Volume 1

1
Introduction to 3D Integration

Philip Garrou

1.1
Introduction

Wafer level 3-dimensional (3D) integration is an emerging, system level integration architecture wherein multiple strata (layers) of planar devices are stacked and interconnected using through silicon (or other semiconductor material) vias (TSV) in the Z direction as shown in Figure 1.1.

The technical and market drivers for such a new architecture are discussed in Chapter 2. Several process sequences have been developed to fabricate such stacks, which are discussed in Chapter 3. All of them depend on the following enabling technologies.

- TSV formation – realization of electrically isolated connections through the silicon substrate. The diameter of the TSV is dependent on the degree of access needed to an individual strata, which differs with application area.
- Thinning of the strata – usually to below 50 μm in memory stacks, 25 μm for CMOS silicon circuits and to below 5 μm for SOI circuits.
- Alignment and bonding – either as die to wafer or wafer to wafer. Several technologies are available (Chapter 3).

The main attribute of these stacked structures are the z axis interconnects which are usually called "through silicon vias" (TSV) but are also described as "through wafer vias" (TWV) and/or "through wafer interconnect" (TWI).

Conceptually, 3D can alleviate interconnect delay problems, while reducing chip area. When the large number of the long interconnects needed in 2D structures are replaced by short vertical interconnects this greatly enhances the performance of logic circuits. For instance, logic gates on a critical path can be placed very close to each other by stacking them and interconnecting in the z direction. Circuits with different voltage requirements and or performance requirements can also be put on different layers [1].

Handbook of 3D Integration: Volumes 1 and 2. Technology and Applications of 3D Integrated Circuits.
Edited by Philip Garrou, Christopher Bower and Peter Ramm
© 2012 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2012 by Wiley-VCH Verlag GmbH & Co. KGaA.

IPR2026-00065
TSMC EX1028 Page 1

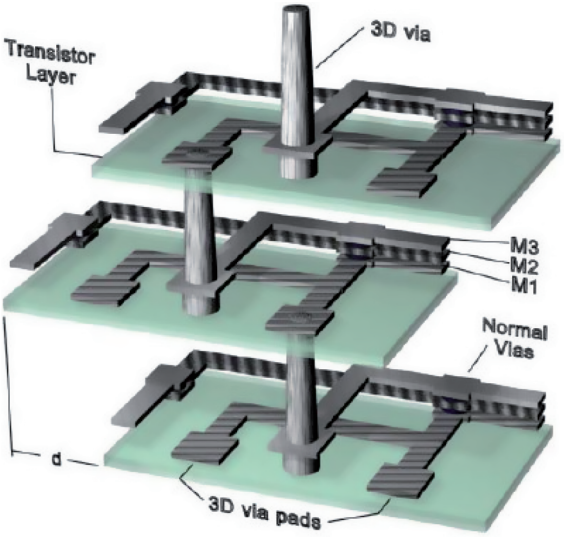

Figure 1.1 3D wafer level stacking using through silicon interconnect vias. Figure courtesy of Professor Duncan Elliott, Department of Electrical and Computer Engineering, University of Alberta.

Figure 1.2 depicts memory to logic interconnections created using a current 2D interconnection scheme, a system on chip (SOC) solution and a 3D integration solution using through silicon vias (TSV).

SOC, system-on-chip, refers to the integration of nearly all aspects of a system design on a single chip. Such chips are often mixed signal and/or mixed technology designs, which include embedded DRAM, logic, analog, RF, and so on. While this technology at first glance looks appealing, integration of such disparate technologies on a single chip dramatically increases the chip area and increases long global interconnect, which can lead to significant signal transmission delays. Since it takes

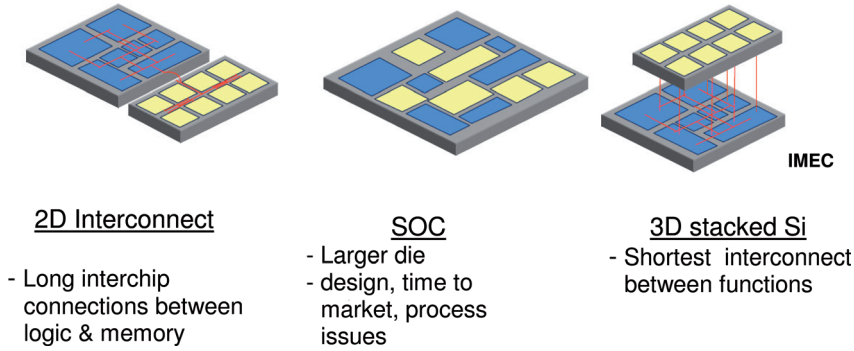

Figure 1.2 2D vs SOC vs 3D. Figure courtesy Professor Eric Beyne, MEC.

