. Well-cooled operation providing peak junction temperatures of 40°C for example.
3. 100% assembly yield, to be further described.
4. Module verification by full power and full speed testing, to be further described.
5. System-level standards for electrical, thermal, and mechanical parameters, potentially leading to high design productivity for heterogeneous systems[2].
6. No requirement for TSVs.

Thermal Model and Performance

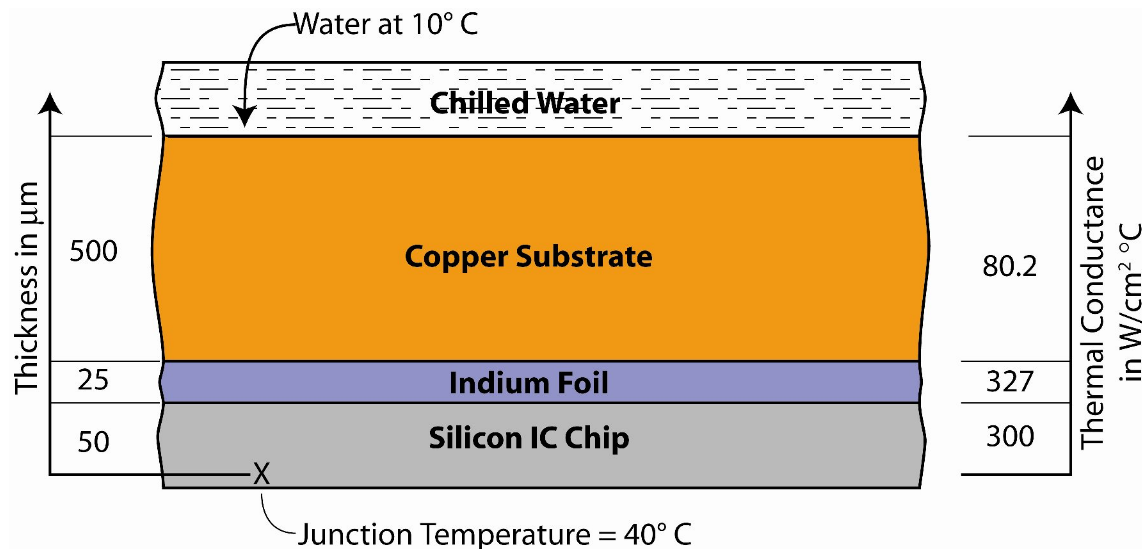

FIG. 2. Thermal model for the highest powered IC chips

FIG. 2 shows an expanded view of the interface between a high powered chip in the top row of a stacked module and the chilled water coolant. The chip has been thinned to 50 µm, and a maximum junction temperature of 40°C is shown as a realistic example. Indium foil is provided between the back side of the IC chip and the copper substrate; the soft indium will form itself to micro features at the interface, filling air pockets and improving thermal coupling. The compression required for the buckled pillars also serves to keep this interface under compression. The copper substrate thickness of 0.5 mm is chosen to support the mechanical stiffness required for processing 18 x 24 inch copper panels, especially with regard to forming the build-up interconnection layers. Since water has a specific heat of 4.186 joules per gram per degree Centigrade, a flow rate of 5 liters/min will support over 3,400W of cooling if the water temperature rises by 10°C.

---

[2] Ironically, this feature is enabled by the elimination of conventional chip packages - which have burdened the design process.

© Salmon Technologies, LLC, 2008

Trade-offs can be made between channel width and flow rate, maximum junction temperature, maximum water temperature, and total power dissipated. The thermal conductance of the total 3-layer sandwich is calculated at 53 W/cm^2 °C. For a ΔT of 30°C between the junction temperature and the peak water temperature, a peak energy density of 1590 W/cm^2 is achievable.

The proposed cooling system does not provide hot-spot cooling as might be achieved with micro channels embedded in the silicon chips. However, the high thermal performance quoted is achievable in a practical arrangement without the expense and delays involved with forming intricate embedded micro channels in the chips, typically 50 µm in width. These dimensions enable only limited flow rates. The need for hot-spot cooling can be eliminated by providing low overall thermal impedance and substantial flow rates, as depicted in FIG. 2.

Assembly and Test Methodology

Modern systems operating at gigahertz signaling rates are difficult to test using conventional testers. The test interface is problematic because the signal speeds keep increasing and the voltage supplies keep decreasing, resulting in reduced test margins. System manufacturers have typically increased built-in-self-test (BIST) capabilities; these usually include boundary scan. In a system assembly, boundary scan can be used to verify that all of the correct IC chips are present, that they have the correct orientation, and that the board-level interconnection circuits are good.

However, it is increasingly important to provide full speed functional tests. Many failures in modern circuit structures cannot be otherwise identified. The recommended test methodology is to provide an embedded test chip on each module layer. The test chip would be an application specific integrated circuit (ASIC), requiring significant investment. However the chip would be of modest size and cost. It would combine support for both scan tests and at-speed functional tests. Functional testing would include streaming of system bus data, at full bus width and at-speed. The test chip would contain sampling circuits and comparators that operate on the streaming data. Masking of test vectors would be supported. The test chip would have a low-speed interface to an external test support computer containing software for diagnostic assistance. Although each module layer now includes the cost burden of a test chip, several important benefits will accrue. Firstly the systems will be well-verified, reducing the high cost of field failures. Secondly, using the test chip support and suitable software, 100% assembly yield can be guaranteed, reducing costs by eliminating yield-related wastage. This solution can be expressed as “solving the known good die (KGD) problem”. Each layer in the stack is tested and reworked prior to module assembly. The stacked system is tested as a complete system and at full clock speed. If a failure occurs at the module level, hot inert gas can be directed at the wax in the defective layer to melt it, allowing the layer to be removed and replaced.

Interoperability

Military system designers have long championed the concept of interoperability, wherein standardized equipment and procedures can be employed across multiple product lines.

Because the proposed test approach is fundamental in nature, the test chip and its supporting computer can be designed to interoperate across broad families of products, thereby reducing bill-of-material (BOM) costs (via increased chip volumes) as well as reducing support costs.

Proposed solution versus TSVs

TSVs provide a compelling solution for stacked chip architectures that require the highest electrical performance and are produced in high volume; high development costs can be amortized over substantial product volumes. The 3D integration method proposed herein can easily incorporate such a specialized structure, along with many other heterogeneous components that may not be required in high volume. i.e., TSVs may be used most appropriately in a limited number of high volume chip-level assemblies, while the proposed 3D integration architecture may be suitable for system-level assemblies that require a combination of high power performance, compact size, and a standardized design environment covering all of the chips (and stacked assemblies) in a system.

Conclusion

A 3D silicon integration method has been proposed that offers a quick path to low cost systems having high thermal and electrical performance. This combination of high performance and low cost is achievable because the solution is provided at the system level, simultaneously addressing parallel issues of structure, density, cooling, testing, and yield. While the BOM cost is slightly increased, manufacturing efficiencies can make the yielded system cost attractive. The method employs proven materials and processes, and can employ current IC chips without modification.

© Salmon Technologies, LLC, 2008