Chiplet-Based Modular Computing Platform

COWOS/ EMIB / MCM Research Analysis

Sprint 3 Research into Encapsulation

We analyze the development of chips from the perspective of packaging technology to explore and develop the Chiplet-Based Modular Computing Platform.

First, we researched the evolution and history of packaging technologies, progressing from 2D to 2.5D and then to 3D.

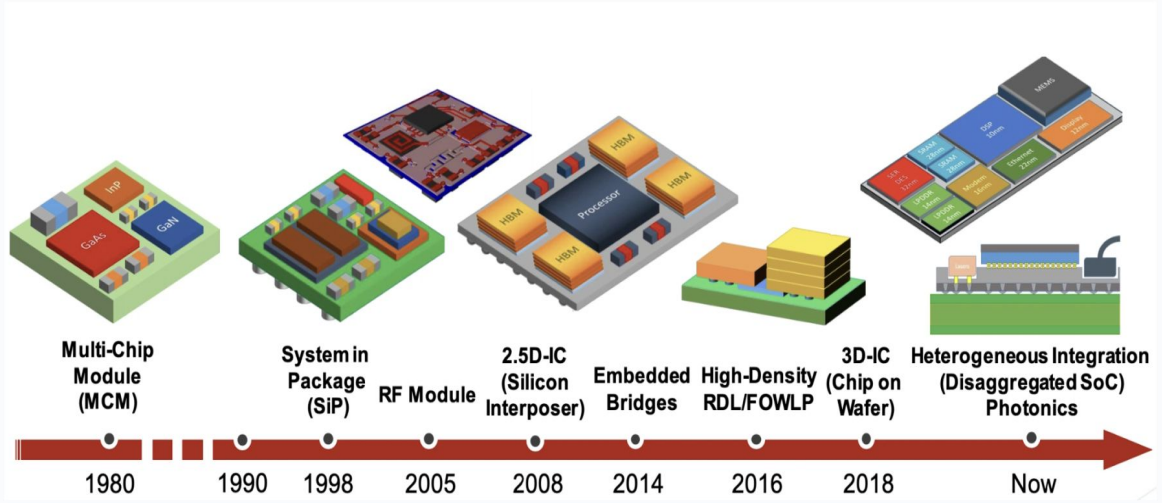
Focusing on the current leading packaging technologies, we explored and studied the performance of Multi-Chip Modules (MCM) and CoWoS, analyzing their transmission rates, compatibility, and scalabilit.

The roadmap of chiplet packaging development

Start of the chiplets:

In his seminal 1965 paper "Cramming More Components Onto Integrated Circuits," Gordon Moore foresaw the limitations of monolithic SoCs.

Besides being the author of Moore's Law, Moore famously stated, "It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected."



Independent Packaging

Defintion of Independent Packaging:

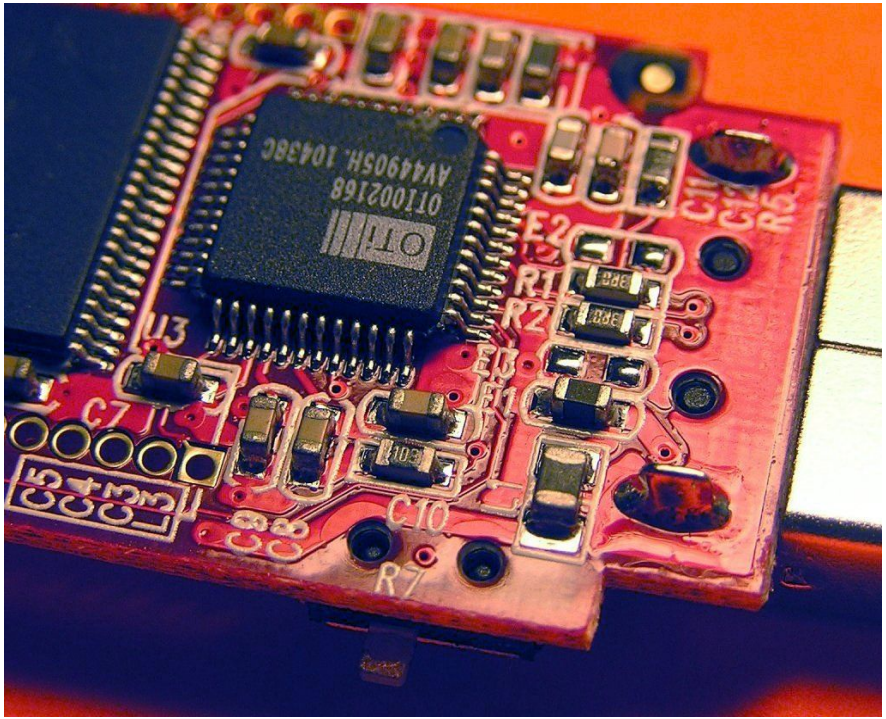
Each chiplet was packaged separately, with communication facilitated via circuit boards.

