3D SYSTEM INTEGRATION TECHNOLOGIES

Eric Beyne
IMEC
Kapeldreef 75, B-3001 Leuven, Belgium
ERIC.BEYNE@IMEC.BE

ABSTRACT

Electronic interconnection and packaging is mainly performed in a planar, 2D design style. Further miniaturization and performance enhancement of electronic systems will more and more require the use of 3D interconnection schemes. Key technologies for realizing true 3D interconnect schemes are the realization of vertical connections, either through the Si-die or through the multilayer interconnect with embedded die.

Different applications require different complexities of 3D-interconnectivity. Therefore, different technologies may be used. These can be categorized as a more traditional packaging approach, a wafer-level-packaging, WLP ('above' passivation), approach and a foundry level ('below' passivation) approach. We define these technologies as respectively 3D-SIP, 3D-WLP and 3D-SIC. In this paper, these technologies are discussed in more detail.

INTRODUCTION: WHY 3D?

As the semiconductor roadmap strides on, packaging and interconnection technologies are required to follow. In order to stay in pace with system demands on scaling, performance and functionality 3D integration is gaining a lot of interest as a solution to this demand.[1] The reasons and requirements for 3D integration are however very diverse and often application specific.[2,3,4,5,6,10]

A basic reason for 3D-integration is system-size reduction. Traditional assembly technologies are based on 2D planar architectures. Die are individually packaged and interconnected on a planar interconnect substrate, mainly printed circuit boards. The area-packaging efficiency (ratio of die to package area) of individually packaged die is generally rather low (e.g. 5x5mm die in 7x7mm package: 50% area efficiency) and an additional spacing between components on the board is typically required, further reducing the area efficiency (for example above e.g. 1mm clearance: 30% area efficiency). If we consider the volumetric packaging density, the packaging efficiency drops to very low levels. If in the previous example, we consider the active area of a die to be about 10 µm, and the combined package and board thickness to be 2 mm, the volumetric packaging density is only 0.15%. There is clearly room for improvement of the packaging density.

A different reason for looking at 3D integration is performance driven. Interconnects in a 3D assembly are potentially much shorter than in a 2D configuration, allowing for a higher operating speed and smaller power consumption. This is of particular interest for advanced computing applications. Due to the rising on-chip clock speeds, only a limited distance may be traveled by a signal in a synchronous operating mode. Using 3D-IC stacking techniques, more circuits may be packed in a single synchronous region. This requires a technology with 3D interconnects with low parasitics; in particular low capacitance and inductance are needed to avoid additional signal delay. The interconnection of circuit elements can be performed at several levels of the on-chip hierarchy. Of particular interest is the 3D stacking at the so-called "tile-level". As shown in figure 1, typical system-on-chip, SOC, devices are constructed of a number of functional blocks. The longest on-chip lines are those that are used to interconnect these tiles. These lines are typically in the top-on-chip interconnect layers and are referred to as 'global ' interconnects in the on-chip wiring hierarchy. Within the tiles, 'local' and 'intermediate' wiring hierarchy levels are mainly used. In a 3D approach, the large die is split in a number of smaller die, using the 3D interconnects as 'global' interconnects between the tiles on both die. As this interconnect goes one or more levels down the traditional IC-pad level, a very high 3D interconnect density is required for such an application.

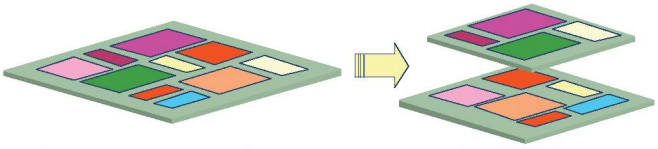

Figure 1: Conceptual view a 3D stacked SOC. Functional 'tiles' on the die are rearranged in multiple die that are vertically interconnected, resulting in much shorter global interconnect lines.

A third, and maybe most important, reason to consider 3D integration is so-called hetero-integration. As silicon semiconductor technologies continue to scale (vertical scaling), the realization of true SOC devices with a large variety of functional blocks becomes very difficult to achieve. Technologies need specific optimization for logic, analog, memory etc. to reach the desired performance levels and circuit density. Furthermore, the substrates used to build active devices may vary significantly between technologies, including non-silicon substrates, e.g. compound semiconductors. Also systems may contain other planar components, such as MEMS and integrated passive devices. Besides the 'vertical' scaling we are also experiencing a 'horizontal' scaling. Realizing the full system on a single SOC die is becoming increasingly difficult and often not economically justified. If however a high-density 3D technology is available, a "3D-SOC" device could be manufactured, consisting of a stack of heterogeneous devices. This device would be smaller, lower power and higher performance than a monolithical SOC approach.

Such an approach is the obvious choice for many sensor-array applications. Many sensor applications use particular substrate materials, such as IR and X-ray sensing, that are incompatible with Si-CMOS processing. These applications require however high-density circuits to read-out the signals from individual sensor pixels, a requirement best met with advanced CMOS technologies. The solution therefore consists in flip-chip (3D) mounting the sensor-array on a read-out electronics chip.

1-4244-0182-8/06/$20.00 ©2006 IEEE

Another possible application for this approach is the combination of logic and memory (see figure 2).

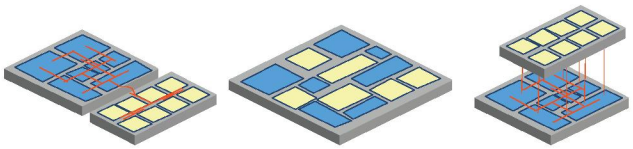

Figure 2: Different approaches for combining logic and memory circuits. Left: 2D interconnect between logic and memory die; Center: (2D-SOC) combined logic and memory device; Right: "heterogeneous 3D-SOC" stacking of a memory and logic device with 3D interconnects between individual logic tiles and memory banks.

