Heterogeneous integration and chiplet assembly – all between 2D and 3D

Peter Ramm[1], Paul Franzon[2], Phil Garrou[2,3], Raja Swaminathan[4], Pascal Vivet[5], Mustafa Badaroglu[6]

[1]Fraunhofer EMFT, [2]NCSU, [3]MCNC, [4]AMD, [5]CEA, [6]Qualcomm

Introduction

"Real 3D" integration - 3DIC Integration in its true definition [1] - has a long history. As early as 1985, Richard P. Feynman expressed this vision [2]: "Another direction of improvement of computing power is to make physical machines **three dimensional** instead of all on a surface of a chip. That can be done in stages instead of all at once - you can have several layers and then add many more layers as time goes on" [2].

Successively, several R&D initiatives started on 3DIC throughout the globe. Towards the end of the 1980´s in a consortium with a.o. Siemens and Fraunhofer Munich, 3D CMOS test devices as e.g. 3D SRAMs were realized based on recrystallization of thin Si [3]. Such "sequential processing" or monolithic concepts are reconsidered today as "ultimate 3D" for stacking at transistor level - in the IRDS More Moore roadmap for the 2030´s [4].

Towards the end of the 1990´s Prof. Mitsumasa Koyanagi´s team at the Tohoku University, for the first time in the world, succeeded in fabricating 3DICs using TSV, explicitly 3D stacked image sensor and 3D stacked memory test chips [5] representing *the* pioneering contributions of today´s two key applications in high volume production (see Fig. 1). At about the same time Fraunhofer in Munich focused already on the key application of heterogeneous systems, consisting of components with different materials/technologies and die sizes. Robust die-to-wafer stacking technologies were developed to achieve what is called today **3D heterogeneous integration** [3, 6]. But despite these early technology demonstrators, it needed decades, until finally Samsung started in 2015 a high volume 3DIC product, the stacked DDR4 and later the HBM2 memory (see Fig. 1, bottom left). The other application that has gone into high volume production is CMOS image sensors (see Fig. 1 top middle). Since 2017, Sony is producing a stacked CMOS image sensor (CIS) for smart phone cameras. On the other hand, there have also been drawbacks. Most significantly, 3D memory-on-logic applications, widely forecasted by many sources, have been postponed several times.

Realization of 3D chips needs specific design methodology. Essential driving forces for 3D integration are performance (speed), power consumption, costs, and form factor. While TSV technologies using advanced IMC bonding or hybrid bonding processes provide very high vertical interconnect densities, the major issue is the high cost of 3DIC manufacturing. Nevertheless, TSV technology shows up as packaging mainstream for high performance 3DICs. But alternative concepts "between 2D and 3D" were in fact very successful for products with no need of such high interconnect performances, i.e. Si interposer technology (see Fig. 1, top right). And moreover, alternative interposer concepts avoiding costly TSV technology are gaining importance, as e.g. Intel´s omni-directional interconnect (see Fig. 1).

In order to pay more attention to such new stacking concepts, the IEEE Technical Committee 3D decided to broaden its objectives correspondingly and include so-called 2D enhanced architectures (see Fig. 2) and also "chiplet" integration (see further down).