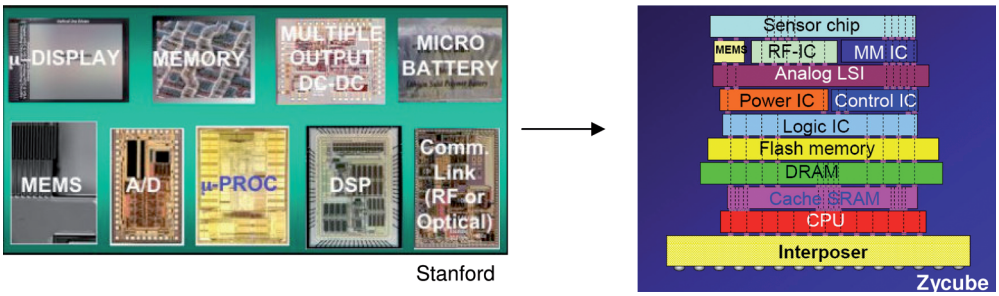

Figure 1.3 The "Holy Grail" – 3D stacking for "heterogeneous integration." Figures courtesy of Stanford University, CA, and Zycube, Japan.

different process technologies to produce these different functions, the complexity of materials and process issues is significant.

In 3D architecture, device fabrication is achieved by the production of full wafers of a specific function, that is, embedded processors, DSPs, SRAM, DRAM, and so on. These are then thinned, aligned and vertically interconnected (chip to wafer or wafer to wafer) to create a functional device. Thus, the 3D concept allows integration of otherwise incompatible technologies, and offers significant advantages in performance, functionality, and form factor. In some sectors this has become known as "heterogeneous integration." This is shown pictorially in Figure 1.3. Other technologies that could be conceivably included in the stack include antenna, sensors, power management and power storage devices.

Such technology requires both a common die size and a common interconnection scheme. We will see in Chapter 3 that the interconnecting vias can be created by fabrication in the IC foundry (FEOL) or fabrication by the assembly and packaging house after the chip is finished. If done by the latter, open areas must be left in cells, or between cells, to accommodate these interconnecting vias. While some Si real estate is consumed by such post chip TSV fabrication, a high interwafer interconnect density can be achieved with a minor area penalty.

Having shorter signal paths between die make it possible to improve the system's performance by permitting the system to run faster, it also wastes less power. Wire length is directly related to power usage, and keeping wire lengths short helps keep power use down. As we shall see later (Chapter 33) one of the concerns about using stacked-die is heat removal, but the use of TSV reduces the overall wire length, which reduces heat generation, somewhat.

1.2 Historical Evolution of Stacked Wafer Concepts

GE started to investigate the possibilities of forming electrical interconnections through semiconductor wafers for NASA in 1981 [2].

During 1986–1990 Akasaka [3] and Hayashi [4] laid out the basic concepts for and proposed technologies for 3D ICs. Later, Hayashi [5] proposed fabrication of separate devices in separate wafers, reduction in the thickness of the wafers, providing front and back leads and connecting the thinned die to each other. This was dubbed CUBIC (CUmulatively Bonded IC) and a two active layer device was fabricated in a top to bottom fashion and tested.

1.3
3D Packaging vs 3D Integration

Over the past few years, die (chip) stacking has emerged as a significant packaging option. Integrating chips vertically in a single package multiplies the amount of silicon that can be crammed in a given package footprint, conserving hand held device real estate. At the same time, it enables shorter chip to chip routing, which speeds communication between them. Another benefit is the simplification of board assembly because there are fewer components to be placed on the board.

Initial applications consisted of two-chip memory combinations such as flash and SRAM and flash plus flash. Today, chip stacking has been extended beyond memories to logic and analog ICs in packages that may also contain surface-mount passives. In addition, chip stacking has evolved to include three or four die stacks and side-by-side combinations of stacked and unstacked die within a package. The die are typically mounted to a substrate, which is bumped to create either a chip scale package (CSP) or ball grid array (BGA) as the final package.

