Design and technology spaces for heterogeneous chiplet integration

Yujeong Shim
Alphabet Inc.
Google Cloud
Mountain View, California
yujeongshim@google.com

Woon Seong Kwon
Alphabet Inc.
Google Cloud
Mountain View, California
wskwon@google.com

Anna Gao
Alphabet Inc.
Google Cloud
Mountain View, California
gaoq@google.com

Abstract*— Heterogeneous chiplet integration is an emerging technology to boost up computing power and build cost effective systems for HPC, AI and ML ASICs. In this paper, we introduce efforts to build chiplet ecosystems as well as design and technology spaces for chiplet integration. We also discuss the decision tree to select appropriate technologies for chiplet implementation.

***Keywords*—Heterogeneous integration, die-to-die interconnect**

I. Introduction

The ASIC silicon size has reached the reticle limitation so the device scaling is no longer aligned with the trend of computing power increase. In order to integrate more transistors in complex CPUs and SoCs for HPC, AI and ML applications, the semiconductor industry has been moving forward to heterogeneous chiplet integration. It boosts up computing capacity and overcomes various challenges such as low yield, reticle size limitation and high cost of advanced process due to large die size.

Various products in the HPC market have already adopted heterogeneous integration to disaggregate based on functions with mixed process nodes and/or to integrate multiple computing chips to scale up computing resources. As there exists more demands for chiplet integration, efforts to build an ecosystem for chiplets have recently emerged by standardization of die-to die-interfaces by ODSA, UCIe, OIF, and so on for plug-in solutions, while the previous applications adopted proprietary die-to-die solutions. Recently, the chiplet industry has added more weight to UCIe. In addition to the development of die-to-die interface IPs and standards, advanced packaging technologies (2.5D/3D packaging) developed by foundries and OSATs have contributed significantly to enable chiplet integration requiring high bandwidth and low latency of die-to-die interfaces in order to accordance with the demands for system scaling.

Since the chiplets require additional silicon area for die-to-die interface, wafer masks and packaging cost, chiplet integration is not always cost beneficial. For balancing the performance and cost, it is very important to analyze PPA (Performance, Power, Area) and packaging costs. In this paper, we introduce design spaces for chiplet applications requirements and present how to select UCIe based die-to-die interface (advanced vs standard), design solutions and packaging technology based on comparative studies and requirements (total bandwidth, bandwidth/area, latency, power consumption, etc). We suggest further technology enhancement in IO circuits, interface design and packaging technology to improve area efficiency and increase scalability.

II. Chiplet Ecosystem

A. Why chiplets

As mentioned in the introduction, SoCs have integrated more transistors to increase computing power. Although the process node has been shrunk, die size increase can not be avoided and the size of ASIC has reached the reticle limitation. For advanced process nodes, wafer yield exponentially drops as the die size increases which leads to significant increase of net die cost. Therefore, die size scaling is no longer economically feasible. One possible approach is to increase the number of integrated chips on a single package instead of building a large single chip. Fig.1 presents the cost vs die area for monolithic die and 2 identical split dies. In this example, for the die size bigger than 350mm^2, chiplets clearly show cost benefits over monolithic die.

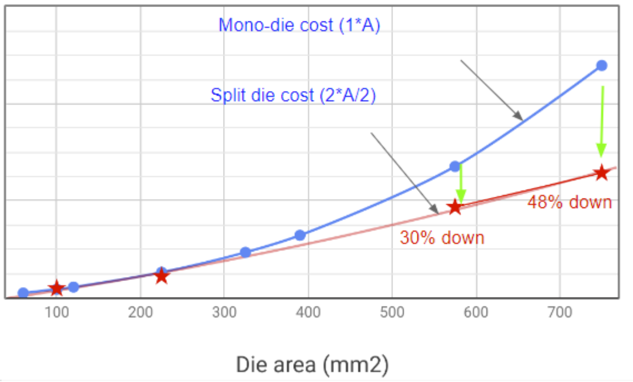


Fig 1. Cost of monolithic die and 2 split dies vs total die area

Chiplets also enable scalability and reusability of IPs based on split capabilities. In order to hit fast time to market, IO blocks can be built in a chiplet with a mature process node and then integrated with high-performance computing chiplets with an advanced process node. This approach has been productized by multiple leading companies. By integrating multiple computing chiplets, the development of multiple derivatives of products is greatly simplified with scalability [1].

