

# REPAIRABLE 3D SEMICONDUCTOR SUBSYSTEM

Peter C. Salmon  
Sysflex, Inc.  
Mountain View, California, USA  
[PSalmon@sysflex.biz](mailto:PSalmon@sysflex.biz)

## ABSTRACT

A conceptual design is described for a stacked 3D electronic subsystem that achieves a miniaturization factor of over 100 compared with current assemblies, yet it can be well-tested, repairable, and adequately cooled. Cooling channels are provided between each module in the stack. All functionality in the subsystem is implemented on IC chips, including digital, analog, RF, integrated passives, optical, and test functions. The chips are attached to copper substrates using re-workable flip chip connectors, and the substrates are stacked using ball grid arrays. The subsystem is repairable down to the chip level; this is a fundamental advantage that can lead to complex systems having advanced performance and high reliability. Reliability is potentially high because conventional cables and connectors are eliminated, the subsystem is mechanically rugged, cooling is improved, and a better integration of heterogeneous chips is provided. The integrated cooling channels support heat dissipations up to  $1\text{W}/\text{mm}^2$  of chip surface, including optimal allocations of cooling resources to hot spots on a die. Since the flip chip connectors have an inductance  $\sim 0.1\text{ nH}$  and high-performance interconnections are provided on the copper substrates, signaling rates of at least 20 Gbps for digital signals and 10 GHz for RF signals should be achievable.

Keywords: 3D electronic subsystem, re-workable flip chip, cooling channels, repair-ability, heat bumps, electro-optic connector, stacked BGA, SiP, copper BGA.

## BACKGROUND AND MOTIVATION

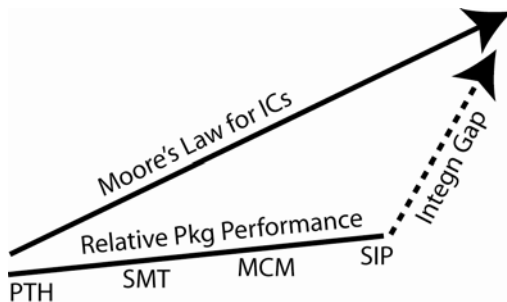


Figure 1. Semiconductor integration gap

Over the last 40 years transistor density in silicon integrated circuit (IC) chips has increased by a factor greater than 100,000 according to the phenomenon known as Moore's Law. Meanwhile, the ability to integrate silicon chips into systems has progressed more slowly. Package development can be traced from printed circuit boards (PCBs) having plated through holes (PTHs) around 1970. Surface mount technology (SMT) followed, multi-chip modules (MCMs), and systems in package (SiPs). The slow development of integration methods has resulted in an integration gap; this gap has dimensions of cost, performance, cooling, and scalability.

The 2003 International Technology Roadmap for Semiconductors (ITRS) shows packaging costs for microprocessor circuits exceeding chip costs in 2010<sup>1</sup>, assuming 2,000 leads per chip. This is predicted because the number of leads per package is increasing faster than the cost per lead is decreasing. It makes sense to address this issue using wafer level processing, wherein the manufacturing cost is largely independent of the number of leads per chip. The proposed flip chip connectors include electroplated bumps on the die (wafer) and wells filled with solder on the substrate (panel); both are fabricated using wafer level processing (WLP).