Time and Inventor:

Early pioneers in IC design, such as Texas Instruments and IBM, began using modular approaches in the 1980s and 1990s.

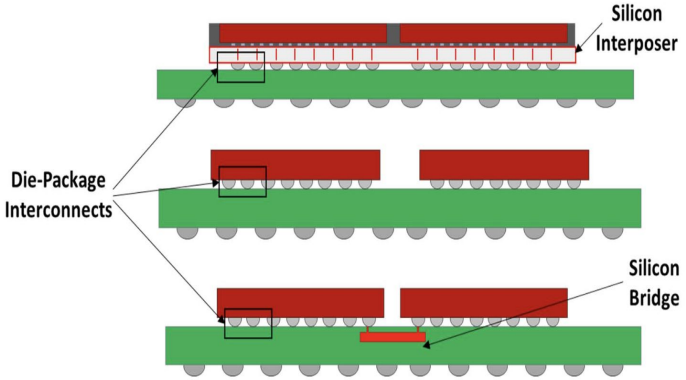
Disadvantage

This approach, while effective initially, suffered from latency and increased power consumption due to long signal paths.



Picture of independent package

Multi-Chip Module (2D)



Defintion of MCM:

A Multi-Chip Module (MCM) integrates multiple semiconductor chips (or die) into a single module, typically mounted on a common substrate, as the demand for higher performance and more compact electronic systems grew.

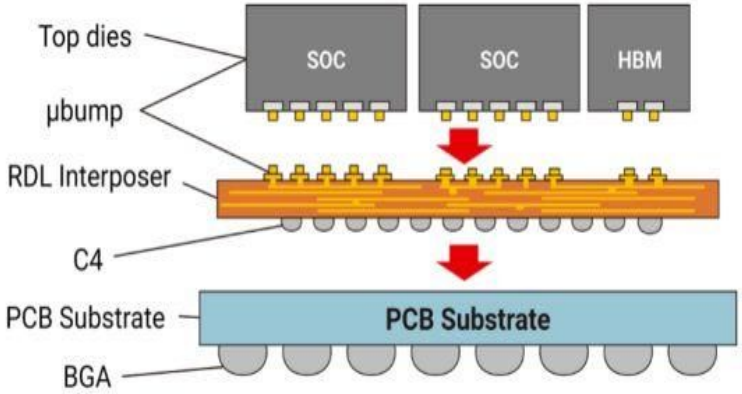
Time and Inventor:

Companies such as IBM and Intel were early adopters and developers of MCM technology in the 1980s.

Disadvantage:

The reliability of interconnects between chips is critical; failure in the interconnects can compromise the entire module and so on.

2.5D Packaging



Defintion of 2.5D Packaging(Cowos)

2.5D packaging allows multiple semiconductor dies (chips) to be placed side-by-side on a shared interposer. The interposer serves as a bridge that connects these chips using high-density interconnects.

Time and Inventor:

In the 2010s, TSMC pioneered the use of silicon interposers with its CoWoS (Chip-on-Wafer-on-Substrate) technology.

