Design and Development of High Density Fan-Out Wafer Level Package (HD-FOWLP) for Deep Neural Network (DNN) Chiplet Accelerators using Advanced Interface Bus (AIB)

Mihai D. Rotaru
Institute of Microelectronics
*A*STAR*
Singapore,
mihaidr@ime.a-star.edu.sg

Wei Tang
EECS Department
University of Michigan
Ann Arbor, MI, USA
weitang@umich.edu

Dutta Rahul
Institute of Microelectronics
*A*STAR*
Singapore
duttar@ime.a-star.edu.sg

Zhengya Zhang
EECS Department
University of Michigan
Ann Arbor, MI, USA
zhengya@umich.edu

2021 IEEE 71st Electronic Components and Technology Conference (ECTC) | 978-1-6654-4097-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ECTC32696.2021.00204

***Abstract*—Emerging applications such as machine learning (ML) and artificial intelligence (AI) require more computing capabilities that ought to be distributed and have access to large memory and storage, while the systems need to be energy efficient and low-cost. The increase in cost of advanced nodes and the difficulties of shrinking analog circuits such as input and output (I/O) to address the computation and communication needs of ML/AI applications have created the opportunity to bring into the mainstream chiplet-based systems. The chiplet based systems enable modularity, scalability and technology partitioning providing a cost and energy efficient solution. The chiplet integration has been enabled by the development of a raft of advanced packaging technologies such as silicon interposer, EMIB, COWoS, high density fan-out wafer level packaging (HD-FOWLP) to name a few. In this work the design, development and electrical characterization of a four-chiplet system integrated using in 2.5D HD-FOWLP platform is discussed. The chiplet accelerators are fabricated in 22 nm CMOS technology, while the package uses a five metal layer HD-FOWLP with dielectric polymer and 2 um width and space as minimum design rules. The Advanced Bus Interface (AIB) die-to-die PHY-level standard is used to interconnect the four chiplets in a ring topology. The AIB bus requires 192 lines between each two chiplets, and a total of 768 2umx2um lines are routed on the top three layers of the HD-FOLWP. The bottom two metal layers of the package are used to distribute the ground and power necessary for all four chiplets. Each chiplet requires seven distinct voltage islands that are separately routed on the bottom metal layer.**

Keywords— High Density Fan-out Wafer Level Package, chiplet, heterogeneous integration, signal and power integrity

I. INTRODUCTION

Applications such as big data, IoT, 5G and AI/ML require to handle large amounts of data and an increasing computing capability while the power consumption should be minimised and the cost of the system ought to be reduced. All these requirements come at the time when traditional semiconductor scaling has slowed down dramatically while the design, fabrication and test cost of large SoC solutions have been exponentially increasing [1,2]. To achieve an efficient and optimised solution for the applications listed above the system needs to integrate generic hardware such as CPU cores, GPU, embedded FPGAs, dense and fast memories together with specialised hardware such machine learning neural networks accelerators. Integrating all this functionality using a SoC approach, although desirable, it has reached a limitation due to the design and fabrication costs of large SoC solutions using advanced nodes such as 14 nm and below. A much more effective methodology in designing and building such systems is the approach based on heterogeneous integration using as enabler advanced packaging technologies. Large modular architectures can be built using heterogeneous chiplets to construct a complex system via integrating and stitching chiplets on organic substrates [4, 5], 2.5D silicon interposers [3], high density-FOWLP [6] or through the use of silicon bridges [7]. Cost reduction due to yield improvement [2], Known Good Die (KGD) strategies [3], and reuse of hard IPs can be readily achieved using a chiplet integration approach while the performance of the system is preserved or improved versus the SoC solution. Although heterogeneous integration has many advantages and its philosophy is straightforward, applying it is not a simple task because of the diversity of chiplets and I/O interfaces. Efforts such as DAHI [8] and CHIPS [9] by DARPA are made to standardise the die-to-die interfaces. The Advanced Interface Bus (AIB) developed by Intel has emerged as a one of the front runner standard that can be used to implement logic to logic communication.