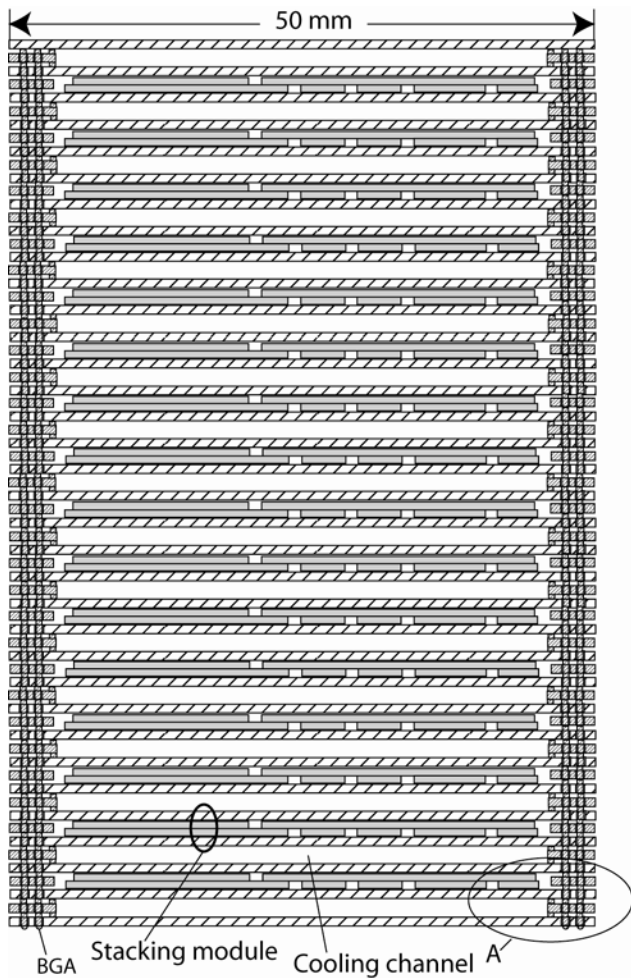
Digital IC chips currently operate at signaling rates of 10 Gbps while many packages do not support speeds greater than around 200 Mbps. Cooling has also become critical although the recent push toward multi-core processors by Intel and others has delayed the impact of this problem for some microprocessor systems. Modern servers typically have bulky finned aluminum heat sinks mounted on each of the processors. This increases the volume of the server units with attendant cost increases and performance decreases. Cooling costs for a 30,000 square foot data center are reported at \$8 million per year.

Scalability has not been much discussed at the system level, apart from providing servers in a blade form factor for higher packaging density and user convenience. Generally, system or subsystem scalability is difficult if multiple component types and packages are employed.

Electrical connections to an IC chip have typically occurred on the front side of the chip where the active circuits and bonding pads are located, while cooling has been provided at the back side. Thermal interface materials (TIMs) such as thermal grease have been used between the chip and its heat sink. When thermal grease is used, it is typically the highest impedance element in the thermal path. The proposed subsystems will provide cooling at the front side of the chip, with no thermal grease.

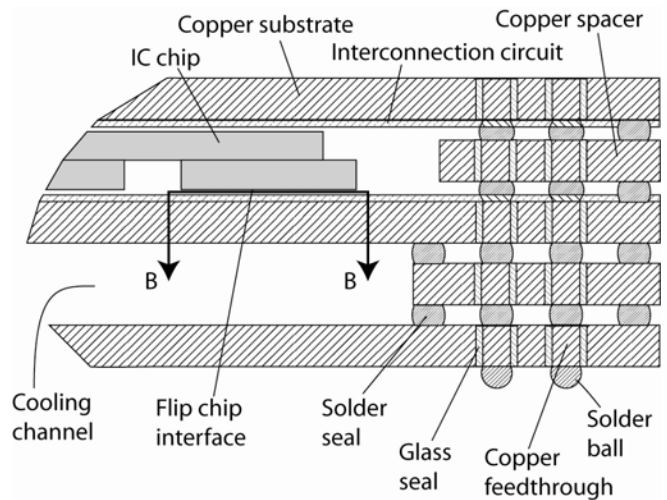
For complex flip chip assemblies it is difficult or impossible to test them at full power and full speed through a cable to an external test box. Also, the large size of a typical test connector and cable effectively negate the miniaturization and performance advantages of flip chip. It is preferable to provide test chips resident in the modules; they can include high speed sampling circuits and comparators and an interface to a test support computer.