B. *Efforts to Build the Chiplet Ecosystem*

For the chiplet based products in the market, chiplets were connected by proprietary die-to-die solutions. However, recently, industrial efforts have been emerging to build chiplet ecosystems. Standardization of die-to-die interface is one of the most important efforts for building the universal chiplet ecosystem. In the past few years, many open standards have come up, e.g. BOW [9], OHBI [8], AIB based open source and UCI express [7]. Table I and Table II present the attributes of various die-to-die interface standards. These attributes are the keys to define the interface based on optimization of PPA (performance, power, area) and cost.

TABLE I. Die-to-Die Standard for Advanced Packaging

	UCIe-advanced	**OHBI 1.0/2.0**
Pin speed (Gbps)	4/8/12/16/24/32	8/16
BW/mm per direction	1.3 Tbps/mm (8G) 2.6 Tbps/mm (16G) 4.0 Tbps/mm (24G) 5.2 Tbps/mm(32G)	1.2 Tbps/mm (8G) 4 Tbps/mm(16G)
Bump pitch	>31 µm for >16G >38 µm for >24G	55 µm (1.0) 40 µm (2.0)
DQ/channel	64 & 4 redundancy	42 & 2 redundancy
IO voltage(V)	0.7/0.5	0.4
Power eff(pJ/bit)	0.6/0.5 for >16G	0.4 for OHBI 1.0
Target Latency	<2ns	<4ns

TABLE II. Die-to-die Standard for Traditional Package

	UCIe-standard	**OHBI-L**	**BOW**
Pin speed (Gbps)	4/8/12/16/24/32	~40Gbps	16
BW/mm Per direction	0.23 Tbps/mm (8G) 0.45 Tbps/mm (16G) 0.65 Tbps/mm (24G) 0.9Tbps/mm (32G)	1Tbps/mm	0.44Tbps/mm
Bump pitch	>110 µm	130 µm	130 µm
DQ/channel	16	40	16
IO voltage(V)	0.7/0.5	0.75V	0.75
Power eff(pJ/bit)	1.0/0.5 for <16G	1.0	0.5~1
Target Latency	<2ns	<4ns	<5ns w/o FEC

Recently, industry has merged standardization efforts into UCI express [7], which has the full protocol stack from the transaction layer all the way to packaging. And this enables an easy software integration path.

In addition to standardization of die-to-die interface, EDA and design methodologies are also critical to enable chiplet designs. There have been many challenges such as multi PDK design and sign-off. Moreover, the design methodology for advanced packaging technology is completely different from traditional package design. It adopts silicon physical design methodology. Collaborations among EDA, foundry and ASIC companies have enhanced design methodologies for heterogeneous chiplet integration. [2] describes details of deliverables, format and requirements for interface design and integration.

III. Advanced Packaging Technology for Heterogeneous Integration

Packaging technologies are the key enablers of chiplet integration. Fig 2.(a) depicts the cross sectional of the traditional organic multi-layer package substrate to connect chiplets through build up layers. The minimum bump pitch requirement for this technology is 110 µm. However, for most HPC AISC requiring large silicon area, 130 µm or higher pitch is typically utilized. This scheme is the most cost effective way while the bandwidth/mm is limited due to the limited number of routing layers.

