FACE-TO-FACE CHIP INTEGRATION WITH FULL METAL INTERFACE

H. HUEBNER*, O. EHRMANN**, M. EIGNER*, W. GRUBER*, A. KLUMPP**, R. MERKEL**, P. RAMM**, M. ROTH*, J. WEBER**,R. WIELAND**
*Infineon Technologies, St.-Martin-Str. 76, D-81541 Munich
**Fraunhofer Institute Reliability and Microintegration, Hansastr. 27d, D-80686 Munich

ABSTRACT

We present a novel process for 3D chip integration. Chips are soldered together in face-to-face orientation with a minimum of wiring length and a maximum of in-area contacts. It is a low cost process with high performance and is well suited as a replacement for embedded technologies. The specific benefits of our new technology stem from a planar interface, made out of a thin metal layer, which can be patterned and used as an additional wiring layer.

INTRODUCTION

3D stacking technologies are under development for a long time [1], but none of these approaches has been widely accepted till now. The progress of planar integration continued and there had not been a necessity for a risky deviation from the main stream. But today the situation is changing dramatically: shrinking of main stream technologies and their related embedded processes becomes more and more expensive, the demand for high integrated solutions for small hand-held products is rapidly growing and the high end / high frequency products need interconnects, which are able to handle frequencies of 100 GHz and above. Last but not least, despite of lateral shrinking, the transmitting frequencies of the interconnect lines are decreasing, due to the growing parasitics („wiring crises“).

All these points are good reasons to think about 3D-stacking - but chip stacking technologies either are still expensive, or their performance is poor.

In our opinion, a competitive 3D-technology must meet the following requirements:

- Parallel wafer scale pre-processing of all chip layers
- Minimum number of process steps
- No processes on stacked chip level – stacking in a single and final step
- Reliable und low cost processing
- Short wiring lengths

With this in mind, we developed a low cost reliable process for the stacking of one or more chips in face-to-face (F2F) orientation onto a substrate-chip which in term should have a somewhat bigger area than the total area of the stacked top chip(s). (Fig. 1).

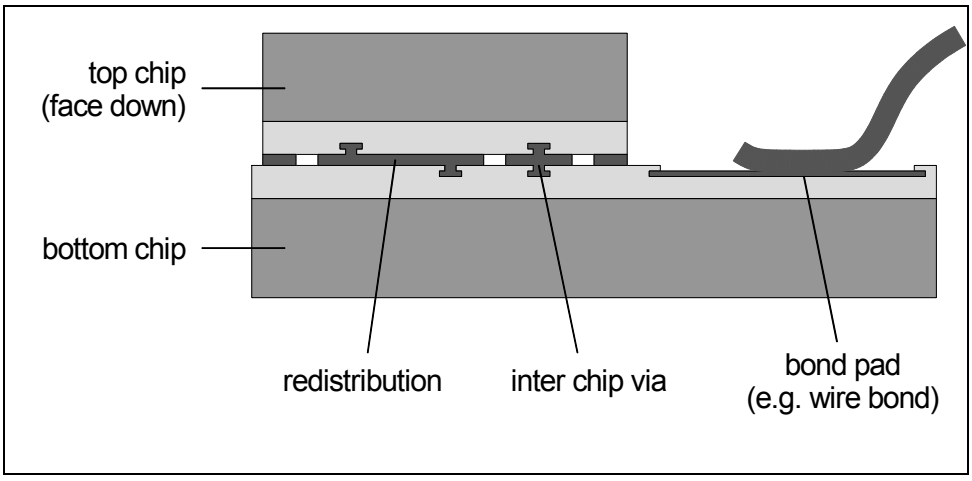


Fig. 1 Twin stack in F2F technology

The mechanical contact and the electrical inter chip vias are performed in one step by a soldering technique, called Solid-Liquid-Interdiffusion (SOLID) [2]. The external connection can be performed by conventional wire bonds or a flip chip technology as well. Despite the fact, that this process is very low-cost, its performance is maximum because the face-to-face orientation needs the minimum wiring legth between two chips. It limits the stacking to just two integration levels, but many stacking applications of interest only need two active layers and thus can be manufactured this way.

EXPERIMENT

Pre processing

We start with wafers of any size and completely processed in any technology, with the final passivation layer opened and (optionally) electrically tested. The final metallization layer can be either standard Al or Cu.

