Electrical design challenges in High Bandwidth Memory and Advanced Interface Bus interfaces on HD-FOWLP technology

Mihai Dragos Rotaru
Institute of Microelectronics
ASTAR Singapore
Mihai_Dragos_Rotaru@ime.a-star.edu.sg

Li Kangrong
Institute of Microelectronics
ASTAR Singapore
Li_Kangrong@ime.a-star.edu.sg

***Abstract*—Wide and slow bus interfaces such as High Bandwidth Memory (HBM) and Advanced Interface Bus (AIB) are the main drivers behind the development and implementation of technologies such as Embedded Multi-die Interconnect Bridge (EMIB) and Chip-on-Wafer-on-Substrate (CoWoS). These technologies can create very dense interconnect structures using back end of line (BEOL) approaches used to interconnect chiplets and to form functional system in package applications. With recent improvements in fabrication and process technologies organic based high density redistribution layer approaches such HD-FOWLP have become popular alternatives to EMIB and CoWoS. Redistribution layers with 2um x 2um cross section and 2um space between adjacent lines or 1um x 1um cross section with a minimum spacing of 1um have become achievable with the current fabrication technology. In this work the electrical design challenges in organic based fine RDL packages have been studied via simulation and electrical measurements.**

Keywords—wide and slow bus, HBM, AIB, multi-layer RDL, FOWLP

I. INTRODUCTION

Leading-edge applications such as artificial intelligence, machine learning, automotive, and 5G, all require high bandwidth, higher performance, lower power and lower latency. They also need to do this for the same or less money. Achieving these products using a monolithic system on chip approach (SoC) is not economic anymore because of the associated large costs of the design, prototyping and process. Also large and complex SoC solution have notorious long design cycles. A faster and a lower cost approach is to rely on heterogeneous integration [1]. Making use of very fine pitch interconnect technologies such as silicon interposers [1,2], silicon bridges [3] or high density fan-out approaches on organic substrates [4,5] the heterogeneous integration at package level provides solutions that meet the scaling demand while reducing the cost and shorten the design cycles. The modular approach of heterogeneous integration on advanced package technologies can address the bandwidth scaling issues while enabling integration of different silicon nodes, including the most advanced nodes if required, through stitching together different chiplets with fine pitch interconnects. 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 have been made to standardise the die-to-die interfaces such that different type of interconnect solution can serve these interfaces [6]. These efforts will eventually enable faster design cycles for the chiplets based modular systems. Two of the most popular interfaces are: HBM2/3 – High Bandwidth Memory and AIB – Advanced Interface Bus. The HBM interface is used to connect CPU/GPU/ASIC to large DRAM memories especially for leading-edge high-performance computing, graphics, and networking applications which demand massive bandwidth and high power efficiency while the AIB is used to interface two or more logic dies (e.g. ASIC/FPGA with an accelerator chiplet [3], ASIC/FPGA with electro-optical chiplet transceiver [7]).

The HBM2 standard has eight channels. Each channel is 128 wide with a data rate per pin of 2 Gbps (Gigabits per second). Higher data rates per pin (3.2 Gbps) are supported in the recent updates of to the HBM standard hence the eight channels could provide an aggregate bandwidth up to 409.6GBps (Gigabytes per second) [8]. The physical routing of the HBM interface requires about 1700 lines to be routed within available routing space – approximately 6mm of die beachfront.

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. Each data signal runs up to 2Gbps, producing an aggregate bandwidth per AIB interface of 1920Gbps Tx and 1920Gbps Rx (3.84Tbps total). The routing of AIB interface will require more than 3900 lines to be routed within roughly 8mm of die beachfront [9].

Both HBM and AIB interfaces can be implemented using different advanced packaging technologies such as silicon interposer [11], EMIB [3] or high density organic substrate [10]. In this paper the electrical design challenges using a polymer based HD-FOWLP RDL to implement these interfaces are discussed and presented.