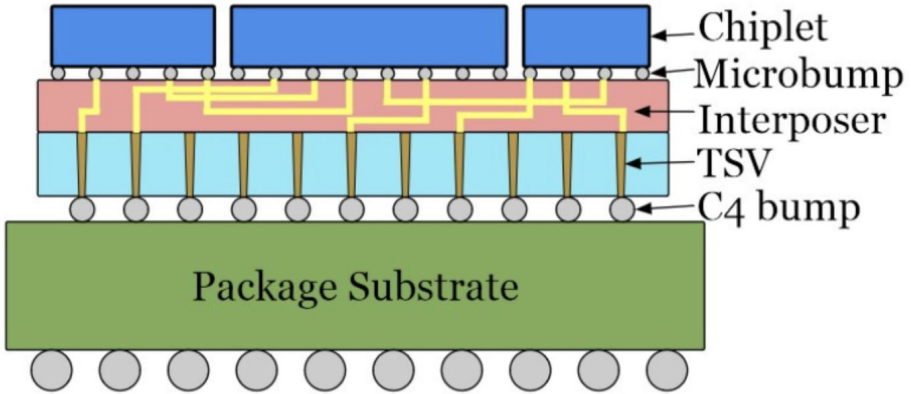
Advantage:

Silicon interposers allowed higher-density wiring and greater performance. This increased interconnect density and reduced signal latency between chiplets.

Disadvantage:

The cost and complexity of silicon interposers were major concerns, limiting the technology's use primarily to high-end applications.

2.5D Packaging



Interposers and TSVs

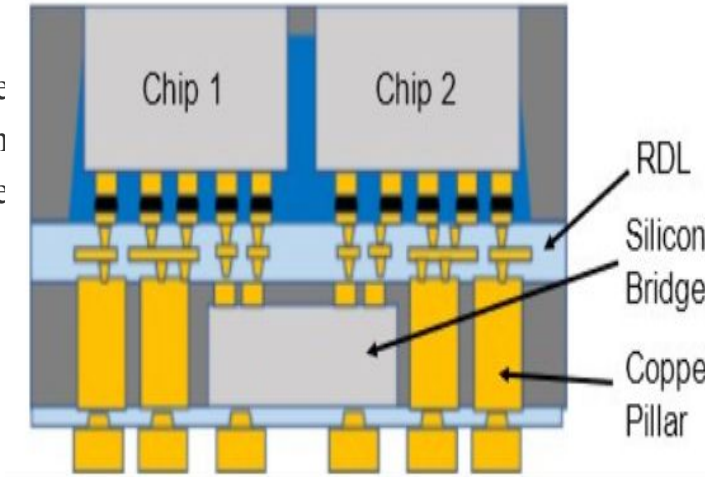
An interposer is a substrate used in semiconductor packaging that acts as a bridge between multiple chips or dies. It provides mechanical support, electrical connections, and facilitates communication among different integrated circuits (ICs) within a package.

Through-Silicon Vias (TSVs) are vertical electrical connections that pass through a silicon wafer or die. They are used to enable communication between different layers of stacked semiconductor devices in advanced packaging technologies.

2.5D Packaging

Defintion of 2.5D Packaging(EMIB)

Embedded Multi-die Interconnect Bridge (EMIB) uses a thin bridge layer to connect different chips, allowing them to communicate at high speeds without the need for through-silicon vias (TSVs) or large interposers



Time and Inventor: Developed by Intel,in 2016

Advantage:

Eliminate the through-silicon vias (TSVs) traditionally required for connecting to an interposer, as well as the size limitations imposed by the interposer itself.

3D Stacking: Vertical Integration

Defintion of 3D Stacking:

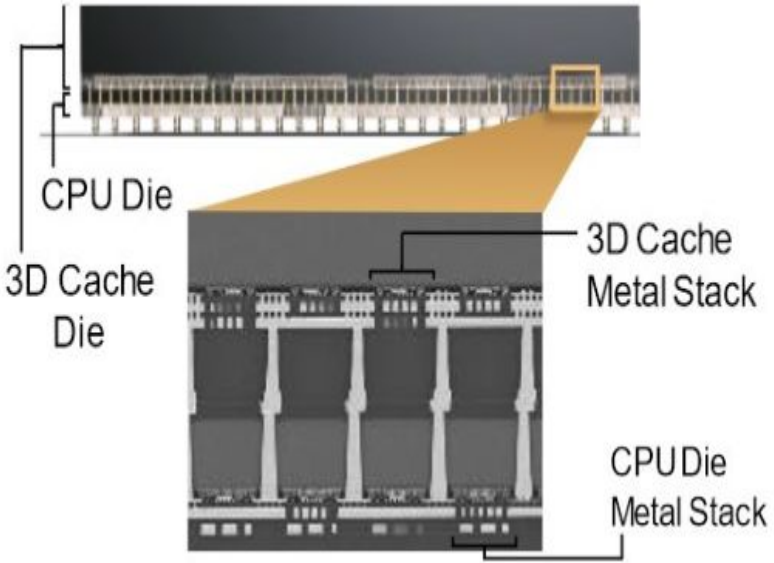
Vertically integrating multiple semiconductor dies (chips) within a single package.