Both chips of the twin stack to be formed can be pre-processed in parallel with the same process flow (Fig. 2): The wafer is sputtered with TiW, serving as a barrier and adhesion layer, followed in-situ by a Cu seed layer. Then a wafer scale proximity lithography step is applied in order to define the future insulation areas. Inside the left open areas of the resist copper is electro-plated. Additionally, the top chips are plated with Sn or SnAg solder alloy onto the previously plated Cu layer. Finally, after resist removal, the seed and barrier layers in the insulation areas are removed using a wet chemical process sequence. If required, the wafers can be thinned now.

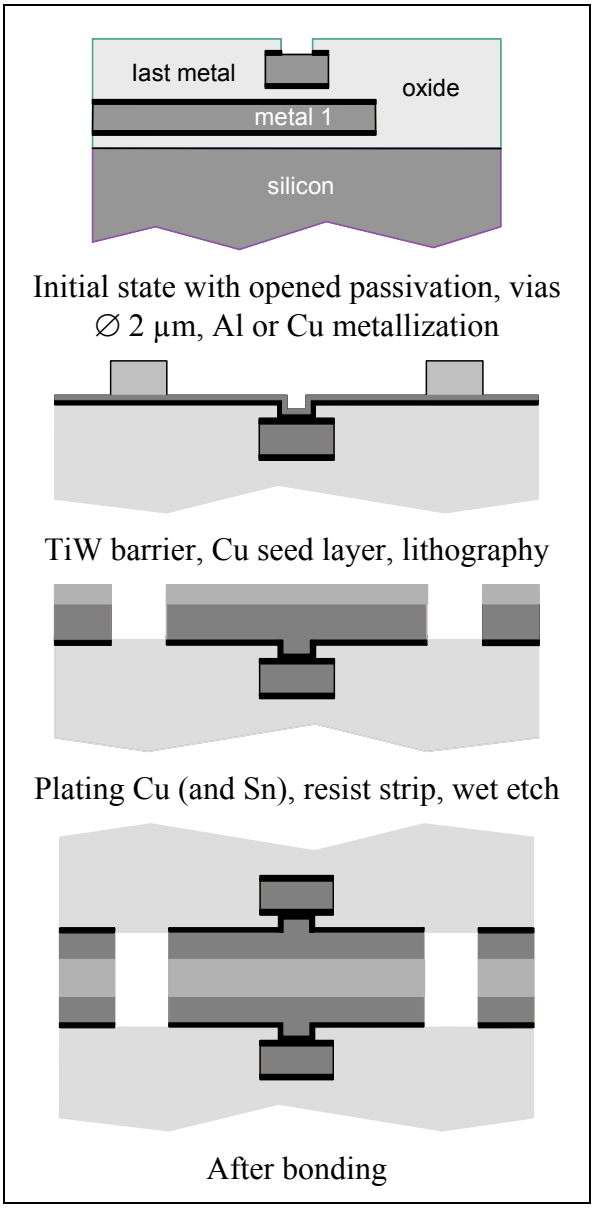


Fig. 2 Process flow

Bonding

The top chips are diced and then bonded to the chips on the non-diced bottom wafer in a two-step chip-to-wafer process. The chips are picked, flipped, dipped, optically aligned and then sticked to the bottom wafer. During the dipping step the top chips are covered with the fixing agent - a kind of a wax - in a dipping station at elevated temperature. As the bottom wafer is held at room temperature, the molten film freezes immediately when the top chips get into contact with the comparably cold wafer and fixes the chips solely by liquid-solid transition. Thus the molten film does not influence the optical alignment system. Due to the low wafer temperature, the oxidation of the Cu-metallization layer is neglible during the assembling sequence.

The pick and place procedure is done in a prototype flip chip die bonder with an alignment accuracy of 620 µm and a speed

of 4s/chip. The future production tool, which is under development by the austrian company DATACON, will achieve 6 10 µm / < 1 s/chip.

In the second step the assembled wafer will be put into a vacuum oven, where all chips are soldered in parallel. Since this batch process can be performed in parallel to the assembly, the stacking will not be delayed by the soldering step.

