(China)

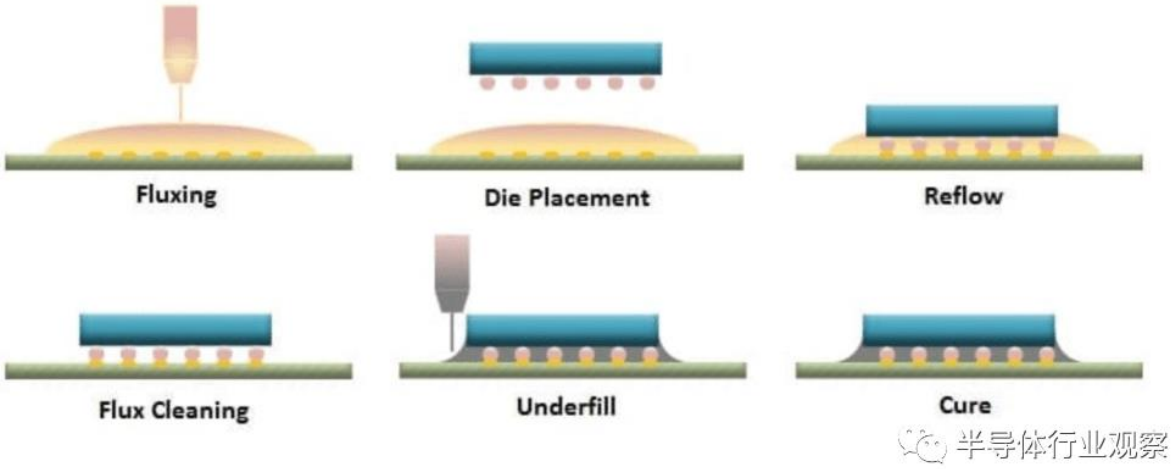
Interpretation of the giants' advanced packaging technology – January 6, 2022

