A 4 × 4 × 2 Homogeneous Scalable 3D Network-on-Chip Circuit With 326 MFlit/s 0.66 pJ/b Robust and Fault Tolerant Asynchronous 3D Links

Pascal Vivet, *Member, IEEE*, Yvain Thonnart, *Member, IEEE*, Romain Lemaire, *Member, IEEE*, Cristiano Santos, Edith Beigné, *Member, IEEE*, Christian Bernard, Florian Darve, Didier Lattard, Ivan Miro-Panadès, Denis Dutoit, Fabien Clermidy, *Member, IEEE*, S. Cheramy, Abbas Sheibanyrad, Frédéric Pétrot, Eric Flamand, Jean Michailos, Alexandre Arriordaz, Lee Wang, and Juergen Schloeffel, *Member, IEEE*

***Abstract*— Future many cores, either for high performance computing or for embedded applications, are facing the power wall, and cannot be scaled up using only the reduction of technology nodes; 3D integration, using through silicon via (TSV) as an advanced packaging technology, allows further system integration, while reducing the power dissipation devoted to system-level communication. In this paper, we present a 3D modular and scalable network-on-chip (NoC) architecture implemented using robust asynchronous logic. The 3DNOC circuit targets a Telecom long-term evolution application; it is composed of two die layers, fabricated in 65 nm technology using TSV middle aspect ratio 1:8, and integrates ESD protection, a 3D design-for-test, and a fault tolerant scheme. The 3D links achieve 0.66 pJ/b energy consumption and 326 Mb/s data rate per pin for the parallel link. Thin die effect is demonstrated by thermal analysis and measurements, as well as the dynamic self-adaptation of the 3D link performances with 3D thermal conditions. Finally, the scalability of the 3DNOC circuit, in terms of power delivery network and thermal dissipation, is demonstrated by using simulations up to a 3D stack of eight die layers.**

***Index Terms*— 3D technology, asynchronous logic, multicore, network-on-chip (NoC), thermal dissipation, through-silicon-via (TSV).**

Manuscript received May 4, 2016; revised July 19, 2016 and September 6, 2016; accepted September 7, 2016. Date of publication October 26, 2016; date of current version January 4, 2017. This paper was approved by Guest Editor Dennis Sylvester. This work was supported by the French National Program Programme d'Investissements d'Avenir, IRT Nanoelec under Grant ANR-10-AIRT-05.

P. Vivet, Y. Thonnart, R. Lemaire, C. Santos, E. Beigné, C. Bernard, F. Darve, D. Lattard, I. Miro-Panadès, D. Dutoit, F. Clermidy, and S. Cheramy are with CEA-LETI, MINATEC Grenoble, 38054 Grenoble, France (e-mail: pascal.vivet@cea.fr).

H. Sheibanyrad and F. Pétrot are with the TIMA Laboratory, 38000 Grenoble, France.

E. Flamand and J. Michailos are with STMicroelectronics, 38920 Crolles, France.

A. Arriordaz is with Mentor Graphics, 38330 Montbonnot, France.

L. Wang is with Mentor Graphics Corporation, Fremont, CA 94538 USA.

J. Schloeffel is with Mentor Graphics Development, 21079 Hamburg, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2016.2611497

I. Introduction and State of the Art

CURRENT semiconductor roadmaps foresee that the CMOS technology node can still be reduced allowing higher integration density for transistors, however, the cost reduction has now reached a limit [1]. In order to pursue technology and design integration, "More than Moore" technologies, such as 3D approach, are becoming more attractive. The 3D technology presents a wide scope and defines a full set of different technology elements [2]: through silicon vias (TSVs), microbumps, and redistribution layer (RDL); different technology options: face-to-back (F2B) and face-to-face (F2F); and different technology steps: wafer bonding, wafer debonding, die-to-die (D2D) assembly, and die-to-wafer assembly. These different 3D technology options offer various integration schemes: 3D integration (where dies are stacked vertically) and 2.5-D integration (where dies are stacked horizontally using silicon interposers). The 3D technology presents numerous advantages: it opens a full scope of new application possibilities, by integrating devices from different technologies (CMOS, MEMs, DRAMs, or even photonic) in order to benefit from the most suitable technology. Regarding manufacturing, the 3D technology introduces the possibility of testing the dies before the final 3D assembly, to identify the so-called Known Good Die (KGD). By integrating ad hoc design-for-test logic [37], and by modeling the different test costs and technology yields for all 3D steps [38], it is possible to define and optimize the 3D test flow and corresponding yield tradeoffs. As a summary, 3D provides shorter communication distance between dies, thus reducing communication power consumption [3], offers the possibility of building heterogeneous systems, and finally improves system yield and cost by partitioning the system in a divide and conquer approach.

The 3D technologies have started to become a reality for some product categories. The 3D technology (TSVs and microbumps) has been early introduced by Xilinx for the Virtex-7 FPGAs as a cost effective approach, using 2.5-D

integration with a passive silicon interposer. A large-area regular design is replaced by four dies (pretested using a KGD approach) organized as slices stacked onto a silicon interposer in order to improve the manufacturing yield; the connections to the package are provided by TSVs [4]. More recently, 3D stacking has been largely adopted to increase DRAM density by taking advantage of die stacking to create "memory cubes" [5]. Standards, such as HMC [6] and HBM [7], [8], rely on multiple (4–16) high-density DRAM layers stacked on a logic die managing I/O access and arbitration. Significant gains have been demonstrated with a three times performance per watt and three times PCB footprint reduction for the AMD Fury system replacing GDDR5 by HBM memories [9].

For higher level of integration, solutions for DRAM stacking directly on the top of logic die have also been introduced following the wide I/O standard [10]–[12]. In these previous 3D systems, memory and logic layers of the system are designed independently according to common JEDEC specifications for the interfaces. The WIOMING circuit [12], integrating a wide I/O (version 1) DRAM on the top of a logic die, is a precursor of the work presented in this paper. Indeed, the previous WIOMING circuit and the proposed 3D network-on-chip (NOC) circuit are sharing the same asynchronous NoC communication infrastructure and the same CMOS and 3D technologies: CMOS 65 nm with TSV middle, aspect ratio (AR) 1:8 (diameter 10 μm and die thickness 80 μm), and pitch 40 μm, using a D2D and F2B 3D stacking. In the meantime, 3D tightly coupled memory-on-logic multicore designs have also been prototyped. Kim *et al.* [13] and Fick *et al.* [14] present 3D-stacked partitioned designs, where SRAM is vertically connected to processing elements enabling high transfer bandwidth combined with limited power consumption. These designs take advantage of fine-pitch F2F copper connections using TSV only for external I/Os, thus preventing more than two dies in the 3D stack. The current TSV pitch (40 μm) is not suitable for fine-grained design partitioning (where pitch below 1 μm is required) with F2B assembly scheme. In this context, on-chip communication infrastructures appear as a promising tradeoff to exploit 3D technologies, since they can benefit from shorter connections while maintaining a coarse-grained partitioning at SoC level (with 3D pitch in the 10–50 μm range), using existing 3D implementation tools [15], [16].

For multicores, NoC has been widely studied, and high-dimensional topologies have demonstrated gains in terms of performance due to shorter connections [17], however, planar silicon technologies have up to now limited actual physical implementation to 2D topologies, such as ring, mesh, or torus. Maturing 3D technologies offer a promising opportunity to revisit NoC implementation in the direction of truly 3D NoC. The 3D NoC has been already well studied [18], with a focus on various aspects: theoretical performances, gains in terms of cross-sectional bandwidth and saturation throughput, router implementation, adaptive routing algorithms in case of defective 3D links, fault tolerance schemes, and so on. A multicore architecture implementing a 3D NoC communication scheme has been shown [19], but is reduced to a proof of concept with a small NoC topology using a single NoC vertical link.

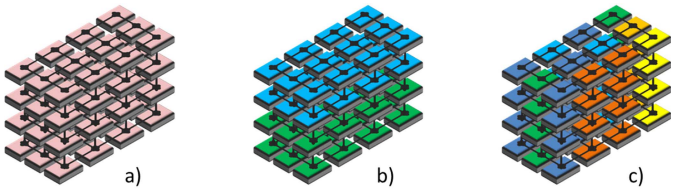


Fig. 1. Homogeneous and heterogeneous 3D NoC architectures examples. (a) Homogeneous cores and technologies (the same cores and the same technology). (b) Heterogeneous cores and technologies (two set of cores and technologies). (c) Heterogeneous cores and homogeneous technologies (different cores, using a single technology).

Regarding clocking strategy, implementing a 3D clock tree using available 3D place and route tools is still infeasible today. The 3D clock implementation has been studied but is still ad hoc design with some limitations [20]. As a consequence, coarse-grained 3D partitioning with system level interconnects using well-identified timing interfaces (asynchronous, mesochronous, source synchronous, and so on), such as NoC, is adapted to 3D, to avoid 3D global clocking. On the other hand, many asynchronous NoCs [21]–[25] have been introduced to help solving timing limitations of 2D global clocking, using a globally asynchronous locally synchronous (GALS) template [26].

In this paper, we propose to extend the existing 2D NoC as a scalable and modular 3DNOC circuit, using asynchronous signaling to implement robust 3D interfaces [27], [28], while preserving a coarse-grained partitioning, which can be implemented using current 3D CAD tools. The outline of this paper is as follows. Section II presents the proposed 3DNOC circuit demonstrator and architecture for Telecom long-term evolution (LTE) application, Section III presents the 3D asynchronous link design, its overview and associated details, Section IV presents the proposed 3D design-for-test and fault tolerant architecture, Section V presents the circuit technology and measured performances, Section VI presents a 3D thermal effect analysis and the self-adaptation of the circuit to the thermal conditions, and finally, Section VII presents the scalability of the 3DNOC circuit architecture up to a 3D stack of eight dies, with regard to the power delivery network and to the thermal dissipation.

II. 3DNOC Circuit Architecture

A 3D NoC using TSVs in an F2B configuration is fully scalable to build large multicore systems. Such 3D NoC can be either homogeneous or heterogeneous, depending on the architecture and technology partitioning (Fig. 1): either full homogenous architecture [Fig. 1(a), the same cores and the same technologies], or different cores using different technologies [Fig. 1(b) like a memory array layer stacked on a top of a processing array layer], or different cores using the same technology [Fig. 1(c)], to build an homogeneous stack using heterogeneous blocks, as the proposed 3DNOC circuit. The objective is to scale the 3D NoC multicore performances

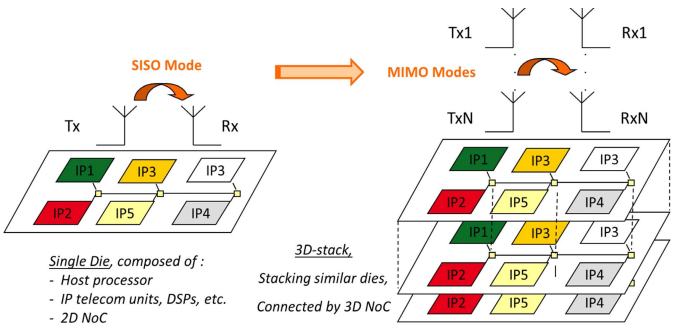


Fig. 2. 3D LTE Telecom baseband, stacking several dies implementing SISO scheme, providing a larger MIMO scheme.

with regard to application constraints, while using the best technology for each layer.

By using regular tile-based design and coarse-grained TSVs in an F2B configuration, the stacking of logic dies through a 3D NoC can be compared at first glance to the stacking of DRAM layers: for the same footprint, and using the same design masks—in the case of homogeneous stacking—the end user gets more memory capacity for a DRAM memory cube, and more computation capacity in the 3D NoC case.