During the heating period the liquid Sn reacts with the copper metallization, forming high melting intermetallic phases, first the metastable Cu_6Sn_5 (η-phase), followed by the thermodynamically stable Cu_3Sn (ε-phase). The joint solidifies isothermally when all the Sn is consumed. With film thicknesses in the range of several µm, the solidification is completed within a few minutes. In the end, the joint reaches a high melting point of more than 600°C, despite the low soldering temperature of 300°C maximum.

Equipment

The wafer processes need eight process steps in three different cluster tools – sputtering, lithography and plating. The chip processes need three tools – wafer saw, bonder and vacuum oven. The process works with inexpensive standard equipment and fast and reliable standard processes.

Characteristics of the process

The process is a planar technology, which makes use of a thin metal interface instead of point contacts. This metal layer may be patterned and used as an additional wiring layer. In consequence, the location of the large inter chip contact areas is shifted from the last metal layer of the chip (LM) to the soldering interface with only small standard vias to LM. The placement of the electrical contacts can be freely designed all over the chip, including active area, but the existing bond pads also can be metallized and used as inter chip contacts.

It is a special feature of this process that areas of the metallization which do not serve as electrical interconnects, are left as a passive layer instead of removing them. This optional dummy layer is not a must, but it increases the mechanical strength of the joint. Therefore it is possible to minimize the diameter of the inter-chip contacts without causing possible mechanical stress to the contacts. These dummy structures can serve as an electrical shield in order to reduce electrical coupling between the two chips. Finally, these dummy structures provide excellent thermal conductivity, since typically 80% of the top chip area is in mechanical contact with the bottom chip. Since the joint is polymer-free, it is intrinsically more reliable than technologies which make use of an adhesion layer, anisotropic adhesives or polymer underfill.

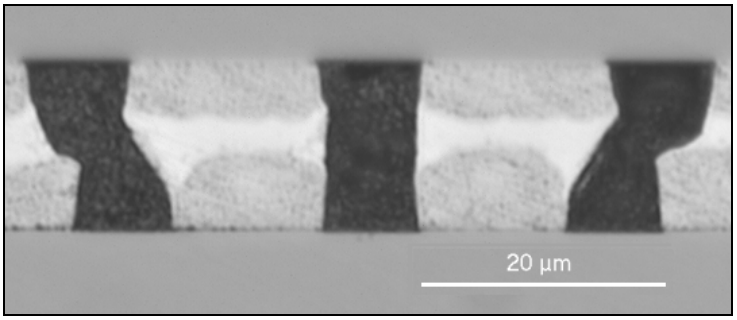


Fig. 3 Demonstration of 10 µm pads with 10 µm spacing

The maximum misalignment of the bonding tool and the resolution of the lithography limit the minimum diameter of the contacts (in order to get an overlap of the pads) and the minimum width of the insulation areas (in order to avoid shorts). Assuming a bonding accuracy of 6 10 µm, a pitch of 30 µm, corresponding to a density of more than $2 \cdot 10^5$ inter chip connects / cm² can be achieved.

RESULTS

Passive structures

We performed the first experiments with passive contact structures (v.d. Pauw structures, daisy chains, ...). The bonds have been made on the obove mentioned prototype bonder. The test contact pads have been designed with a diameter of 25 µm, in order to guarantee an minimum overlap of 5 µm. The achieved results are summarized in Tab.1.

Minimum contact dimension	10 µm with 20 µm pitch (cf. Fig. 3)
Shear strength	> 120 N @ 5 x 5 mm² chip area
Peel strength	> 40 MPa @ 5 x 5 mm² chip area
Contact resistance	7.5 mΩ @ 15 x 15 µm²