Most applications require a combination of logic and memory. When large amounts of memory are needed, the memory is realized as a separate die, using a high density, optimized memory technology. Due to the use of large busses on the logic and memory die and the use of off-chip interconnects, only a relatively slow and power-hungry interconnect between memory and logic is possible. To overcome these limitations, e.g. for real-time data processing applications, a SOC approach is typically used. Although not optimal for the integration of high-density memory, the IC logic technology is used for integrating large amounts of memory. This allows for allocating smaller pieces of memory (memory-banks) to specific logic blocks. Distance between logic and memory is short, resulting in the required performance. The integrated memory is however of the same performance as dedicated memory technologies would offer. In particular, a much larger die area is consumed by the memory cells, resulting in a die are that is significantly larger than the case with 2 die solutions. 3D interconnect technology may solve this problem, by allowing for logic 'tiles' on a first die to directly access memory banks on a memory chip. In this case the number of 3D connections required from the memory die to the logic die will increase by an order of magnitude compared to the I/O count of standard memory devices. Similarly as for the example shown in figure 1, this approach uses 3D interconnects as "global-on chip" interconnect layers to realize a "heterogeneous 3D-SOC" structure.

3D TODAY

Currently, one particular type of 3D-IC packaging is highly popular, the so-called stacked die package, shown schematically in figure 3. [7] In this technology, standard BGA package technology is used to create a 3D-die stack. The individual die are thinned down aggressively – down to about 50 µm – and are glued on top of each other. The die are connected to the base interposer substrate by wire bonding. This method is used to stack as many as five functional die in a single package. It is particularly popular for portable applications where a processor die is combined with several types of memory die (ROM/RAM/Flash…) in a very small volume. Packaging area efficiencies of 100 to 300% may be obtained. The volumetric packaging density (considering 10 µm of "active" Si-thickness) is 0.5 to 1.5% (up from only 0.15% for CSP packages), still relatively low.

The BGA interposer substrate has typically only a very limited capability for routing signal lines, other than the traditional bond pad-to-solder ball connections. Increasing the interposer wiring capabilities requires the use of additional wiring layers and finer line/space board technologies, both resulting in a higher cost package. In fact, only in the case of memory die a relatively simple interposer can be used.

Stacking many die on each other also results in very long wire bonds at the top layers, typically several mm in length. This results in connections with several nH of inductance. Furthermore, the closely spaced, but long, wire bond busses result in significant cross talk. This method is therefore not very well suited to high-speed circuits (including fast memory). Another limitation is that the technique does not allow for area-array contacts to the die, which is also often combined with high speed die.

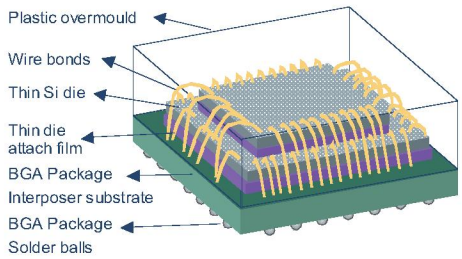

Figure 3: Schematic drawing of a BGA package with multiple stacked die and wire bond interconnects

A final issue with the wire-bond stack is the requirement for "known-good-die", KGD, to avoid compound yield problems. This is why some companies have altered this technique to also include pre-packaged and tested die in such a die and package stack. This will however increase the package cost and lower its volumetric density.

3D DIE STACKING NEEDS

It is clear that the wire-bonded die-stack will not offer the generic 3D packaging solution for advanced systems and scaled semiconductor devices. A number of requirements for 3D packaging and interconnection technologies can however be put forward:

The technology should allow for high-density 3D interconnects. Peripheral interconnects, such as wire-bonds, are inherently limited in density as the pitch has to decrease linearly with the wiring demand, whereas in an area-array situation, the pitch has only to decrease with the square root of the wiring demand.

High speed and low power applications both demand shorter interconnects with low parasitic capacitance. For high speed, also low inductance is of high importance.

Any 3D-technology that crosses through the Si-die area should have minimal impact on the FEOL (Front-end-of-line, the active die area with the transistors) and the BEOL (Back-end-of-line, the on-chip interconnect layers). Large numbers of large 3D-via connections may block large die-areas, where active circuits and interconnects must be excluded. This will cause the die to require a large Si-area on the wafer, increasing the die cost and defeating the purpose of 3D stacking, which is to miniaturize the system and shorten the circuit interconnect lengths. Actually, the loss in die area may be considerably larger than the actual via size, as the routing of the circuit cells becomes more complicated when large areas of the BEOL are blocked.

A 3D stacking technology should allow for different die sizes. In general, when assembling die made in different technologies, using combinations of existing and newly designed die, it is very unlikely that the die sizes will match. Furthermore, when using heterogeneous integration of various technologies, the wafer size of the different die may not match (300mm, 200, 150 mm and even smaller diameter sizes will continue to co-exist for different technologies such as memory, logic, analog, rf, high voltage and compound semiconductors). Wafer-to-wafer bonding for 3D stacks will therefore be limited to 3D IC-stacks where all layers are realized in the same or similar technology. The main applications for this are memory stacks with a single type of memory and high performance logic wafers, where advanced CMOS-wafers are vertically stacked to allow for packing more transistors in a synchronous region of space.

A significant treat to 3D packaging is the so-called "known-good-die" (KGD) problem. When combining n untested die from wafers with a die yield Y_i, the compound yield of the structure will be $Y_m=Y_s \times Y_i^n$ (with Y_s the yield of the interconnect and packaging process). As an example, combining 5 die with a processing yield Y_i of 80% and an assembly yield Y_s of 95%, results in a module yield of only 31%. Lower wafer yields, result in exponentially smaller module yields, e.g. Y_i=70%: Y_m=16%. Such yield loss can not be avoided when using wafer-to-wafer 3D bonding techniques. Technologies that allow for die-to-die or die-to-wafer bonding may introduce a component-screening test to increase the confidence level in the die to a "Good-enough-die" level, e.g. 95 or 97%. This can be done using relatively simple test schemes. For the example above, raising the die yield level to 97% would result in a more acceptable 82% module yield.

3D stacking schemes should also consider the thermal management of the module. The key issue for thermal management in an electronic system is how to transfer the heat generated by a localized heat source (the active silicon die area) to the ambient environment, generally the air surrounding the devices. By stacking the die in a small volume the total heat dissipation of the system may be reduced, due to the shorter interconnect lengths. However the local heat density in the system will dramatically increase. The thermal problem becomes twofold: getting the heat out of the stack to the 'package' boundaries and getting the heat from the small package to the environment. The first problem requires the use of highly thermal conductive materials in the package and the use of thin layers of in the package build-up, in particular for the electrically insulating and 'gluing' layers which generally posses a poor thermal conductivity compared to metals or silicon.