As described in the specification documentation [10] AIB is a die-to-die PHY-level standard, that uses a clock forwarded parallel data transfer mechanism. The AIB channel has an array of AIB I/O buffers for collection of data, control, clock and asynchronous signals. Half of the AIB I/Os are configured in a master setting and the other half are configured as slave. The AIB is using a 1GHz clock and DDR hence each AIB I/O can support data transfers up to 2 Gbps. The AIB standard can be configured to use up to 24 channels. Each data channel can have up to 80 transmitters (Tx) and 80 receivers (Rx) data signals. At

2377-5726/21/$31.00 ©2021 IEEE
DOI 10.1109/ECTC32696.2021.00204

2 Gbps per data line, an aggregate bandwidth of 1920Gbps Tx and 1920Gbps Rx (3.84Tbps total) can be achieved. The routing of a full AIB interface will require more than 3900 lines to be routed within roughly 8 mm of die beachfront.

In this work a package using HD-FOWLP RDL with a polymer dielectric technology has been developed to integrate four chiplets in a ring topology. The chiplets are deep neural network accelerators that communicate via the AIB interface. The HD-FOWLP technology offers a very flexible packaging platform that allows routing complex bus structures and package chiplets that have already been hardened and designed to be integrated using other advanced packaging technologies.

II. System Arhitecture

A. Chiplet and system arhitecture

High-performance and energy-efficient ASIC chips can be designed to accelerate the processing of large DNN models. However, both the model size and the input data are growing larger, making it infeasible to build a fixed ASIC chip that remains useful to new generations of applications.

Instead of building a DNN chip, we opt for building a modular DNN chiplet, and use many copies of the chiplet to compose a system to meet the requirements of new applications. A 4mm x 4mm 22nm chiplet was designed. It consists of an engine to compute convolution (CONV) and an engine to compute fully connected (FC) layers as shown in Fig. 1. Weights are stored in on-chip SRAM arrays. Inputs are streamed into the chiplet. The processing elements in the chiplet perform multiplications of inputs and weights, and the outputs are accumulated and streamed out.

The fixed chiplet size limits the processing capacity and the weight storage. The processing capability limits the maximum frame rate in processing DNNs, and the weight storage limits the model size that can be stored on-chip. To support a large DNN model, multiple copies of the DNN chiplet are needed.

In our proposed system, the chiplets are connected in a ring via high-bandwidth IOs as shown in Fig. 2. Each chiplet is assigned a set of DNN layers. The high-bandwidth IOs are placed on one side of the chiplet. We make use of the AIB interface to enable 80 Gbps input and output per channel that consume sub-pJ/b. 55um-pitch microbumps are used to enable a bandwidth density of 256 Gbps/mm-shoreline. Each chiplet uses 4 AIB channels for input and 4 AIB channels for output. Standard 130um-pitch bumps are used for low-speed GPIOs. A chiplet contains several power domains for the compute engines, the AIB IOs, as well as the GPIOs.

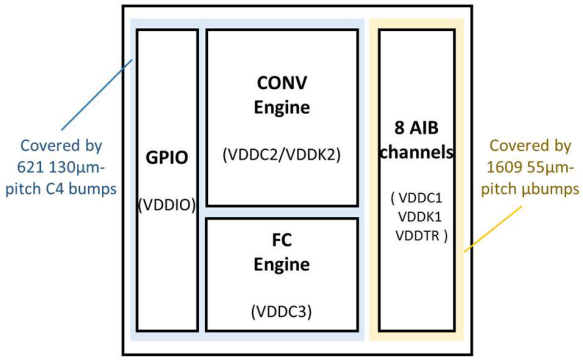
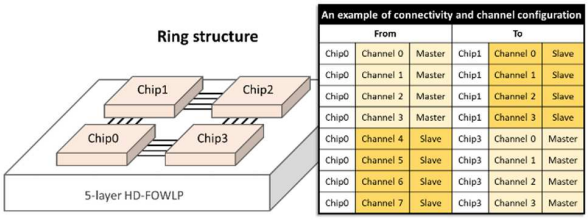


Fig 1. Chiplet functional block diagram.