巨头们的先进封装技术解读

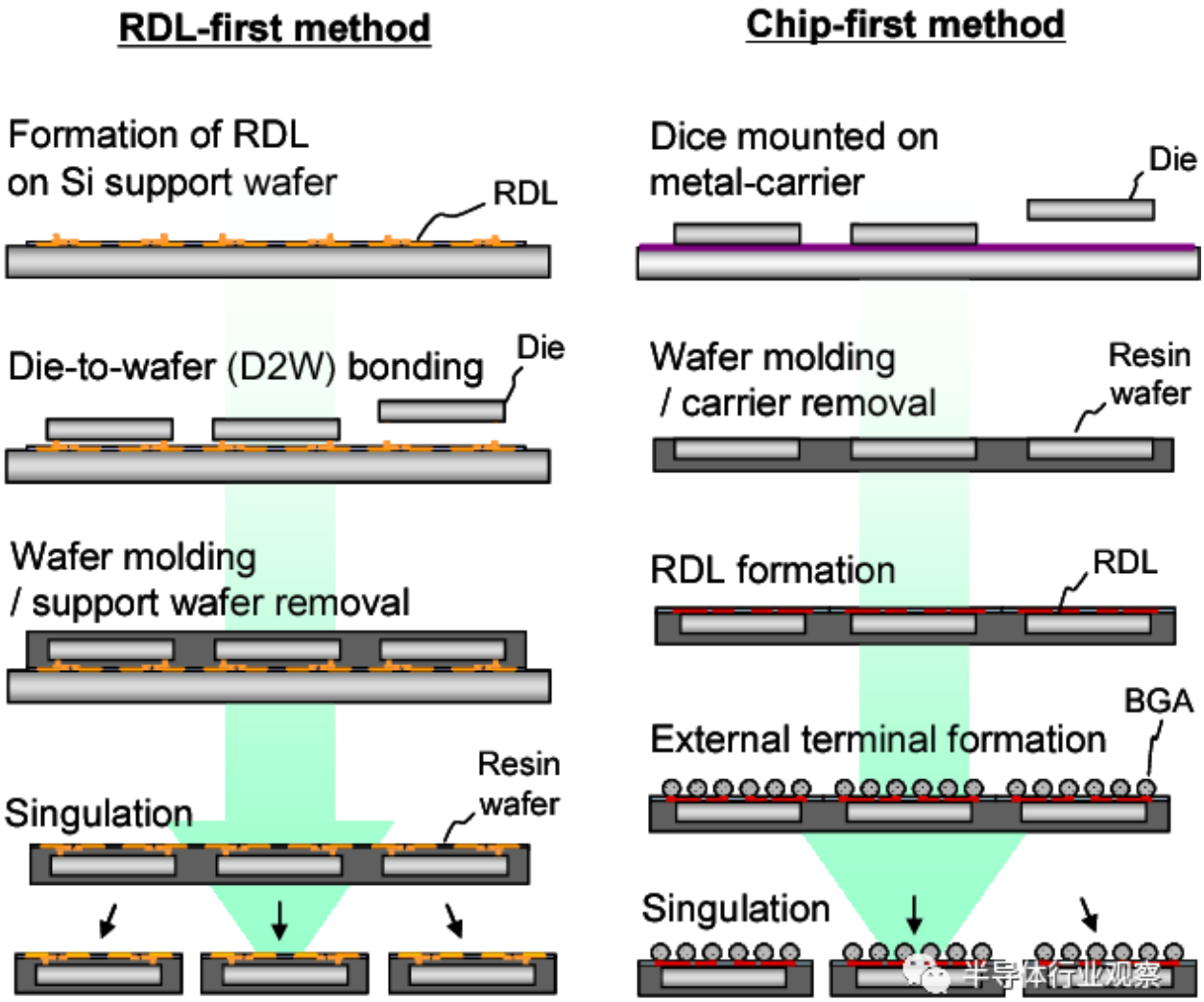
本文来自格隆汇专栏：半导体行业观察

各种公司和封装类型使用工具的多样性非常广泛. 但赢家并不像看起来那么明显。

在上《先进封装最强科普》中，我们对市场上的先进封装需求进行了一些讨论。但其实具体到各个厂商，无论是英特尔（EMIB、Foveros、Foveros Omni、Foveros Direct）、台积电（InFO-OS、InFO-LSI、InFO-SOW、 InFO-SoIS、CoWoS-S、CoWoS-R、CoWoS-L、SoIC）、三星（FOSiP、X-Cube、I-Cube、HBM、DDR/LPDDR DRAM、CIS）、ASE（FoCoS、FOEB）、索尼（ CIS)、美光 (HBM)、SKHynix (HBM) 还是YMTC (XStacking)，他们的封装的各不相同，而且这些封装类型也被我们所有最喜欢的 AMD、Nvidia 等公司使用。在本文中，我们将解释所有这些类型的封装及其用途。

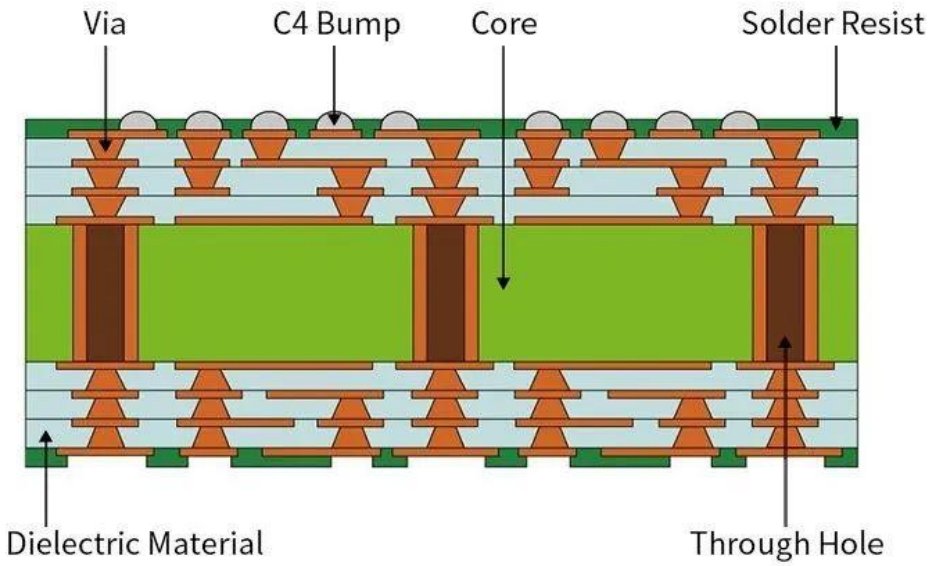


倒装芯片是引线键合后常见的封装形式之一。它由来自代工厂、集成设计制造商和外包组装和测试公司等众多公司提供。在倒装芯片中，PCB、基板或另一个晶圆将具有着陆焊盘。然后将芯片准确地放置在顶部，并使用凸块接触焊盘，之后芯片被送到回流焊炉，加热组件并回流焊凸点以将两者结合在一起。焊剂被清除，底部填充物沉积在两者之间。这只是一个基本的工艺流程，因为有许多不同类型的倒装芯片，包括但不限于fluxless。