Finally, last-but-not-least, the process for realizing a 3D stacked device should be cost effective. The main implication of this factor is that no single generic 3D-packaging solution will be possible. The 3D technology should be chosen to fit with the requirements set forward by the application. A technology with a very high 3D wiring density may be 'overkill' for an application requiring only a moderate number of interconnects.

In order to achieve the goal of a cost-effective 3D process, a number of technology process requirements may be put forward:

- The technology should maximize collective processing. This favors a wafer-level approach. Although at some stage in the process individual die will need to be handled, because of compound yield issues (which also significantly impact cost).

- The process should maximize the amount of parallel processing:

  - Wafers (die) should be prepared separately for 3D stacking
  - The process should allow for die screening to obtain "good-enough-die", e.g. using self-test and IDDQ testing methods.
  - Preferably a Die-to-Wafer placement is performed (with KGD on KGD), followed by a collective bonding step of the individual die at the wafer level.
  - The 3D stack is build by repeating this process with a minimum of sequential processing steps.

Processes that are fully sequential (e.g. wafer-to-wafer bonding, followed by a contact formation process, followed by additional sequential wafer-to-wafer bonding and contact formation processes) suffer from additional yield loss inherent to processes with a large number of sequential steps and also due to the large process strain put on the die that are placed first in the stack.

TECHNOLOGIES FOR 3D

In literature one can find a very large number of different approaches to 3D stacking of die. Generally solutions are sought, starting from an available technology. Apart from the large variety of techniques proposed, also a large variety of achievable package density and 3D interconnectivity are observed.

In an attempt to categories these technologies, one can start from the technology platform ("factory-type") used to create the 3D interconnect structures. We identify 3 major technology platforms for 3D integration, based on the underlying infrastructure:[2]

- 3D-SIP: Packaging infrastructure
- 3D-WLP: Wafer-level packaging infrastructure
- 3D-SIC: IC-foundry infrastructure

The 3D-SIP technology encompasses the packages with wire-bond die-stacks, but also involves package-on-package 3D stacks. It is currently the most mature technology and in high volume production. A relatively low packaging density characterizes 3D-SIP.

The 3D-WLP technology is based on wafer-level packaging infrastructure, as used for flip chip bumping and redistribution metallisations. Using additional technology elements developed for MEMS-technology, such as deep anisotropic Si-etching, 3D electrical connections can be realized at the wafer level. This technology allows for higher integration densities than 3D-SIP. In order to realize cost-effective 3D-WLP stacks, a die-to-wafer and parallel processing route should be explored. Many approaches in this field are proposed and several are being applied in products today.

The interconnects between the die in a 3D-WLP can be at the traditional chip I/O boundaries, or at the global on-chip interconnect layer. In the latter case we can, from a system point of view, consider the 3D-WLP as a heterogeneous "3D-SOC", as discussed in the introduction.

The 3D-SIC approach uses the Si-foundry technology to create very high density vertical interconnects. Many technologies are proposed in this area, but so far they are still in the R&D phase. This type of 3D stacks can be divided in two classed.

- A first class consists of wafer stacks where relatively large circuit blocks "tiles" are interconnected in a 3D fashion. The 3D interconnects mainly correspond to global and possibly intermediate BEOL on-chip interconnects. In this case we can also consider the 3D-SIC as a "3D-SOC". Stacking is in this case also preferably realized using die-to-wafer bonding.
- A second class of high-density wafer stacks aims at connecting small circuits, logic gates and even transistors in a 3D manner. This requires interconnects at the local BEOL wiring hierarchy. Such a device could be considered a true 3D-IC.

As the number of local interconnects exceeds the number of global interconnects on a die with multiple orders of magnitude, an extremely large 3D wiring density must be achieved. This requires extremely small and narrow pitch 3D interconnects. At e.g. a 45nm node, the via pitch should be below 1 µm. Furthermore, the area blocked by these connections will be very large compared to the available die area, significantly reducing the active device density.

The 3D-IC technology requires a wafer-to-wafer bonding approach and is sequential in nature. Wafers with FEOL layers will be stacked first with local 3D interconnects realized after bonding each individual wafer level. Only when all "local" layers are finalized, the BEOL intermediate and global interconnect layers will be added. These layers are common for all layers. Considering the fact that the BEOL layers are one of the significant bottleneck for the success of future die-shrinks, it can be anticipated that this will also be the case for 3D-IC's, therefore requiring a large numbers of interconnect planes.

For the reasons outlined above, a "3D-SOC" approach to 3D-SIC is more likely to be successful and economically viable than a pure 3D-IC approach. The 3D-IC approach is restricted to special applications, such as the artificial Retina chip from Tohoku university[12], where the 3D-IC approach offers unique possibilities, however at a high cost.

3D TECHNOLOGIES AT IMEC

3D-SIP FOR BUILDING HIGHLY MINIATURIZED SYSTEMS

3D-SIP is of particular interest when it is used as a stacking technology of SIP packages. Consider a system composed of a number of clearly defined sub-systems. Each sub-system could be integrated in a system-in-a-package fashion, using the appropriate packaging technology for that particular subsystem. At the end, the SIP sub-systems can be stacked in the 3rd dimension by a collective process, creating a 3D-SIP system solution.

As the layers of the 3D stack are SIP's by themselves, only a modest 3D interconnect density is required. Also testing of the different SIP layers is greatly simplified.

An example of such 3D-SIP integration scheme, realized at IMEC, is a fully integrated low power rf transceiver shown in figure 4. This device measures only 7x7 mm. It consists of two CSP-type devices (CSP="Chip-Scale-Package"). The top CSP is realized using IMEC's rf-MCM-D technology with integrated passives.[11] The bottom CSP is a double-sided high-density printed circuit board with a high density flip chip die on the bottom side and several discrete passive components mounted on the topside. The connection of this bottom part to the top part is obtained by using solder balls on the topside of the bottom laminate and encapsulation of the topside devices.