An example of connectivity and channel configuration

From			To		
Chip0	Channel 0	Master	Chip1	Channel 0	Slave
Chip0	Channel 1	Master	Chip1	Channel 1	Slave
Chip0	Channel 2	Master	Chip1	Channel 2	Slave
Chip0	Channel 3	Master	Chip1	Channel 3	Slave
Chip0	Channel 4	Slave	Chip3	Channel 0	Master
Chip0	Channel 5	Slave	Chip3	Channel 1	Master
Chip0	Channel 6	Slave	Chip3	Channel 2	Master
Chip0	Channel 7	Slave	Chip3	Channel 3	Master

Fig 2. Integration of four chiplets, and an example of channel connection and configuration.

B. Package Design and Layout

The HD-FOWLP package structure built in this work had a 5-layer stack-up. The top metal layer, that is closer to the four chiplets, consists of UBM pads on which the chiplets are flipped and C4 bumps and micro-bumps are created for the electrical interconnects. The next two metal layers immediately below the UBM layer are used for the AIB bus interconnect. The metal lines used to route all the AIB channels have a 2um x 2um cross section and minimum spacing between neighbouring lines of 2um. The following two metal layers are used to distribute the power and ground network of the system. The bottom layer is routed out to solder balls that are used to connect to the next packaging level. Fig 3. illustrates the cross section of the package.

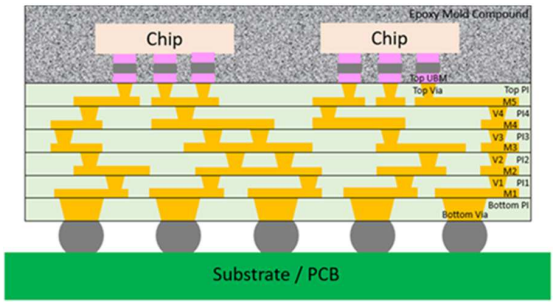
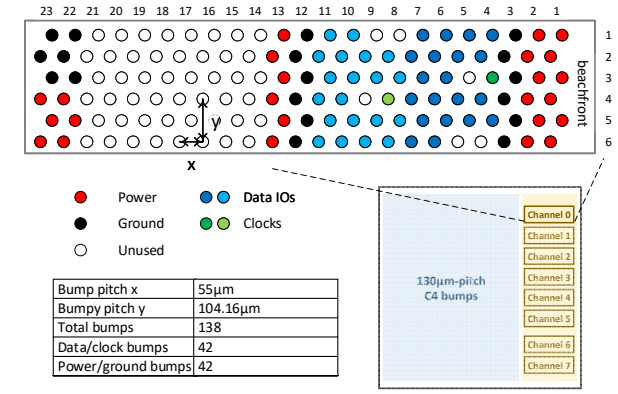


Fig. 3. Illustration of the five layer RDL FOLWP package

Leveraging the 5-layer HD-FOWLP, four chiplets were connected by high-bandwidth IOs as shown in Fig. 2. Each chiplet used 8 AIB channels, where four channels were grouped and configured as either master or slave channels. Following the legacy bump design in AIB specification documentation [10], each channel is organised in banks of 138 bumps (6 rows of 23 bumps) with a 55um pitch. The first two columns (closest to the beachfront) are used to distribute the power for the AIB IOs. The third column is used for the ground. From column 4 to column 11 the bumps are used for the AIB IOs (20 Tx data and 20 Rx data) as well as other signals such as clock, control and asynchronous. Column 12 is assigned for ground followed by another column used for power. From column 14 to column 21 there are another 48 bumps allocated for data (40 IOs) and 8 for other type of signals. Column 22 and 23 are used again for power and ground. The map of the IOs in one master channel is shown in Fig. 4. In this implementation only 42 of the IOs from the master channel need to be connected to the slave channel. Note that in the slave channel configuration the Tx signal IOs are in the back of the bank away from the beachfront while the master channel has the Tx signal IOs allocated to the front of the array (Fig. 4).

In all the AIB implementations that we are aware of, connection between the master and slave channels was done through a straight point to point interconnection, hence the beachfronts where the AIB interface are located are facing each other. The current chiplet implementation (Fig. 4) has the AIB IOs situated on one side only and connecting four chiplets will result in connections that will have be routed sideways.