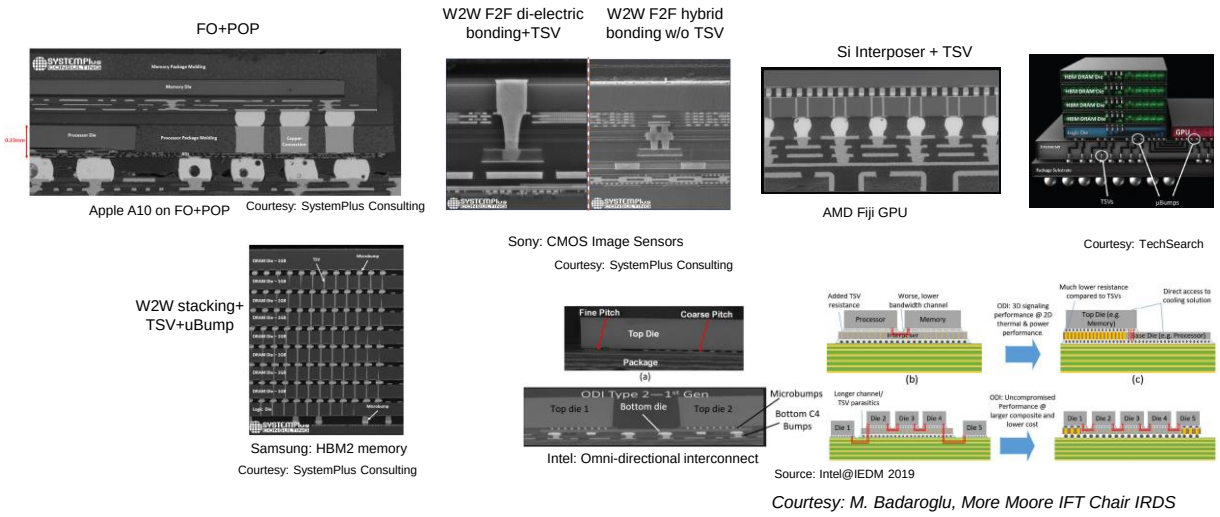

Figure 1: Fine-pitch 3D stacking technologies today ([4]).

The **Heterogeneous Integration Roadmap** [7] has categorized corresponding architectures between 2D and 3D as follows:

2DO (Organic) multi-chip package (MCP) Architecture: Side by side active Silicon die interconnected at very higher densities on the package using organic materials-based approaches. These can be further sub-categorized into *Chip First 2DO* achieved using a redistribution based approach with fan-out architectures (wafer or panel level: fan-out wafer-level packaging (FOWLP) or fan-out panel-level packaging (FOPLP), (examples: Infineon´s embedded Wafer-Level Ball-grid array (eWLB) and ASE`s Fan-out Chip on Substrate (FoCoS)) and *Chip Last 2DO*.

2DS (Inactive Silicon) MCP Architecture: Side by side active Silicon interconnected at extreme higher densities using inactive Silicon integrated into an organic package. These can be further sub-categorized into *Inactive Si with TSV* (example: TSMC´s Chip-on-Wafer-on-Substrate (CoWoS) architecture in Nvidia Tesla) and *Inactive Si without TSV* (Example: Intel Embedded Multi-die Interconnect Bridge (EMIB) based products [8].

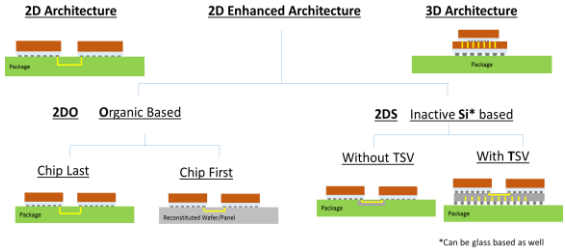

Figure 2: 2D to 3D architectures (source: HIR [7])

Dis-Integration is Underway

We have known for some time that with lateral scaling is slowing down the industry would need to find another way to continue to move forward. One of the options being implemented is to actually "disintegrate" SoCs into their functional parts and then connect these "chiplets" back together on high density interposers.

This was first done by Xilinx on their FPGAs in 2010. Chiplets, as they are now called, are not simply small chips. Chiplets cannot be used by themselves; they're specifically intended to be interconnected together to build complete functionality. Thus, it is better to think of chiplets as a silicon IP (intellectual property) subsystem, designed to integrate with other chiplets through advanced package interconnect (usually micro bumps) and standardized interfaces.

Building complete circuits from pre-verified chiplets is beginning to gain traction as a way of cutting costs and reducing time to market for heterogeneous designs. Chiplets allow us to use the latest node only where needed which in turn results in reduced silicon cost. These silicon savings, in turn, can be allocated for more expensive packaging solutions.