Though chip stacking began with mounting smaller dies onto larger ones to enable wirebonding of both, packaging vendors have developed techniques for stacking same-size die or for stacking a larger die on top of a smaller one such as placing a spacer (a dummy piece of silicon) between the two. The spacer lifts the top die just enough to allow wirebonding to the bottom die. While standard wirebonding might have a loop height of 150–175 μm, die stacking could require loop heights under 100 μm. Figure 1.4 shows typical wire bonding in such 3D stacked structures.

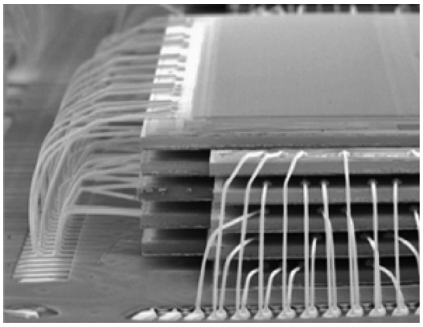

Figure 1.4 Wire bonded chips stacked in 3D package.

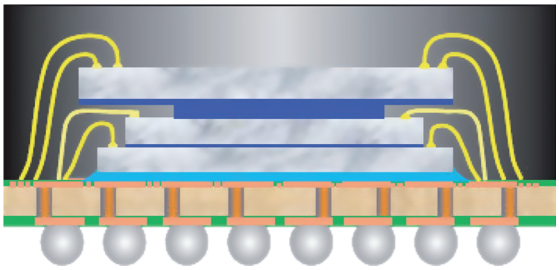

Figure 1.5 3D stacked die BGA package.

Such variations have helped expand the number of stacked-die package options, creating whole portfolios of what vendors commonly call 3-D packages. A variety of these 3D BGA packages are now in high-volume production (e.g., Figure 1.5).

The number of dies that can be stacked depends on the required thickness of the final package and the thickness of each layer (substrate, die, spacers, and BGA ball diameter) within the package. Typical ball diameters range from 0.75 mm for 1.27-mm pitch down to 0.2 mm for 0.35-mm pitch. Package height of 1.4 mm was the standard for stacked-chip packages in portable applications. Demand has recently shifted to 1.2- and 1.0-mm high packages, and even 0.8 mm is a possibility. It is currently possible to build three- and four-die stacks in 1.4-mm packages.

Another 3D packaging alternative is called package stacking or PoP (package on package). While package stacking increases material costs per package and overall package height, it provides higher yields per stacked device, which lowers cost. Package stacking needs thin, flat, high-temperature, moisture-resistant packages to handle the multiple reflows and rework associated with SMT. Vendors like Amkor have been developing processes to stack CSPs and BGAs (Figure 1.6).

Digital camera and cell phone applications are currently stacking two packages for logic + memory architectures. High density DRAM and Flash memory modules are stacking up to four packages high, with this capability demonstrated to eight high stacks.

3D packaging technology does:

- Thin die to save weight and volume.
- Stack die to save x–y space (wire bonding).

3D packaging technology does not:

- Minimize interconnect or enhance electrical performance (C and L parasitics).

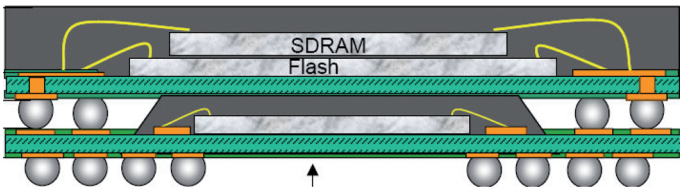

Figure 1.6 Typical Amkor PoP.

1.4 Non-TSV 3D Stacking Technologies

1.4.1 Irvine Sensors

Irvine sensors started delivering stacked Si memory from the IBM Burlington factory in 1992. In their first generation process, gold rerouting metallization was added to bring all signals to an outer die edge, and then the wafer was diced. The die were stacked, and the stack lapped to expose the ends of the gold rerouting metal. Bus metallization was deposited on the side of the stack, interconnecting the dice and a ceramic top cap substrate was added, which allowed signals into and out of the stack [6, 7]. Limitations to this technology included:

- All dice must be the same size, limiting the stack to a single die type.
- Frequent die shrinks required substantial retooling.
- The trend in commercial wafers is for street widths to shrink, which made the process more and more difficult.