Bump pitch x	55µm
Bumpy pitch y	104.16µm
Total bumps	138
Data/clock bumps	42
Power/ground bumps	42

Fig. 4. AIB channel bump map specifications.

Several chiplet placement trials were conducted, in order to understand and resolve various routability induced signal and power integrity challenges. There are a variety of chiplet placement and pin assignment choices that may be used to minimize the electrical impact, where the key contributor to electrical uncertainty are the package interconnect parasitics including complex capacitive coupling between neighbouring AIB wires. In order to execute a complete chiplet placement, routing, extraction and analysis design flow, an advanced HD Fanout package design kit has been developed. The development task involved generation of test keys/structures, measurement and calibration of simulation models of basic building blocks such as (micro bumps, RDL, C4 bumps).

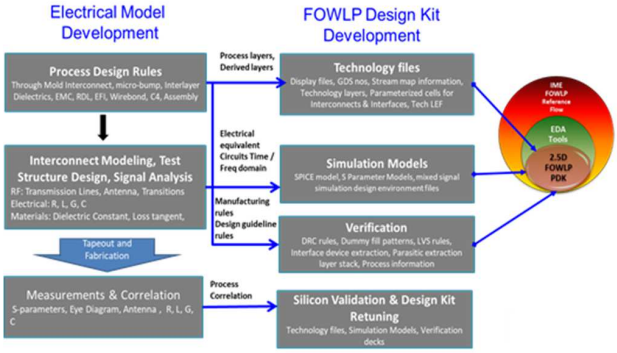


Fig. 5. HD FOWLP PDK development process, supporting upstream design tools from Cadence, verification checks using Mentor (Siemens) tools and system analysis tools from Keysight

Referring Fig. 5, an open access based design kit and a reference design methodology is built to support NN accelerator die data import, design view generation (abstract, layout, symbol, CDF), placement planning, automatic routing, and parasitic extraction framework. We adopted a routability aware design flow, which first improves signal assignment between NN accelerator chiplets by guiding the choice of master and slave channels based on global routing analysis and later implementing with required number of signal / power RDL layers. Fig. 6 depicts several of our physical implementation trials.

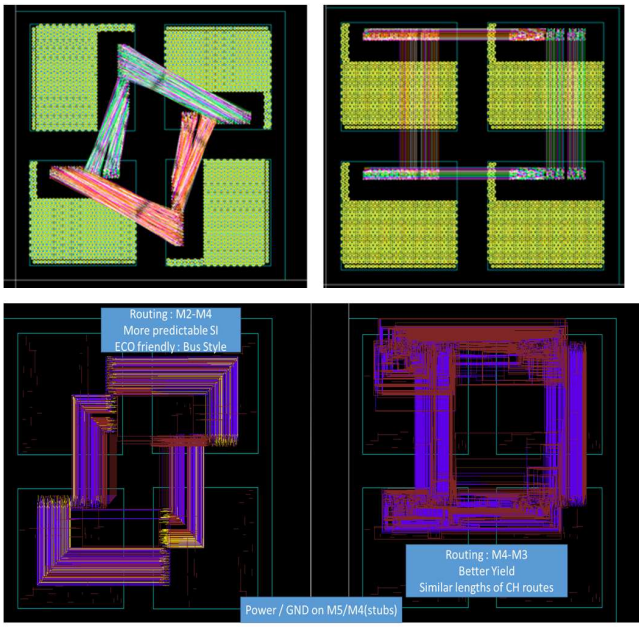


Fig. 6. Four NN accelerator chiplet configurations with different chiplet placement trials

Since the AIB specifications require that maximum skew between data edges within a channel is 0.02*UI and between clock and data edges within a channel is 0.01*UI, we have

adopted the same length signal routing (Fig. 6, R.H.S.) for each channel. Due to the constraints explained above, positioning the chiplets with the AIB interface facing each other did not result in a topology that allowed routing such that the skew specification could have been met. The complete SIP design reference flow is depicted in Fig. 7, which is centred around interconnect analysis. Here the design kit provides essential automation to import physical and logical information of the chiplets and system connectivity. This