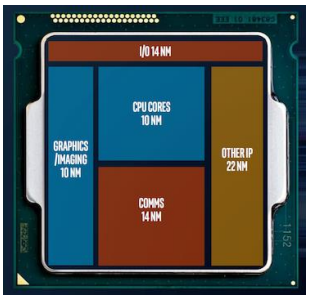

Figure 3: Example of chiplet configuration in a single package

AMD, Intel and TSMC have all introduced or announced chiplet based products and/or technologies. It is also widely accepted that for us to be able to mix and match chiplets produced at different foundries we will need to have standard interfaces and communication protocols. This is presently the most important thing that we can do to stabilize and broaden the chiplet infrastructure. Currently Intel is recommending their AIB (advanced interface bus) interfaces such as AIB and AIB2 while TSMC is offering Lipincon (low voltage in package interconnect) to their advanced customers. It is presently unknown whether these two interfaces can be made compatible.

It is hoped that these functional chiplets will create a library and in the future, we can combine these tested chiplets from multiple foundries, to devise future circuits. DoDs DARPA is in year 4 of their CHIPS program which has been looking at chiplet based solutions for the military.

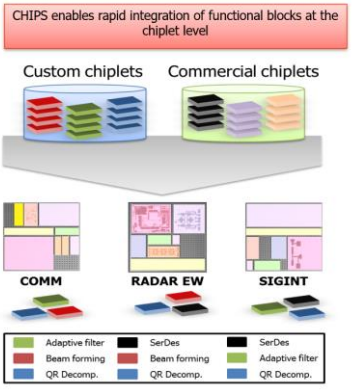

Figure 4: DARPA´s CHIP concept

Chiplet Physical Interfaces

The central idea behind chiplets is to enable new systems to be designed from a set of existing small parts, possibly combined with a small value-add part (or parts) and integrated using advanced interposer technologies. One goal behind chiplet technology is to enable the fast and low-cost design of new systems – systems enabled by unique combinations of chiplets – and thus avoid the high costs of designing a single SOC.

A key enabling technology is the chiplet to chiplet interface. There are several layers to such an interface including protocol and physical layers. This article will briefly discuss the physical layers. The ideal physical layer interface would achieve the power and area footprint of a long range on-chip SOC driver/receiver pair while enabling a high aggregate bandwidth, being able to drive a wide range of wire lengths (with the attendant range of line losses), and support standardized DFT. Key decisions include voltage swing, serialization, clock management, bus-widths, etc.

In a recent experiment Paul Franzon´s group at NCSU designed a RISC-V chiplet with an AIB interface. The layout and die photo are shown below. The area cost of using a standardized interface is self-evident. However, note this includes area (FIFOs etc.) required to support clock domain crossing.

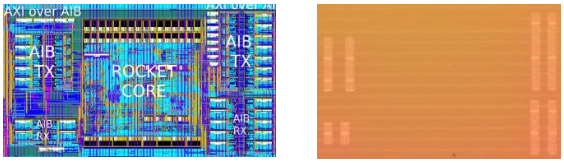

Figure 5: Layout and photo of a RSIC-V chiplet with an AIB interface (source: NSCU)

A number of of chiplet interfaces have been proposed. These proposals and some salient features are summarized in the Table below.

Standard	Source	Bandwidth density/edge	Throughput / lane	Delay	PHY Energy/bit
Advanced Interface Bus (AIB)	Intel [23]	504 Gbps/mm	Up to 2 Gbps	<5 ns	0.85 pJ
Multi-Die I/O (MDIO)	Intel	1600 Gbps/mm	Up to 5.4 Gbps		0.5 pJ
High Bandwidth Memory (HBM3)	JEDEC [26]		4.8 Gbps		0.37 pJ
XSR/USR	Rambus/ OIF		112 Gbps		
Lipincon	TSMC [21]	536 Gbps/mm	2.8 Gbps	<14 ns	0.486 pJ
Bunch of Wires (BoW)	OCP/ODSA [22]	1280 Gbps/mm	Up to 16 Gbps	<5 ns	0.7 pJ
Bandwidth Engine	Mosys [24]		Up to 10.3125 Gbps	<2.4 ns	
Infinity Fabric	AMD [25]		10.6 Gbps	< 9 ns	2 pJ

Chiplet Integration onto Active Interposers

Circuit and system designers need a more affordable, scalable and efficient way of integrating heterogeneous functions, to allow more reuse, at circuit level, while focusing on the right innovations in a sustainable manner. Due to the slowdown of advanced CMOS technologies (7nm and below), with yield issues, design and mask costs, the innovation and differentiation through a single die solution is not viable anymore. Mixing heterogeneous technologies using 2.5D/3D is a clear alternative.