**PROPOSED 3D ARCHITECTURE**



**Figure 2.** Stacked copper modules with cooling channels.

Each stacked module includes two copper substrates, each having multiple chips attached using flip chip connectors, to be described. The modules are hermetically sealed to exclude water and create cooling channels as shown in Figure 2. A ball grid array (BGA) interface is shown at the bottom, and also between each stacked element. Although the stacked elements may differ in function and chip set, they can also replicate standard modules. In Figure 2 each module is a 4-way server comprising approximately 80 chips. These include processors, memory of various types, legacy controllers, power distribution chips, test chips, and integrated passives. The entire stack implements a 256-way server that is fully repairable and can dissipate up to 10 kW with an appropriate fluid (air or water) circulating through the cooling channels. This subsystem is over 100 times smaller and lighter than current state-of-the-art blade servers, and its small form factor will lead to at least a 20% speed advantage. The portion marked as “A” is expanded in Figure 3.

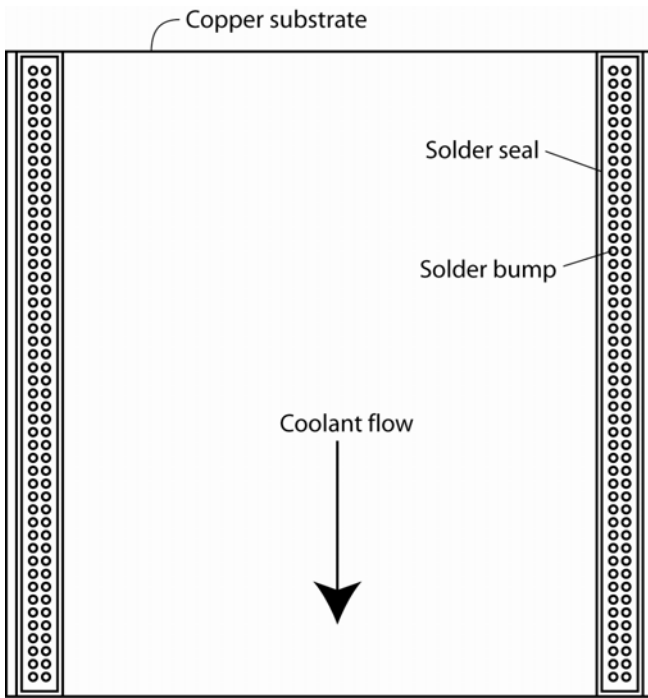


**Figure 3.** Detail A of Figure 2.

In Figure 3 the IC chips are mounted back-to-back, and heat extraction occurs through the front face of each chip, to be further described. Hermetic seals are created using solder and glass as shown. Input/output (I/O) within the stack is supported by BGA connections between each copper element, including both substrates and spacers.

**BGA INTERFACE**

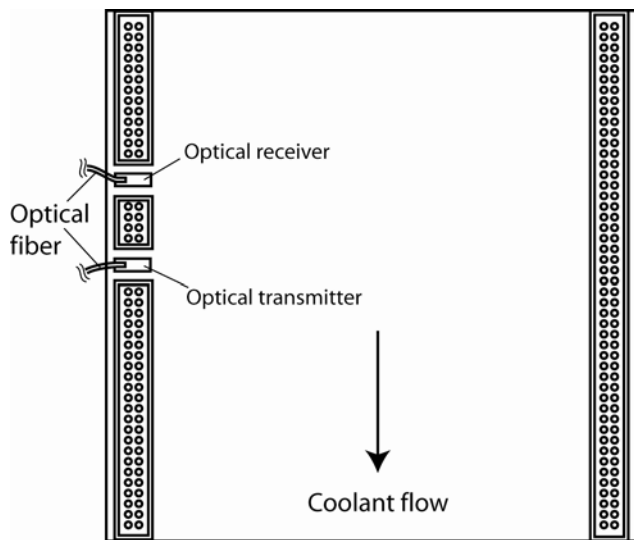
Figure 4 shows a BGA interface between a substrate and a spacer of Figures 2 and 3. Two edges are kept clear for unimpeded flow of coolant. A solder seal is provided as a line feature as shown. Using a 50 mm square substrate and 1mm BGA pitch, 192 connections are available for power and I/O in the figure.