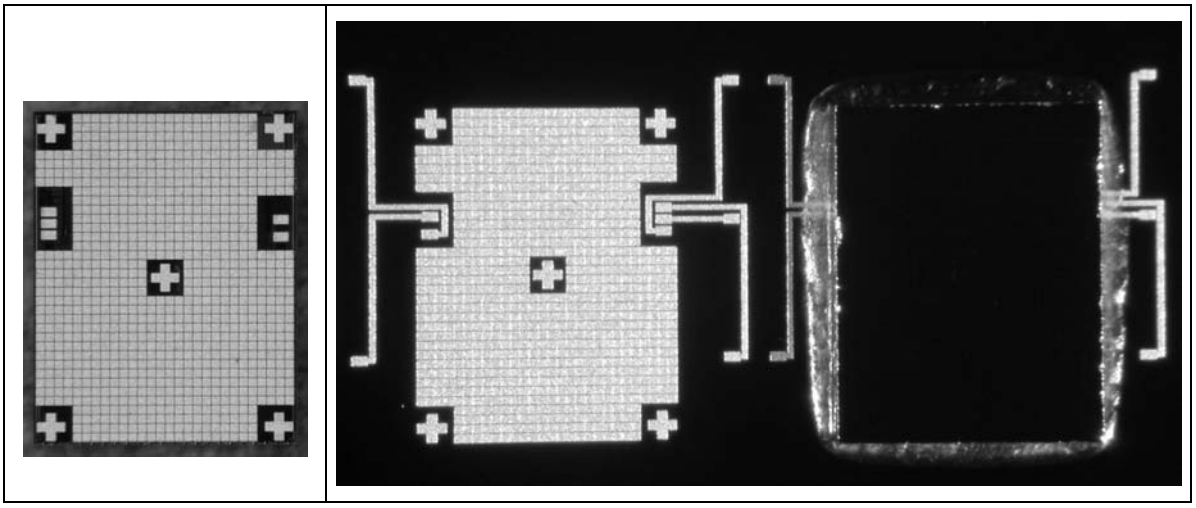
Tab.1 First results

The contact dimensions are only limited by the alignment accuracies. The process works stable and reliable. No detrimental influence of particles has been observed. This is particularly interesting, since we worked in a laboratory environment without precautions against particles. The contact resistance corresponds well to the resistivity of the copper metallization.

The parasitic stray capacity of a 20 x 20 µm² contact calculates to 55 fF, a value which is noticably better than competitive interconnect technologies. It should allow transmitting frequencies of 100 GHz and above. The heat conductance of the 10 µm thick metal interface calculates to 0.25 mW/K. This value indicates that the heat dissipation of the top chip will be dominated by the heat conductance of the inter metal oxides (IMOX). Given a total thickness of the IMOX layers of 2 x 10 µm and a power consumption of 1 W/cm² in the top chip, the temperature difference between top chip and bottom chip should not exceed 1°C.

Feasibility

In order to demonstrate early feasibility, we stacked an active top chip onto a passive bottom "chip", which solely serves as a passive redistribution layer (Fig. 4 shows the stack after assembly). The top chip is an Infineon SLE66 chip card controller, made in 200 mm / 0.25 µm embedded flash (C9-FL) technology. Although this assembly is different from



Top chip metallization (Cu + Sn) | Passive bottom redistribution with mirrored design, (Au) | Top chip pinned (squeezed-out sticking agent)

Fig. 4 Partially assembled wafer

the final product, it is well suited to demonstrate feasibility for two reasons: 1) the chip undergoes the complete processing and 2) we use a standard chip with a well known performance. The bottom metallization has been realized in gold instead of copper, enabling wire-bonding of the stack for testing. The functional and DC testing was in spec and did not show any detrimental effect of the 3D processing.

Test chip and demonstrator

Additionally, we designed a special test chip for detailed reliability tests. A complete set of test structures monitor the performance of the interconnects and the influence of the processing on the front end technology. Further, a product demonstrator has been designed, consisting of a chip card controller in C9-FL on top and a 160k memory extension on the bottom. This circuit has 52 inter chip contacts and five external bond pads. Fig. 5 shows a cross section through one of the inter chip contacts. The tests are in progress.