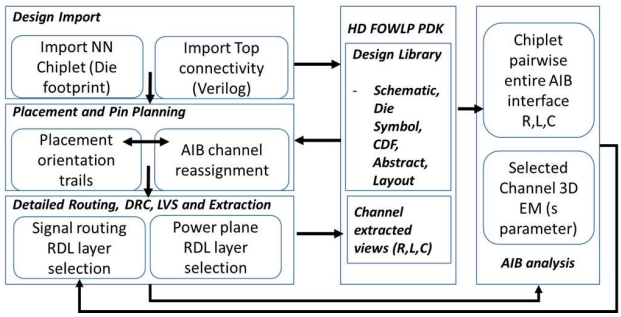
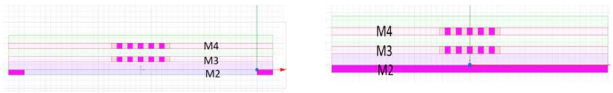


Fig. 7. Interconnect aware SIP reference design flow.

generates essential design views needed for the placement trials and pin assignments, where the global routing statistics are analysed to make decisions on the possible placement and master slave channel configurations. Progressive trial route is performed with extraction and back annotation of interconnect parasitics using extraction CAD tools, while critical portion of the channel is analysed within the 3D EM full wave simulator using S-parameter model to understand the coupling effects. Routing quality feedback with more trial routes are conducted to fully close the interconnect design loop. Referring to Fig. 6, we could fully route signals using 2 layers M4 and M3 while the ground and PDN plane Power and Ground is implemented on M2 and M1 respectively.

III. Electrical Response of the High Density Bus

As discussed in the previous section several possible arrangements of the accelerator chiplets relative to each other have been explored. For all these possible scenarios the wires were assumed to be routed on M4 and M3 layers (Fig. 3.). In order to understand the signal integrity performance of these structures several simulations and measurements of test structures have been done.



a. b.

Fig. 8. High density bus cross section structure: a. the return path is constrained; b. the return path is a full metal plane.

The length of the bus structures varies from 4.4 mm to 5.8 mm, however as explained earlier the length was kept the same as long as the wires were part of the same channel. With these assumptions two scenarios were simulated and analysed. The two situations are graphically shown in Fig. 8. In the first case the return path for the signals routed on layers M3 and M4 was on M2 but it was constrained on certain areas and it had very limited width, in the second case the return path routed on M2 was a full metal plane.

Full wave simulations were done for different lengths of the bus. The routing information as well pads and vias used to connect different layers and to the bumps of the chiplets were included in the models, however their effects have been found to be minimal. This can be understood if the parasitic capacitance and resistance of the pads and vias are compared with the distributed capacitance and resistance of the 2um x 2um wires. As shown in Table I one millimetre of line will introduce a capacitance eight time larger than the capacitance of the pads and almost twenty times larger than the capacitance of the via while the resistance is hundreds of time larger than the resistance of pads or vias. Therefore, the signals traveling along these wires will be less affected by the vias or the pads. The values of the parasitics shown in Table I have been confirmed via frequency domain measurements.

Pads (55um pich)	Vias (3um diam, 3um height)	Wires (2umx2um)
C_{pads}=24 fF; R_{pads} = 11.7 mOhms,	C_{via} = 10.2 fF; R_{via} = 6 mOhms	C_{line}= 192 fF/mm; R_{line}= 4.31 ohm/mm

TABLE I. PARASITICS

To understand the signal integrity effects and the performance of the two different topologies shown in Fig. 8., a transient domain solver using a convolution engine and statistical simulation [11] was used to calculate the eye diagrams for low BER (<1e-16). The bus interconnects were represented by 20x20 S parameters data files, which included the coupling between 10 wires in the bus of interest. The drivers in the