**Figure 4.** BGA interface.

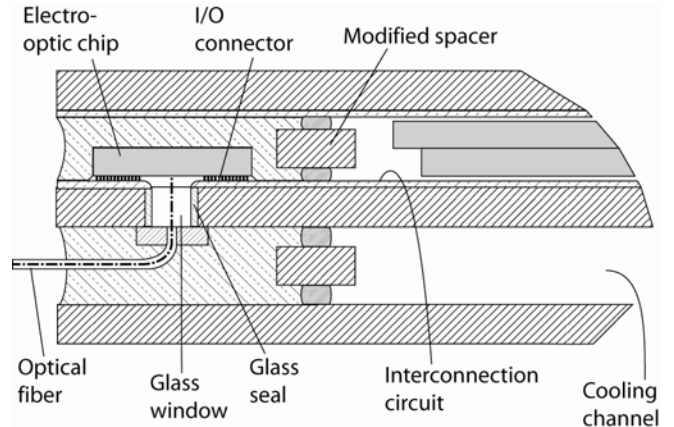
In many applications it will be advantageous to expand the number of leads for power and I/O beyond the typical 192 shown in Figure 4. More columns of bumps and reduced BGA pitch could be used. However, this would reduce the available area for mounting chips or complicate the I/O routing options near the bumps.

### I/O USING FIBER OPTIC CONNECTIONS



**Figure 5.** BGA plus optical interface.

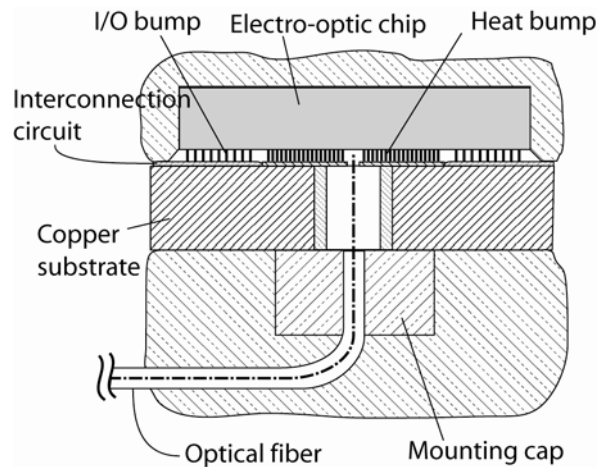
Figure 5 shows replacement of some solder balls at the BGA interface with fiber optic connectors. New semiconductor-based lasers, modulators, and sensors are currently under development at several universities and companies. In particular, use of the Raman effect has enabled silicon-based lasers at reduced power levels<sup>2</sup>. The designs illustrated in Figures 6-9 anticipate the integration of an optical transmitter or receiver alongside supporting electrical circuits on the same chip.



**Figure 6.** Electro-optic connection.

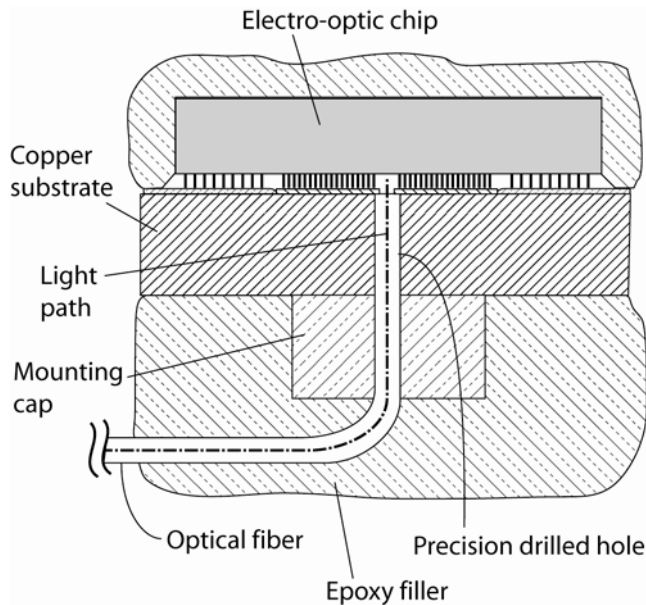
In Figure 6 an electro-optic chip has been mounted using I/O connectors to an interconnection circuit on a copper substrate. The I/O connectors and interconnection circuit will be further described. A fiber optic data feed is shown, employing a clear glass window in the copper substrate. The optical interface hardware is compact and can be integrated with a stacked architecture like that of Figure 1.