Their newer technology, known as neo-stack, addresses previous limitations. In the Neo-stacking approach KGD are bumped using a gold wire bonder. A new wafer, or "Neo-wafer," is constructed using many of the bumped dice in a potting compound matrix. A standard Neo-die size, slightly larger than the largest die in the stack, is used for all dice in the stack. This feature allows the stack to be heterogeneous. Blank silicon is added to open areas on layers where smaller dice are used to enhance thermal conduction between layers. The Neo-wafer is metallized and thinned before dicing into individual Neo-die. Other die types are similarly fabricated into Neo-die of the same dimension. All of the necessary dice are then laminated into a single stack, with all signals to be interconnected brought out to two sides of the stack. On the top of the stack is a cap, with metallization on both sides, connected through vias. Metallization is added to the two sides of the stack to complete the interconnection between dice, bringing all input/output signals to the cap chip.

Figure 1.7 shows a cross section of the "Neostack" is. A Flash neo-stack and its composite layers are shown in Figure 1.8.

1.4.2 UTCS (Ultrathin Chip Stacking) IMEC, CNRS, U. Barcelona

Similar, though not identical technology has been proposed by IMEC, CNRS, U. Barcelona [8–10]. This technology named Ultra Thin Chip Stacking (UTCS) can be fabricated by the following sequence: chips are thinned down to 10 μm, interspaced with BCB dielectric layers and the vertical interconnection is achieved with metallized vias. The final stack is significantly thinner than the individual silicon chip.

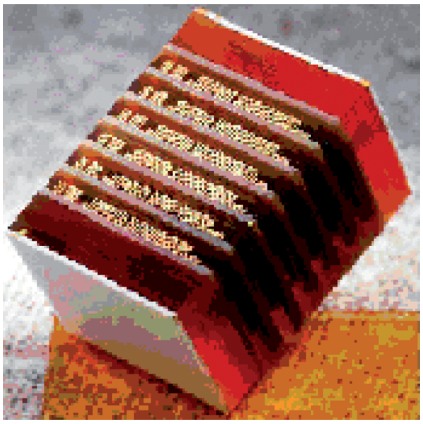

Figure 1.7 Irvine sensors neo-stack wafer concept [7].

BCB, the adhesive and planarization layer, has poor thermal conductivity, which degrades the heat extraction efficiency through the vertical path. Heat extraction is vastly improved by the use of copper grids or full metal plates to remove heat from the thinned chips.

Figure 1.9 shows schematically how the chips are interconnected. The detailed procedure is shown in Figure 1.10.

1.4.3 Fujitsu

In the summer of 2002 Fujitsu introduced its similar CS Module based on wafer-thinning, chip stacking and re-distribution technologies [11]. The technology stacks chips into two layers and redistributes signal circuitry between them. Figure 1.11

Figure 1.8 Flash neo-stack and its composite layers.

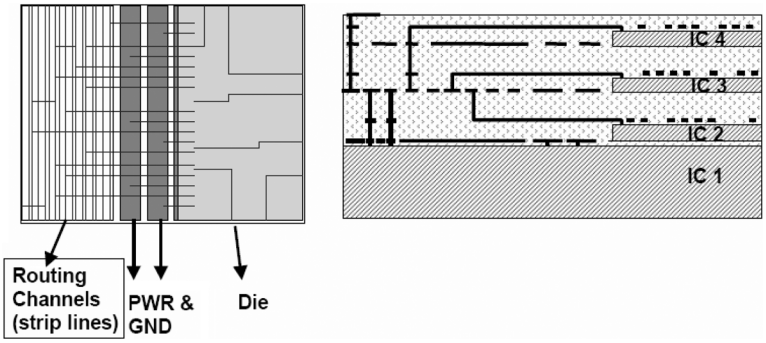

Figure 1.9 UTCS 3D routing in thin film dielectric [8].

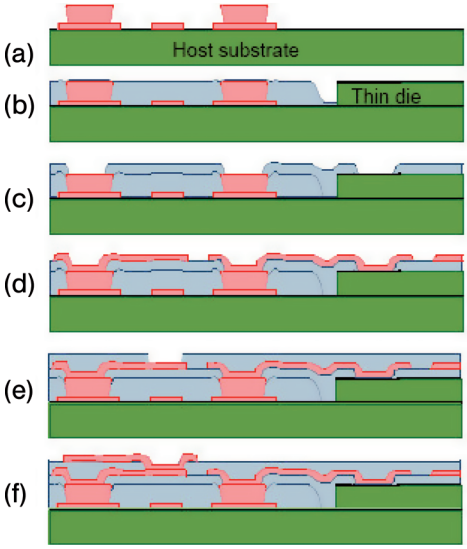

Figure 1.10 UTCS process sequence: (a) Patterning of the first interconnection level and growth of studs; (b) transfer of the thin die, deposition of a thick photo-BCB layer, opening of the cavity around the chip and vias on studs; (c) deposition of BCB planarization layer, opening of contacts and dry etching to remove BCB residues; (d) patterning of the second metal layer; (e) deposition of an insulating and planarization BCB layer; (f) patterning of the contact metal layer for pad definition. [8].