The HD-FOWLP can allow up to six layers in the stack-up which combined with the size of the wires provide a versatile packaging platform to achieve very dense and wide bus structures. Using a 2um x 2um cross section wire with a 2 um spacing between adjacent wires, two metal layers will be

sufficient to route the HBM interface within the 6 mm beachfront using a HD-FOLWP implementation. Similarly, the AIB interface can be fully routed within a width of 8 mm and two layers using the same 2 um width and spacing wire arrangement. Wire density can be theoretically doubled using a 1um x 1um cross section and a 1um spacing, however consideration to the electrical performances need to be addressed. In this work two possible stack-up configurations were considered. First topology assumes that the top and third metal layers are used for routing, the second and fourth metal layers are used for distributing the ground plane, while the fifth metal layer is used for the VDD (Fig. 1). The second topology uses the top four layers to distribute the signal and ground within the same layer using Ground-Signal-Signal (GSS) pattern that is repeated across the four layers. The fifth metal layer is used to for the power as shown in figure 2.

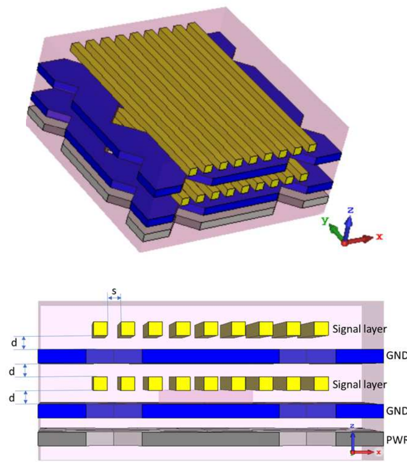

Fig. 1. First topology - 3D view and cross section view of the 5 layer HD FOLWP stack-up; s is the distance between adjacent signal; d is the distance between the signal layer and ground.

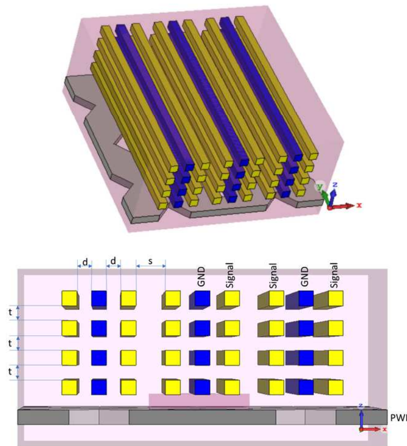

Fig. 2. Second topology - 3D view and cross section view of the 5 layer HD FOLWP stack-up; s is the distance between adjacent signal lines; d is the distance between the signal and ground; d is the distance between adjacent signal and ground layers. The top four layers are signals and ground in GSS pattern. The fifth layer is the power.

The cost of the HD-FOWLP is directly proportional to the number of metal layers hence minimising the number of layers should be considered in the design of such substrates. For example, if topology one (Fig. 1) is considered for a HBM interface with the signal wire cross section 2um x 2um and the gap between the adjacent wires 2um, the wire density per mm is 250 signal lines/mm. Hence in a 6 mm beachfront 1500 lines can be routed. Therefore, to rout all 1700 lines required by the HBM interface two signal layers are necessary. The full stack-up will require 5 layers – two layers for ground and one layer for power, which is the structure shown in Fig. 1. On the other hand, if the signal wire cross section is reduced to 1um x 1um with a 1um gap the wire density is doubled to 500 signal lines/m. In this situation all the 1700 lines required can be routed within the 6 mm width available in one layer. Therefore, the stack-up will require only 3 layers – one signal layer, one ground layer and one power layer. If the topology two (Fig. 2) is considered for the HBM example, for the 2um x 2um wires a density of 142 signal lines/mm is obtained. This value is obtained assuming that the gap between the signal lines and ground line is 2um, while the distance between two adjacent signal lines is 4um. In this configuration 852 lines can be routed within 6mm beachfront, hence 2 signal layers are required to rout all 1700 HBM lines. This stack-up will require three metal layers – two for signals and ground and one for power. Once again if the 1um x 1um cross section is considered the wire density is doubled and all the 1700 signal lines and the required ground lines can be routed in one layer, hence the stack-up will only require two layers. A similar analysis can be done for the AIB case. From the above discussion it is apparent that the topology two has the potential to provide structures that have the lowest number of layers, hence reducing the cost of such RDL based solutions. However, the signal integrity of the two possible topologies will have to be considered before the correct and most economical solution is selected.

II. Frequency Domain Analysis of Signal Lines

A. Performance regions of the signal line

The first aspect that has to be considered is that these very small cross section interconnects have a very large DC resistance. For example, a 2um x 2um cross section copper interconnect will have a DC resistance of about 4.31 ohms/mm while a 1um x 1um wire will have a four times larger DC resistance. It is instructional to study the attenuation curves of these interconnects and their variation with frequency.