Chiplet partitioning is raising new interests in the research community [9], in large research programs as DARPA CHIPS [19] and in the industry. It is actually an idea with a long history in the 3D technology field [1]. Motivation for chiplet-based partitioning is driven by cost, modularity and heterogeneity. By dividing a single circuit in various chiplets and sub-modules, a large system is achieved with acceptable yield and cost using Know-Good-Die (KGD) sorting [10]. "Chipletization" allows building modular systems from elementary blocks and circuits, more focusing on function than on technology constraints.

Finally, proper technology is selected for the right function (advanced CMOS for computing, mature technology for analog and IOs, etc), while 3D integration is used for the overall system assembly.

Several technologies have been assessed for chiplets assembly: organic substrates as a low-cost solution adopted by AMD [11], passive interposers as a high-performance solution adopted by TSMC [12] or silicon bridges as an intermediate solution adopted by Intel [13]. These technologies are mature, economical benefits and performances are achieved, but they still raise limitations. Due to wire-only interconnects, inter-chiplet communication are still limited to side-by-side communication reducing the number of connected chiplets; the passive interposers cannot carry less scalable functions such as analog and IOs; co-integration of chiplets with incompatible interfaces is impossible.

Homogeneous Chiplets

To tackle these issues, the concept of active interposer is introduced that enables integration of some active CMOS circuitry on a large-scale interposer (Fig. 6). The active interposer can be seen as a generic bottom die infrastructure which integrates i) flexible and distributed system interconnect topologies between chiplets for scalable communication traffic, ii) energy efficient 3D-plugs for dense inter-layer communication, iii) fully integrated voltage regulators for efficient power supply close to the cores and iv) memory-IO controller and PHY for socket communication. Finally, the active interposer integrates system infrastructure: clock, low speed interfaces, thermal sensors and 3D design-for-test (DFT) to enable KGD strategy.

As an active interposer large prototype [14], the INTACT circuit demonstrator from CEA (Fig. 7) is composed of 6 chiplets (28nm FDSOI) each integrating 4 clusters of 4 cores (16 cores per chiplet), 3D stacked with 20 µm pitch µbumps on an active interposer (65nm CMOS) with 40µm pitch TSV middle (Fig. 8) [15]. In terms of technology partitioning, there are two technology node differences, ensuring enough performances in the bottom layer, while preserving system costs. The active interposer integrates numerous distributed interconnects for long distance low latency communication, 3D-plug interfaces achieving 3 Tbit/s/mm^2 bandwidth density, and fully integrated switched capacitor voltage regulators with up to 82% power efficiency. The circuit implements a total of 96 cores with a scalable cache coherent architecture, delivering a peak 220 GOPS. As a result, users will get more GOPS at the same power budget – or a reduced energy footprint for the same task – and will benefit from an increased memory-computing ratio along the memory hierarchy. These are main drivers to address big data applications.