### HEAT BUMPS AND I/O BUMPS



**Figure 7.** Flip chip interface showing mixed array of I/O bumps and heat bumps.

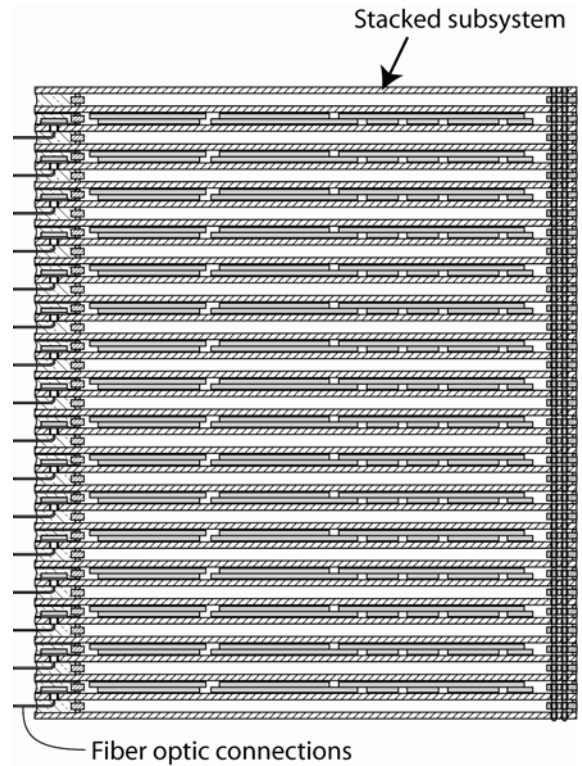
Figure 7 introduces the concept of a mixed array of heat bumps and I/O bumps at a flip chip interface. Hot spots on the die may require extra cooling; they can be provided with a dense array of “heat bumps” which are simply a closely-spaced version of “I/O bumps”. Both types of bumps employ copper columns to be further described; the copper columns bend like fine wires to relieve mechanical stress at the interface. In Figure 7 the electro-optic chip may include a laser as part of an optical transmitter. The laser may be embedded in the silicon chip and may require the extra cooling provided by the heat bumps.



**Figure 8.** Simple fiber interface without glass window

Figure 8 shows that the transparent window can be eliminated by providing a precision drilled hole in the copper substrate that supports the end of the optical fiber and aligns it with the optical circuits on the chip. The alignment can be fine-tuned by monitoring link performance during a reflow/adjustment cycle subsequent to initial assembly. Drilling tools having a repetition accuracy of 1µm are available for creating the precision drilled feature; see for example reference<sup>3</sup>.

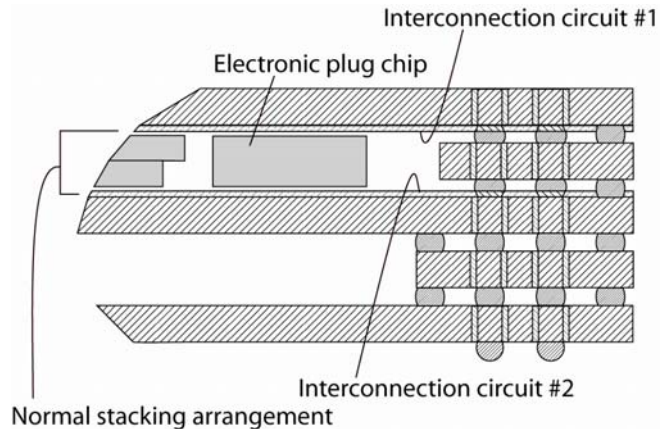
Figure 9 illustrates a stacked subsystem employing fiber optic connections at each level in the stack.