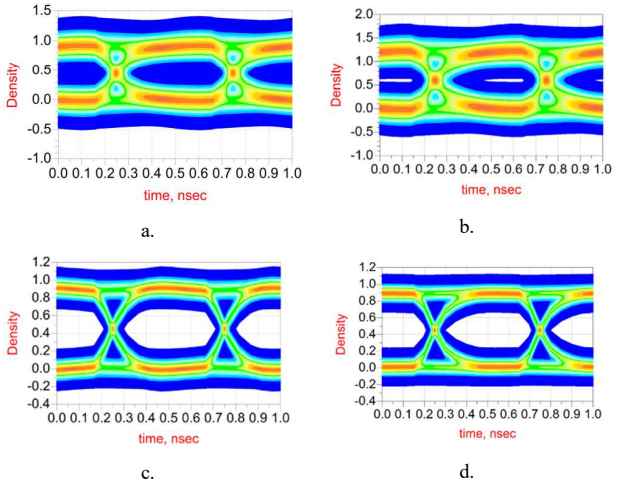
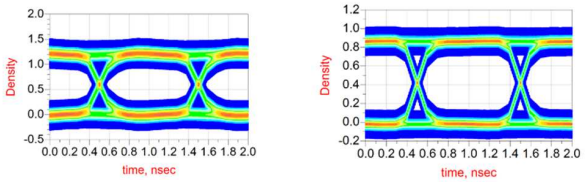


Fig. 9. Eye diagrams for a 6 mm long high density bus at 2 Gbps for different bus configurations.

transmitter side were represented by a source with a max voltage swing of 0.9V, with a complex source impedance that had the real part equal with 50 ohms and the imaginary part was represented by a capacitance that was set-up to 500 fF; the lines were terminated on the receiver side with a capacitance of 500

fF [10,12]. The results of the time domain simulation for 2 Gbps data rate are presented in Fig. 9.

From the above results (fig 9 c. and d.) it can be noted that providing a good return path is critical in achieving a good eye opening at 2 Gbps data rate, that meets the AIB specifications [10]. Further simulations have shown that by reducing the data rate to 1 Gbps the structure depicted in Fig. 8a. can also meet the AIB timing specifications (Fig.10.)



a. b.

Fig. 10. Eye diagrams for a 6 mm long high density bus at 1 Gbps with the return path constrained (a.), with the return path a full plane (b).

This improvement in the response of the bus can be explained if the frequency response of the 2um x 2um wires is considered. As described and discussed at length in [13], these wires can be loosely described as working in the RC regime, however this depends on the size of the wires as well as the Nyquist frequency of the data signal. For 2 Gbps signal the Nyquist frequency is 1 Ghz, and for the 2um x 2um wires, the boundary between the RC and LC regime is around 1 Ghz. From this frequency upwards the inductance that appears between the signal line and the return path becomes important and it will affect negatively the behaviour of the bus if the inductance is too large. On the other hand, when the Nyquist frequency is 0.5GHz (for the 1 Gbps data rate), the most important parameters of the signal lines are the R and C and the inductance is less important and has a much smaller effect on the signal integrity of the bus. It is important to recognise that by only considering these interconnects working in the RC it may not be sufficient especially in the case of HD-FOLWP with a thicker metallisation, 2 um and above. It is suggested that the high density bus structures should be modelled when possible as S-parameters with enough bandwidth such the full electromagnetic behaviour of these structures is captured and used for the time domain simulations.

IV. Power Delivery Network

The power delivery network for each chiplet requires seven separate power islands. As explained in the previous section the return path immediately below the layers where the AIB bus was routed is important from the signal integrity point of view, hence the M2 was implemented as a full ground plane. For the seven power islands the only option left available was layer M1, hence they were routed on M1. The voltages and the estimated current drawn by the chiplet through this PDN are listed in Table II.