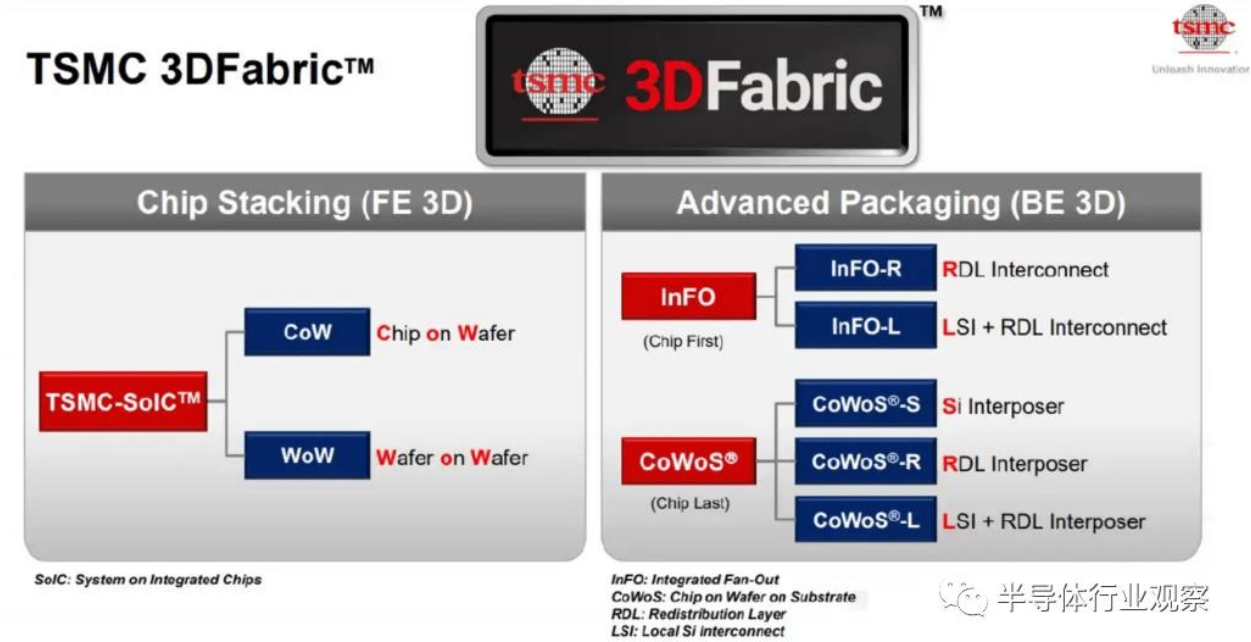
虽然倒装芯片非常普遍，但间距小于 100 微米的高级版本则不太常见。关于我们在第 1 部分中建立的先进封装的定义，只有台积电、三星、英特尔、Amkor 和 ASE 涉及使用倒装芯片技术的大量逻辑先进封装。其中 3 家公司也在制造完整的硅片，而另外两家公司则是外包组装和测试 (OSAT)。

这个尺寸就是大量不同类型倒装芯片封装类型开始涌入的地方。我们将以台积电为例，然后扩展并将其他公司的封装解决方案与台积电的封装解决方案进行比较。台积电所有封装选项的最大差异与基板材料、尺寸、RDL 和堆叠有关。