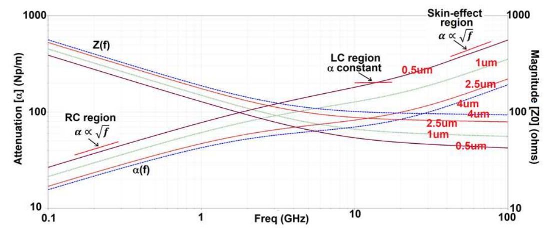

Fig. 3. Frequncy variation of attenuation curves (left axis) and characteristic impedance (right axis) for 1um x 1um wires routed on the top layer of the first topology (Fig. 1); d = 0.5, 1, 2.5 and 4 um dielectric thickness between the signals and return path are shown.

The curves in Fig. 3 can be used to explain the behaviour of the 1um x 1um lines. Unlike the thicker structures used in PCBs and packages, the 1um x 1um line have very strong and clear RC behaviour in the lower parts of the frequency band. The RC region is affected by the thickness of the dielectric layer between the signals and the return path. The structures with a thinner dielectric layer have a larger attenuation and the boundary between the RC and LC regimes are at a higher frequency. This is due to the proximity effect that is stronger for thinner dielectric forcing the return current in a smaller region of the return path, hence increasing the total resistance. The results in Fig. 3 were obtained via simulation with the assumption that the polymer dielectric material has a 3.2 dielectric constant and 0.02 tangent loss. The characteristic impedance of these structures also vary a lot in the RC regime. At 1GHz, which is the Nyquist frequency for 2Gbps HBM2 and AIB interfaces, the 1um x 1um wires will have a characteristic impedance magnitude of more than 120 ohms for the thinnest (d=0.5 um) dielectric case and almost 190 ohms for d=4 um, thick dielectric case (Fig. 3).

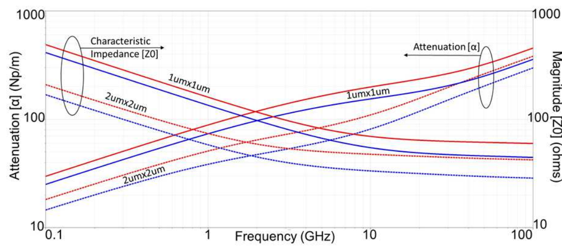

Fig. 4. Variation with frequency for α and |Z0| for 1um x 1um (full lines) and 2um x 2um (dotted lines) wires routed in toppology one configuration. Top layer wires (microstrip) - red lines for α, blue for |Z0|; third layer (strip line) – blue lines for α, red for |Z0|.

The effect of increasing the size of the wire cross-section for both topologies is illustrated in Fig 4 and Fig. 5 respectively. It can be noted that both the attenuation as well as the magnitude of the characteristic impedance of these wires are decreasing. The change in the attenuation can be explained by the decrease in the wire resistance for the larger cross section lines, while the drop in the characteristic impedance can be attributed in the increase of the capacitance of the wires with a larger cross section. Fig. 4 & 5 also present the effect of different layers to the attenuation and characteristic impedance. A much larger change can be noted for the topology one (Fig. 1) where the return path has a stronger effect on the total resistance and capacitance. The top layer in topology one, which is a microstrip structure, has overall a smaller attenuation but the characteristic impedance is larger. The effect of the different layers is less evident for the topology two (Fig. 2) as the return path of the signal for both layer is almost the same. The proximity of the power plane to the second layer has a minor effect on the attenuation and impedance that can be noted better in the 2um x 2um wire cross section case. The results presented in Fig. 5 assumes that distance to the power plane is the same for both the 1um x 1um and the 2um x 2um case.

Another important difference that can be noted in Fig. 4 & 5 is the position of the RC and LC regions for the thinner and thicker wires. For the 2um x 2um wire the boundary between the RC and LC region shifts to a lower frequency, which can be as a low as 1GHz for a thick dielectric layer between the signal and the return path. Although the characteristic impedance is now lower (at 1GHz - 73 ohms – microstrip and 58 ohms - strip line respectively for topology one - assuming a d=1 um distance to the return path, and 82 ohms for the lines in the topology two), magnitude of Z0 still varies and it will only settle at frequencies above 10 GHz for topology one and frequencies above 6 GHz for topology two. This behaviour will make impossible a meaningful broad band matching of these kind of lines.