The information on Table II was used to determine the number of solder bumps that are required to carry the maximum current for each VDD. Depending on the size of the bump and UBM at T_j=150ºC the typical values of current that can be safely handled by solder bumps range from 45 mA to 120 mA. For this package a 500 um pitch was chosen hence a 255 um ball and UBM will be used. Such a solder ball can carry safely 82.5 mA. The number of balls necessary for each VDD was calculated by dividing the maximum current with 82.5 mA. For the VDDIO, VDDK1 and VDDK2 where the current is small only two balls are used (Table II – third column). For each of the VDD ball a ground ball was added to the package, this brought the total

TABLE II. Voltages, Currents and Number of VDD Balls

	Voltage (Min/Typ/Max, V)	Estimated Current (Min/Typ/Max, A)	Number of Solder balls
VDDTR	0.81/0.9/0.99	0.3/0.5/1.0	13
VDDC1	0.68/0.85/1.02	0.3/0.5/1.0	13
VDDC2	0.68/0.85/1.02	0.5/1.0/2.0	25
VDDC3	0.68/0.85/1.02	0.05/0.1/0.2	3
VDDK1	0.68/0.85/1.02	0.05	2
VDDK2	0.68/0.85/1.02	0.05	2
VDDIO	1.62/1.8/1.98	0.05	2

number of balls used for the PDN to 120. Each of the chiplet required also 27 IOs to be routed out to the second level of packaging, hence the total number of bumps each chiplet requires is 147. With a 500 um pitch 147 solder balls can be distributed within a square area of 6.75mm x 6.75mm, hence the total area of the package for four chiplets is 13.5mm x 13.5mm (Fig. 11). The 28 separate power islands that constitute the VDD for all four chiplets are routed within the 13.5mm x 13.5mm area.

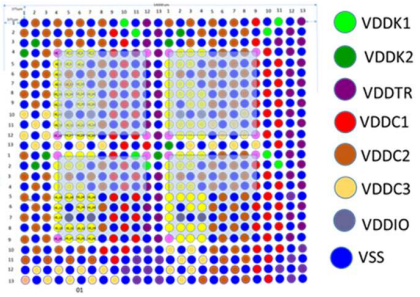


Fig. 11. Bump layout of the 4 chiplet SiP – yellow represent the IOs, while the rest of the bumps are used for distributing the PDN

The layout of the power islands as designed in layer M1 is presented in Fig. 12. It can be noted that the four quadrants of the package are not symmetrical. This is due to the chiplet layout which was fixed and hardened before the package design. As expected the largest islands are for VDDC2, VDDC1 and VDDTR. Models of the PDNs consisting of unit cells of 100um x 100um that were connected to realise the full size of the different islands were built and simulated to evaluate the response of the PDNs (Fig. 13). Each cell was represented by an equivalent circuit with a parallel capacitance between the VDD and the ground, and a series resistance and inductance representing the flow of current on the VDD and return path. The values for the unit cell capacitance, resistance and inductance were extracted from measured data on a test vehicle that was fabricated to characterise the electrical performances of

a multilayer FOLWP package. Based on the measured data the capacitance of a 100um x 100um square was 70fF while the inductance is 5.02 pH. The resistance has been estimated to be 17.25 mOhms.

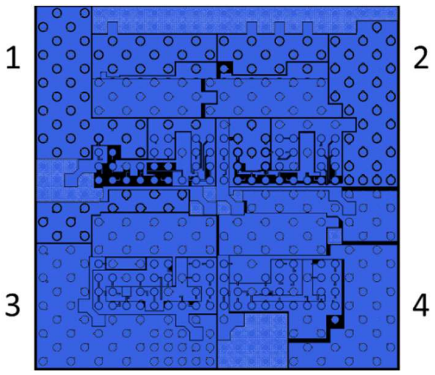


Fig. 12. Power islands layout in M1, the four quadrants of the package are labelled as above.

The computed self-impedance of the PDN for the VDDC2 in each of the four quadrants of the package is presented in Fig. 13. The VDDC2 in quadrant 1 (VDDC2-q1) has the lowest self-impedance which is mainly due to the lower inductance of this power island. VDDC2-q1 is a wider structure which minimises the inductance. The other three VDDC2 have similar area however, due to their routing (narrower structures) they have larger inductance. VDDC2-q2 has the largest inductance, which introduces the resonance noted at 2 GHz. These structures also have a relative large resistance, due to the thin metal (2um), hence reducing the PDN self-impedance through decoupling capacitors schemes could have limited success.