Time and Inventor:

TSMC: SoIC(System of Integrated Chips) in 2010s.

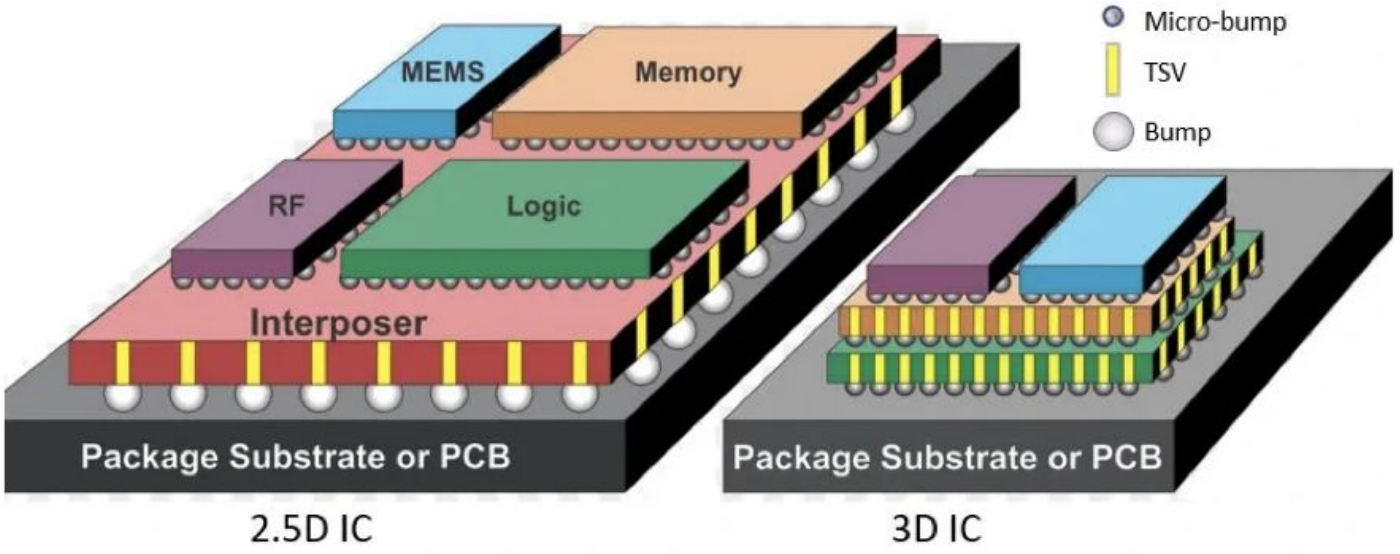
Advantage and Disadvantage:

It provides higher interconnect density, smaller and simpler circuitry, greater bandwidth, lower capacitance, and reduced power consumption. However, The technical barriers and complexity are both higher.

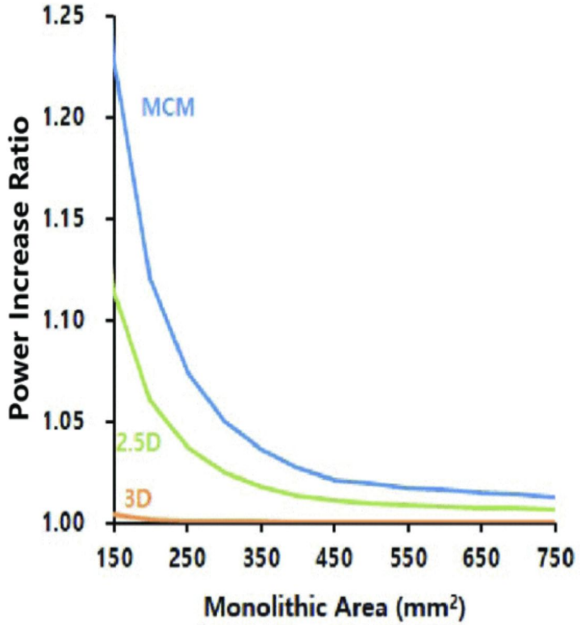


Difference between 2.5D and 3D

In 2.5D packaging, the logic chip and other stacked memory components are arranged side by side on a silicon interposer, whereas in 3D packaging, the logic chip and memory components are directly stacked on top of each other.