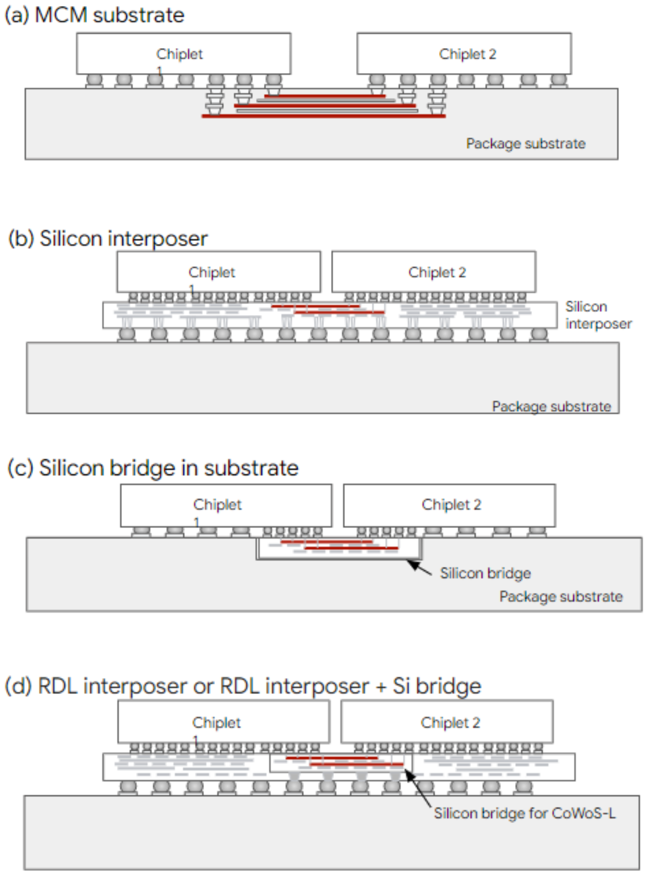


Fig 2. Packaging technology (a) traditional organic multi layer substrate (b) silicon interposer (c) silicon bridge embedded in substrate (d) RDL interposer or silicon bridge embedded in RDL interposer

Advanced packaging technologies enable much higher bandwidth/mm by using a finer bump pitch (<55 µm depending on the foundry) compared to the standard organic packaging. Silicon interposer depicted in Fig 2.(b) is the popular technology to integrate chiplets with HBMs (High Bandwidth Memory). This technology offers fine pitch bumps and line width/space with a few mm of channel length to achieve high bandwidth. As the per pin speed increases to achieve higher bandwidth of interconnection, thicker metal solution has been offered in order to reduce resistive loading. This technology is limited by the interposer size and the number of metal layers for deeper IO connections. It may also have some challenges or retractions due to the top die placement rule.

The silicon bridge embedded in the substrate shown in Fig 2.(c) can be another way to integrate chiplets with lower manufacturing and assembly cost. The bump pitch of the silicon bridges is similar to the silicon interposer, while the bump pitch of the area outside of silicon bridges is the same as the pitch of regular C4 bumps. A big benefit of this technology is integrating large and multiple dies since there is no silicon interposer which has a reticle limit. Since other external interfaces such as Ethernet and PCI express don't go through lossy TSVs (Through Silicon Vias), the channel loss is better than silicon interposer. However, without the TSVs, the power network design can be challenging due to no vertical connections from the µbump to the substrate. [3]. In order to deliver the power, C4 bumps must be placed outside of the silicon bridge and connected horizontally through surface routings which have a longer inductive path than direct vertical connection. In order to compensate this higher inductive path, additional area may need to be allocated for more on-chip decaps and C4 bumps.

Another technology to be introduced is RDL (re-distributed layer) interposer shown in the Fig.2 (d). The RDL interposer has no limitations in the form factor compared to silicon interposer so that integration of multiple chiplets and memories are possible. Although the width/space of the metal trace is wider than the silicon interposer, it is fine enough to ensure signal integrity. RDL interposer has good signal integrity performance by low insertion loss from the thick metal and low loss dielectric material. Signal integrity of the external signals such as high-speed SerDes is also guaranteed because the RDL interposer doesn't have TSVs which have undesired high capacitance and shunt resistance. However, it also has restrictions of integration density and routings feasibility due to large via and pad size. In order to overcome these two issues, a new technology combining RDL interposers and silicon bridges have been introduced as well. Unlike the silicon bridge embedded in the substrate, the silicon bridge embedded in the RDL interposer allows TSVs offering direct power delivery paths.

Other than the technologies introduced in this section, many other 2.5D/3D packaging technologies have been under development and enhanced to achieve better performance and lower cost.

IV. Design Spaces for Chiplet Integration

In this section, we'd like to discuss the decision trees for chiplet technologies and design solution spaces in IO, full chip implementation, package design and system integration.

A. *Decision Tree for Chiplet Technologies and Design Factors*

Fig.3 demonstrates the decision tree for chiplet technologies. Disaggregating schemes are determined by chip architecture depending on computing power and interface requirements. As shown in Fig.1, there is a break point where chiplet becomes more cost and performance effective. And if the leading edge interface is required as well as adoption of the most advanced process node is required, IO functions can be disaggregated and stay in the mature process node [1]. Scalability over generations of products could be a motivation for chiplet integration as well. By integrating multiple computing and IO chiplets, performance can be scalable within a shorter development time. Since the die-to-die interface has an area overhead, precise performance per cost analysis is required at the architecture stage.