**Figure 9.** Stacked subsystem with optical connections

**INCREASED BANDWIDTH USING “ELECTRONIC PLUG”**

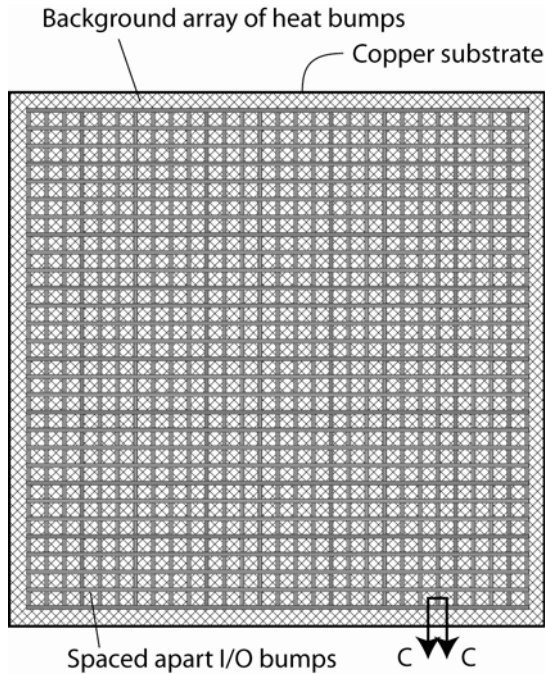
New processes have been developed by Allvia Inc.<sup>4</sup> and others for through wafer interconnects. This provides an alternative way to increase I/O bandwidth in a stacked subsystem, as shown in Figure 10, where the plug chip can provide multiple high speed signal paths between adjacent interconnection circuits on the copper substrates.



**Figure 10.** Electronic plug chip

## FLIP CHIP INTERFACE INCLUDING HEAT BUMPS AND I/O BUMPS

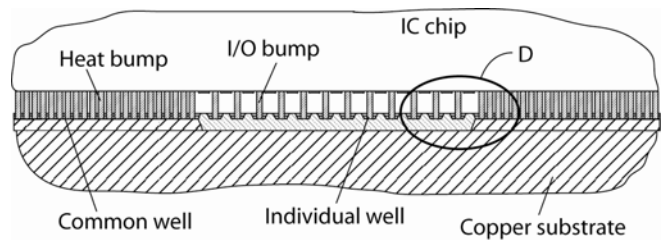
The best thermal access to active junctions on a chip occurs at the front face of the chip. It is possible to provide closely spaced copper bumps at the front face, wherein the copper bumps include a good metallurgical bond to the semiconductor surface. By this means, and by providing additional heat bumps at hot spots, excellent thermal performance can be achieved. This includes heat dissipation rates as high as  $1\text{W}/\text{mm}^2$  or  $100\text{W}/\text{cm}^2$ . A mixed array of heat bumps and I/O bumps is shown in Figure 11.



**Figure 11.** Default layout of heat bumps and rows and columns of I/O bumps

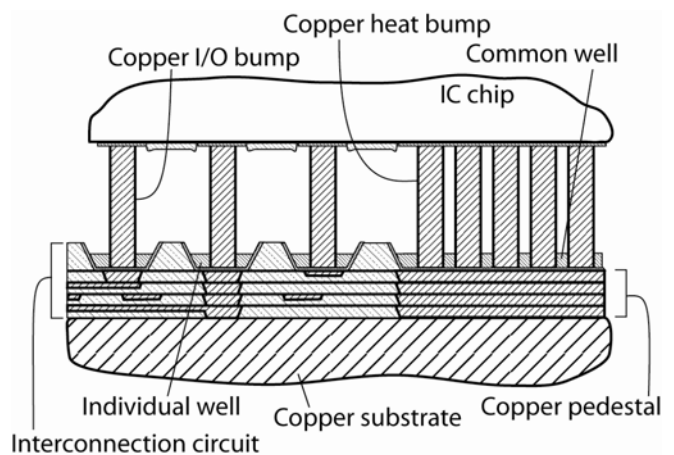
FIG. 11 corresponds to section BB of FIG. 3; it is a cross-section representing an interface between a chip and a substrate. A background array of heat bumps is shown; it is comprised of copper columns that are closely spaced for maximum heat conduction and that bend individually to relieve stress at the interface. I/O bumps are arrayed in rows and columns; they are spaced apart and connect to substrate nodes individually, as will be further described. The layout shown represents a default or starting condition; it can be adjusted to allow for localized hot spots, where a greater concentration of heat bumps can be provided. Note that the default layout shown in FIG. 11 provides a vertical connector within a millimeter or two of any location on the chip; this means that signal path lengths can be generally

short, aiding high frequency operation. Section CC in Figure 11 is expanded in Figure 12.