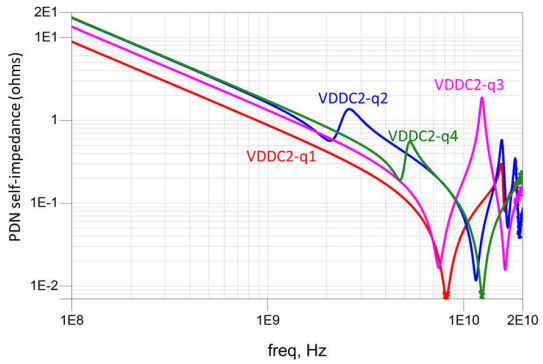


Fig. 13. Simulated PDN self-impedance for the VDDC2 in the four quadrants of the package.

At the time of preparing this manuscript the five-layer package described throughout this document is being fabricated. We expect that details of the functional testing of the four chiplet system will be available for presentation during the conference.

V. Conclusions

This paper discusses the conception, design and simulation of a system in package that integrates four DNN chiplets using the high density FOWLP technology. HD-FOLWP technology offers a very flexible platform to integrate chiplets that have been harden and initially designed to be integrated using different packaging approach. Also this packaging technology offers the opportunity to create novel systems, such as the DNN ring topology which will be very difficult or impossible to integrate using other approaches.

The developed design flow and PDK are very important enablers in for designing and optimising such chiplet based systems. The simulation and measured data have shown that the signal integrity and power integrity of our five layer RDL FOLWP package fulfils the specification of AIB based chiplet system.

Acknowledgment

The University of Michigan effort is supported in part by the Defense Advanced Research Projects Agency CHIPS program and the Office of Naval Research under grant N00014-17-1-2992. We thank Intel for the chiplet fabrication and bumping.

References

[1] T. Coughlin, “A road map for technologies that drive consumer storage,” IEEE Consum. Electron. Mag., vol. 8, no. 2, pp. 97–99, Mar. 2019.

[2] T. Coughlin, “New electronic arhitectures,” IEEE Consumer Electronics Magazine vol. 9, Issue: 2, March 1 2020.

[3] P. Vivet et al,. “ INTACT: A 96-Core processor with six chiplets 3D stacked on an active interposer with distributed interconnects and integrated power management,”, IEEE Journal of Solid-State Circuits, vol. 56, no. 1, Jan 2021.

[4] N. Beck, et al,. “Zeppelin’: An SoC for multichip architectures,” in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2018, pp. 40–41.

[5] M.-S. Lin et al., “A 7nm 4GHz Arm-core-based CoWoS chiplet design for high performance computing,” in Proc. Symp. VLSI Circuits, Jun. 2019, pp. 28–32.

[6] M.-F. Chen et al., “System on Integrated Chips (SoIC(TM) for 3D Heterogeneous Integration,” in Proc. IEEE 69th Electron. Compon. Technol. Conf. (ECTC), 2019, pp. 1–5.

[7] Mark Wade et al., “TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical I/O”, IEEE Micro vol: 40, issue: 2, March-April 1 2020.

[8] T. M. Hancock and J. C. Demmin, “Heterogeneous and 3D integration at DARPA,” 2019 International 3D Systems Integration Conference (3DIC), 8-10 Oct. 2019, Sendai, Japan.

[9] DARPA Microsystems Technology Office, “Broad Agency Announcemnt Common Heterogenous Intergation and IP Reuse Strategies (CHIPS),” DARPA Microsytem Technology Office, Sept 2016 [online], Availble : https://www.darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies

[10] “Advanced Interface Bus (AIB) Specifications” Revison 1.2. September 2019, [online], Available: https://github.com/chipsalliance/AIB-specification.

[11] ADS – User Manual.

[12] N. Pantano et al., ” Technology Optimization for High Bandwidth Density Applications on 3D Interposer”, 6th Electronic System-Integration Technology Conference (ESTC), 13-15 Sept., 2016, Grenoble, France.

[13] M.Rotaru and K.Li, ”Electrical characterization and design of hyper-dense interconnect on HD-FOWLP for die to die connectivity for AI and ML accelerator applications”, IEEE 22nd Electronics Packaging Technology Conference (EPTC), 2-4 Dec, 2020, Singapore.