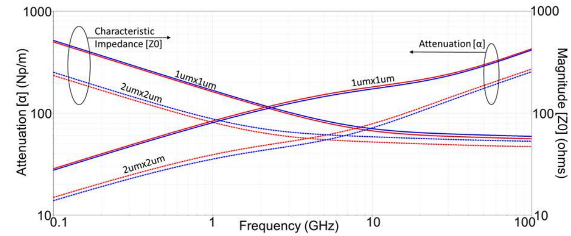

Fig. 5. Variation with frequency for attenuation and magnitude of characteristic impedance for 1um x 1um (full lines) and 2um x 2um (dotted lines) wires routed in toppology two configuration. Top layers wires are shown in blue while second layer are represented by the red lines.

B. S parameter response of the signal lines

The simulated insertion loss and return loss for 5 mm long lines routed in topology one and in topology two are presented in Fig. 6 and Fig. 7 respectively. As expected and in correlation with the data presented in previous section, the 1um x 1um structure have larger insertion loss and higher return loss due to the larger impedance mismatch in the lower frequency band. The return loss is rather high and modulated for both of the topologies.

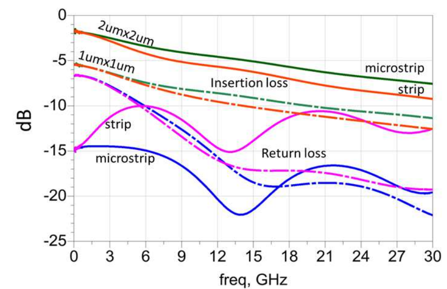

Fig. 6. Return and insertion loss of the 2umx2um and 1umx1um wires routed as in topology one. Both the microstrip and strip line structure are shown.

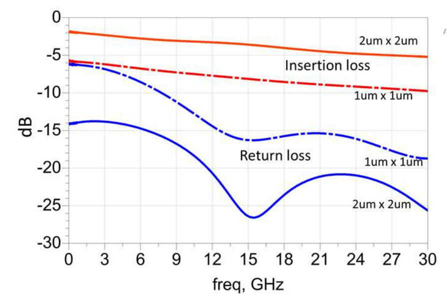

Fig. 7. Return and insertion loss of the 2umx2um and 1umx1um wires routed as in topology two.

Identify applicable funding agency here. If none, delete this text box.

III. MEASURMENT DATA OF 4 LAYER RDL

A topology one test vehicle with four metal layers was built and characterised via high frequency measurements. The signal lines built in the test vehicle were 2um x 2um cross section. Coupled lines with different lengths (1,3 and 5mm) and different distance between lines (2,3 and 4um) were measured and compared with the simulation models. The measurements were done with a four port VNA from 0.1 GHz to 20 GHz, using 100um pitch GSSG probes. Good agreement between the simulated and measured data as presented in Fig. 8&9 was noted.

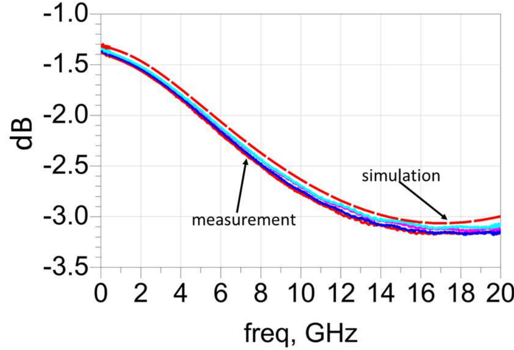

Fig. 8. Insertion loss of 3 mm long lines with a 2umx2um cross section – measured and simulated data.

The good agreement between the simulated and measured data allowed further analysis of more complex structures to be confidently done with the simulation models.

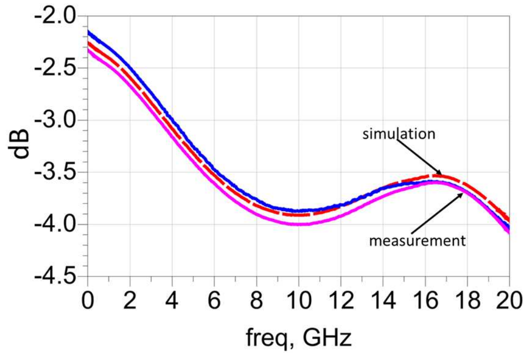

Fig. 9. Insertion loss of 5 mm long lines with 2um x 2um cross section – measured and simulated data.