A. 4G Telecom LTE Application

As an application target, in the context of wireless telecommunications, such as LTE, it is required to handle multiple antennas in a multiple-input-multiple-ouput (MIMO) approach (Fig. 2). The MIMO scheme offers space-time diversity to handle multipath propagation using several transmit and receive antennas, and is generally associated with orthogonal frequency-division multiplexing (OFDM). As proposed earlier in [29], this can be achieved by 3D integration of identical chips: a single die controls one single antenna [providing a single input single output (SISO) mode, equivalent to an 1 × 1 MIMO], while the stacking of several of these dies allows to control more antennas (2 × 2 MIMO, 4 × 2 MIMO, and so on). Each circuit layer is composed of several dedicated Telecom processing units interconnected by a 2D NoC and controlled by a host processor. The same design and associated masks are shared between layers. These tile-based layers are interconnected by a 3D NoC using 3D links: by stacking more logic layers, the application achieves more wireless bandwidth and a number of radio channels, increasing the number of supported MIMO antennas.

B. 3DNOC Circuit Architecture: Single Layer

The proposed 3DNOC circuit is composed of two layers of the same die. Each circuit layer supports a 2 × 2 RX/TX MIMO scheme, based on a previous low power NOC circuit for MIMO 4G Software-Defined-Radio [30]. By stacking two identical dies, the proposed 3D circuit supports a 4 × 2 MIMO scheme. Each circuit layer (Fig. 3) integrates an ARM1176 host processor and 18 specific computing units: DSP engines, OFDM engines, memory engines for frame manipulation, and turbo decoders.

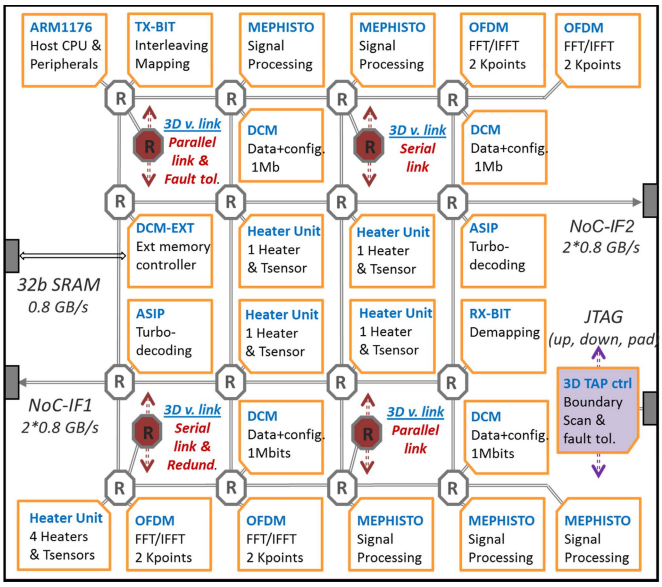


Fig. 3. 3DNOC circuit single-layer architecture, integrating a 4 × 4 router topology, an ARM1176 host processor, and 18 computing units.

Within each layer, the communication interconnect infrastructure is based on a 32 b asynchronous NoC, offering a GALS scheme [31]. Each computing unit is implemented using synchronous design, and its own locally generated clocks (using programmable delay lines), while all the system communications are implemented using asynchronous logic, providing robustness to any timing variations. The NoC presents the following characteristics: packets composed of 32 b words called flits, packet switching, source routing using a path-to-destination, two virtual channels for best-effort and guaranteed-service traffic, five ports routers (North, East, South, West, and Local), and pipelined links. The NoC is implemented using quasi-delay insensitive (QDI) asynchronous logic, as presented in [31]. The low-level handshake signals are encoded using a QDI four-phase four-rail asynchronous protocol, also called 1-of-4. The four data rails encode a two bit value, with one acknowledgment signal. Whatever the delays in the gates and wires are, the asynchronous handshake protocol enforces the correct sequence of request and acknowledge events, which leads to the correct behavior of the asynchronous logic, providing robustness and self-adaptation of the performances to any source of process voltage temperature (PVT) variations [31].

The single die layer also offers system communication with two 100 Mb/s 32 b bidirectional off-chip communication links using a synchronous NoC protocol version for compatibility with off-chip FPGA. Overall, the asynchronous NoC provides a 4 × 4 topology, and integrates four 3D NoC routers and ports, for 3D stacking of the proposed die, to build a 3D NoC topology, as presented in more detail in Section III.

C. 3DNOC Circuit Architecture: Multiple Layers and 3D Cross Section

The 3DNOC circuit (Fig. 4) is homogeneous, i.e., composed of two layers of the same die. The overall

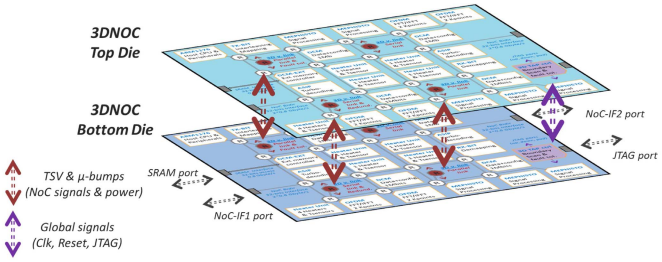


Fig. 4. 3DNOC circuit multiple-layer architecture, integrating two logic layers interconnected by four 3D NoC links.

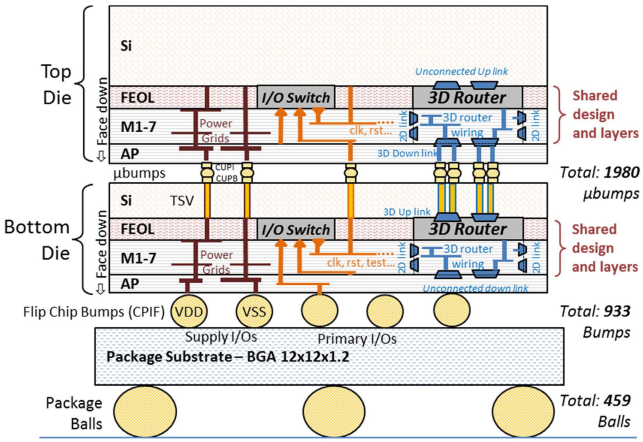


Fig. 5. 3DNOC circuit cross section, including details of signal wiring and power distribution.

3D interconnect is composed of a 4 × 4 × 2 topology: 32 NoC routers in two layers, with four asynchronous 3D links, providing a 7.4 GB/s 3D cross-sectional aggregate bandwidth. The 3D NoC extends the GALS scheme to the 3D direction, providing robustness and self-adaptation of 3D NoC link throughput under PVT variations (especially thermal gradients, which are challenging in 3D), dynamic frequency scaling decoupling (each circuit unit and each die layer are using local clocks, allowing the decoupling of the timing domains while offering dynamic optimization of the power consumption with regard to application constraints), and finally, avoiding any clocking issues at 3D interfaces.

As shown in the 3D cross section of Fig. 5, design and masks are shared between the bottom and top dies to optimize cost, except for the specific 3D masks: TSV, micropillars, flip-chip bumps, and the last metal layer (aluminum, defined as the layer AP in Fig. 5). The cost constraint leads to the following logical and physical design: the 3D NoC input–output connections are mirrored and aligned between up/down links and corresponding circuit faces; power supplies are replicated and transmitted from backside of bottom die to the front side of top die through TSVs; a few global nets (reference clocks, reset, and test) are transmitted through the stack and multiplexed between the 2D pad version (for the bottom die) and the 3D port version (for the top die). A homogeneous scalable 3D stack is obtained in an F2B assembly scheme, which could be implemented on more than two layers. The current 3DNOC circuit integrates only two layers, but

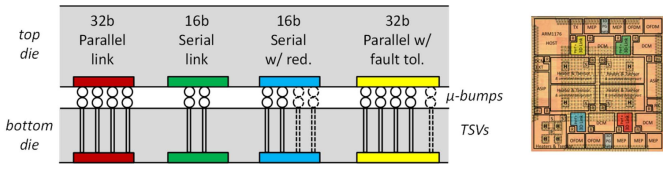


Fig. 6. Four different types of 3D NoC vertical links (principle and floor plan location).

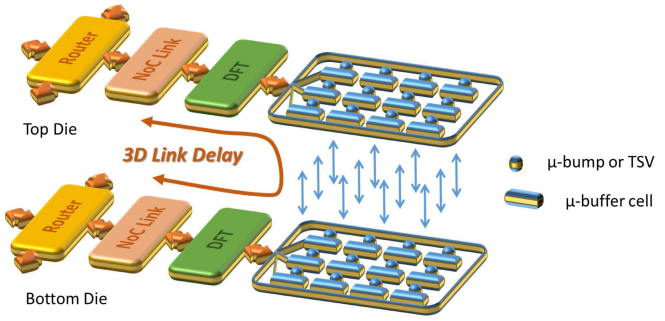


Fig. 7. 3D NoC vertical links architecture overview.

could be extended using the existing design and die to additional layers. The limitation to the size of the 3D stack would come from the power distribution network and from the thermal dissipation. Section VII will discuss in detail the scalability of the 3D stack with regard to its power delivery network and thermal dissipation. Finally, the 3D NoC topology has unconnected 3D up links at the top of the stack (respectively, unconnected down links at the bottom of the stack), which are cleanly terminated using pull-down logic. In Section III, the logical and physical designs of the 3D NoC vertical links is presented in detail.

III. 3D NoC Link Design

A. Overview

For circuit experiments, four different vertical links versions are implemented (Fig. 6):

1) a full parallel 32 b link (*the red link version, located bottom right*), providing the highest throughput;
2) a 2:1 serial 16 b link (*the green link version, located top right*), as a tradeoff between number of TSVs (gain TSV area) and performances (reduction of throughput and latency);
3) a 2:1 16 b serial link with 2:1 TSV redundancy (*the blue link version, located bottom left*), as a simple redundancy scheme by duplicating each serialized link signal;
4) a parallel 32 b link with a 1:8 fault-tolerance scheme (*the yellow link version, located top left*), providing a test-and-repair fault-tolerance scheme with 12% additional TSVs.

For each of these 3D asynchronous links, the baseline parallel link design is composed of the following elements (Figs. 7 and 8 for the 3D link overview and details): a NoC routing stage, an asynchronous pipeline stage, a design-for-test stage that integrates boundary-scan registers and corresponding data-path multiplexers, and the final stage

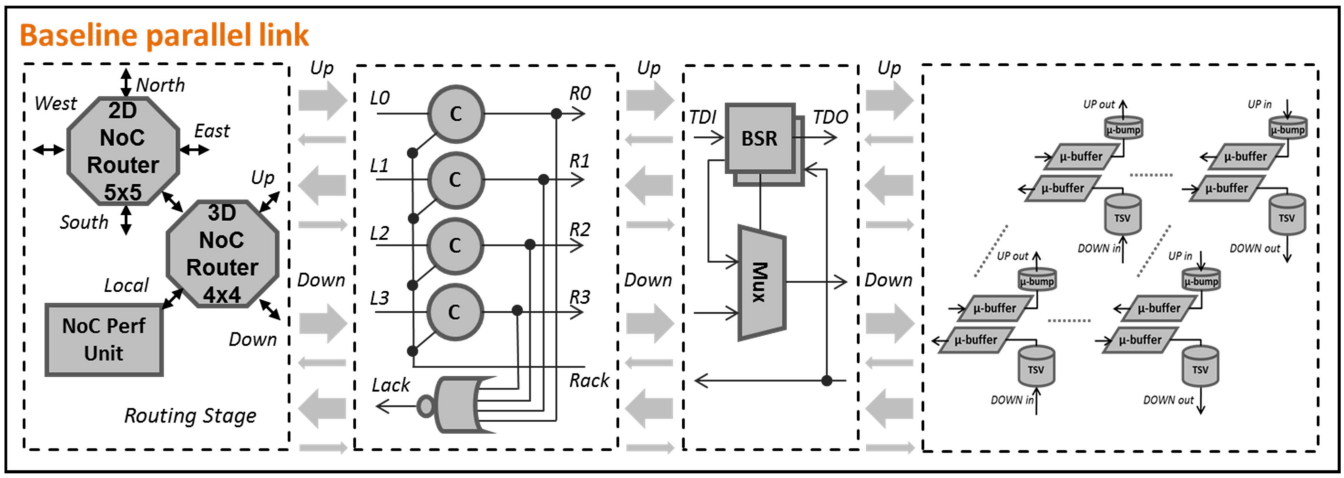


Fig. 8. 3D NoC vertical links architecture details, including routing stage, pipeline stage, DFT stage, microbuffer cell, and TSV stage.