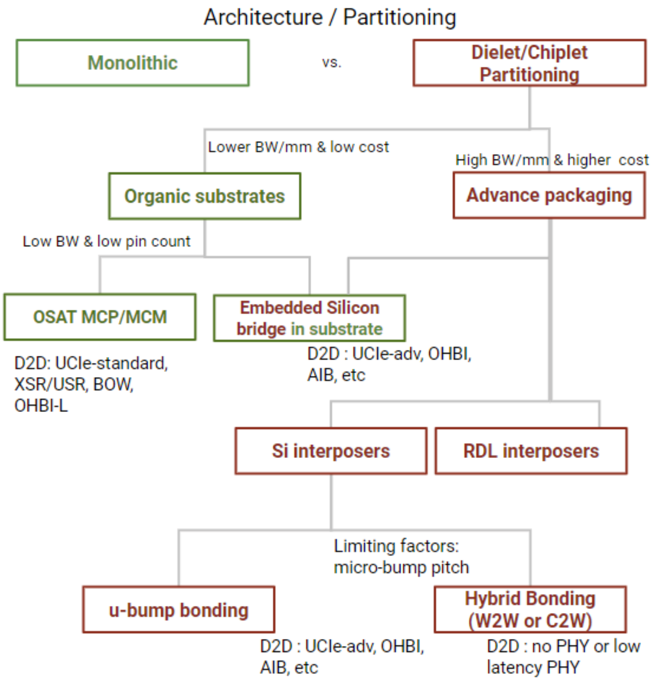

Architecture / Partitioning
Monolithic
vs.
Dielet/Chiplet Partitioning
Lower BW/mm & low cost
High BW/mm & higher cost
Organic substrates
Advance packaging
Low BW & low pin count
OSAT MCP/MCM
Embedded Silicon bridge in substrate
D2D: UCIe-standard, XSR/USR, BOW, OHBI-L
D2D : UCIe-adv, OHBI, AIB, etc
Si interposers
RDL interposers
Limiting factors: micro-bump pitch
u-bump bonding
Hybrid Bonding (W2W or C2W)
D2D : UCIe-adv, OHBI, AIB, etc
D2D : no PHY or low latency PHY

Fig 3 Decision tree for chiplet technologies

Once the architecture of chiplets is chosen, packaging type needs to be defined. The major criteria of standard organic packaging vs advanced packaging is the bandwidth/mm requirement. As of today, the rule of thumb is higher or less than 0.5Tbps/mm. If the cost is critical, the standard organic package can be the right solution by keeping the bandwidth requirement <0.5Tbps/mm per direction (aggregating bandwidth <1.0Tbps/mm for both TX and RX). In this case, to gain higher total bandwidth, the width and depth of the die-to-die interface can be increased while resulting in a larger chip size.. As a result, chip area increases. Moreover, increasing the depth requires adding more number of the package substrate layers in order to route the deeper IOs. For example, double stacked x32 UCIe-standard requires 4 routing layers while x16 UCIe-standard does only 2 routing layers. One positive side of the standard package based chiplet integration is flexibility of floor planning and channel length (<30mm). For standard applications, both parallel IOs and serial IOs can be adopted as the die-to-die interface. Parallel IOs (<40Gbps) have a simpler analog front-end, better power efficiency and lower latency than serial IOs (>56Gbps USR/XSR), while parallel IOs are more sensitive to supply fluctuations. Due to these benefits, the die-to-die interfaces have been converged to parallel IOs by multiple standardization efforts including UCIe-standard. If power efficiency is critical, reducing IO power by removing or disabling ODT (on-die termination) at RX and driving short channels can be a solution.