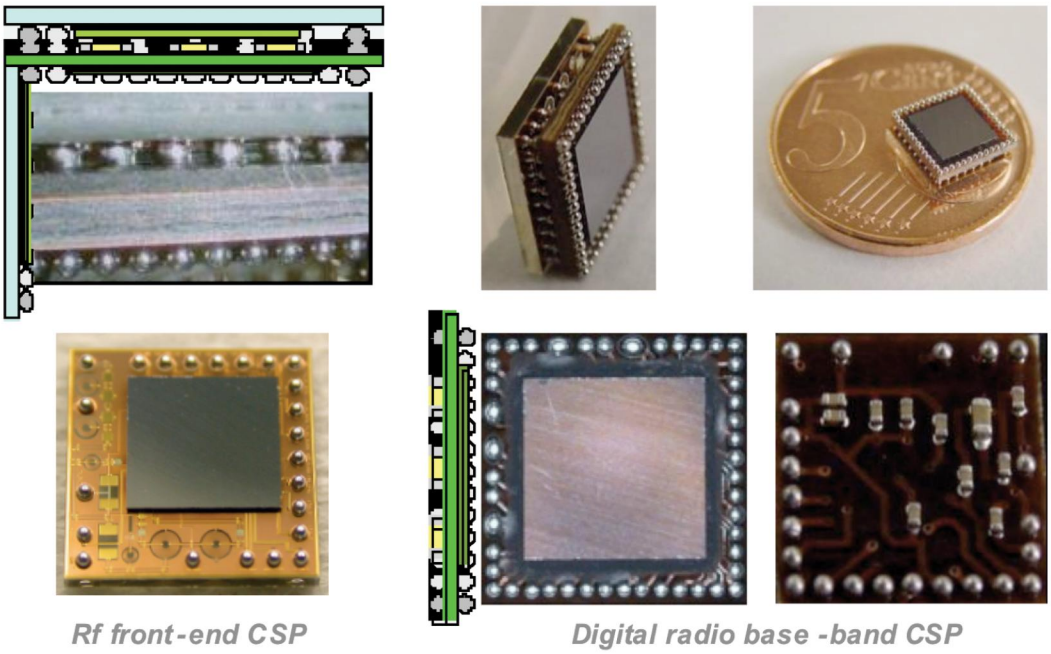

Figure 4: Fully integrated low power rf radio, measuring 7x7x2.5 mm, realized by 3D stacking of CSP packages.3D joining using micro-bumps

A particular interesting application area for 3D-SIP is the realization of distributed, fully autonomous systems for realizing so-called "ambient intelligence" systems. These are sometimes referred to as smart-dust, e-grains or e-cubes. As shown in figure 5, such systems can be divided into clear subsystems: the radio (antenna, rf-front-end base band), the main application (processor, sensors, actuators) and the power management (regulation, storage, generation). Each of these functions can be realized as a SIP-subsystem. These may be very small (a few mm to a few cm), enabling its realization using wafer level processing technologies. These 2D-subsystems can be stacked on top of each other, realizing a dense 3D-SiP system. Figure 6 shows such an "e-cube" module with a volume of only one $1cm^3$ realized at IMEC.

A further evolution of 3D-SIP technology is the embedding of Si-die and SMD components in the 3D-SIP interposer substrate using sequential build-up board technologies.

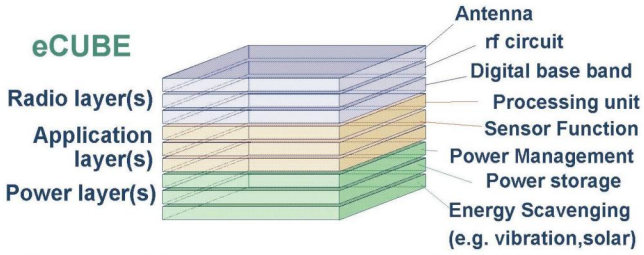

Figure 5: Schematic representation of a 3D-SIP concept "eCube", for the realization of distributed, fully autonomous "ambient intelligent" systems. Each layer in the stack is a fully integrated SIP sub-system.

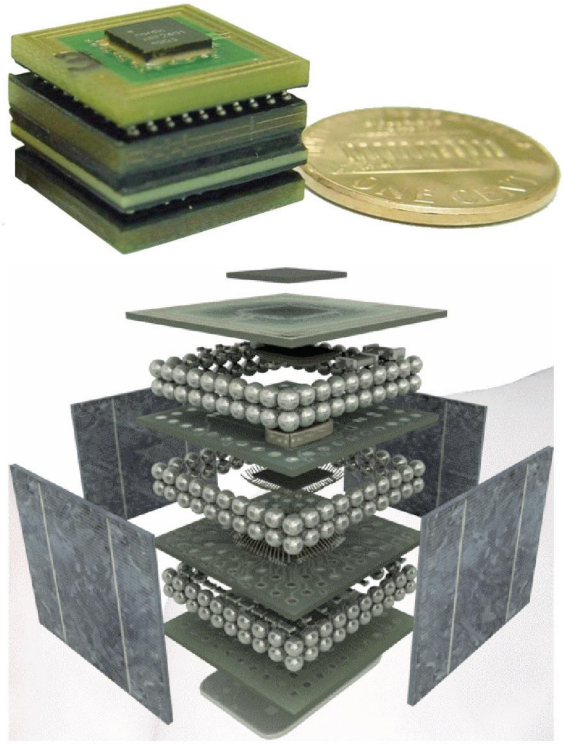

Figure 6: Photograph (top) and Schematic view of $1cm^3$ eCube (14x14mm), developed at IMEC for a medical application. The 3D-SIP consists of a rf-SIP, with integrated antenna, a low-power DSP SIP, a 19 channel EEG/ECG sensor die, a power SIP. Small solar cells power scavengers, to provide for long time autonomy, may be added on the module sidewalls.

3D-WLP

Wafer level packaging technologies may be very cost effective to realize 3D stacks. The simplest implementation is the face-to-face bonding using flip-chip or micro-bump connections.

Two different 3D-WLP options can be chosen:

- Realization of 3D-through silicon via connections, followed by stacking by a method similar to flip chip mounting.
- Stacking thin die on-top of other die or substrates and contacting both die using a multilayer thin film technology.