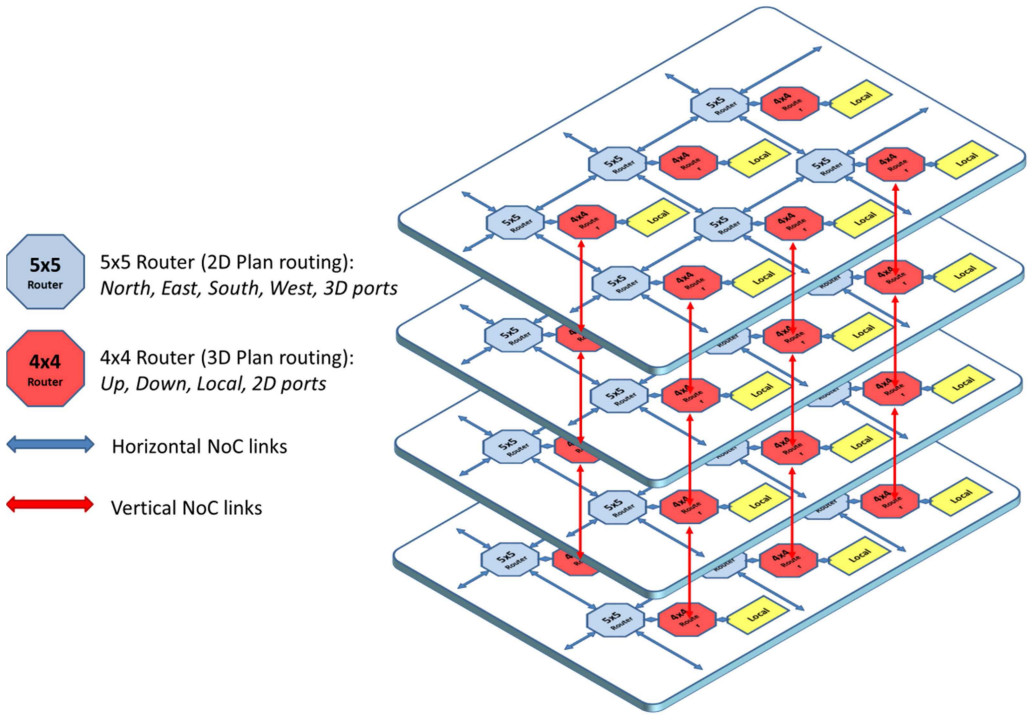


Fig. 9. 3D NoC topology and routing with hierarchical routers.

providing electrical buffering using microbuffer cells within the TSV matrix layout. The bidirectional up and down asynchronous links are composed of their respective data and acknowledge signals. The NoC vertical link delay between two circuit layers is the sum of the delays of all these elements. The corresponding stages are explained in more detail in section III.B to section III.G.

B. 3D NoC Router

For routing through the 3D NoC mesh topology, it is required to provide routing between the seven directions: North, East, South, West, Up, Down, and Local Resource. Due to design constraints, an asynchronous 3D NoC 7 × 7 router cannot be efficiently implemented in a single stage. The 3D router is then implemented hierarchically, as proposed in [32], by splitting the initial 7 × 7 router in the respective 2D plan and 3D plan (Figs. 8 and 9): the 5 × 5 router handles the intradie 2D communications, while the 4 × 4 router handles the interdie 3D communications and the access to the local resource. As a result, the communication latency is one hop for 2D communication only, one hop for 3D direction only, and two hops when changing the direction between 2D and 3D. Finally, the 4 × 4 and 5 × 5 asynchronous router throughputs are preserved, with a total area smaller than a single large 7×7 asynchronous router [32].

C. Traffic Generator

A dedicated unit for on-chip 3D NoC traffic generation and performance analysis is integrated and attached as the local resource of each 3D NoC 4 × 4 router: the unit generates

on-chip traffic with configurable packet length and destination, and measures both throughput and latency of received traffic. This unit is implemented in mixed synchronous and asynchronous logic, and uses a 100 kHz external clock as a time reference. Generated packets are sent at best effort of the asynchronous handshaking, potentially up to 1.2 GHz in typical conditions. The duration of the four-phase handshake loop across the link determines the actual performance for each link. Measurement of this performance is done by monitoring the acknowledgment signal period (toggling twice per data word), by means of a frequency division by 2 to 16 depending on the local clock, and counting the occurrences between events of the 100 kHz real-time clock. This setup provides traffic performance measurement accuracy below 1 MHz.

D. Asynchronous Pipeline Link

A 32 b four-phase four-rail QDI bidirectional link requires 160 data and acknowledge signals [31]. The asynchronous pipeline stage (Fig. 8) decouples the timing between the two circuit layers: from the bottom die router up to the top die router through the DFT stage and logical/physical interfaces. While similar asynchronous pipeline stages are used along the long 2D horizontal links (Fig. 3), here, the 3D pipeline stages are used for "short distance" 3D vertical links, which still present large interface delays of about 720 ps (see Section V-B). The asynchronous logic is implemented using C-element cells: a specific library of about 50 cells has been developed in the target CMOS 65 nm technology, with all required views and models (schematic, layout, timing and power characterization, and back-annotable gate level simulation) [33], [34]. For timing constraints and physical implementation, the design flow proposed in [35] breaks the asynchronous logic timing loop and optimizes efficiently all pipeline stages using a dummy clock method.

E. 3D Serial Link

To optimize the 3D link connection and associated TSV area cost, an asynchronous serial link offers a tradeoff between the number of TSV connections per link and the 3D link performance [36]. The serial link is fully implemented using asynchronous logic. It is composed of serialization logic, which transfers NoC data flits to serial subwords. The serial link, Fig. 10, is composed of asynchronous self-controlled muxes/demuxes that serialize/deserialize the inbound/outbound packets (design details can be found in [36]).

In a first option, a 2:1 serial link (32 b down to 16 b) is implemented in asynchronous logic, which reduces the TSV area impact by 2, at the cost of half throughput. In the last option, a serial link with 2:1 TSV redundancy is implemented to mitigate TSV or assembly defects, with nearly the same TSV count as the full parallel version.

F. Electrical Interface

The final stage of the 3D link is composed of microbuffer cells (Fig. 11), which is providing a 3D I/O capability.

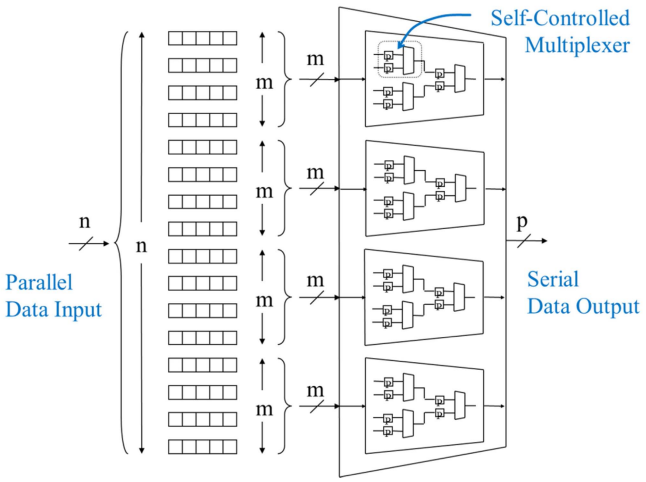


Fig. 10. Serial link microarchitecture [36].

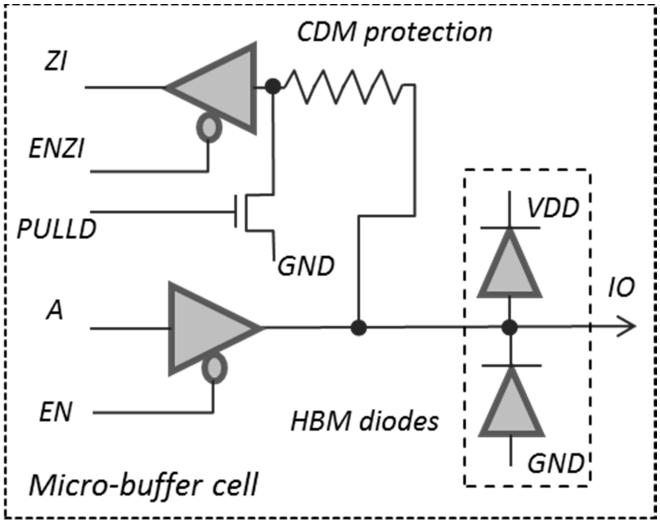


Fig. 11. 3D IO cell: microbuffer cell schematic.

The microbuffer cells are arranged in an array within the matrix of TSVs (for down connections) and microbumps (for up connections). The microbuffer cell is the 3D IO bidirectional cell, which provides: 1) a buffer of 2.5 mA drive strength to handle a 250 fF 25 mΩ TSV; 2) two gated diodes (70 fF junction capacitance) providing 200 V human-body model and 10 V machine model ESD protection for a robust interface during the 3D assembly process; 3) bidirectional capability for prebond test of the cell; and 4) a pull-down resistor for correct termination of unused 3D ports. The microbuffer cell is developed as a classical standard cell, four row height, to be integrated with direct abutment of its power stripes to the surrounding cells.

G. Physical Interface

Finally, as presented in the 3D cross section of Section II-C, a careful layout design is performed to offer direct 3D stacking of the same die on top of itself. Two overlapping 3D ports for the up and down in/out connections (Figs. 7 and 12)

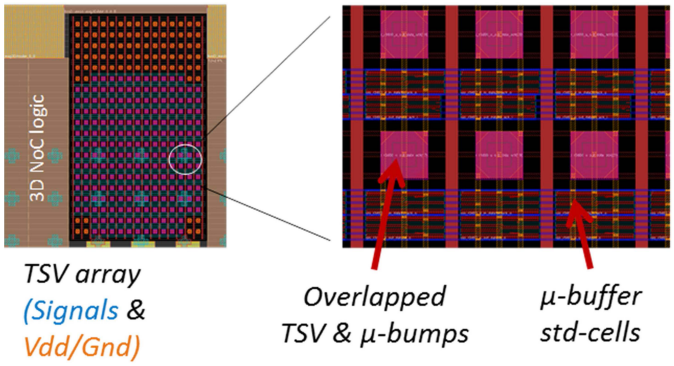


Fig. 12. TSV array, including signals (blue), P/G nets (orange), and interleaved microbuffer standard cells.

are defined, using aligned and mirrored connections for direct assembly of the top and bottom dies. The microbuffer cells of all 3D NoC input and output signals are interleaved within the TSV array of 40×50 μm pitch, while respecting a keep-out zone of 10 μm. The TSV array not only contains TSVs for the 3D NoC signals but also additional power and ground (P/G) TSVs for transmission of the power mesh through the 3D stack (orange TSVs in Fig. 12). The proposed physical design has been performed using the cadence encounter tool with TSV feature enabled allowing 3D IO mirroring within a full 3D netlist [41], while the physical 3D verification and corresponding layout-versus-schematic have been performed using the Mentor Graphics Calibre 3D STACK tool to check correct TSV and microbump alignment and connections [42].

IV. 3D Design-for-Test and Fault Tolerance

A. 3D Design-for-Test Architecture

For testing the proposed 3D asynchronous NoC, two kinds of tests must be addressed: the test of the NoC infrastructure itself and the test of the 3D connections. For the asynchronous NoC itself, functional test patterns are generated manually to cover the NoC topology. Because of the QDI asynchronous logic property, a minimal number of patterns are required, corresponding to the four-rail/four-phase encoding (NoC packets with 0×0000, 0×5555, $0\times$ AAAA, and $0 \times$ FFFF token values). For testing the synchronous IPs within the NoC, a test wrapper compatible with IEEE 1500 standard has been developed and integrated within the network interface of each IP. The synchronous IP scan patterns are generated using standard ATPG tools, and IP test data are then transported from the off-chip NoC interfaces through the NoC.