**Figure 12.** Expanded view of heat bumps and I/O bumps

In Figure 12, an expanded view of heat bumps and I/O bumps is shown. The heat bumps terminate in a common well and the I/O bumps terminate in individual wells, further described in Figure 13.



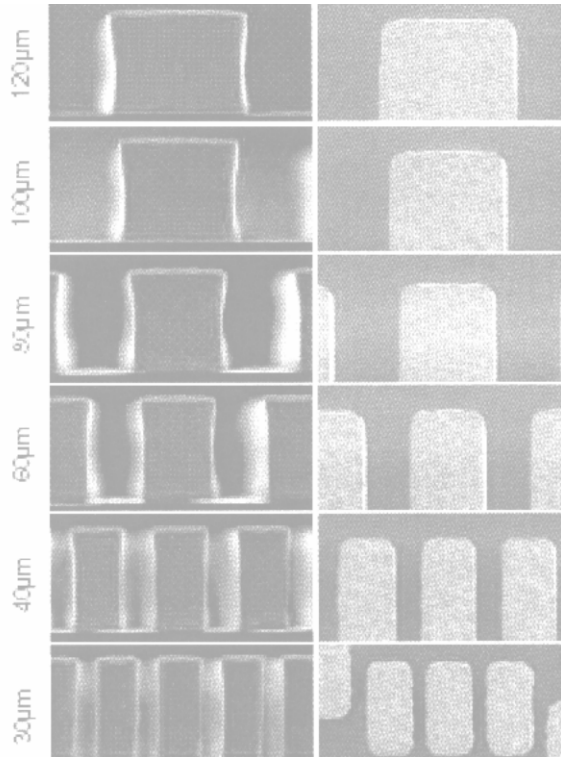
**Figure 13.** Further expanded heat bumps and I/O bumps

Figure 13 shows an expanded view of portion D of Figure 12. It shows an I/O bump paired with an individual well, providing a connection for a power supply or a signal. I/O connectors typically will have a pitch of  $80\ \mu\text{m}$ , providing over 15,000 I/O per square centimeter. This density can be used to surround each signal connector with power and ground connectors to reduce cross talk. Heat bumps are closely spaced at a typical pitch of  $30\ \mu\text{m}$ , providing a density of over 100,000 bumps per square centimeter. They terminate on a copper pedestal for maximum heat conduction. Both types of bumps are preferably  $100\ \mu\text{m}$  in height, providing adequate flexibility to relieve expansion stresses and also shock stresses for large die as well as high-powered die, without requiring an epoxy under layer. As previously mentioned, the I/O connectors have an inductance of around  $0.1\ \text{nH}$ , enabling high speed operation. The interconnection circuit is fabricated using high-

resolution build-up techniques that may include either photolithography or imprinting methods.

### MANUFACTURING THE BUMPS

The key to manufacturing the bumps is the photo resist used for the plating step. A new positive resist is well-suited to this application, Clariant Exp 100XT<sup>5</sup>. An example of achievable process definition is shown in Figure 14, as produced by the manufacturer. Resist features are on the left and plated features are on the right. It can be seen that the sidewalls are almost perfectly vertical. This resist is also easy to strip after the bumps are formed.



**Figure 14.** Plating resist and copper bump examples

### BUMP/WELL METALLURGY

It is desirable to use a well-proven solder having a higher melting point than commonly used solders; this will ensure that a stacked subsystem as described herein can be mounted to a conventional PCB using standard reflow temperatures, without compromising the integrity of the subsystem. Such a well-proven solder is 80Au20Sn, having a reflow temperature of approximately 320°C. The ability of the copper bumps to relieve stresses allows the use of a higher temperature solder without inducing unacceptable thermal stresses.