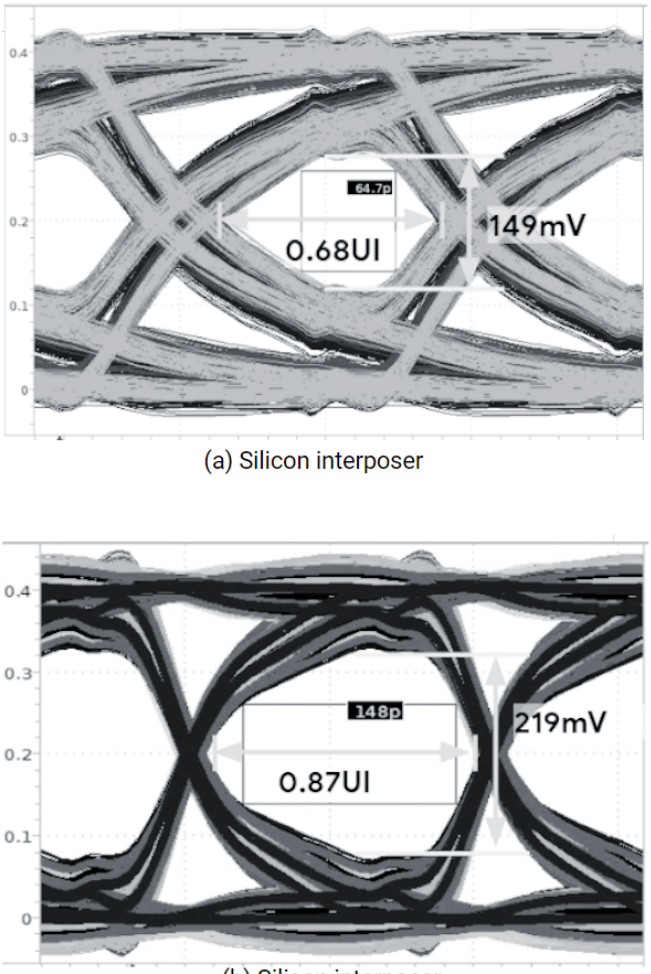
TABLE III. Comparison between Advanced Packaging and Standard Packaging

PKG type	Advanced	Standard
Bump pitch	<55um	>110um
Min L/S	<1um/1um	>10um/10um
L/S for performance	2.5um/2.5um	18~25um/30um depending on build up material and thickness
Channel length	~2mm	<10mm Design target ~30mm and depending on floorplanning
IO type	Parallel IO	Parallel IO or Serial IO
Per pin speed	16Gbps/32Gbps (UCIe, OHBI2.0)	16Gbps/32Gbps (UCIe-std) 40Gbps for single-ended Up to 112Gbps PAM4 56Gbps (w/o FEC)
BW/mm per direction	2.6 Tbps or 5.2Tbps/mm	0.5 Tbps or 1 Tbps/mm

Advanced packaging solutions introduced in Section III are able to achieve higher than 0.5Tbps/mm^2 per direction. As of today, the maximum pin speed is below 16Gbps while the UCIe-advanced 1.0's EOL target speed is 32Gbps. Since the typical bump pitch of advanced packaging is 36um ~ 55um, which differs among foundries and OSATs, higher bandwidth density is feasible. This fine pitch has 5 ~ 14 times of pin density compared to standard packaging [4]. Each technology provides different pros/cons for design. This paper compares three advanced packaging technology introduced in the Section III based on 5 design factors, bandwidth/mm, silicon area (or bandwidth/mm^2), signal integrity, power integrity and form factor.

First, silicon interposers can provide the highest bandwidth/mm due to high the pin density. The minimal bump pitch of silicon interposer is less than embedded silicon bridge in the organic substrate. In order to increase bandwidth/mm for the embedded silicon bridge, it is possible to tighten the shoreline pitch. But this will increase the depth so that bandwidth/mm^2 is reduced. RDL interposer's minimal bump pitch is the same as silicon interposer. However, the sizes of vias and via pads are very large compared to the silicon interposer and minimal line width/space rule is bigger so that the bump pitch is limited by routing and pin depth.

Second, the RDL interposer can provide great signal quality as long as routing is feasible. Although placement of shielding lines is not as dense as silicon interposers or embedded bridges, channel loss is remarkably low due to thick metal (low R), thick dielectric thickness (low C) and low Df (dielectric loss). Fig. 4 shows an example. Fig 4(a) shows the eye diagram with a silicon interposer and Fig.4(b) is the eye diagram with a RDL interposer. Silicon interposers also have multiple metal options (thickness and number of layers), which can be chosen based on signal integrity analysis.