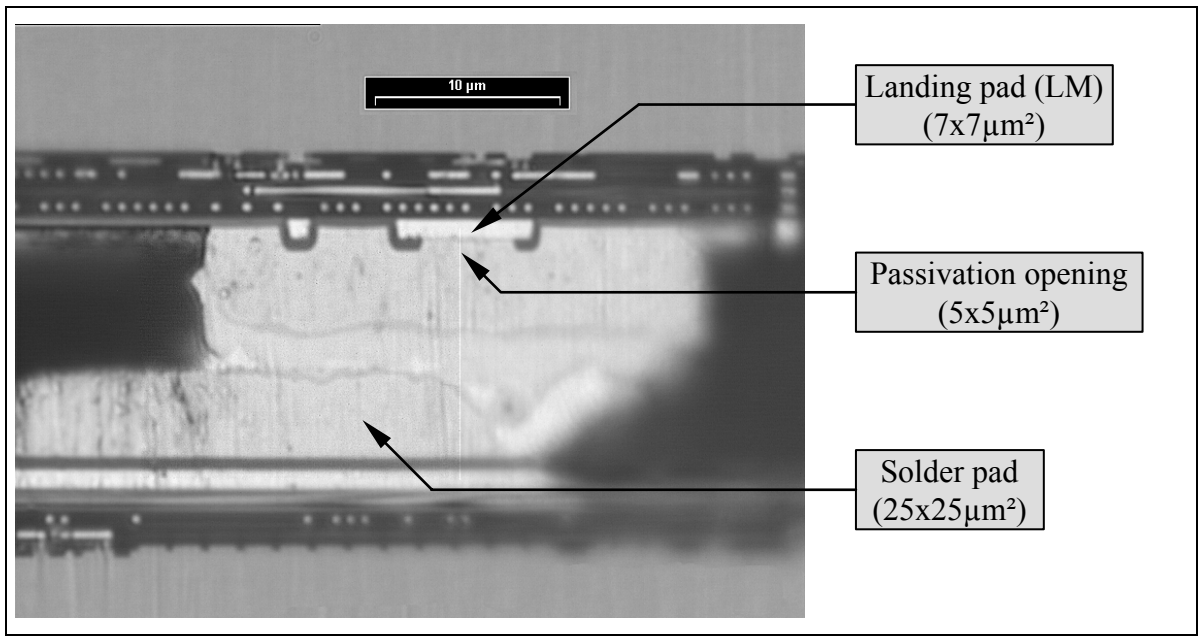


Fig. 5 Cross section of an inter chip contact (the bottom via is not visible due to lateral displacement of the chips)

CONCLUSIONS

Due to its high mechanical strength, excellent heat conduction of the interface and the superior electrical performance and high density in-area contacts, this stacking process is considered to be a chip integration technology rather than an assembly technology.

The most convincing benefit of F2F-stacking is the cost reduction in comparison to embedded processes: chip layers of different technology can be manufactured in parallel with their own dedicated processes and with maximum wafer yield. The saving of board area and of at least one chip package, the use of known-good dies only and maybe the saving of the last metal layer – if its applied only to generate the aluminum bondpads –, all these points contribute to the cost savings of the stacking process. Thus planar F2F stacking is best suited as a replacement for planar integration as e.g. Embedded Flash and DRAM, SmartPower or BICMOS. It allows to integrate quickly new technologies, as e.g. Low-k or FeRAM. Generally, planar F2F stacking helps to speed-up the time-to-market as it saves development time.

The face-to-face placement of the chips with a minimum wiring length and high performance inter chip contacts helps to overcome the wiring crises.

Addditionally, all kinds of pads – inter chip contacts as well as external IOs and testpads – can be build in the copper interface, thus saving the area of IO pads in LM, which in term can be used for additional wiring or for shrinking purposes.

The inter chip contacts could be placed without restrictions all over the chip area and, the vias of the top and bottom chips can be placed at different locations because the metal interface could be used as a redistribution layer. This allows to match different chip designs with the aid of a low cost contact lithography, making the stacking technology insensitive to future redesigns.

In summary, planar F2F-stacking is a very versatile chip integration technology with the potential to be extended to a modular chip integration process for a non-limited number of chip layers by combining with 3D integration methods using inter-chip vias [3], [4].

ACKNOWLEDGMENTS

This report is based on a project which is supported by the German Bundesministerium für Bildung und Forschung under support-no. 01M 2999 A.

The authors wish to thank the companies DATACON Semiconductor Equipment GmbH and ATV Technologie GmbH for support and helpfull inputs.

REFERENCES

[1] P. Ramm, *Trends in 3D Integration*, PLUS **2**, p. 811 (May 2000)

[2] L. Bernstein, H. Bartolomew, *Applications of Solid-Liquid Interdiffusion (SLID) Bonding in Integrated-Circuit Fabrication*, Trans. Met. Soc. AIME **236**, p. 404 (1966)

[3] P. Ramm, D. Bonfert, H. Gieser, J. Haufe, F. Iberl, A. Klumpp, A. Kux, R. Wieland, Proc. International Interconnect Technology Conference IITC 2001, p. 160

[4] J-Q. Lu et al., Proc. Advanced Metallization Conference AMC 2001, p. 151