For testing the 3D connections, full control and observability are proposed, in line with the recent on-going 3D test proposals [37]. The four 3D NoC individual links are instrumented using boundary scan and standard JTAG (Figs. 8 and 13). Each 3D link integrates boundary scan registers and its corresponding multiplexers. Since the same die is used through the stack, a combinational JTAG switch is proposed using a die detection mechanism, so that a coherent JTAG tdi/tdo chain is composed, wherever the die is within the stack. The die-detection mechanism is implemented using the pull-down feature of the microbuffer cell. The 3D JTAG

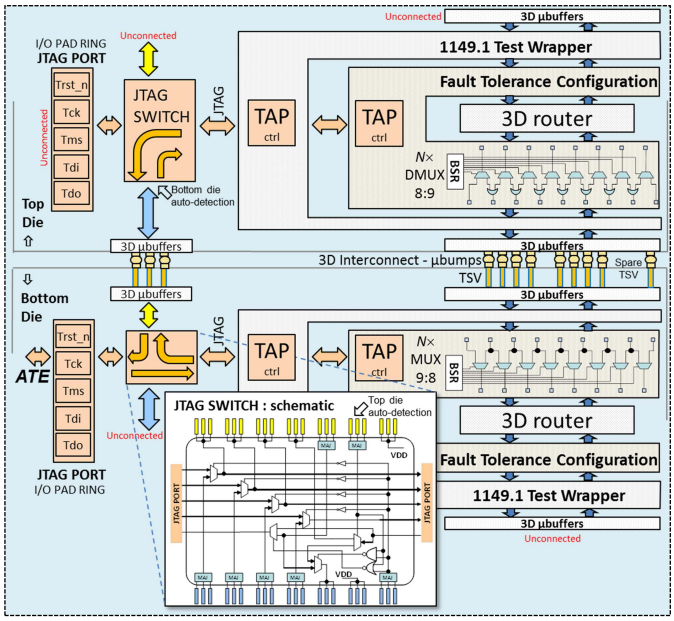


Fig. 13. 3D design-for-test architecture, including a JTAG switch and a configurable fault tolerance scheme.

connections use triple-modular redundancy for robustness of JTAG test access. The proposed JTAG switch avoids additional test registers to configure the vertical path through the 3D stack: the tdi/tdo chain is automatically configured, from bottom die to top die, to load/unload the 3D boundary scan shift registers. The proposed 3D-DFT architecture has been implemented using Synopsys DFT compiler for TAP and boundary scan register insertion, while the 3D test patterns have been generated and simulated using the Mentor Graphics Tessent test flow using the ICL and PDL languages applied to 3D DFT [39].

B. 3D Fault Tolerance Scheme

To increase yield, a 3D fault tolerant architecture is proposed similar to [40] (Fig. 13). One of the four 3D link versions integrates 1:8 TSV redundancy, where TSVs can be individually tested and repaired by bundle using configurable multiplexers. The fault tolerant architecture is fully integrated within the JTAG infrastructure (Fig. 13), offering a two-step test and repair architecture.

The yield of the 3D link, with or without correction scheme, is then defined by the following equations:

$$Y_{\text{link}}(\text{No Correction}) = (Y_{\text{TSV}})^N$$

$$Y_{\text{link}}(\text{With Correction}) = \left(\sum_{i=0}^{R} \binom{S+R}{i} (1 - Y_{TSV})^i (Y_{\text{TSV}})^{S+R-i} \right)^{N/S}$$

where Y_{TSV} is the yield of the 3D connection (TSV or microbump), N is the total number of wires of the link, and R and S are, respectively, the number of redundant and useful TSVs in each TSV bundle.

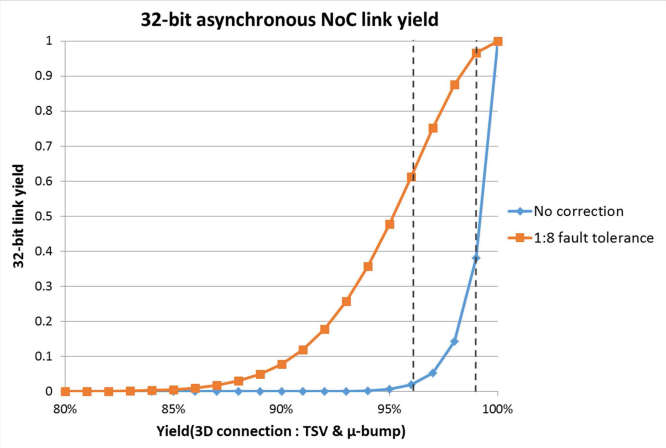


Fig. 14. Achieved 32 b NoC link yield using proposed fault tolerance scheme.

For the 32 b link, we have $N = 80$ (32 b link using 1-of-4 encoded data and acknowledge signals), and $R = 1$ and $S = 8$ (1 spare TSV every 8 TSVs).

With the proposed scheme and using 1 spare TSV every 8 TSV (12.5% additional spare TSVs), the NoC link yield can be increased from 2% yield to 61% yield in the case of 96% 3D yield, and from 38% to 96% in the case of 99% 3D yield (see Fig. 14).

V. 3DNOC Circuit Performances

A. Circuit Description

The 3DNOC circuit has been designed and fabricated in an STMicroelectronics 65 nm technology, embedding a 1:8 AR TSV middle (10 μm diameter, 80 μm die thickness, 10 μm keep out zone, and 40 μm pitch), and using a CEA-LETI 3D technology with 40 μm pitch microbumps. Dies are 3D stacked using an F2B D2D assembly and are finally flip chip reported on a BGA package. The 72 mm^2 circuit integrates 1980 TSVs and as many microbumps between dies, and 933 flip-chip bumps to the BGA package. Fig. 15 shows the circuit micrograph, floor plan, and the 3D cross section.

B. 3DNOC Circuit Link Performances

The performance measurement of the four 3D NoC links is performed using the on-chip traffic generators, described in Section III-C, providing traffic performance measurement accuracy below 1 MHz. Performances are given in megabit per second, equivalent to 32 b word MFlit/s. The measurement results are presented Fig. 16. The baseline full-parallel link version has a cycle time measured at 3.06 ns at 1.2 V and 25 °C, achieving 326 Mb/s. The fault-tolerant link achieves 268 Mb/s, while the serial link achieves 153 Mb/s without redundancy and 148 Mb/s with redundancy.

Overall power consumption numbers and energy efficiency of the 3D traffic are extracted by differential measurements using the NoC power supply and the microbuffer power supply. The measurement setup was the following: 1) measurement of different 2D NoC traffic patterns in the bottom die to extract 2D router power consumption and 2) then, measurement of different 3D NoC traffic patterns to extract the 3D link power consumption. The 3D microbuffer is using a separate power supply, which allows extracting directly its power consumption. The measured power numbers were confirmed by power simulations using Synopsys PrimePower. The measured power for the 3D link in typical conditions is 0.66 pJ/b for the parallel link and 0.85 pJ/b for the fault-tolerant link, of which 0.32 pJ/b is dissipated on the microbuffers power supply.

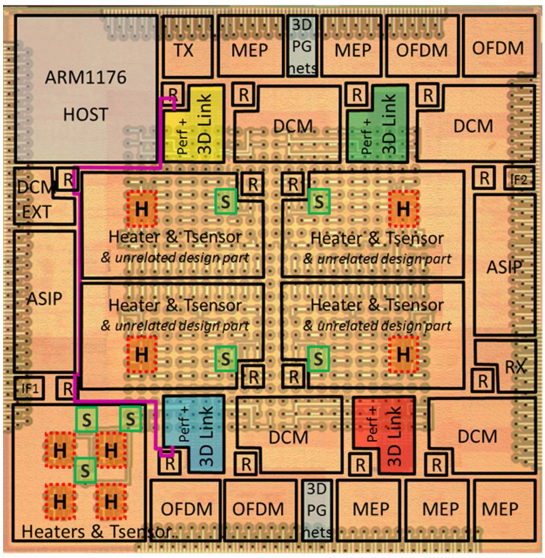
The latency breakdown (Fig. 17) has been refined by simulation. The sum of the contributions of asynchronous logic (200 ps), boundary scan (70 ps), microbuffers, and TSV (450 ps) leads to a forward link delay of around 720 ps between asynchronous pipeline stages on both dies for the full-parallel link version. The backward acknowledgment logic is about 40 ps slower, and due to the four-phase protocol, the complete cycle time is the result of two forward phases and two backward phases, providing the 326 Mb/s. Fault tolerance only adds 40 ps, while serialization by two adds 180 ps latency for a half word, and needs two four-phase handshakes per word, roughly dividing the throughput by two.

The performance bottleneck of the asynchronous 3D link is clearly the microbuffer cell loading the TSV and including ESD protection, which creates a large parasitic capacitance. It is also interesting to compare the relative performances of the 2D NoC traffic versus the 3D NoC traffic. The 2D NoC traffic has been measured at 890 Mbps in same conditions, to be compared with the 3D NoC link, measured at 326 Mbps. The 3D link achieves 2.7 times slower bandwidth compared with the 2D traffic, due mostly to the microbuffer 450 ps delay (including ESD and TSV capacitance), which implies a large impact due to four-phase handshake protocol. Nevertheless, even if 3D link throughput is reduced within the 3D NoC topology, the latency and power consumption of a 3D link is very close to a regular 2D router (respectively, 720 ps and 0.64 pJ/b). The aggregate bandwidth of the 32 b bidirectional NoC link is then 20.8 Gb/s. As a conclusion, the 3D NoC link provides energy efficient, low latency, and relatively high traffic bandwidth for future 3D many-core systems.

Self-adaptation thanks to asynchronous logic has been verified by performance measurements at different (statically defined) temperatures using a thermal chamber. Operation of the circuit is triggered after 15 min for each temperature. As seen in Fig. 18, the highest performance is achieved for a parallel link at 344 Mb/s at −25 °C. Performance drops by around 50 Mb/s within the whole temperature range: −25 °C–100 °C. This self-adaptation of the asynchronous logic allows removing the timing margins required for a synchronous link between dies. Compared with similar 3D stacked circuits [10], [11], [13], [14] (Fig. 19), the 3DNOC circuit provides the highest bitrate on the 3D link, with the smallest power consumption of the 3D link, while offering ESD protection, design-for-test, and fault tolerant features.

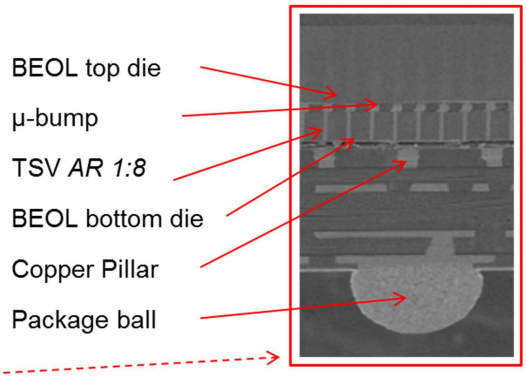
VI. 3D Thermal Impact on Performances

The 3D technology has strong impact on thermal effects, due to increased power density and thinned die [43].