Both technologies are studied at IMEC and described in more detail below.

3D-WLP USING SI-THRU VIAS

The most common approach used for realizing Si-through vias consist of the following steps:

- Etching of a "blind" via hole in the Si-wafer using the Bosch RIE-ICP etching method.
- Dielectric isolation of the Si-hole: Using of CVD oxide or nitride passivation
- Metallization of Si-holes by realizing a solid metal via-"plug". Typically Cu electroplating is used for the via filling, followed by a CMP polishing step to remove the excess Cu plated on the wafer.
- Back grinding of the wafer, exposing the Cu plug, finalizing the 3D via-process.

This process has been shown to be effective for realizing high-density 3D via connections. However a number of important issues remain:

- Only a thin insulation layer is used between the Si substrate and the Cu-plug. This results in a rather high electrical capacitance of the through hole connection, exceeding the capacitance of standard wire-bond pads.
- A rather thick Cu plug is used in the Si-via hole. Due to the large CTE mismatch between Si and Cu this will cause significant thermo-mechanical stresses upon thermal cycling.
- Electroplating of Cu to fully fill the Si-via is a complex process that requires a long process time for each wafer. The use of a CMP polishing step further increases the cost of this technology.

As a solution to these shortcomings, we propose a modified 3D-via build-up [15], shown in figure 7:

- The thin CVD insulating layer is replaced by a 2-5 µm thick polymer isolation layer, deposited by spin or spray coating
- The via is only partially filled with electroplated Cu, similar to a build-up PCB board via, but with smaller dimensions.
- A polymer coating is used to fill the remaining hole in the copper plating.

This method has some significant advantages:

- Lower cost by simplifying the processes, reducing the process time and reducing the required equipment capital investment.
- Strongly reduced capacitance through the use of thicker low-k isolation layers, allowing for high speed and rf 3D-via feed-throughs.

- Strongly reduced thermo-mechanical stresses by using "open" copper metallisations and the use of lower modulus dielectric materials: "compliant" through-hole structure.[9]
- Compatible with common wafer-level packaging redistribution and bumping technologies

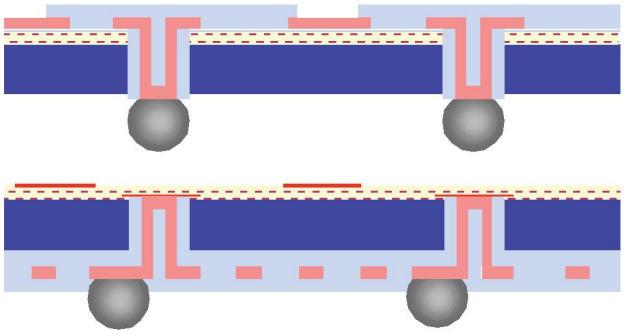

Figure 7: Schematic representation of IMEC's 3D-WLP process with Si through-hole electrical connections. Top: approach with thinning after via formation. Bottom: thinning-first approach: the vias are processed after the wafer thinning process.

3D-WLP BY ULTRA-THIN CHIP STACKING (UTCS)

A different approach to 3D stacking consists in stacking thin die on active device wafers and using multilayer thin film technology[11,16] to interconnect the thin die with the host wafer. Such approach allows for a high level of system level flexibility and integration density. This technique also allows for stacking of die with largely varying dimensions, as well as the integration of thin film passive components in a 3D interconnect stack.

Several approaches of this type of stacking are under investigation.[1]. IMEC´s ultra-thin-chip-stacking, UTCS, approach [13] uses very thin (10-20µm) Si die, embedded in a redistribution technology. In order to realize a multilayer die UTCS stack, four basic processes are used.

- Ultra-thin chip-on-die, UTCD, process
- Ultra-thin chip embedding, UTCE, process
- Ultra-thin chip-in-flex, UTCF
- UTCF-stacking process for more than two die layer UTCS structures

The UTCD process is shown schematically in figure 8. A wafer with active, tested die is bonded to a temporary silicon carrier wafer. Using a combination of coarse and fine grinding, the active wafer is thinned to a thickness of 15-20 µm). Plasma etching is used to remove any remaining Si damage and to obtain the desired final thickness. Plasma etching is then used to etch the scribe lanes of the die. The next step consists in dicing the carrier wafer to obtain the UTCD chips for further processing.

For the UTCE process, shown schematically in figure 9, either an active wafer or another dummy carrier wafer is used as host substrate. A polymer glue-layer, such as BCB, is spun on the wafer. A flip chip bonder is used to place with high alignment accuracy KGD-UTCD chips on KGD host wafer die. The actual bonding of these die (polymer glue curing under pressure) is collectively performed at the wafer level. The next step is the collective removal of the sacrificial layer and the temporary carrier chips of the UTCD die (e.g. thermal or chemical). The next steps consist of depositing thin film dielectric layers for isolation and thin film, electroplated copper patterns for electrical connection. At the end of this process a 2-layer UTCS stack is obtained, as shown in figure 10. An example of a two-die layer UTCE structure is shown in figure 11.

If, for the UTCE process a sacrificial substrate is used, this substrate may be removed, effectively resulting in a very thin flex foil (10 to 30 µm) with embedded active die, the UTCF foil shown schematically in figure 12.

For realizing an n-layer UTCS stack, two different approaches could be used. The UTCE process could be sequentially repeated on another UTCE stack, as shown in figure 13. However, it is also possible to stack UTCF films on UTCE stacks, using micro-bump flip chip connections. This results in a more cost effective parallel process flow.