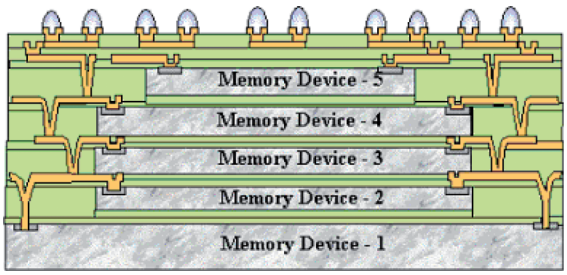

Figure 1.11 Fujitsu chip scale module memory stack [11].

shows a schematic of a five-layer, high-capacity memory product with four chips stacked on top of the base-level memory device.

1.4.4 Fraunhofer/IZM

Fraunhofer/IZM have a similar technology they call "Chip in Polymer" [12]. It is based on embedding of ultrathin chips into build up layers on a printed circuit board.

Working with Nokia, Philips, AT&S Datacon and IMEC under a European STREP (specific targeted research project) they have attempted to determine whether such processes are suitable for manufacturing [13]. Figure 1.12 shows the process sequence they developed.

The structures consist of a double layer core fabricated from high Tg FR4. The die are bonded to the top surface of the core and then high Tg RCC (resin coated copper) is bonded to both sides of the laminate substrate. The RCC has a Cu thickness of 5 µm and a dielectric thickness of 70 µm. Vias are laser drilled and then plated after typical desmear and electroless seeding.

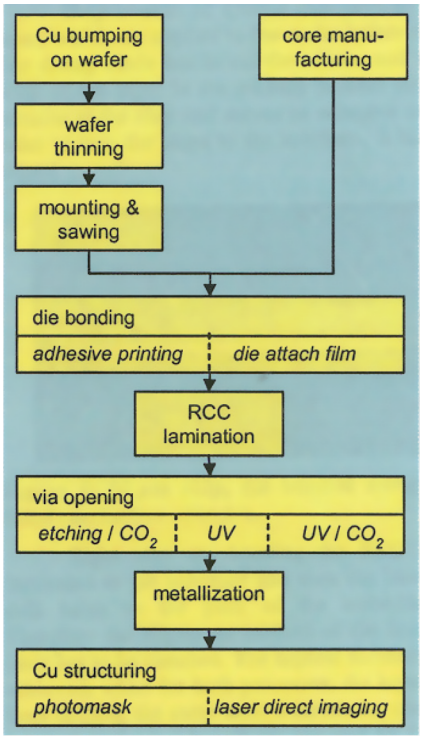

Figure 1.12 Fraunhofer/IZM chip in polymer process flow [13].

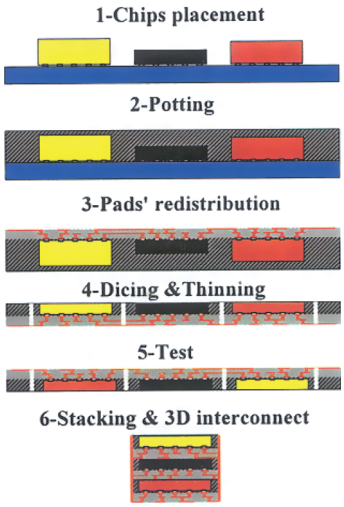

Figure 1.13 Leti – 3D plus "re-built wafer" technology [14].

1.4.5
3D Plus/Leti [14]

Leti and 3D Plus have proposed the 3D structure shown in Figure 1.13 which they call "re-built wafer." In this approach chips and passives of various sizes are imbedded into a resin matrix (active side down). The pads are redistributed, the