Bottom die photo (72 mm²)

Cmos Process	**65nm** *STMicroelectronics*, Low-Power, Multi-VT
3D Tech.	**1980 TSV middle** (AR 1:8), *CEA-LETI* **µ-bumps,** 50µm x 40µm pitch **Face2Back** stacking, **Die2Die** assembly
Package	12x12x1.2 **flip-chip** package, 4 layers substrate
Complexity	1.63 Mbyte SRAM, 228 Mtransistors, 276 IOs
Die Size	8.5mm x 8.5mm = **72.2 mm²**
Power Supply	Core = **1.2 Volts**, I/O pads = 2.5 Volts



3D cross section

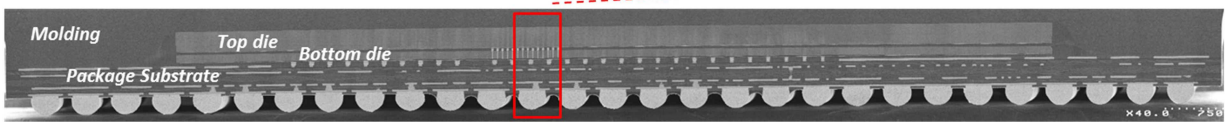
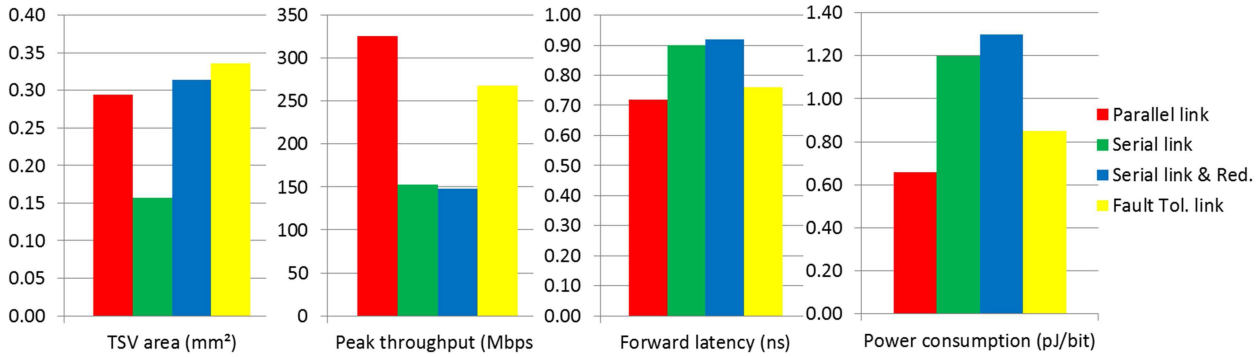


Fig. 15. 3DNOC circuit microphotograph, overlapped floor plan, technology description, and 3D cross section.



	Number of TSV	TSV area (mm²)	Peak throughput (Mbps)	Forward latency (ns)	Power consumption (pJ/bit)
Router 2D/3D	—	—	**890**	**1.10**	**0.69**
Parallel link	2x 92	0.294	326	0.72	0.66
Serial link	2x 49	0.157	153	0.90	1.20
Serial link & Red.	2x 98	0.314	148	0.92	1.30
Fault Tol. link	2x105	0.336	268	0.76	0.85
3D I/F signals	**688**	**1.101**	**7.4 GB/s**	—	**55.7 mW**
3D I/F power delivery network	**1292**	**2.067**	—	—	**deliver up to 800 mW**
2-layer circuit total	**1980**	**3.168**	—	—	**1.3 W**

Fig. 16. 3D NoC link performance measured at 25 °C 1.2 V.

The 3DNOC circuit has been extensively used for 3D thermal characterization (Section VI-A), while dynamic adaptation of 3D NoC link performances has been demonstrated on an application board (Section VI-B).

A. 3DNOC Circuit Thermal Behavior

The 3DNOC circuit is instrumented with eight resistive heaters and seven thermal sensors per layer to enable the use

	Asynchronous pipeline	*Optional Ser/Des*	*Optional Redundancy*	*Optional Fault Tolerance*	Boundary Scan cells	Micro-buffers & TSV
Latency contribution (ns)	**0.20**	*0.18*	*0.02*	*0.04*	**0.07**	**0.45**
Power contrib. (pJ/bit)	**0.23**	*0.54*	*0.10*	*0.19*	**0.11**	**0.32**

Fig. 17. 3D link latency and energy breakdown per stage.

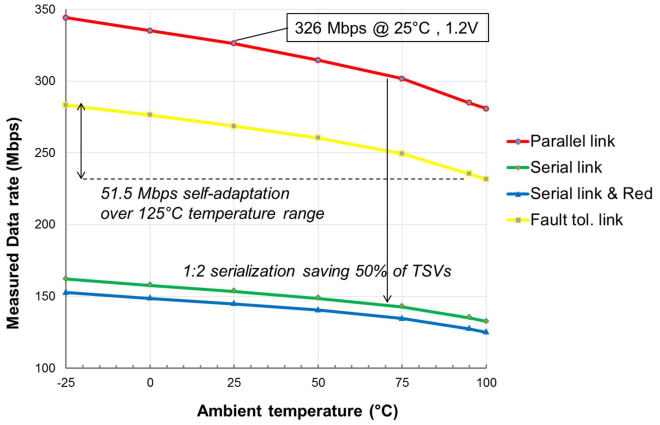


Fig. 18. Asynchronous link performances self-adaptation versus temperature (performance measurements within a thermal chamber).

of very accurate power dissipation patterns and the characterization of the resulting thermal responses. Each heater is supplied by a dedicated off-chip dc–dc converter controlled via embedded software and can dissipate up to 1 W, which results in a max power density of 0.56 kW/cm^2. The integrated bandgap thermal sensors are monitored in real time with a temperature resolution of 1 °C and an accuracy of ±1 °C for the calibration temperature (25 °C) and ±4 °C at 100 °C. As shown in Fig. 20, four heaters are placed in the bottom-left corner to emulate a strong hot spot behavior, such as, for instance, the one produced by a quad-core processor, while the other four heaters are evenly distributed in the center area of the circuit.

Thermal modeling and simulation has been carried out using a thermal analysis prototype provided by Mentor Graphics Calibre [44], [45], which relies on the FloTHERM simulation engine. The developed thermal model includes the detailed geometry representation and thermophysical properties of the complete system: the 3D NoC chip, its package, and the application board. The effective thermal property extraction feature in the thermal analysis tool extracts the equivalent thermal properties for complex geometries in GDS layout files, thus allowing an accurate modeling of the fine-grain vertical interconnections used in 3D technology, such as TSVs and μ-bumps. Thermal measurements and simulations with the calibrated thermal model are used for a detailed understanding of the thermal impact of the 3D integration technology.

Fig. 20 shows the correlation between measured temperature and steady-state simulations and also the resulting temperature maps for a strong hot spot application scenario with four heaters in the bottom die dissipating a total power of 2.54 W. The implemented thermal model presents a very good accuracy, the worst case difference between simulation and measured data is equal to 3.75% while the average difference considering all thermal sensors is lower than 2%.

Heat dissipation in hot spots is primarily diffused through the silicon substrate and usually spreads in a semispherical direction, rapidly decreasing the heat density and lowering the peak temperature. In the case of the 3DNOC circuit, the thinned bottom die (80 μm) presents reduced lateral heat spreading capability, while the underfill layer (25 μm) acts as a thermal barrier between layers due to its poor thermal properties. These combined effects result in exacerbated hot spots thus explaining the difference of peak temperature between the bottom and top dies (15 °C) which can be observed in Fig. 20. A more systematic study of the thermal impact of the 3D technology can be found in [46].

B. 3DNOC Circuit Performances Dynamic Adaptation to Thermal Behavior

To exercise further the 3D NoC links with 3D temperature effect, a complete demonstration has been setup with a laptop controlling the application board (Fig. 21). As shown in Fig. 22, embedded software is executed on the ARM1176 core, which activates heaters injecting a power scenario, measures the 14 thermal sensors of the bottom and top dies, and measures the on-chip NoC traffic under thermal variations. The reconstructed thermal maps of the bottom and top dies are computed and displayed by the laptop, as well as the performances of the four 3D vertical link versions (parallel, serial, serial with redundancy, and parallel with fault tolerance), and the performance of a 2D path of both bottom and top dies (*pink link on the floor plan*). As presented in Section II-B, whatever the gate and wire delays are, the QDI asynchronous handshake protocol enforces the correct sequence of request and acknowledge events, which leads to the correct behavior of the asynchronous logic, providing robustness and self-adaptation of the performances to any source of PVT variations. Contrarily to setup and hold margins requiring complex clock adaptation schemes for synchronous design, the performance curves show the self-adaptation of the asynchronous link to dynamic temperature variations (up to 15 Mb/s variation for the 3D links and up to 50 Mb/s variation for the 2D links), and its robustness to such environment constraints.

	3D-MAPS ***[D. Kim, ISSCC'2012]***	***Centip3de*** ***[D. Fick, ISSCC'2012]***	***512b Wide-I/O*** ***[J.S. Kim ISSCC'2011]***	***4096b Wide-I/O*** ***[S. Takaya, ISSCC'2013]***	***This Work*** ***[P. Vivet, ISSCC'2016]***
Application	Manycore CMP	Manycore CMP	2-stack DRAM	DRAM+Logic	Telecom Baseband
Technology	130nm	130nm	50nm	90nm	65nm
3D partitioning	Heterogeneous CPU/Cache	Heterogeneous CPU/SRAM	Homogeneous Shared TSV	Heterogeneous Active Int.	Homogeneous NoC
3D technology	Face-to-face Cu-Cu	Face-to-face Cu-Cu	Face-to-back TSV	Face-to-back TSV last	Face-to-back TSV mid.
3D scalability	no	no	yes	N/A	**yes**
Inter-die 3D pitch	5 µm	5 µm	50 µm	50 µm	40 µm
Inter-die 3D signal count	50000	25456	512	4224	688
Total 3D bandwidth	70.9 GB/s	254 GB/s	12.8 GB/s	100 GB/s	7.4 GB/s
Inter-die 3D Data rate	277 Mbps	80 Mbps	200 Mbps	200 Mbps	**326 Mbps**
Inter-die 3D I/O power	— (no VDDio)		0.78 pJ/bit	0.56 pJ/bit	**0.32 pJ/bit**
Inter-die 3D supplies	42,752	N/A	252	~3200	1292
Inter-die 3D test	no	no	dedicated boundary scan	N/A	**Boundary Scan + JTAG elevator**
Inter-die 3D fault-tol.	no	no	no	1:32 ratio	**1:8 ratio**
Inter-die 3D ESD prot.	no	no	N/A	no	**200V HBM**

Fig. 19. 3DNOC circuit performance summary and comparison table.

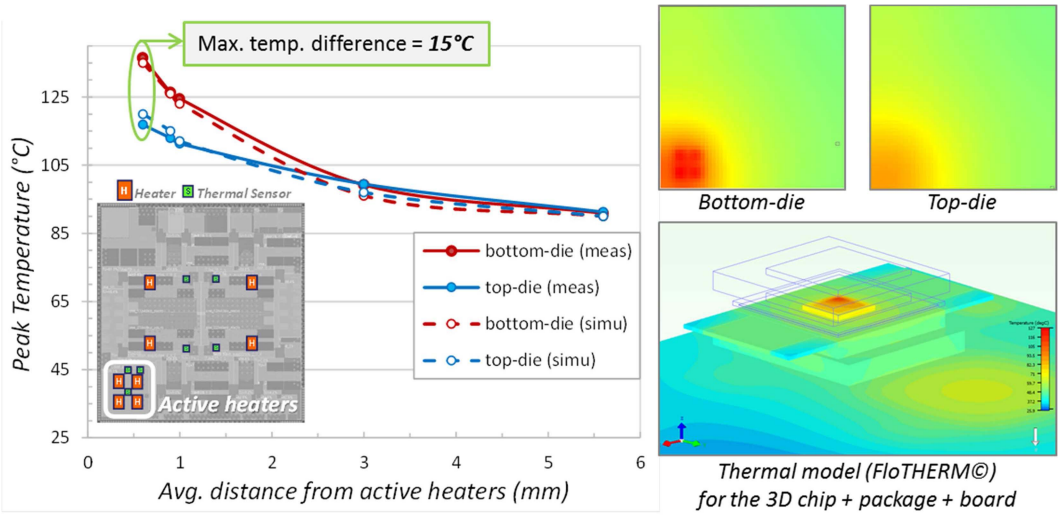


Fig. 20. Measured thermal effect, thermal simulations, and thermal maps, using active heaters.