Performances Analysis



1.Power Overhead

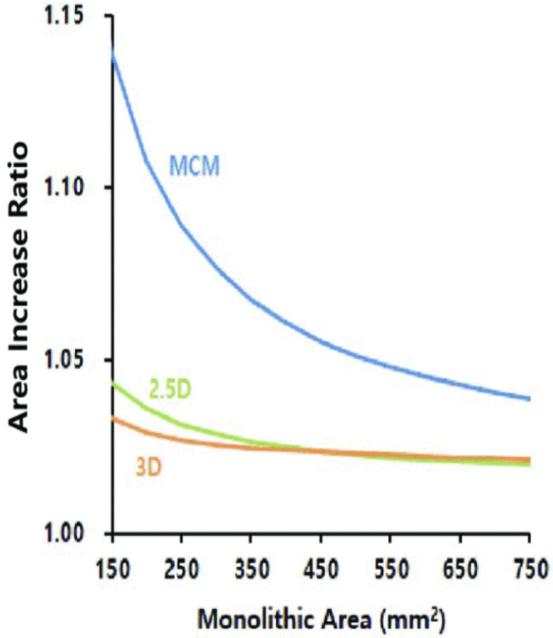
The figure shows the trend of the power overhead increase ratio (total power of divided die / power of die) as the monolithic die area increase, the power overhead due to die split decrease.

3D package: Vertically stacked chips reduce the overall interconnect path and the power consumption impact of TSVs is relatively reduced as the area increases.

2.5D package: utilizes intermediary layer layout to reduce signal delay and power consumption for high performance applications.

2D package: simple and reliable, suitable for lower bandwidth requirements, larger power gains with larger chip splits.

Performances Analysis



2.Area Overhead

As the die is divided the area overhead increases due to the addition of interface IP and as the die area increases it can be seen that the proportion of IP area becomes smaller and the overhead increase ratio is getting smaller.

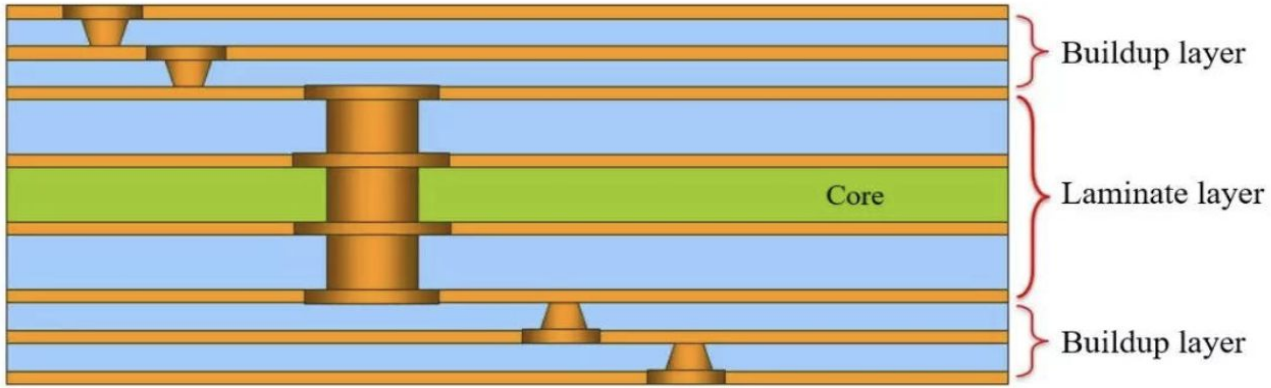
3D Packaging: As monolithic area increases, the relative ratio becomes smaller and the area increase decreases.

2.5D package: as the area increases, the relative proportion of IP decreases and the area gain is smaller.

2D package: more reliant on horizontal connections, larger relative change in area after.