Flip-Chip Package Substrate（DLL®）

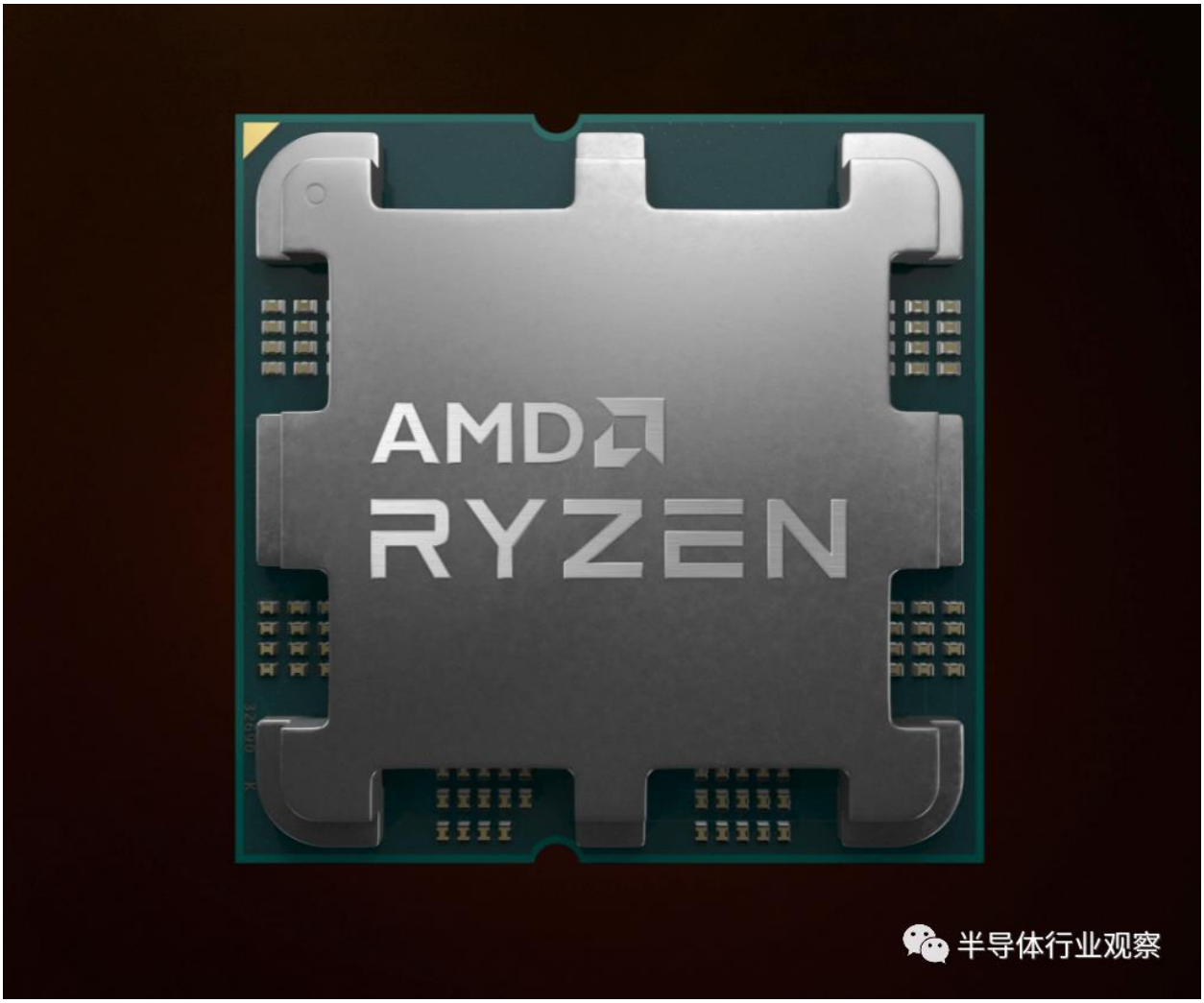
在标准倒装芯片中，最常见的基板通常是有机层压板，然后覆以铜。从这里开始，布线围绕核心两侧构建，讨论最多的是 Ajinomoto build-up films （ABF)。该内核在顶部构建了许多层，这些层负责在整个封装中重新分配信号和功率。这些承载信号的层是使用干膜层压（dry film lamination）和使用 CO_2 激光或紫外线激光进行图案化构建的。