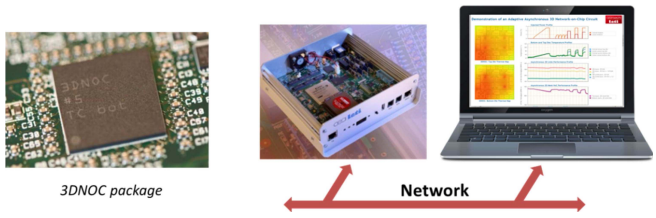


Fig. 21. 3DNOC circuit package, mother board, and laptop for circuit demonstration.

VII. Scalability Evaluation of a Multiple Layer 3DNOC Circuit 3D Stack

This section investigates the feasibility, from the power delivery and thermal points of view, of stacking multiple layers, up to eight layers, using the circuit architecture and layout proposed in the 3DNOC circuit. This scalability study of the 3DNOC circuit is performed by using simulations. For this purpose, every layer in the 3D stack presents the same power budget of 800 mW, corresponding to a worst case scenario as estimated by PrimeTime simulations when all NoC computing units are active, considering a typical corner, 1.2 V and 25 °C, with local clocks at 400 MHz, resulting in an average power density of 1.1 W/cm^2 per layer. Fig. 23 shows the details of the instance power map used for static IR drop and thermal analyses.

A. Evaluation of the 3D Power Delivery Network

As detailed in Section III-C (Fig. 5), power is supplied from the BGA package to the 3D stack through flip-chip C4-bumps

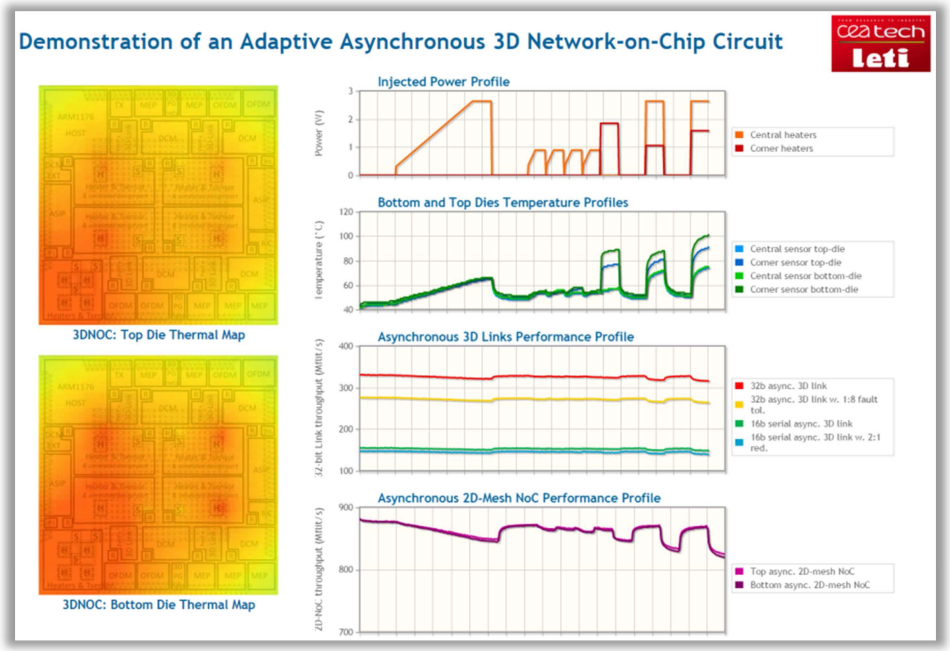


Fig. 22. 3D NoC demonstration: circuit performances dynamic adaptation to thermal behavior (*laptop screen capture*).

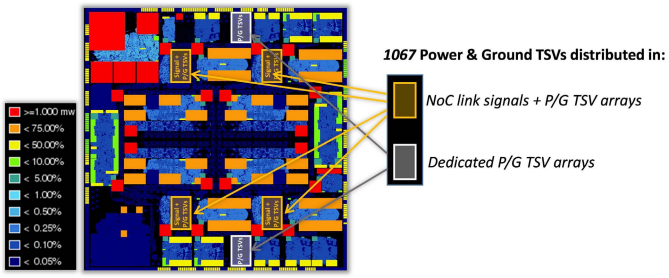


Fig. 23. Instance power map with the location of the P/G TSVs.

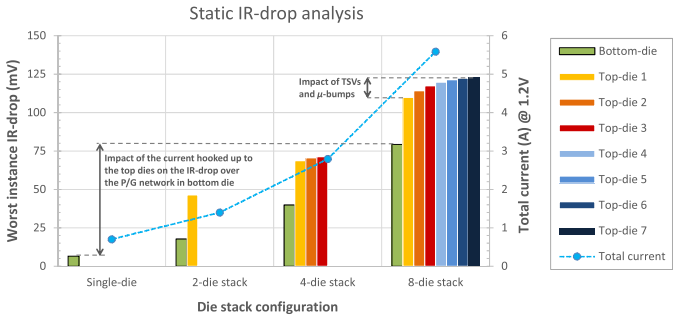


Fig. 24. Static instance IR-drop results and total current consumption for different die stack configurations (single die, two die, four die, and eight die).

connected to the bottom die. Out of a total 933 C4-bumps, 76 power and 172 ground C4-bumps are used to connect the 1.2 V digital power domain to the P/G mesh in the bottom die. In addition to two dedicated P/G TSV arrays, each of the four TSV arrays used for the physical interface of the 3D NoC link also contains TSVs used to deliver power to the top die (Figs. 12 and 23), hence providing a total of 1067 P/G TSVs for the 1.2 V power domain. It is important to notice that the dies are stacked in an F2B configuration, where the TSVs are connected to aligned μ-bumps without any additional routing using RDL, hence minimizing the voltage drop. Static IR-drop analyses have been performed using ANSYS RedHawk for 3D stacks with two, four, and eight layers, and also for the case of a single die.

Fig. 24 shows the worst instance IR drop per layer for the different die stack configurations, as well as the total current within the 1.2 V digital power domain. Fig. 25 presents the resulting current and instance IR-drop maps for the case of an eight-die stack. These static instance IR-drop results report the combined voltage drop and ground bounce, thus meaning the reduction of the instance voltage swing. It is possible to observe that the IR drop in the bottom-most die is higher in regions close to the TSV arrays, since such TSVs are responsible to supply power to the stacked top dies and thus could be seen as greedy current sinks. Therefore, every additional stacked die provokes extra voltage drop over the power delivery network in the bottom die and through the TSV and μ-bump arrays. In the case of the top dies, the IR drop is higher in regions containing instances with high power consumption and which are away from the TSV arrays, such as the ARM core located at the top-left corner of the die (Fig. 23). Considering the 3D NoC topology and that every layer exhibits the same power distribution, the difference of the worst instance IR drop between the top dies in a given stack configuration (up to 10 mV) reveals the impact of the resistance of the TSVs (25 mΩ each) and μ-bumps (15 mΩ each) on the 3D power delivery network. Finally, the simulated worst case instance IR drop is acceptable even for the eight-die stack configuration where the peak instance IR drop (123 mV) is just

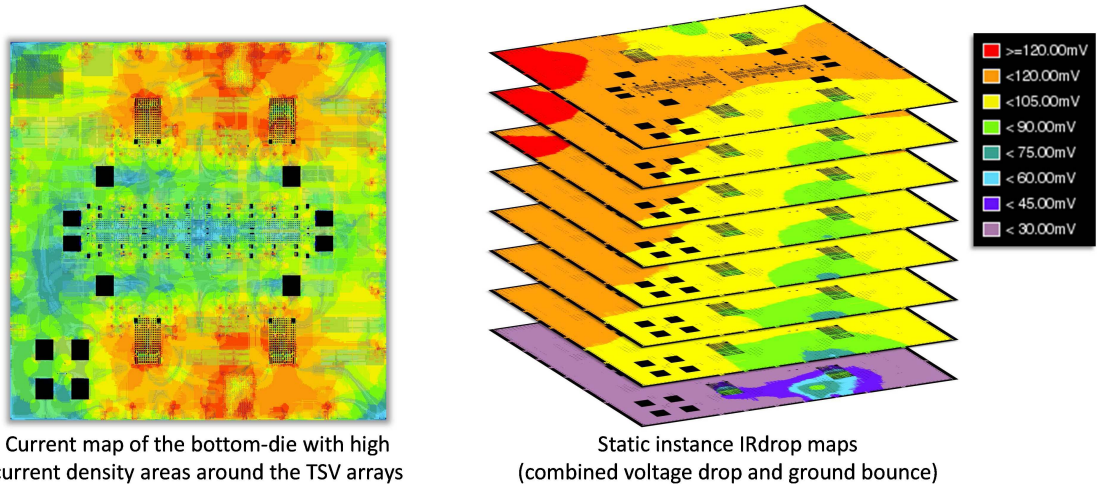


Fig. 25. Current and instance IR-drop maps for an eight-die stacked 3DNOC circuit.

over 10% of the 1.2 V voltage domain. Further optimizations of the 3D power delivery network and TSV allocation could be envisaged to reduce the IR drop [47]. Alternatively, thanks to the GALS 3D NoC, the exceeding voltage drop could be compensated within each layer and each NoC computing unit by using independent clock configurations.

B. Evaluation of the Thermal Dissipation

The same thermal analysis tool and methodology described in Section VI-A have been used to perform steady-state thermal analysis of the 3DNOC circuits with up to eight stacked dies. However, instead of using the integrated heaters, an applicative power scenario (Fig. 23) was used to generate heat. The resulting peak temperature per layer is shown in Fig. 26 considering that the packaged circuit is under boundary conditions typically found in low-power applications. In this case, heat transfer coefficients (HTC) of 20 and 200 W/K-m^2 are applied on the top and bottom surfaces of the package, respectively. For the considered power scenario, the temperature distribution is rather uniform over each layer and, contrarily to the case of hot spots, the heat flows mostly perpendicular to the silicon substrates toward the bottom and top package surfaces, which act as heat sinks. This results in an increase of the peak temperature, which is nearly linear with regards to the number of stacked dies and to the total power consumption. These results show that a 3D NoC stack containing four layers is still acceptable with the current BGA package (max temperature of 117 °C), but in the case of eight active layers where the peak temperature reaches 204 °C, it requires the use of a thermal packaging with improved heat removal capabilities. Fig. 26 (blue dashed line) reports additional simulation results, where an HTC of 600 W/K-m^2 is applied on the top surface of the package to emulate the use of an enhanced thermal package featured with a copper lid and a graphite sheet as heat spreader [46]. This brings a considerable reduction of the peak temperature (down to 93 °C), allowing the possibility of integrating up to eight active layers in a 3D NoC configuration.

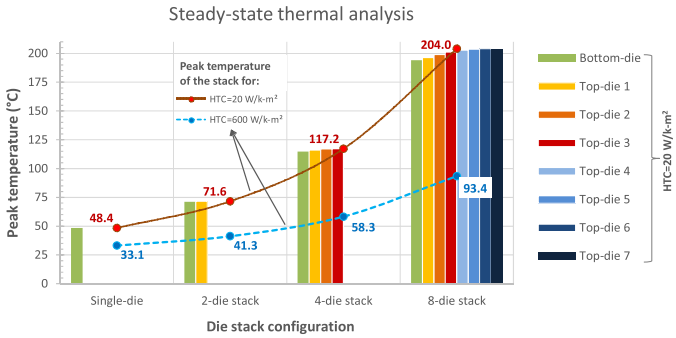


Fig. 26. Steady-state thermal analysis results for different die stack configurations. Peak temperature per stack configuration is also reported for the case of a higher HTC applied on the top of the package to emulate an improved heat transfer conditions.