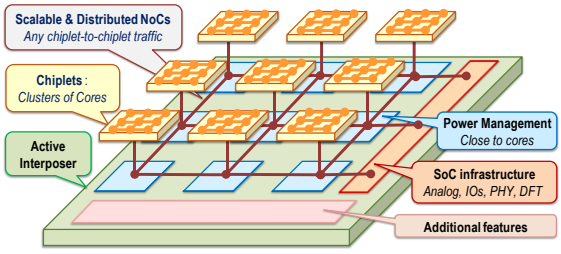

Figure 6: Active Interposer concept and main features

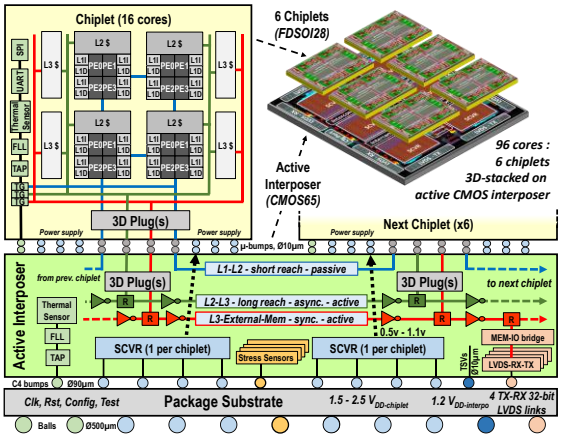

Figure 7: INTACT circuit architecture [14]

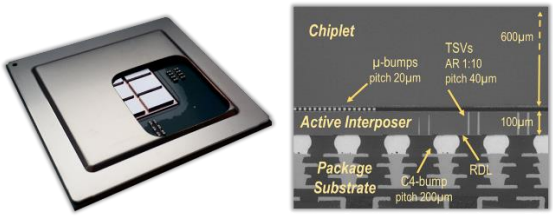

Figure 8: INTACT package and 3D cross-section [15]

Heterogeneous Chiplets

For active interposer, chiplets can be either identical, with similar functions as presented in this first example, or with distinct functions. In another example, the EXANODE CEA prototype integrates 2 bare FPGA dies, 2 chiplets onto an active interposer, all integrated onto a large substrate MCM module, as presented Fig. 9 [16]. This circuit targets ultra-wide range of workloads for next generation scalable and high-performance compute nodes.

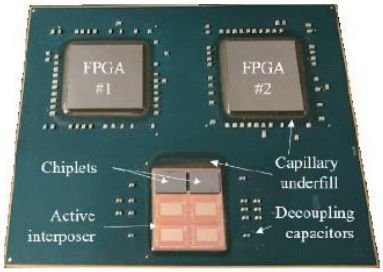

Figure 9: EXANODE heterogeneous Multi-Chip-Module [16]

For 3D integration, the technologies are still evolving to provide more advanced chiplet integration, with reduced pitches, for improved energy efficiency, die-to-die parallelism, and thermo-mechanical behavior. Hybrid bonding technology initially devoted to Wafer-to-Wafer (for BSI Imagers) are also appearing for Chip-to-Wafer assembly, with reduced pitches (down to 10µm and targeting below) [17].

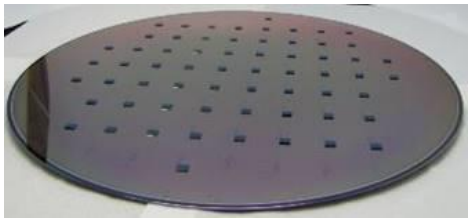

Figure 10: Chip-to-Wafer Hybrid Bonding integration

The INTACT AND EXANODE circuit demonstrators are a first step towards larger scale and heterogeneous chiplet-based systems, showing the benefits of 3D connectivity and the smart features of active interposers.

Interfacing Chiplets, Die-to-Die Standards, and Testability

As presented in previous circuit demonstrators, the so-called 3D-plug IP solution, integrating both the logical and physical aspects are used for energy efficient parallel die-to-die interfaces. For true chiplet compatibility, it is currently complex to integrate chiplets from different sources, due to missing standards. Strong standardization initiatives are on-going from ODSA group [18] and CHIPS program [19]. As of today, with passive interposers, wire-only connectivity prevents the integration of chiplets using incompatible protocols. On the contrary, active interposer is a solution enabling to bridge incompatible chiplets by ad-hoc logic. This solution has been proposed by zGLUE Inc. for instance as a generic connectivity solution for medium performance devices.

Lastly, testability is also a strong concern for chiplet based design, where testability and Know-Good-Die strategy must be available to yield the overall 2.5D or 3D system. Strong progresses have been achieved recently, since the on-going standardization effort of the IEEE1838 Working Group has led to the approval of the IEEE1838 3D test standard in March 2020 [20].