IV. HBM AND AIB BUS IMPLEMENTATION ON HD-FOWLP

As discussed in the previous sections the HD-FOWLP interconnects will be working in the RC regime and at the limit between the RC regime and LC regime depending on the size of their cross section. In these regime the characteristic impedance of these wires varies widely and any impedance match design in a broadband sense is impossible. Another aspect that need to be considered when these interconnects are to be used for HBM or AIB interfaces is that the drivers connected at one end of the interconnect will have a complex value impedance while their load is usually capacitive. In this kind of environment using only the S parameters to optimise the interconnect response will not be as useful as it is for the regular transmission lines used on the regular packaging substrates.

As discussed in [10] these kind of interconnects can display in certain situation a ringing behaviour, particularly, such ringing is observed only if the line is driven with an actual I/O circuit. The ringing does not appear if the driver has a simple real output impedance valued. It requires a complex driver output impedance to excite the ringing. To observe such behaviour a transient analysis is required where the S parameter data generated from measurement or simulation are used in a time domain analysis coupled with driver models and loads that include the complex driving impedance mentioned above. The simplest driver model consists in a voltage source with a series resistance and a shunt capacitance to represent total output parasitic capacitance including the ESD protection, micro-bumps and transceiver capacitances. The series resistance represents the driver strength. The typical values used in this work are summarised in Table I.

TABLE I. DRIVER AND LOAD EQUIVALENT CIRCUIT PARAMETERS

	Voltage Swing [V]	*Resistance [ohms]*	*Output Source Capacitance [fF]*	*Input Load Capacitance [fF]*
HBM	1.2	66.6	500	500
AIB	0.9	30	500	500

In order to understand which is the most suitable topology for routing HBM and AIB interfaces a transient time domain simulation using a recursive convolution algorithm was performed and the eye diagrams obtained from this simulation were used to qualify the physical structure. The recursive convolution algorithms allow simulation of long bit patterns in very short time, however because it is using an algorithm based on vector fitting, it requires S parameters data that have a relative large bandwidth [12]. The S parameter data used for these simulations need to be valid from DC to frequencies above at least the first pole – as shown in Fig. 6 and 7 the structures studied here have the first pole between 14 and 15 GHz. Therefore, a bandwidth from DC up to 20GHz has be found to be appropriate.

To qualify the suitability of the different configurations the HBM2 and AIB specifications were used. The HBM2 specifications (8) require an eye mask with a width of 0.7 UI and a height of 480 mV (+ 240 mV and – 240 mV from 600 mV, the maximum swing is 1.2 V). On the other hand, the AIB specifications (9) require an eye opening which has to be wider than 0.4 UI while the height of the eye should be larger than 180 mV (+ 90 mV and – 90 mV from 450 mV, maximum swing is 900 mV).

As discussed in section 1, a bus structure based on topology two with a 1um x 1um line will minimise the number of layers required to layout the HBM interface. In this situation only two metal layers are necessary, one for the signal and ground and one for the power. The length of the simulated bus is 5 mm. The eye diagram of this topology is shown in Fig. 10. It can be noted

that the HBM2 requirements in terms of the eye opening are not met.

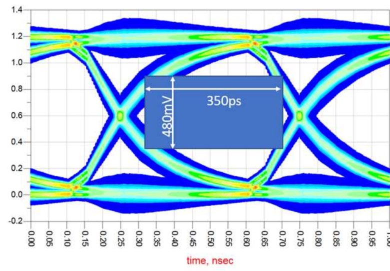

Fig. 10. Eye digaram for HBM interface implemented in topology two with 1um x 1um wires, spacing between the adjacent signal lines is s=2um, distance to the ground is d=1um and the distance between signal layers is t=2.5um.

For the AIB interface case the minimum number of metal layers to distribute all 3900 lines within 8 mm beach front is three. Using topology one with a 1um x 1um cross section wires and spacing between adjacent signal lines s=1um the AIB interface can be distributed within three layers. Topology two also needs three layers when a 1um x 1um wires are used and s= 2um while d=1um (Fig. 2.) Fig. 11 shows the eye diagrams for the two topologies. These results were obtained assuming the bus length is 10 mm, which is the maximum length stipulated in the AIB specifications (9).