VIII. Conclusion

Compared with the previous works on 3D-stacked logic circuits, which are mostly memory-on-logic [10], [11], [13], [14], or a single 3D NoC vertical link [19], this paper presents an homogeneous logic-on-logic circuit partitioning, with a large 4 × 4 × 2 3D NoC topology, modular and scalable in an F2B configuration. The 3DNOC circuit has been successfully designed, fabricated, tested, and demonstrated on board. The asynchronous 3D link design is robust, self-adaptive to PVT conditions, integrates ESD protection, and integrates an extensible test and fault-tolerant architecture. It achieves low energy consumption on its 3D I/O power supply (0.32 pJ/b), for an overall 0.66 pJ/b per link and a high data rate (326 Mb/s) for the parallel link, and offers a 7.4 GB/s 3D aggregate bandwidth. A study of 3D thermal effect shows the thermal impact of thinned 3D layers, and finally, the self-adaptation of the 3D NoC asynchronous links to such dynamic thermal variations. The proposed 3DNOC circuit is fully modular and scalable. As studied by simulations for the targeted application power budget, the IR drop is sustainable up to eight layers using the implemented power delivery network, while the thermal dissipation is sustainable up to four layers using a classical package, and up to eight

layers using a more advanced package technology to optimize thermal dissipations. As a conclusion, large 3D multicores can be designed using the proposed asynchronous 3D NoC infrastructure, targeting either heterogeneous architecture as the Telecom baseband presented here, or more general purpose multicore architecture, for high performance computing. As a next step, software development and characterization of the multilayer 3D MIMO Telecom baseband will be carried on, which will provide application-based system level performance numbers of the overall 3D NoC architecture.

REFERENCES

[1] (2015). *International Technology Roadmap for Semiconductors*. [Online]. Available: http://www.itrs.net/

[2] J. U. Knickerbocker *et al.*, "2.5D and 3D technology challenges and test vehicle demonstrations," in *Proc. IEEE 62nd Electron. Compon. Technol. Conf. (ECTC)*, May/Jun. 2012, pp. 1068–1076.

[3] T. Thorolfsson, S. Melamed, W. R. Davis, and P. D. Franzon, "Low-power hypercube divided memory FFT engine using 3D integration," *ACM Trans. Design Autom. Electron. Syst. (TODAES)*, vol. 16, no. 1, Nov. 2010, Art. no. 5.

[4] R. Chaware, K. Nagarajan, K. Ng, and S. Y. Pai, "Assembly process integration challenges and reliability assessment of multiple 28nm FPGAs assembled on a Large 65nm passive interposer," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2012, pp. 2B.2.1–2B.2.5.

[5] T. Kirihata *et al.*, "Three-dimensional dynamic random access memories using through-silicon-vias," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, to be published.

[6] J. Pawlowski, "Hybrid memory cube," in *Proc. IEEE Hot Chips Symp.*, vol. 23. Aug. 2011, pp. 1–24.

[7] U. Lee *et al.*, "25.2 A 1.2V 8Gb 8-channel 128GB/s high-bandwidth memory (HBM) stacked DRAM with effective microbump I/O test methods using 29nm process and TSV," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2014, pp. 432–433.

[8] K. Cho, H. Lee, and J. Kim, "Signal and power integrity design of 2.5D HBM (high bandwidth memory module) on SI interposer," in *Proc. Pan Pacific Microelectron. Symp. (Pan Pacific)*, Jan. 2016, pp. 1–5.

[9] J. Macri, "AMD's next generation GPU and high bandwidth memory architecture: FURY," in *Proc. IEEE Hot Chips Symp.*, Aug. 2015, pp. 1–26.

[10] J. S. Kim *et al.*, "A 1.2V 12.8GB/s 2Gb mobile wide-I/O DRAM with 4×128 I/Os using TSV-based stacking," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2011, pp. 496–498.

[11] S. Takaya *et al.*, "A 100GB/s wide I/O with 4096b TSVs through an active silicon interposer with in-place waveform capturing," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2013, pp. 434–435.

[12] D. Dutoit *et al.*, "A 0.9 pJ/bit, 12.8 GByte/s wideIO memory interface in a 3D-IC NoC-based MPSoC," in *Proc. IEEE Symp. VLSI Circuits (VLSI)*, Jun. 2013, pp. C22–C23.

[13] H. Kim *et al.*, "3D-MAPS: 3D massively parallel processor with stacked memory," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2012, pp. 188–190.

[14] D. Fick *et al.*,"Centip3De: A 3930DMIPS/W configurable near-threshold 3D stacked system with 64 ARM cortex-M3 cores," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2012, pp. 190–192.

[15] S. K. Lim, "TSV-aware 3D physical design tool needs for faster mainstream acceptance of 3D ICs," in *Proc. Design Autom. Conf. (DAC)*, 2010, pp. 2–11.

[16] ESD Alliance. *3D Design Guide*. [Online]. Available: http://esd-alliance.org/sites/default/files/downloads/papers/Multi-die_IC_Design_Guide.pdf

[17] V. F. Pavlidis and E. G. Friedman, "3D topologies for networks-on-chip," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 15, no. 10, pp. 1081–1090, Oct. 2007.

[18] A. Sheibanyrad, F. Pétrot, and A. Jantsch. *3D Integration for NoC-Based SoC Architectures*. Springer, 2011.

[19] G. Beanato *et al.*, "Design and testing strategies for modular 3D-multiprocessor systems using die-level through silicon via technology," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 2, pp. 295–306, Jun. 2012.

[20] V. F. Pavlidis, I. Savidis, and E. G. Friedman, "Clock distribution networks in 3D integrated systems," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 19, no. 12, pp. 2256–2266, Dec. 2011.

[21] J. Bainbridge and S. Furber, "A delay-insensitive chip area interconnect," *IEEE Micro*, vol. 22, no. 5, pp. 16–23, Sep./Oct. 2002.

[22] D. Rostislav, V. Vishnyakov, E. Friedman, and R. Ginosar, "An asynchronous router for multiple service levels networks on chip," in *Proc. IEEE Int. Symp. Asynchronous Circuits Syst. (ASYNC)*, Mar. 2005, pp. 44–53.

[23] T. Bjerregaard and J. Sparso, "A router architecture for connection-oriented service guarantees in the MANGO clockless network-on-chip," in *Proc. IEEE Design Autom. Test Europe Conf.*, vol. 2. Mar. 2005, pp. 1226–1231.

[24] J. J. H. Pontes, M. T. Moreira, F. G. Moraes, and N. V. Calazans, "Hermes-AA: A 65nm asynchronous NoC router with adaptive routing," in *Proc. IEEE Int. SOC Conf.*, Sep. 2010, pp. 493–498.

[25] A. Ghiribaldi, D. Bertozzi, and S. M. Nowick, "A transition-signaling bundled data NoC switch architecture for cost-effective GALS multi-core systems," in *Proc. IEEE Design Autom. Test Eur. Conf. (DATE)*, Mar. 2013, pp. 332–337.

[26] M. Krstic, E. Grass, F. K. Gýrkaynak, and P. Vivet, "Globally asynchronous, locally synchronous circuits: Overview and outlook," *IEEE Des. Test Comput.*, vol. 24, no. 5, pp. 430–441, Sep./Oct. 2007.

[27] P. Vivet, C. Bernard, E. Guthmuller, I. Miro-Panades, Y. Thonnart, and F. Clermidy, "Interconnect challenges for 3D multi-cores: From 3D network-on-chip to cache interconnects," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Jul. 2015, pp. 615–620.

[28] P. Vivet *et al.*, "A 4×4×2 homogeneous scalable 3D network-on-chip circuit with 326 MFlit/s 0.66 pJ/bit robust and fault tolerant asynchronous 3D links," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Feb. 2016, pp. 146–147.

[29] W. Lafi, D. Lattard, and A. Jerraya, "A stackable LTE chip for cost-effective 3D systems," *IPSJ Trans. Syst. LSI Design Methodolo.*, vol. 5, pp. 2–13, Feb. 2012.

[30] F. Clermidy *et al.*, "MAGALI: A network-on-chip based multi-core system-on-chip for MIMO 4G SDR," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers (ISSCC)*, Jun. 2010, pp. 74–77.

[31] Y. Thonnart, P. Vivet, and F. Clermidy, "A fully-asynchronous low-power framework for GALS NoC integration," in *Proc. IEEE Design Autom. Test Eur. (DATE)*, Mar. 2010, pp. 33–38.

[32] W. Lafi, D. Lattard, and A. Jerrya, "An asynchronous hierarchical router for networks-on-chip-based three-dimensional multi-processor system-on-chip," *J. Softw., Pract. Expererience*, vol. 42, no. 7, pp. 877–890, Jul. 2012.

[33] P. Maurine, J. B. Rigaud, F. Bouesse, G. Sicard, and M. Renaudin, "Static implementation of QDI asynchronous primitives," in *Proc. Int. Workshop Power Timing Modeling, Optim. Simulation (PATMOS)*, Sep. 2003, pp. 181–191.

[34] M. T. Moreira, M. Arendt, A. Ziesemer, R. Reis, and N. L. V. Calazans, "Automated synthesis of cell libraries for asynchronous circuits," in *Proc. 27th Symp. Integr. Circuits Syst. Design (SBCCI)*, Sep. 2014, pp. 1–7.

[35] Y. Thonnart, E. Beigne, and P. Vivet, "A pseudo-synchronous implementation flow for WCHB QDI asynchronous circuits," in *Proc. IEEE Int. Symp. Adv. Res. Asynchronous Circuits Syst. (ASYNC)*, May 2012, pp. 73–80.

[36] F. Darve, A. Sheibanyrad, P. Vivet, and F. Petrot, "Physical implementation of an asynchronous 3D-NoC router using serial vertical links," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Jul. 2011, pp. 25–30.

[37] C. Papameletis, B. Keller, V. Chickermane, S. Hamdioui, and E. J. Marinissen, "A dft architecture and tool flow for 3D SICs with test data compression, embedded cores, and multiple towers," *IEEE Des. Test*, vol. 32, no. 4, pp. 40–48, Aug. 2015.

[38] M. Taouil, S. Hamdioui, and E. J. Marinissen, "Quality versus cost analysis for 3D Stacked ICs," in *Proc. 32nd IEEE VLSI Test Symp. (VTS)*, Apr. 2014, pp. 1–6.

[39] Y. Fkih, P. Vivet, M. L. Flottes, B. Rouzeyre, G. D. Natale, and J. Schloeffel, "3D DFT challenges and solutions," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Jul. 2015, pp. 603–608.

[40] I. Loi, S. Mitra, T. H. Lee, S. Fujita, and L. Benini, "A low-overhead fault tolerance scheme for TSV-based 3D network on chip links," in *Proc. Conf. Comput. Aided Design (ICCAD)*, Nov. 2008, pp. 598–602.

[41] [Online]. Available: http://www.cadence.com/solutions/3dic/Pages/default.aspx

[42] [Online]. Available: http://go.mentor.com/calibre-3dstack

[43] C. Torregiani, B. Vandevelde, H. Oprins, E. Beyne, and I. D. Wolf, "Thermal analysis of hot spots in advanced 3D-stacked structures," in *Proc. 15th Int. Workshop Thermal Invest. ICs Syst. (THERMINIC)*, Oct. 2009, pp. 56–60.

[44] White Paper Mentor Graphics. (Jul. 2014). *7 Key Considerations for Effective Chip-Package Thermal Co-Design a High-Level 'How to' Guide*. [Online]. Available: http://www.mentor.com/products/mechanical/resources/overview/7-key-considerations-for-effective-chippackage-thermal-co-design-7fde3df0-b1d6-47d9-82c1-72ff9e3d8af2

[45] T. R. Harris, P. Franzon, W. R. Davis, and L. Wang, "Thermal effects of heterogeneous interconnects on InP/GaN/Si diverse integrated circuits," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, Dec. 2014, pp. 1–3.

[46] C. Santos, P. Vivet, J. P. Colonna, P. Coudrain, and R. Reis, "Thermal performance of 3D ICs: Analysis and alternatives," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, Dec. 2014, pp. 1–7.

[47] B. S. Wang, F. Firouzi, F. Oboril, and M. Tahoori, "P/G TSV planning for IR-drop reduction in 3D-ICs," in *Proc. Design, Autom. Test Europe Conf. Exhibit. (DATE)*, Mar. 2014, pp. 1–6.