Architectures between 2D and 3D – best-tailored for the different specific applications of heterogeneous integration

The different applications, as e.g. memory, CMOS image sensor, GPU, RFIC and chiplet based products need best-tailored technology solutions for their specific performance, power consumption, costs, and form factor requirements. Besides 3DIC integration in its strong definition, a variety of architectures between 2D and 3D are potentially well-suitable for cost-effective production. This is especially true for the growing market of heterogeneous 3D sensor/IC systems with the need for robust die-to-wafer stacking of components of *significant* different device technologies, as CMOS, sensors, actuators and MEMS [27, 28] rather than extremely small TSV/pad pitches.

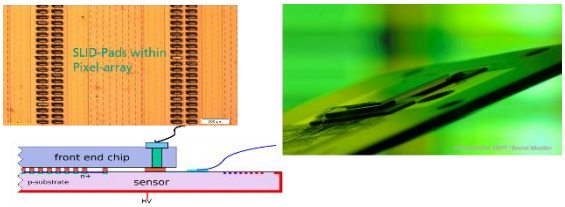

Figure 11: Heterogeneous integration of sensor and readout circuit (Fraunhofer EMFT in co-operation with Max-Planck MPP)

A corresponding example of **heterogeneous 3D sensor integration** is shown in Fig. 11. The photon detector and appropriate read-out IC are 3D integrated by intermetallic compound (IMC) bonding with mechanically stable Cu/Sn Solid-Liquid-InterDiffusion (SLID) interconnects [28].

In order to pay more attention to the described fine-pitch stacking concepts and new architectures "between 2D and 3D", the **IEEE EPS Technical Committee 3D** decided to broaden its objectives from 3DIC and monolithic integration on one side, to non-TSV 3D technologies, chiplet assembly, Si interposer and alternative interposer concepts (including TSV less), on the other side of the fine-pitch interconnect "spectrum". In the last decade we have seen a true success story of 3D and enhanced 2D technologies. Nevertheless, there is still a huge amount of related problems, as e.g. thermal issues, design and test issues, materials optimization, robustness of the processes, thermal-mechanical reliability of the systems and last not least high production costs, which can only be solved by significant development efforts [7, 29].

Visit our website for information on TC 3D/TSV:

https://eps.ieee.org/technology/technical-committees/technical-committee-3d-tsv.html

References

[1] Handbook of 3D Integration "Technology and Applications of 3D Integrated Circuits", edited by Philip Garrou, Christopher Bower, Peter Ramm, Wiley & Sons (2008).

[2] Richard P. Feynman «The computing machines in the future», Nishina Memorial Lecture at Gakushuin University (Tokyo) (1985).

[3] Peter Ramm et al. «3DIC: Past, Present and Future - a European Perspective», Plenary Talk at 2020 IEEE 70th Electronic Components and Technology Conference (2020).

[4] International Roadmap for Devices and Systems (IRDS™) https://irds.ieee.org/editions/2020

[5] Mitsumasa Koyanagi et al., IEEE IEDM Tech. Digest, pp. 879-882 (1999).

[6] Peter Ramm et al., Japanese Journal of Applied Physics 43 (7A), L 829 (2004).

[7] https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2019-edition.html

[8] Ravi Mahajan et al. «Embedded Multi-Die Interconnect Bridge (EMIB)», 2016 IEEE 66th Electronic Components and Technology Conference (2016).

[9] Puneet Gupta, Subramanian S. Iyer, « Goodbye, motherboard. Bare chiplets bonded to silicon will make computers smaller and more powerful: Hello, silicon-interconnect fabric", IEEE Spectrum, Year: 2019, Volume: 56, Issue: 10.

[10] James Quinne, Barbara Loferer, "Quality in 3D assembly - Is "Known Good Die" good enough?", IEEE 3D System Integration Conference (3DIC 2013).

[11] Samuel Naffziger et al., "AMD Chiplet Architecture for High-Performance Server and Desktop Products", ISSCC, 2020.