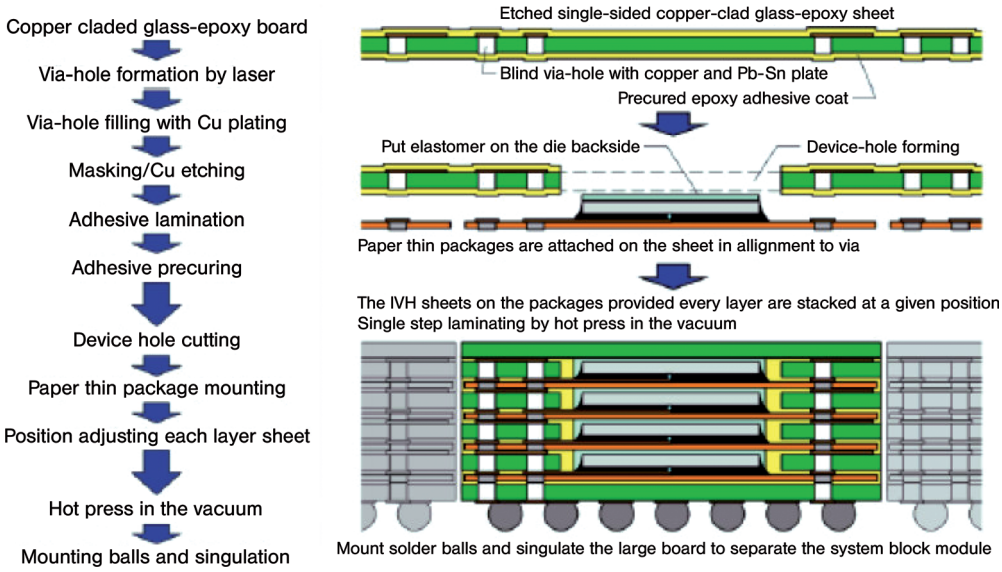

Figure 1.14 Toshiba system block module process flow [15].

imbedded substrate is thinned and the devices are re diced to equal size. After testing, the devices are stacked and connected using 3D Plus edge connecting technique. Two processes were studied for the redistribution: BCB/Cu and laminated film/Cu. The BCB/Cu process was reportedly complicated by "substrate" warpage after thinning.

1.4.6
Toshiba System Block Module [15]

The fabrication process proposed by Toshiba for their "System Block Module" is shown in Figure 1.14 [15].

References

1 Banerjee, K., Souri, S., Kapur, P. and Saraswat, K. (2001) 3D ICs: a novel chip design for improving deep sub-micron interconnect performance and subsystems-on-chip integration. *Proceedings of IEEE*, **89**, 602.

2 Anthony, T. (1981) Forming electrical interconnection through semiconductor wafers. *Journal of Applied Physics*, **52**, 5340.

3 Akasaka, Y. and Nishimura, T. (1986) Concept and basic technologies for 3D IC structure. *IEDM Technical Digest*, 488.

4 Kunio, T., Oyama, K., Hayashi, Y. and Morimote, M. (1989) 3D Ics having 4 stacked active device layers. *IEDM Technical Digest*, 837.

5 Takahashi, S., Hayashi, Y., Kunio, T. and Endo, N. (1992) Characteristics of thin film devices for a stacked type MCM. *Proceedings IEEE MCMC*, 159.

6 Bertin, C., Perlman, D. and Shanken, S. (1993) Evaluation of a 3-D memory cube system. *IEEE Transactions CHMT*, **16**, 1006.

7 Gann, K. (1998) High density packaging of flash memory. Proceedings IEEE Int. Non Volatile Memory Technology Conference, p. 96.

8 Pinal, S. Tassleei, J. Lepinois, F. *et al.* (2001) Ultra thin chip verticle interconnect technique. *Proceedings IMAPS Europe*, p. 42.

9 Pinel, S. *et al.* (2002) Thermal modeling and management in ultrathin chip stack technology. *IEEE Transactions CPMT*, **25**, 244.

10 European Patent UTCS EP 992011061.

11 Fujitsu press release July 15th 2002 at Semicon West.

12 Reichl, H., Ostermann, A., Weiland, R. and Ramm, P. (2003) The 3rd dimension in microelectronic packaging. Proceedings 14th European Micro & Packaging Conf. Friedrichshafen GR, p. 1.

13 Ostmann, A. *et al.* (2005) Technology for embedding active die. Proceedings European Microelectronic Packaging Conference, Brugge BE, p. 101.

14 Souriau, J.C, Lignier, O., Charrier, M. and Poupon, G. (2005) Wafer level processing of 3D system in package for Rf and data applications. Proceedings Electronic Component and Technology Conference, p. 356.

15 Imoto, T. *et al.* (2001) Development of 3-dimensional module package, system module block. Proceedings 51st Elect. Component Technology Conference, Orlando, p. 552.