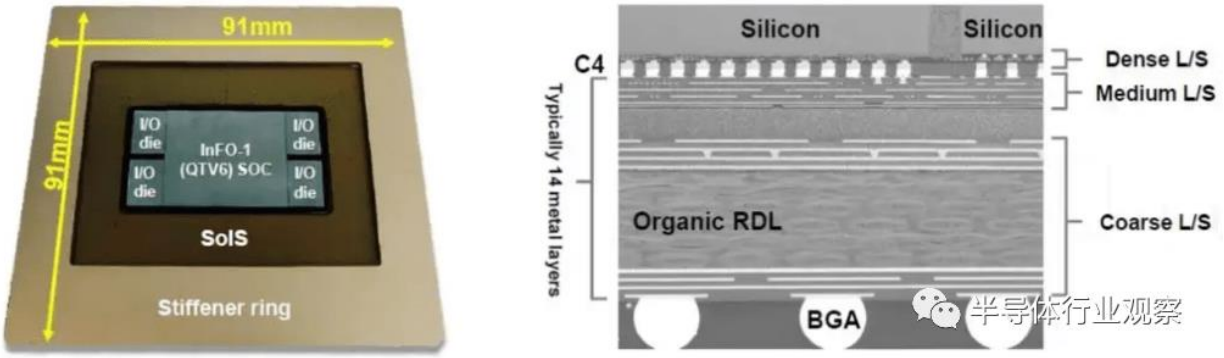
这就是台积电的专业知识开始发挥其集成扇出（InFO）的地方。台积电没有使用 ABF 薄膜的标准流程，而是使用与硅制造更相关的工艺。台积电将使用东京电子涂布机/显影剂、ASML 光刻工具、应用材料铜沉积工具以光刻方式定义再分布层。重新分布层（RTL）比大多数 OSAT 可以生产的更小、更密集，因此可以容纳更复杂的布线。此过程称为扇出晶圆级封装（FOWLP)。

ASE 是最大的 OSAT，他们提供 FoCoS（基板上的扇出芯片），这是 FOWLP 的一种形式，它也利用了硅制造技术。三星还有他们的扇出系统封装 （FOSiP），主要用于智能手机、智能手表、通信和汽车。

使用 InFO-R (RDL)，TSMC 可以封装具有高 IO 密度、复杂路由和/或多个芯片的芯片。使用 InFO-R 最常见的产品是 Apple iPhone 和 Mac 芯片，但也有各种各样的移动芯片、通信平台、加速器，甚至网络交换机 ASIC。三星还凭借 Cisco Silicon One 在网络交换机 ASIC 扇出市场中获胜。InFO-R 的进步主要与扩展到具有更多功耗和 IO 的更大封装尺寸有关。