MCM Organic Substrate Base（2D）

This standard 2D packaging is cost-effective and widely used for applications with lower IO density. It has higher yields due to the absence of delicate microbumps. Each die is connected to a substrate, and the MCM connects to the larger board through a BGA.

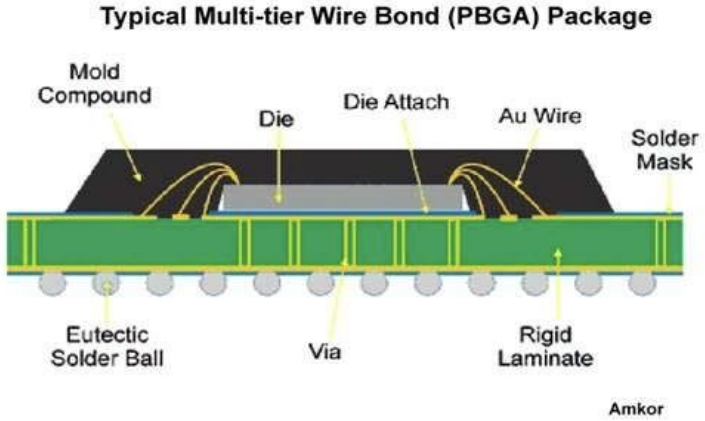


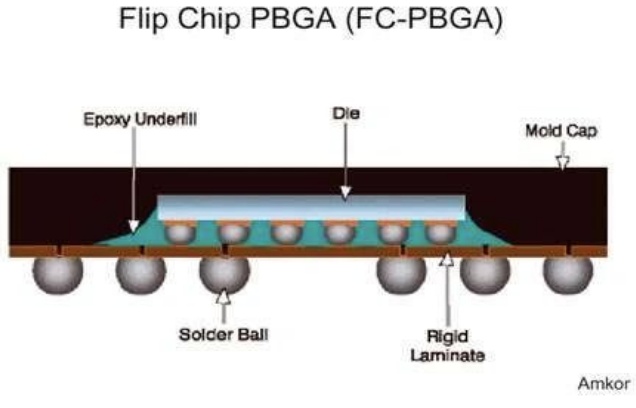
MCM Organic Substrate Base（2D）

Characteristics: In a 2D package, components are typically side by side on a flat surface, connected using wires or traces on the package substrate or printed circuit board (PCB).

The transmission rates for 2D Multi-Chip Module systems typically range from a few Gbps per channel.

More advanced high-speed PHY designs support rates up to 112 Gbps for aggregate throughputs, making them feasible for specific high-performance applications





MCM Organic Substrate Base（2D）

Some problems faced in 2D packaging

Limited Integration: In 2D IC packaging, discrete devices for different functions such as high-performance logic, lower-performance logic, memory, and analog/RF each existed in their own chip packages.

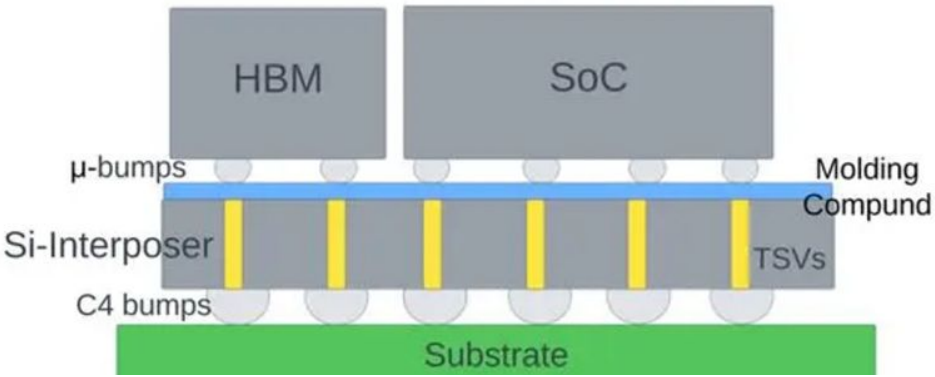
Size and Weight: The resulting circuit board from 2D IC packaging will be larger, heavier, and consume more power.

Reliability: Every soldered joint on the board is a potential point of failure.