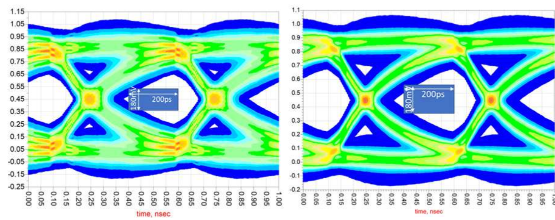

Fig. 11. Eye diagram for AIB interface implemented in topology one (left) and topology two (right) using a 1um x 1um wire cross section.

It can be noted from Fig. 11. that when a 1um x 1um cross section wire is used the bus interface implemented in both topologies will not meet the AIB requirements for maximum bus length. Simulation data have shown that shorter bus structures will be able to fulfil the timing specifications (e.g.: 7mm).

The two HBM and AIB bus structures were also implemented using a 2um x 2um wire. For this situation topology one requires 5 metal layers for both the HBM 1700 lines as well as the AIB 3900 lines. The second topology (Fig. 2) allows the HBM structure to be implemented using three metal layers, while 5 metal layers are required for a full AIB implementation. The eye diagrams for these case are shown in Fig. 12 and Fig. 13. The simulation results showed that the eye diagram for the 2um x 2um wire cross section is larger than in the case of 1um x 1um cross section which is a result of reducing the attenuation of the wires when the cross section is increased as discussed in section 2. However, the HBM specifications cannot be met as clearly seen in Fig. 12. On the other hand, the AIB interface implemented with topology one can meet the timing specifications without any problems when implemented using a 2um x 2um cross section signal lines (Fig. 13).

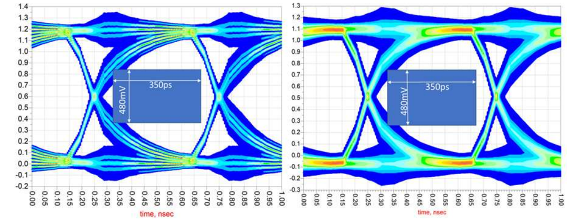

Fig. 12. Eye diagrams for HBM interface implemented in topogy one (left) and topology two (right) using a 2um x 2um wire cross section.

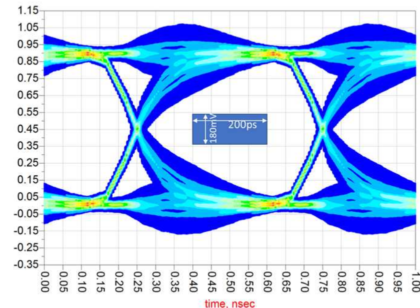

Fig. 13. Eye diagram for AIB interface implemented using topology one with 2um x 2um wires; s=2um between adjacent wires, distance to ground d=1um.

So far all the structures have assumed minimum design rules to achieve maximum wire density within a given beachfront. However, for the HBM case implemented using topology one, with a 2um x 2um cross section wires, the spacing between wires can be relax from the minimum 2 um spacing and still be able to implement the full 1700 wires within 6 mm beachfront within two signal layers.

The high density bus behaviour is fundamentally controlled by two mechanisms – one is the high attenuation, and the second one is the high coupling between the closely routed wires. While the attenuation can be reduced by increasing the cross section of the wires and increasing the distance from the ground, the coupling can be reduced by increasing the distance between adjacent signal lines and reducing the distance to ground. The variation in coupling between adjacent 2um x 2um wires when the distance to the ground (topology one) is varied while keeping the spacing between the lines 2 um is shown in Fig. 14 (left). The same figure (right) shows the coupling variation when the spacing between adjacent wires is changed while keeping the distance to ground constant. In Fig. 14 the data shown in the left figure was obtained via simulation while the data in right hand graph was obtained via measurements. Although the distance to ground is a useful variable to balance the attenuation and coupling in high density bus structures it is not the easiest variable to change and control in practical designs. The minimum and maximum thickness of the polymer dielectric layers depend of the spin coating process and the material properties such as viscosity. Also the uniformity of very thin polymer dielectric layers (below <3um) is not easy to control.

For the above reasons the practical dielectric thickness used in FOWLP packages is 3 to 4 um thick.