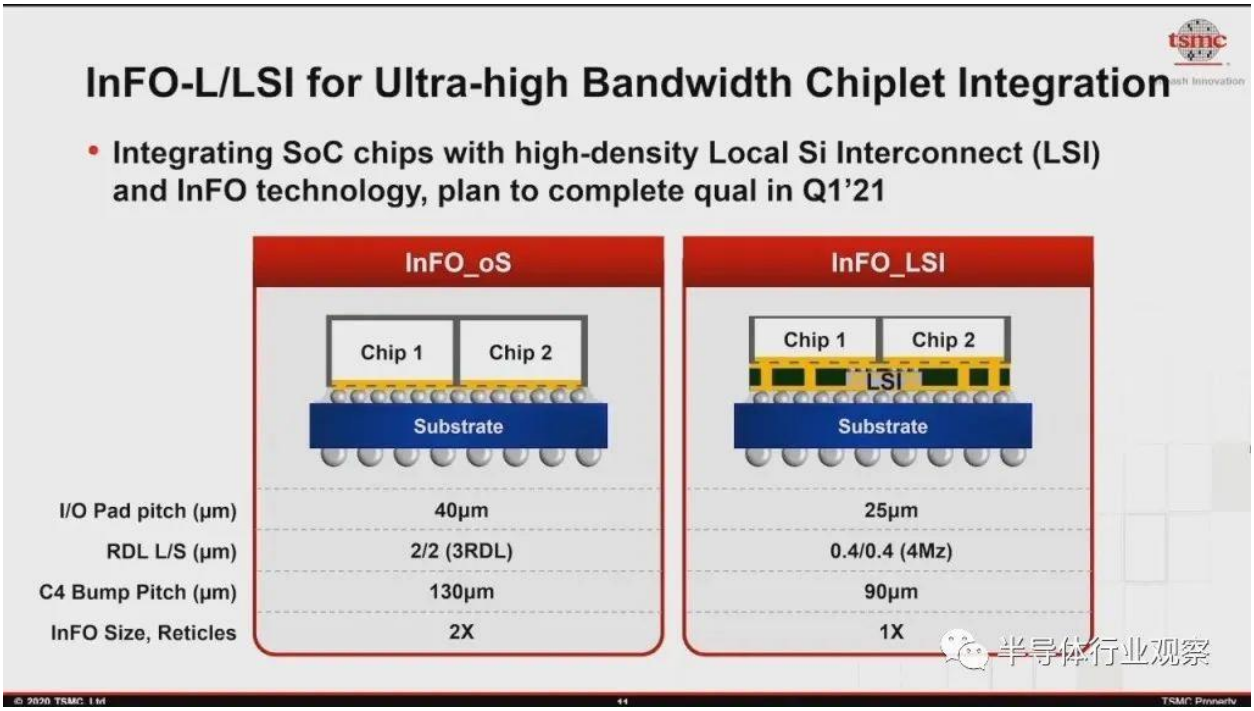


有不少传言称 AMD 将为其即将推出的 Zen 4 客户端（如上图所示）和服务器 CPU 采用扇出封装。SemiAnalysis 可以确认基于 Zen 4 的桌面和服务器产品将使用扇出。然后，该扇出将传统地封装在标准有机基板的顶部，该基板的底部将具有 LGA 引脚。包装这些产品的公司和转向扇出的技术原因将后面揭晓。



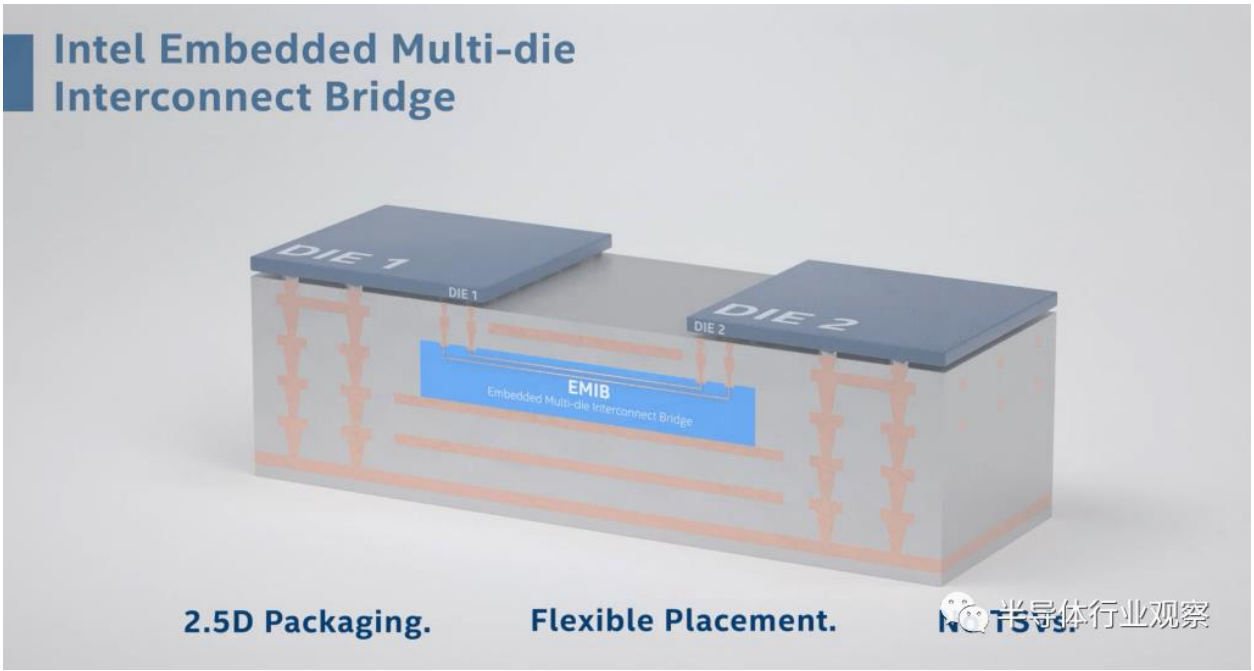
标准封装将具有核心基板，每侧有 2 到 5 层重分布层 (RDL)，包括更高级的集成扇出。台积电的 InFO-SoIS（集成基板系统）将这一概念提升到一个新的水平。它提供多达 14 个重新分布层 (RDL)，可在芯片之间实现非常复杂的布线。在靠近管芯的基板上还有一层更高密度的布线层。