CoWoS

Structural characteristics and transmission rate

There are three technical variants of cowos: COWOS-S, COWOS-R, COWOS-L

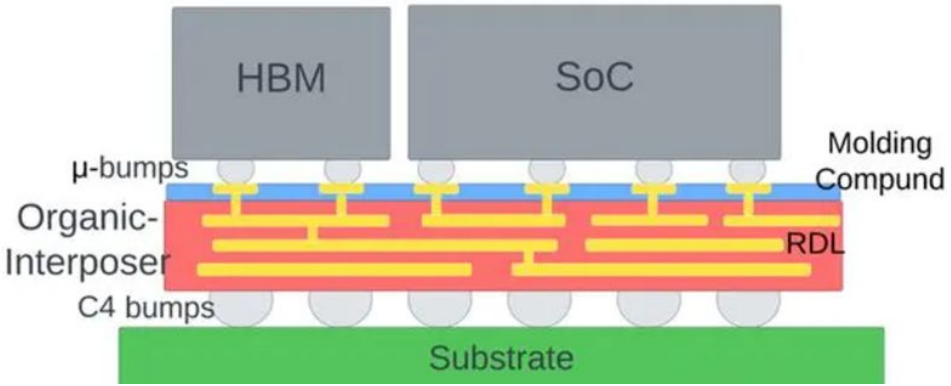


CoWoS-S

Cowos-s: it connects each chip module on a complete silicon intermediate layer, uses silicon wafer as the support material, copper as the wire, and connects through TSV technology to realize vertical signal transmission. The line length is reduced, and the bandwidth and delay performance are greatly improved, which can provide the highest information density and power density among several advanced packaging technologies.

CoWoS

Structural characteristics and transmission rate

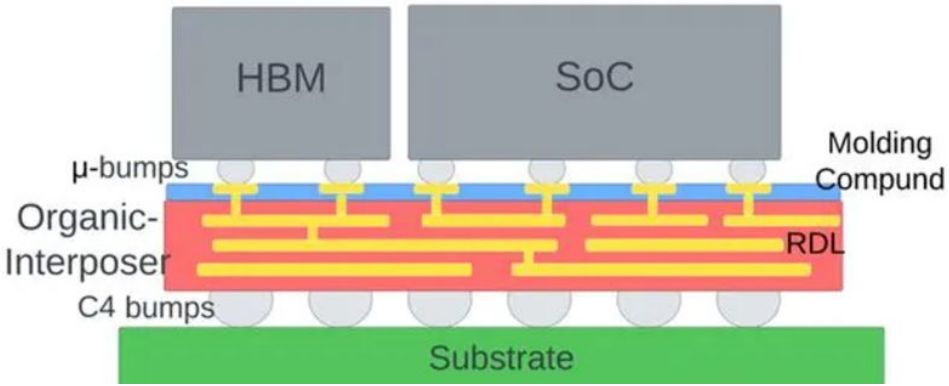


CoWoS-R

Cowos-r: uses RDL as the interconnect layer for different chip modules, organic materials instead of silicon single crystals as the material to support the stereoscopic circuit, and metallic copper as the wire. It is malleable and can alleviate the thermal expansion problem. However, due to the long RDL transmission line, the delay becomes larger and the transmission rate is reduced relative to cowos-s.

CoWoS

Structural characteristics and transmission rate



CoWoS-L

Cowos-l: mixed silicon intermediary layer and organic intermediary layer, using TSV technology at key nodes to achieve faster information transmission.

CoWoS

Scalability and stability

Cowos-s: The silicon substrate has the problem of uneven thermal expansion coefficient, which easily leads to reliability problems. The thermal management is complex, and the fabrication cost of large-area silicon interlayers is high and the yield is low.