[12] Mu-Shan Lin et al., "A 7nm 4GHz Arm®-core-based CoWoS® Chiplet Design for High Performance Computing", VLSI Conference, 2019.

[13] David Greenhill et al., "A 14nm 1GHz FPGA with 2.5D Transceiver Integration", ISSCC , 2017.

[14] Pascal Vivet et al., "A 220GOPS 96-Core Processor with 6 Chiplets 3D-Stacked on an Active Interposer Offering 0.6ns/mm Latency, 3Tb/s/mm2 Inter-Chiplet Interconnects and 156mW/mm2@ 82%-Peak-Efficiency DC-DC Converters", ISSCC, 2020.

[15] Perceval Coudrain et al., "Active Interposer Technology for Chiplet-Based Advanced 3D System Architectures", 2019 IEEE 69th Electronic Components and Technology Conference (2019).

[16] Pierre-Yves Martinez et al., "ExaNoDe: combined integration of chiplets on active interposer with bare dice in a multi-chip-module for heterogeneous and scalable high-performance compute nodes", VLSI Conference, 2020.

[17] Amandine Jouve et al., « Die to wafer direct hybrid bonding demonstration with high alignment accuracy and electrical yields », IEEE 3D System Integration Conference, Sendai (3DIC 2019).

[18] Open Compute ODSA project, https://www.opencompute.org/wiki/Server/ODSA

[19] CHIPS program, https://www.darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies

[20] https://standards.ieee.org/standard/1838-2019.html

[21] Mu-Shan Lin et al., "A 16nm 256-bit wide 89.6GByte/s total bandwidth in-package interconnect with 0.3V swing and 0.062pJ/bit power in InFO package," 2016 IEEE Hot Chips 28 Symposium (HCS), Cupertino, CA, 2016, pp. 1-32, doi: 10.1109/HOTCHIPS.2016.7936211.

[22] S. Ardalan et al., "Bunch of Wires: An Open Die-to-Die Interface," 2020 IEEE Symposium on High-Performance Interconnects (HOTI), Piscataway, NJ, USA, 2020, pp. 9-16, doi: 10.1109/HOTI51249.2020.00017.

[23] David Kehlet, "Accelerating Innovation Through A Standard Chiplet Interface: The Advanced Interface Bus (AIB)," Intel White Paper retrieved from https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/accelerating-innovation-through-aib-whitepaper.pdf
[24] M. J. Miller, "Bandwidth engine® serial memory chip breaks 2 billion accesses/sec," 2011 IEEE Hot Chips 23 Symposium (HCS), Stanford, CA, 2011, pp. 1-23, doi: 10.1109/HOTCHIPS.2011.7477493.
[25] B. Kleveland et al., "An Intelligent RAM with Serial I/Os," in IEEE Micro, vol. 33, no. 6, pp. 56-65, Nov.-Dec. 2013, doi: 10.1109/MM.2013.7.
[26] H. Ko et al., "A 370-fJ/b, 0.0056 mm2/DQ, 4.8-Gb/s DQ Receiver for HBM3 with a Baud-Rate Self-Tracking Loop," 2019 Symposium on VLSI Circuits, Kyoto, Japan, 2019, pp. C94-C94, doi: 10.23919/VLSIC.2019.8778082.
[27] Josef Weber, Montserrat Fernandez-Bolanos, Adrian Ionescu and Peter Ramm, « 3D Integration Processes for advanced sensor systems and high-performance RF components », ECS transactions 86 (8), 2018.
[28] Peter Ramm, Armin Klumpp, Christof Landesberger, Josef Weber, Andy Heinig, Peter Schneider, Guenter Elst, Manfred Engelhardt, « Fraunhofer´s Initial and ongoing contributions in 3D IC Integration», IEEE 3D System Integration Conference, Sendai (3DIC 2019).
[29] Handbook of 3D Integration, Volume 4: “Design, Test, and Thermal Management”, edited by Paul D. Franzon, Erik Jan Marinissen, Muhannad S. Bakir, Philip Garrou, Mitsumasa Koyanagi, Peter Ramm, Wiley & Sons (2019).