(a) Silicon interposer

(b) Silicon interposer

Fig 4 Eye diagram comparison

For power integrity, silicon interposers and RDL interposers provide direct vertical paths through TSVs (through silicon vias) and vias. Additionally, depending on foundries (type, density and DRC rules differ across foundries), silicon interposers have embedded capacitance solutions. As mentioned in the Section III, the embedded silicon bridges in the substrate don't support vertical connection for power delivery. Hence, the u-bumps in the bridge areas are connected to the C4 bumps through surface routings on the package substrate [4][5]. This causes two issues. One is high inductance paths of the power delivery as shown in [5] and Fig.5.

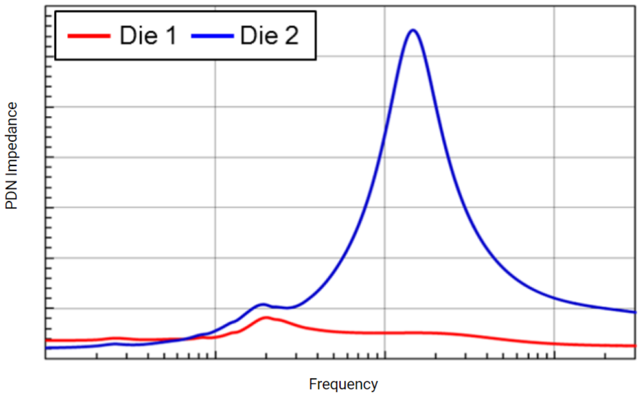


Fig 5.Impedance curves of die 1 and die2 when they are connected through an embedded silicon interposer

The other issue is that C4 placement and on-chip connection must be done outside of PHY & IO area, which is outside of the bridge area.

However, the embedded silicon bridge technology doesn't have limitations in the size or number of chiplets since it doesn't mount multiple chiplets on a single silicon large interposer which has reticle limitations. The RDL interposer offers much larger limitation than the silicon interposer.

As described in this paper, there are many decision criteria and design factors to optimize chiplet integration.

V. Summary

To summarize, heterogeneous chiplet integration is an emerging technology to boost up computing power and build cost effective systems for HPC, AI and ML ASICs. Industry has been moving forward to build the chiplet ecosystem in various areas such as IO technology, end to end standardization, packaging technology, EDA, and so on. We introduce efforts to build chiplet ecosystem as well as design and technology spaces for chiplet integration. We also discuss the decision tree to select appropriate technologies for chiplet implementation. While we focus on 2D/2.5D heterogeneous integration in this paper, 3D packaging technologies have been under development to disaggregate Cores, IOs, VRs, SRAMs and passive components in order to unleash computing capability.

References

[1] Samuel Naffziger;Noah Beck;Thomas Burd;Kevin Lepak;Gabriel H. Loh;Mahesh Subramony;Sean White, “Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product,” 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)

[2] Anthony Mastroianni;Benjamin Kerr;Jawad Nasrullah;Kevin Cameron;Hockshan James Wong;David Ratchkov;Joseph Reynick, “Proposed Standardization of Heterogenous Integrated Chiplet Models,” 2021 IEEE International 3D Systems Integration Conference (3DIC)

[3] Changwook Yoon; Guang Chen; David Greenhill; Hyo-Soon Kang; Ashkan Hashemi; Wendem Beyene, “Analysis of Noise Coupling and Timing Error in Silicon Bridge Application,” 2018 IEEE 27th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)

[4] Yoonjae Hwang;Sungwook Moon;Seungki Nam;Jeong HoonAhn, “ Chiplet-based System PSI Optimization for 2.5D/3D Advanced Packaging Implementation,” 2022 IEEE 72nd Electronic Components and Technology Conference (ECTC), 2022

[5] Yujeong Shim, Dan Oh, “Impacts of Dynamic Noise in Multi-Core or SOC Designs,” 2016 Design Con

[6] MD Arafat Kabir;Dusan Petranovic;Yarui Peng, “A Scalable In-Context Design and Extraction Flow for Heterogeneous 2.5D Chiplet-Package Co-Optimization,” 2021 IEEE 30th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)

[7] UCIe 1.0 standard

[8] OHBI 1.0/2.0 standard

[9] BOW 0.9 standard