The preferred material lining the wells is copper, so both bumps and wells have a copper base. These surfaces are

preferably coated with Ti/Ni/Au. The titanium provides an adhesion layer, the Ni provides a diffusion barrier, and the gold provides compatibility with the Au:Sn solder. The gold coating is preferably at least 1000 Angstroms thick<sup>6</sup>.

### RE-WORKABILITY AND REPAIR-ABILITY

Any copper layer in the stacked construction of Figure 1 can be separated using well-directed flows of hot inert gas that melt the BGA solder bumps selectively at the layer to be removed. Re-work of a separated layer requires use of embedded test chips, in communication with a test support computer. This technique is described in a published patent application<sup>7</sup>. Once a defective chip has been identified, it can be replaced by melting the solder in the wells. This is done by heating the copper substrate to an intermediate temperature, and selectively applying hot inert gas to the defective chip using a shroud. The defective chip is removed, the remaining solder is sucked from the wells, the wells are re-filled using a miniature squeegee, and a replacement part is aligned, inserted, and the solder reflowed. The bump/well construction also supports improved testing of known good die (KGD) using liquid metal in the wells.

### MANUFACTURING COST

Costs are reduced by eliminating components such as conventional packages and PCBs and cables and connectors. More efficient testing of KGD using wells filled with liquid metal can lead to less rework and lower cost. Effective rework itself lowers cost by reducing the number of assemblies rejected due to yield factors.

Other cost advantages are methodological and more difficult to prove. For example, the proposed performance may be adequate for a wide range of applications across all of the dimensions of mechanical, thermal, and electrical design, without requiring special cases and customized design work. This is the same as saying that a standardized set of components can be used to implement a wide variety of systems, and these components follow standard design rules. If this is true, then a new technology platform is possible. This new platform can have a unified set of design rules covering all of the design aspects, and also encompassing multiple design regimes such as digital plus RF plus integrated passives for example. Such simplifications would lead to shorter time to market, fewer human errors, and lower overall cost. In addition, it is well known that computer aided design and verification tools are lagging, especially at the integration boundaries between chips and packages, packages and boards, and boards and systems. If a standard technology platform and a unified set of design rules are possible, then CAD vendors can focus on lucrative markets that serve large numbers of users. The

logical conclusion is that the tools will improve and the costs will go down.

## CONCLUSION

A repairable 3D subsystem can be constructed using a new bump/well connector for attaching IC chips to interconnection circuits on copper substrates. Such a 3D subsystem can have improved ruggedness compared with existing subsystems because conventional cables and connectors are eliminated and the chip attachments are tolerant to CTE mismatch, shock, and vibration because of their inherent flexibility. The 3D subsystem can also be well cooled using channels provided between the stacked elements. Short interconnect traces for power and signals will lead to high speed. These improved design elements can be applied to build more reliable subsystems that will typically occupy less than 1% of the volume, and weigh less than 1% of conventionally packaged designs. The cost of the subsystems will depend on the degree to which a standardized methodology will be applicable. It is believed possible that a new technology platform could evolve that would service a broad spectrum of product areas including: high-rel applications; size and weight-sensitive applications; computer servers; digital + RF systems; digital + RF + optical systems; and after the costs come down the learning curve, consumer products as well.

---

<sup>1</sup> International Technology Roadmap for Semiconductors, 2003 edition, Exec. Summ. p51 leads/chip and p53 cents/lead.

<sup>2</sup> <http://optics.org/articles/news/10/10/21/1>.

<sup>3</sup> LPKF Laser & Electronics, Wilsonville, Oregon, USA.

<sup>4</sup> Allvia Inc., Sunnyvale, California, USA

<sup>5</sup> Clariant AZ Electronic Materials, *Trends in Thick Resist for Solder Bumping – Negative & Positive Resists Compared*, presented at the APiA Seminar, Semicon West, 2004.

<sup>6</sup> K. N. Tu, *Solder Reactions in Flip Chip Technology*, IEEE/CPMT meeting, Santa Clara, California, USA, 7/21/05.

<sup>7</sup> Peter C. Salmon, *Apparatus and Method for Testing Electronic Systems*, Published U.S. Patent application, 20040176924.