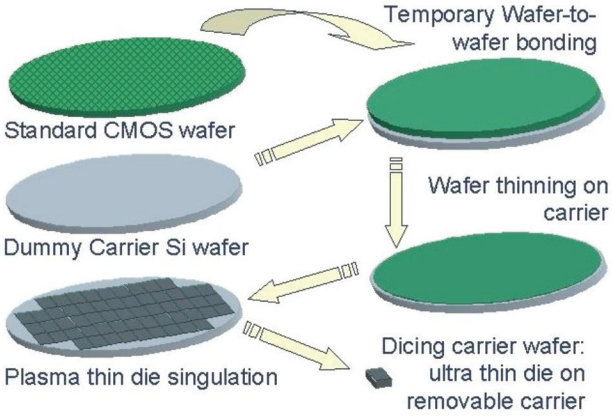

Figure 8 : Schematic representation of the ultra-thin chip-on-die, UTCD, process

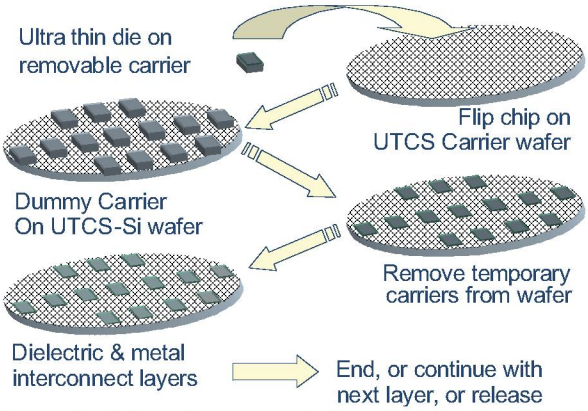

Figure 9: Schematic representation of the ultra-thin chip embedding, UTCE, process

By using thin film lithography, a very high number of interconnects can be realized between the die. This allows for additional interconnects between sub-sections of the die. In the case of the proposed UTCS structure, via connections in the 3rd dimension are realized in the area around the chips. As these thin film vias are realized with pad sizes smaller than 50 µm, a very high interconnect density in the third dimension is obtained.

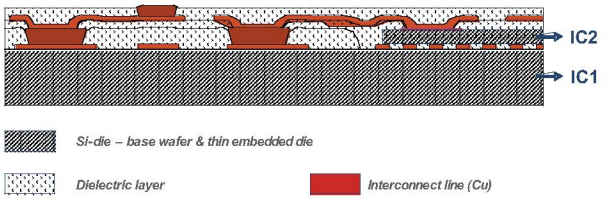

Figure 10: Schematic cross-section of a UTCE stack.

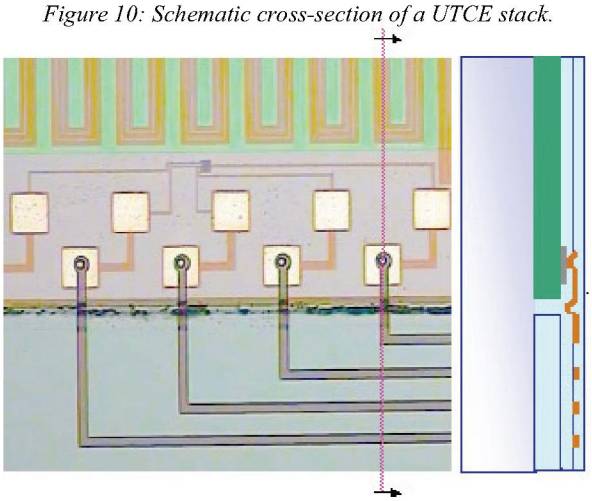

Figure 11: Example of a 15 µm thin Si-die, transferred to a host substrate and electrically connected to that substrate using the UTCE die embedding technique.

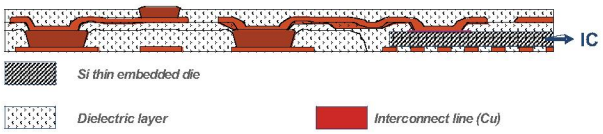

Figure 12: Schematic cross-section of the UTCF foil after removal of the sacrificial substrate.

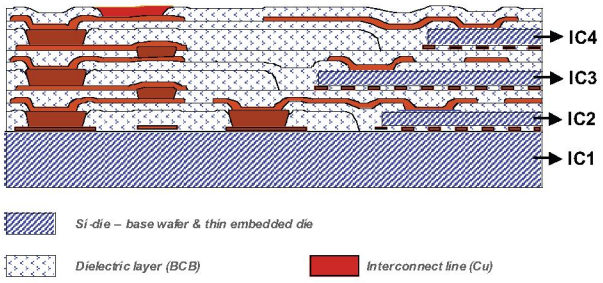

Figure 13: n-layer UTCS stack, realized using a sequence of (n-1) UTCE process steps.

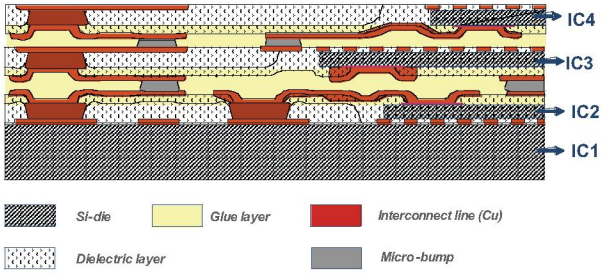

Figure 14: n-layer UTCS stack, realized using parallel processing combining one UTCE and n-2 UTCF layers.

In contrast to the traditional die-stacking technologies, the dies that are stacked can have arbitrary sizes at any layer and do not need any particular process or design adaptation.

A further possible evolution of the UTCS technology is to combine it with the 3D-WLP Si via process discussed in 5.2.1, resulting in a very high 3D interconnectivity with large degrees of freedom.

3D-SIC

As explained in section 4, 3D-SIC technology uses IC-foundry infrastructure to realize through-Si via connections. Most approaches in this field realize these through-vias after finalizing the IC process. [6,12]. Our approach differs from this by introducing a small, Si-via and Cu plug, the so-called "Cu-nail". This Cu-nail is processed after the FEOL process (transistors), but before the BEOL process (multilayer damascene interconnect layers), as shown in figure 15. The Cu-nail is realized by plasma etching a ±15µm deep a Si hole with a diameter of 3-5µm. A modified Cu single damascene process is used to fill the hole. A CVD oxide layer is used as thin dielectric insulating layer and a CMP stop layer. A TaN barrier is then deposited. The via hole is then filled with electroplated copper. CMP is used to remove the Cu "overburden". After this process, standard BEOL is used to finalize the Si-die.

After finalizing the wafer process and wafer test, the wafer is mounted on a temporary carrier and thinned down to a Si-thickness of only 10 µm. In the process, the "Cu-nails" are exposed on the wafer backside (see figure 15).