Cowos-r: Use of organic intermediate layers for larger area encapsulation. RDL supports more complex circuits, is more scalable, and is suitable for more complex chiplets implementation. cowos-r is commonly used for large-scale packaging of multi-chip modules

Cowos-l: Between the first two technologies

CoWoS

Scalability and stability

character	CoWoS-S	CoWoS-R	CoWoS-L
structural features	Using a silicon intermediary layer, the chip is integrated with HBM memory	Use RDL (Redistribution layer) to support longer distance signal distribution	A mix of silicon intermediary layers and RDL is used
Cost	High, silicon interlayers are complex and expensive to fabricate	Medium to high, the RDL layer design is complex and the cost is on the high side	Medium, organic intermediate layer cost is lower than silicon intermediate layer
stability	Low stability, Susceptible to uneven cold and heat	Medium, RDL layer stability depends on wiring and design	Decided by application
communication rate	High rate	Low rate	Decided by design
delay	Low delay	High delay	Medium delay
typical applications	High-performance Computing (HPC), Artificial Intelligence (AI)	Integration of multi-chip modules	Chips with large packaging areas, such as servers and data centers

CoWoS

Case Study

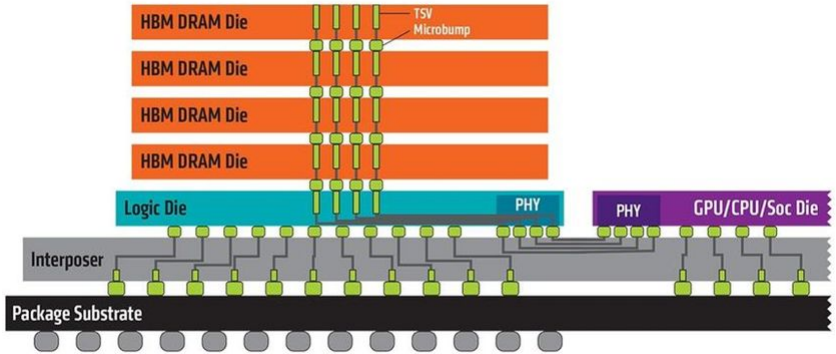
Then, the advantages and applications of cowos and other advanced packaging technologies are introduced through an example analysis.

Cowos has been applied on a large scale in many enterprise chiplets products, the most representative of which are MI300 and H100, which have obtained the vast majority of enterprise orders in the field of large models.

The key to the success of Cowos in the field of large model training is that it can realize the high information density connection between HBM3 video memory and logic operation unit, and satisfy the frequent data exchange operation during large model training and inference.

CoWoS

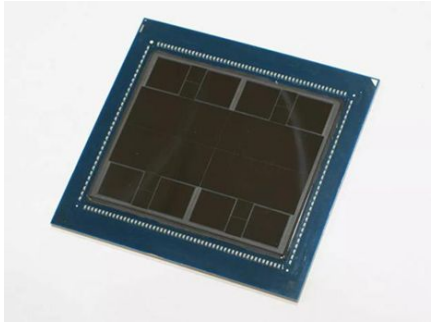
Case Study



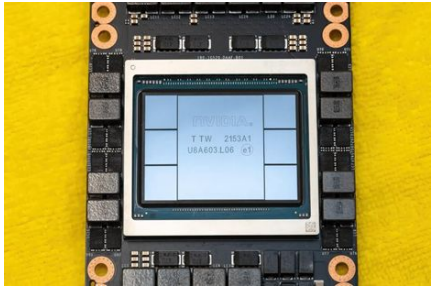
HBM3 video memory itself is also the product of advanced packaging technology, it will be the traditional dram chip through TSV technology vertical connection, usually each internal package 4-8 layers, the same scale of lower latency, bandwidth up to thousands of GB/s, is applied in high IO devices

CoWoS

Case Study



MI300



H100

H100 uses Cowos-l version of packaging technology, TSV technology is used to achieve high transmission rate in the connection of video memory and GPU, and other parts use organic intermediary layer. This is a hybrid packaging scheme that combines the low latency advantage of silicon intermediary layer and the high scalability of organic intermediary layer, and achieves a balance in terms of cost, productivity and communication density.

CoWoS

Problem of chiplets

As a chiplets product with a complete intermediary layer, the damage of any module will lead to the overall scrap of the chip, which has no maintenance value. HBM3 is more likely to be damaged than traditional memory due to its higher thermal density. This has been reflected in a large number of graphics cards in the previous Ethereum boom, and their damage is mainly due to the failure of video memory and the aging of motherboard components. The gddr memory of the traditional graphics card is packaged independently and can be replaced separately, while the h100 packaged with cowos can only be directly scrapped. This is not an environmentally friendly solution, and it also indirectly raises the cost of use for enterprises. chiplets that now use mediation layers as connections all have similar problems, but are most pronounced in larger, higher-load enterprise products.