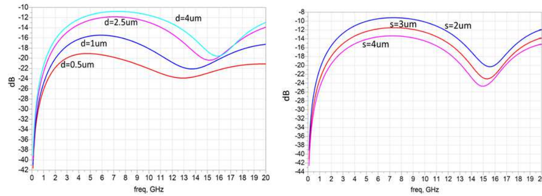

Fig. 14. Coupling variation for a 2um x 2um cross section wire routed using topology one top layer; left (simulation) – when the spacing between ajacent wires is kept constant s=2um and the distance to ground, d is varied; right (measurments) – when the spacing s is varied while the distance to the ground is kept constant.

A much practical variable that can be used to reduce the coupling is the spacing between the adjacent lines. Fig. 15 presents eye diagrams generated from measured S parameter data for 2um x 2um lines, with a d=3um to the ground plane, where the spacing between the lines was varied from 2 to 4 um. It can be noted in Fig. 15 that a 5 mm long bus with 2 um spacing between the lines will not meet the HBM specifications (Fig. 15 left), when the spacing is increased to 4 um the timing specifications are met (Fig. 15 right). Using a 4 um spacing with a 2um x 2um wire will result in a wire density of 166 lines/mm which will require two metal layers to distribute a HBM signal within the 6 mm beachfront, hence making this approach viable and implementable using topology one.

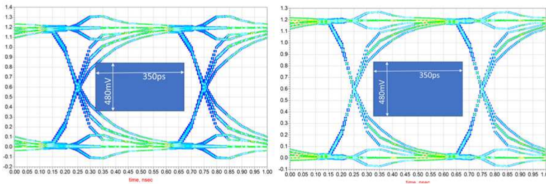

Fig. 15. Eye digrams for HBM interface of a 5 mm long 2um x 2um wires when the spacing s between the adjacent wires is increased from 2um (left) to 4um right. The above diagrams have been generated using measured S parameters data.

V. Conclusions

In this paper we discuss the electrical challenges implementing HBM2 and AIB wide interfaces on polymer based HD-FOLWP technology. We have analysed and demonstrated through simulation and measurements two possible package topologies which can be used to implement such interfaces. Unlike regular interconnects on classical packaging substrates the RDL of FOWLP are working at the limit of RC and LC regimes or just purely in the RC regime depending on the wire cross section hence designing such bus structures solely based on S parameter data information is difficult and not practical. We have shown that we can achieve practical interface designs based on the AIB and HBM2 specifications and both explored package topologies can be used. We have discussed, analysed and suggested possible design variables that can help in practical implementation of such bus structure on the polymer based HD-FOWLP technology.

References

[1] A.C. Durgub, Z. Qian, K. Aygun, R. Mahajan, T.T. Hoang, S. Y. Shumarayev “Electrical performance limits of fine pitch interconnects for heterogenous integartion,” IEEE 69th Electronic Components and Technology Conference (ECTC) 2019

[2] S. Abdennadher, M. Altman and B. Xue,“ Challenges and Emerging Solutions in Testing HBM IO & Systems 2018 IEEE 19th Latin-American Test Symposium (LATS), 12-14 March 2018.

[3] E. Nurvithadhi et al.,“In-Package Domain-Specific ASICs for Intel Stratix 10 FPGAs:A Case Study of Accelerating Deep Learning Using TensorTile,” IEEE 2018 International Conference on Field-Programmable Logic and Applications, pp. 106-110.

[4] Shu-Rong Chun et al., “InFO_SoW (System-on-Wafer) for High Performance Computing,” 2020 IEEE 70th Electronic Components and Technology Conference (ECTC) 2020

[5] Joe Lin et al., “Scalable Chiplet package using Fan-Out Embedded Bridge,” 2020 IEEE 70th Electronic Components and Technology Conference (ECTC) 2020

[6] 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

[7] M. Wade et al., “TeraPHY: A Chiplet Technology for Low-Power, High-Bandwidth In-Package Optical I/O,” IEEE Micro, Volume: 40, Issue: 2, March-April 1 2020.

[8] JESD235A, JESD235B, JESD235C – High Bandwidth Memory JEDEC Standard https://www.jedec.org/.

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

[10] V. Heyfitch et al., “High Bandwidth Memory Interface on Organic Substrate: Challenges to Electrical Design,” 2018 IEEE 68th Electronic Components and Technology Conference (ECTC) 2018.

[11] K. Cho, H. Lee and J.Kim, “Signal and power integrity design of 2.5D HBM (High bandwidth memory module) on SI interposer”, 2016 Pan Pacific Microelectronics Symposium (Pan Pacific), 25-28 Jan. 2016.

[12] ADS Keysight - User manual