Pascal Vivet received the Ph.D. degree from the Grenoble Polytechnic Institute in 2001, where he was involved in designing an asynchronous microprocessor.

After four years with STMicroelectronics, he joined the Digital Design Laboratory, CEA-LETI, Grenoble, France, in 2003, where he is currently a Senior Expert. He has authored or co-authored over 80 papers, and holds several patents in the field of digital design. His current research interests include asynchronous design, network-on-chip, energy efficient multicores, 3D architectures, and related CAD aspects.

Dr. Vivet was a member of the Organizing Committee of the 3D Workshops at DATE from 2013 to 2015, and has been serving as a member of the D43D Workshops since 2011. He participates in various TPCs, such as ASYNC, NOCS, DATE, ICCAD, and 3DIC conferences.

Yvain Thonnart (M'16) was born in Paris, France, in 1980. He received the Engineering degree from École Polytechnique, Palaiseau, France, in 2003, and the Engineering Diploma degree in electrical engineering from Télécom ParisTech, Paris, in 2005.

Since 2005, he has been a Researcher with CEA-LETI, Grenoble, France, where he is an Expert on communications and synchronization in system-on-chips (SoCs). He is currently leading a project on silicon photonics interposers for optical communications in massively parallel SoCs. His research interests include asynchronous logic, network-on-chip architectures, and physical implementation of energy-efficient SoCs.

Romain Lemaire received the engineering degree from the Superior School of Electricity, Supelec, in 2002, and the Ph.D. degree in microelectronics from the Grenoble Institute of Technology, Grenoble, France, in 2006.

He joined CEA-LETI, Grenoble, as a Research Engineer in Digital Design and System Architecture, in 2006. He contributed to the design of advanced on-chip communication and control architectures for heterogeneous system-on-chips. He is currently involved in various projects, including multicore architecture and 3D-stacking integration.

Cristiano Santos received the B.S.E. degree in electrical engineering from the Federal University of Santa Maria, Santa Maria, Brazil, in 2003, and the M.Sc. and Ph.D. degrees in computer science and microelectronics from the Federal University of Rio Grande do Sul, Porto Alegre, Brazil, in 2005 and 2014, respectively.

He was a Physical Design Engineer with Freescale Semiconductor Inc. from 2008 to 2012. He is currently a Research Engineer with CEA-LETI, Grenoble, France. His current research interests include heat dissipation, thermal analysis, power integrity methodologies for advanced 3D integration, and packaging technologies.

Edith Beigné received the Engineering and the HDR degrees from the Grenoble Polytechnic Institute in 1998 and 2014, respectively.

She joined CEA-LETI, Grenoble, France, in 1998. Since 2009, she has been a Senior Scientist with the Digital and Mixed-Signal Design Laboratory, where she is involved in low-power and adaptive circuit techniques. She was leading complex test-chip design dedicated to low power and variability management, exploiting asynchronous design, and advanced technology nodes. She has authored or co-authored over 120 papers, and holds several patents in the field of low power and adaptive digital circuits.

Ms. Beigné serves on the ISSCC Committee as the Digital Circuits Chair, since 2014, and the VLSI Symposium since 2015.

Christian Bernard received the engineering degree from the Grenoble Polytechnical Institute in 1979.

After four years with Thomson, he was with Bull, Paris, France, on mainframe HW design of CPU cores, multiprocessing, and cache coherency aspects. He joined the Digital Design Laboratory, CEA-LETI, Grenoble, France, in 2001, where he is currently a Senior Designer. He contributed to the design of system-on-chips in the laboratory, covering domains such as 4G mobile baseband, space mission, and many core architectures.

Florian Darve received the Technological Research Diploma degree from Grenoble Joseph Fourier University, Grenoble, France, in 2011.

He accomplished his technological research at CEA-LETI, Grenoble, France, in partnership with the TIMA Laboratory, Grenoble, where he was involved in 3D architecture with contributions to the design of NoC fast serial link, and to testability and fault tolerance mechanisms for the 3D architecture. From 2011 to 2013, he was with Dolphin Integration as a Physical Designer for the implementation of many SoC circuits. From 2014 to 2016, he was with CEA-LETI, where he was involved in the physical implementation of 3D many-core architecture and ultralow-power circuits for Internet of Things.

Didier Lattard received the Ph.D. degree in microelectronics from the National Polytechnic Institute of Grenoble, Grenoble, France, in 1989.

In 1990, he joined the Center for Innovation in Micro and Nanotechnology, CEA-LETI Laboratory, Grenoble. He was involved in the design of image and baseband processing circuits as a Senior Research and Development Engineer and a Project Leader. From 2003 to 2006, he led the development of the FAUST NoC-based telecom platform. From 2006 to 2014, he led a project in performance computing applications. In 2014, he was with the Technology Department and is currently involved in the development of 3D stacking technologies. He has authored over 60 papers in books, refereed journals, and conferences. He holds 19 patents in the fields of baseband processing and NoC architectures.

Ivan Miro-Panadès received the M.S. degree in telecommunication engineering from the Technical University of Catalonia, Barcelona, Spain, in 2002, and the M.S. and Ph.D. degrees in computer science from University Pierre and Marie Curie, Paris, France, in 2004 and 2008, respectively. His Ph.D. work dealt with the DSPIN network-on-chip and multisynchronous interfaces for globally asynchronous and locally synchronous circuits.

He joined CEA-LETI, Grenoble, France, in 2008, where he is currently a Research Engineer in Digital Integrated Circuits. His current research interests include Internet of Things architectures, energy-efficient system design, multicore SoC design, and Fmax tracking methodologies on advanced CMOS technology nodes.

Denis Dutoit received the Engineering degree from the École Nationale Supérieure d'Électronique et de Radioélectricité de Grenoble, Grenoble, France, and the Ph.D. degree in signal processing from the École Nationale Supérieure des Télécommunications de Paris, Paris, France.

He joined CEA-LETI, Grenoble, France, in 2009. He has been involved in system-on-a-chip architecture for computing and 3D integrated circuit projects. He is currently a Strategic Marketing Manager, in charge of defining Leti's roadmap of technologies and solutions for advanced computing.

Dr. Dutoit is a co-recipient of the Jan Van Vessem Award for Outstanding European Paper at ISSCC 2005.

Fabien Clermidy received the Ph.D. degree in microelectronic from INPG, Grenoble, France, in 1999, and the Supervisor degree in 2011.

He is a pioneer in designing network-on-chip-based multicore. He was the Leader of the second generation of network-on-chip-based multicore dedicated to 3GPP-LTE. At this period, his team elaborated one of the first 3D multicore prototypes embedding a WIDE-IO DRAM memory called WIOMING. He is currently managing the Digital Circuit Laboratory implied in the development of new architectures using emerging technologies, such as 3D TSV, 3D monolithic integration, and emerging memories. He has authored or co-authored two books and over 75 journal and conferences papers. He holds 15 patents.

S. Cheramy received the Engineering degree in material science Polytech Orleans,Orléans, France, in 1998.

She has spent over eight years with GEMALTO, a leading smart-card company developing technologies for secure solutions, such as contactless smart cards and electronic passports. In 2008, she joined as a 3D Project Leader with the CEA-LETI, Grenoble, France, where she is currently the Head of the 3D Integration Laboratory. This group develops technology and integration for 3D IC, in strong relationship with 3D design, model, and simulation teams. She is also the Director of the Institute of Technological Research 3D program.

Abbas Sheibanyrad was born in Dezful, Iran. He received the bachelor's degree in computer hardware engineering from Tehran Polytechnic University, Tehran, Iran, in 2000, the master's degree in microelectronics and integrated system architectures from the University of Pierre et Marie Curie, Paris, France, in 2004, and the Ph.D. degree in computer science, telecommunications, and electronics from the Computer Lab De Paris 6, University of Pierre et Marie Curie .

He held a post-doctoral position for a year, and then he was a Research Fellow with the TIMA Laboratory, Grenoble, France, up to 2014, and since then, he has been with the ALSOC Team of the LIP6, Paris. His current research interests include future SoC/NoC architectures.

Frédéric Pétrot received the Ph.D. degree in computer science from Université Pierre et Marie Curie (Paris VI), Paris, France, in 1994.

He was an Assistant Professor with Université Pierre et Marie Curie until 2004. From 1989 to 1996, he was one of the main contributors of the open source Alliance VLSI CAD system. Since 1996, he has been involved in the specification, simulation, and implementation of multiprocessor systems on chip architectures, including circuits, software, and CAD aspects. He joined the Grenoble Institute of Technology, Grenoble, France, as a Professor in 2004. Since 2006, he has been the Head of the System Level Synthesis Group, TIMA Laboratory, where he is currently a Deputy Director.

Eric Flamand received the Ph.D. degree in computer science from INPG, France, in 1982.

He was as a Researcher with CNET and CNRS, France, involved in architectural automatic synthesis, design, and architecture, compiler infrastructure for highly constrained heterogeneous small parallel processors. He then held different technical management in the semiconductor industry, first with Motorola, Toulouse, France, where he was involved into the architecture definition and tooling of the StarCore DSP, and then with STMicroelectronics being in charge of the software development of the Nomadik Application Processor and then responsible of the P2012 corporate initiative aiming at the development of a many core device. He is currently a Co-Founder and a CTO of Greenwaves Technologies, Grenoble, France, a French-based startup, developing an IoT processor derived from Pulp. He is a Part Time Consultant with ETH-Z, Zurich, Switzerland.

Jean Michailos received the Ph.D. degree in physics from the CNRS, Grenoble University, in 1989.

From 2007 to 2009, he started the first TSV 300 mm line dedicated to CMOS Image sensors that was a precursor of the future 3D developments. He joined the Thales Group in 1990, where he was involved in solid-state X-ray sensors for medical digital imaging. Since 1997, he has been with STM, Crolles, France, among the Silicon Technology Development Team. He is currently in charge of 3D technologies developments, dedicated to image sensors, RF, photonics, and advanced logics. He is also a 3D Integration Senior Program Manager with STMicroelectronics. His current research interests include the technology developments of differentiated technologies (RF/analog, passives, and image sensors).

Alexandre Arriordaz was a RTL Verification Engineer with Infineon Technologies France (now Intel Mobile), Sofia, France, from 2003 to 2005. After this first experience, he was a Full-Custom Design Engineer and a Project Manager with Freescale, Crolles, France, from 2005 to 2007, first, and then in Austin, TX, USA, from 2007 to 2010. Since 2010, he has been with Mentor Graphics, Grenoble, France, where he was involved in advanced Research and Development features for the Calibre Division (parasitic extraction, electrical rule checking, and physical verification for 3D-IC or photonic circuits.). He is currently leading the Local Calibre Technical Marketing Team and interfacing with local partners to support and promote Calibre tools in the scope of European projects (like IRTNanoelec).

Lee Wang received the B.Sc. degree (Hons.) in electrical and electronics engineering, and the M.B.A. degree (Hons.) in management of technology.

She was with the semiconductor industry as an IC Designer and progressing on to the management of a team with the Microelectronics Design Center, Large Multinational Corporation. She was with the Electronic Design Automation Industry for 16 years working with customers on the application and product requirements of physical verification for IC design. In her most recent undertaking, he has been working with customers to develop a thermal analysis solution for 2.5-D/3D IC design. She is a Technical Marketing Engineer with the Design to Silicon Division, Mentor Graphics Corporation.

Juergen Schloeffel (M'04) received the Diploma degree in physics from the University of Göttingen, Göttingen, Germany.

He is a Program Manager in the area of EDA and DFT with Mentor Graphics Development, Hamburg, Germany. He holds several patents and has authored and co-authored over 60 conference papers and journals. His current research interests include advanced testing techniques, IJTAG, design automation for DSM technologies, and 3D test.

Mr. Schloeffel is a member of VDE, and a Board Member of the German ITG Working Group for test and reliability. He has served on program committees of several conferences and workshops.