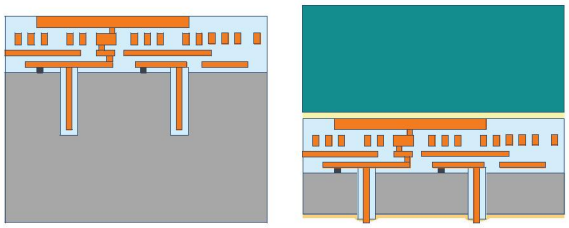

Figure 15: Schematic representation of the "3D-SIC" Cu-nail via process. Left: Standard CMOS wafer with "Cu-nail" before the BEOL process. Right: Thinned CMOS chip on carrier chip with exposed Cu-nails.

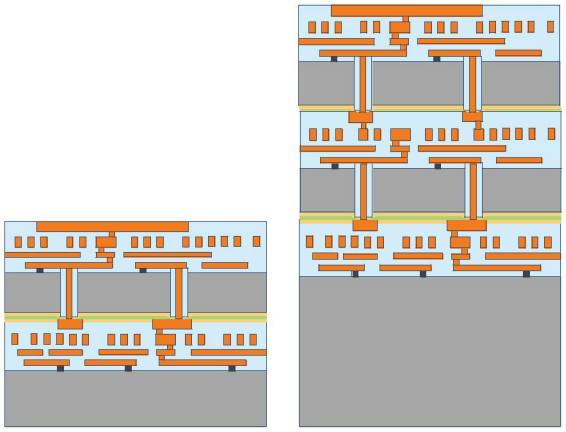

Figure 16: Schematic representation of a 2 and 3-layer 3D-SIC stack using Cu-Cu bonding

The 3D stacking is performed as a die-to-wafer bonding process, similar to the 3D-WLP process mentioned above. However, in this case Cu-Cu direct bonding is used, rather than a solder joint or micro bump connection.[14] The stacking process consists of a fast die-to-wafer alignment and placement, followed by a collective wafer-level Cu/Cu bonding process. It can easily be repeated to obtain multi-die stacks. (See fig. 16)

The main advantages of this 3D-SIC approach are:

- Minimal impact on the CMOS wafer design and processing:
  - Only a small exclusion area on the FEOL (small via holes)
  - No impact on the BEOL wiring
  - Small number of additional process steps needed, only one additional litho step, resulting in a low process cost.
- Parallel processing route : wafers are prepared for 3D stacking and only KGD die are stacked involving a minimum number of processing steps. This is required to achieve high 3D-module Yields and low cost processing.
- A very high density 3D-interconnect is possible, as the Cu-nail size is only a few micrometer in diameter. Densities up to, and exceeding, $10^4/mm^2$ are feasible.

One of the possible draw-backs of this 3D-SIC, as well as other 3D stack technologies, is the need for routing 3D-contacts through intermediate die in the stack when going from a connecting two arbitrary die from the stacks. The die, or at least the 3D-contacts, must be arranged in a "wedding-cake" fashion.

A possible solution to this drawback is to combine the UTCS technology, described above in 5.2.2, with the 3D-SIC Cu-nail and stacking technology. This is illustrated schematically in figure 17.

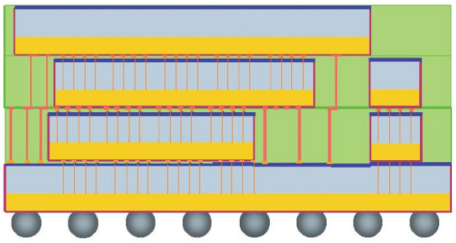

Figure 17: Schematic representation of a combined UTCS and 3D-SIC technology.

3D RESEARCH ROADMAP IMEC

In the previous section, IMEC's approaches to 3D-SIP, 3D-WLP and 3D-SIC technologies were discussed. In Table 1, a comparison of these different approaches is presented. A common aspect of all technologies is the need for die thinning technologies down to 50µm and less. For the 3D-SIC and UTCS technologies, die will be thinned down even further. A schematic representation of IMEC's R&D roadmap is given in figure 18.

CONCLUSION

The further evolution of microelectronic technology and electronics systems will increasingly require 3D interconnect technologies. Different applications will require various 3D-interconnect densities. Therefore, a variety of technologies can be used to reach a cost-effective solution.

In order to be cost-effective, a 3D technology should meet a number of criteria. Parallel processing of the 3D-stack layers should be maximized. Also processing should be performed as much as possible on a wafer or panel level. Also of crucial importance is to provide a solution for the compound yield problems that are inherent to multi-die systems. This favours a die-to-wafer bonding of KGD die to KGD die locations on a wafer.

We propose a classification of the different 3D technologies, based on the underlying manufacturing infrastructure: 3DSIP; 3D-WLP and 3D-SIC.. The approach and roadmap of IMEC to these technologies was presented in detail.

REFERENCES

[1] Proc. of the 1st and 2nd conf. on "3D Architectures for Semiconductor Integration and packaging", RTI international, Burlingame, California, April 13-15, 2004 and Tempe, Arizona, June 13-15, 2005

[2] E.Beyne, "3D Interconnection and packaging: impending reality or still a dream?" proceedings of the IEEE International Solid-State Circuits Conference, ISSCC2004, 15-19 February 2004; San Francisco, CA, USA, IEEE, 2004, pp.138-145.

[3] Phil Garrou, "3D Integration: A status report", proceecings "3D Architectures for Semiconductor Integration and packaging", RTI international, Burlingame, Tempe, Arizona, June 13-15, 2005.

[4] Scott List, "The third dimension: fact or fiction?" idem [3]

[5] Albert Young, "Perspectives on 3D-IC technology" idem [3].

[6] A.Klump et al. "3D Integration of CMOS Transistors with ICV-SLID Technology" idem [3].

[7] M. Karnezos. "3-D Packaging: Where All Technologies Come Together". IEEE/SEMI 29th International Electronics Manufacturing Technology Symposium, July 2004, pp. 64–67.

[8] N.Ranganathan et.al, "High aspect ratio through-wafer interconnect for three dimensional integrated circuits, Proceedings of the 55th ECTC, , Orlando, Florida, May 31-June 3, 2005, pp. 343-348

[9] M.Gonzalez et al, "influence of dielectric materials and via geometry on the thermomechanical behaviour of silicon through interconnects" ". Proc. of 10th Pan Pacific Microelectronics Symposium, SMTA, Hawaii, January 25-27, 2005.

[10] Manuba Bonkohara, "Technologies for 3D assembly and chip level stack" Proceedings of 2nd International Symposium on Microelectronics and Packaging, ISMP2003, IMAPS-Korea, Seoul, Korea, September 24-25, 2003,, pp.85-90.

[11] E.Beyne, "Multilayer thin film technology as an enabling technology for System in Package (SiP) and "above-IC" Processing", idem [7] pp.91-99.

[12] Misma Koyanagi, "3D LSI Technology and Wafer-level Stack", idem [7], pp.101-108.

[13] E.Beyne, "Technologies for very high bandwidth electrical interconnects between next generation VLSI circuits", IEEE-IEDM 2001 Technical Digest, December 2-5, Washington, D.C., S23-p3, 2001.

[14] J.H.McMahon et al., "Wafer bonding of damascene-patterenced metal/adhesive redistribution layers for via-first 3D Interconnect", Proc. of the 55th ECTC, , Orlando, Florida, May 31-June 3, 2005, pp. 332-336.

[15] patent US 10817763

[16] patent EP 0100014, US 6,506,664

Table 1: Classification and comparison of different 3D interconnect technologies at different levels of the interconnect hierarchy.

	3D-SIP	3D-WLP		3D-SIC
Technology	Package interposer	WLP, Post-passivation		Si-foundry, Post FEOL
3D interconnect	Package I/O	UTCS Embedded die	Si-through vias	Si-through "Cu nail" vias
Interconnect density *Peripheral* *Area-array*	'package-to-package' 2 - 3 /mm 4 - 11/mm^2	'around' die 10 - 50 /mm 100 -2.5k/mm^2	'through' die 10 - 25 /mm 16 - 100/mm^2	'through' die 25 -100 /mm 400-10k/mm^2
3D Si Via pitch 3D interconnect pitch	- 300 – 500 µm	- 20 – 100 µm	40 – 100 µm -	< 10 µm -
3D Si Via diameter	-	-	20 - 40 µm	1 - 5 µm
Die thickness	> 50 µm	10 - 50 µm	40 - 100 µm	10 - 50 µm

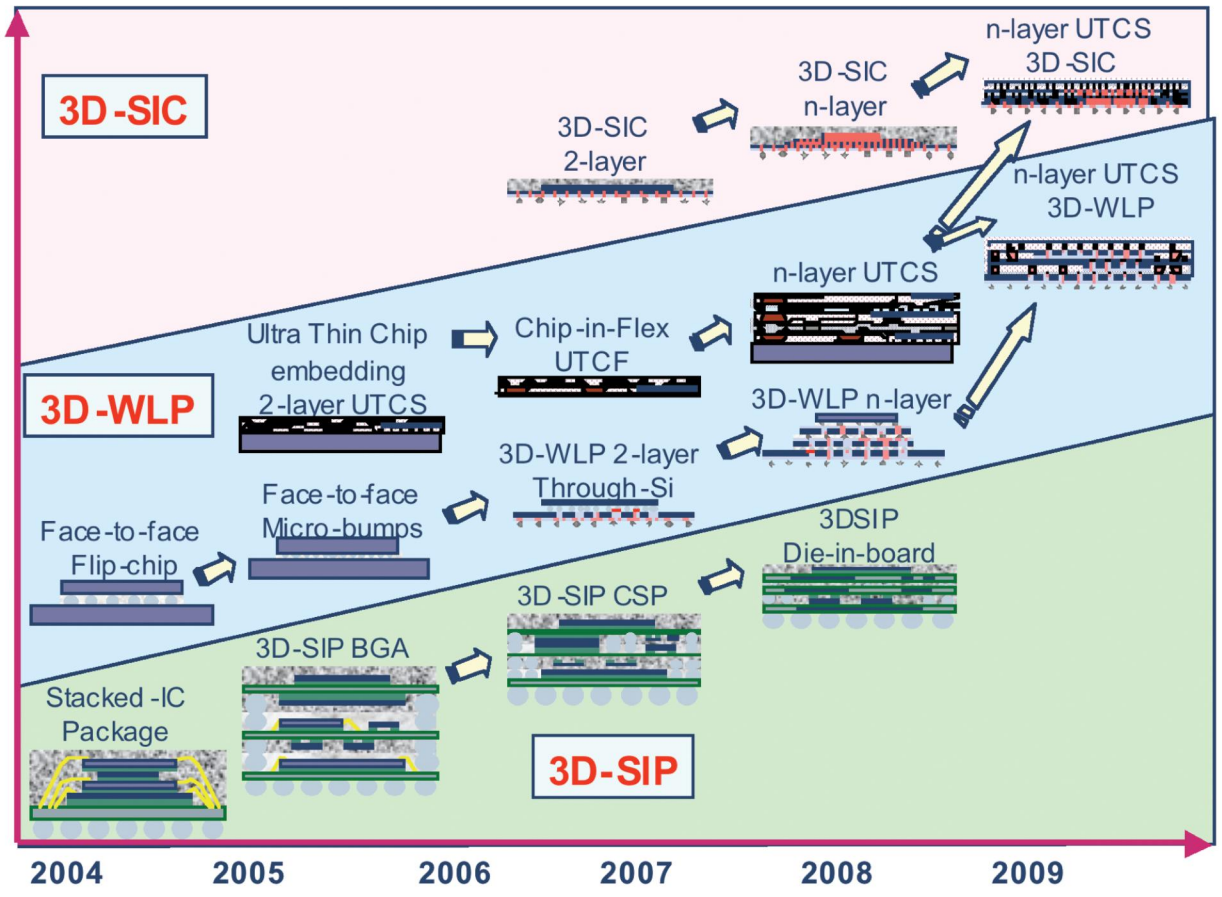

Figure 17: IMEC's 3D packaging and interconnection roadmap for 3D-SIP, 3D-WLP and 3D-SIC technology families.