TSMC 还提供InFO-SOW（晶圆上系统），它允许扇出包含数十个芯片的整个晶圆的大小。我们撰写了有关使用这种特殊包装形式的 Tesla Dojo 1 的文章。我们还在特斯拉去年的 AI 日公布这项技术的几周前独家披露了该技术的使用情况。特斯拉将在 HW 4.0 中使用三星 FOSiP。



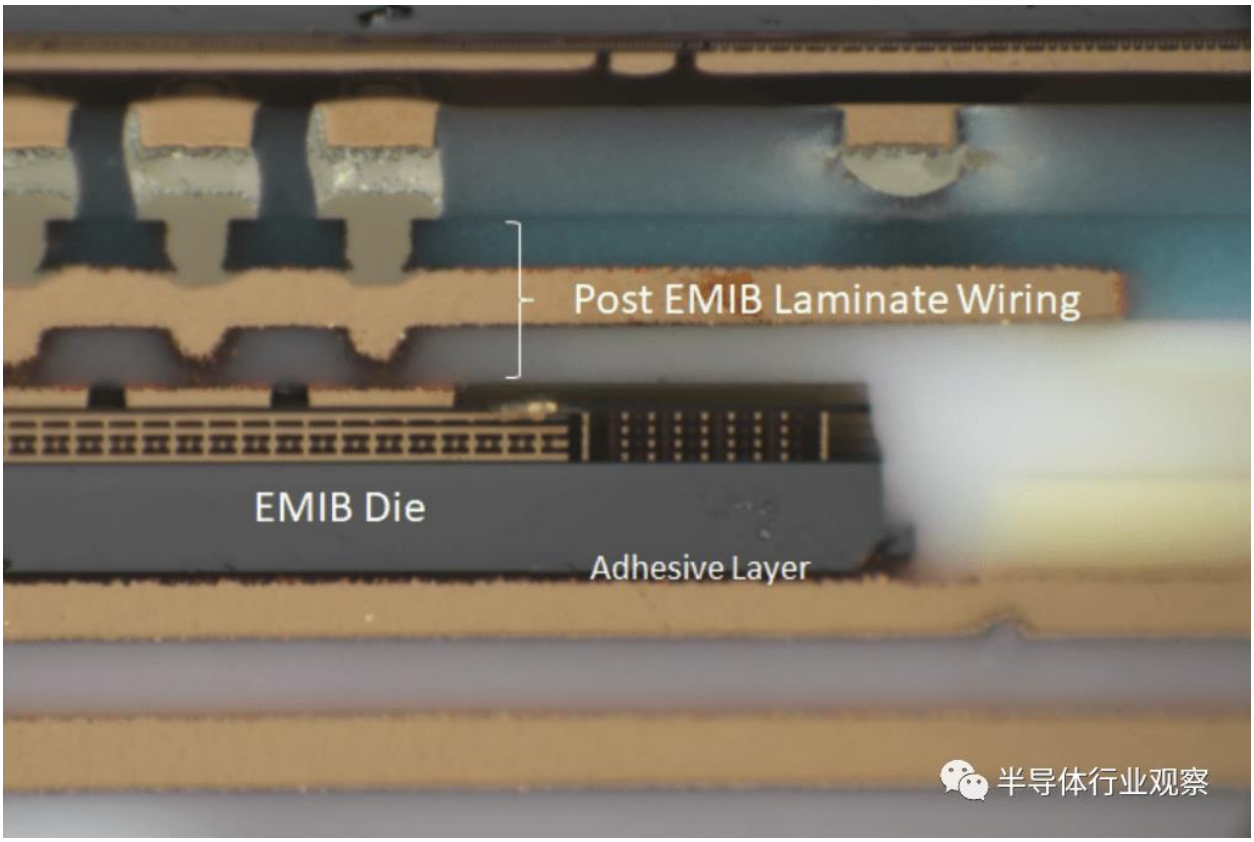
最后，在台积电的集成扇出阵容中，还有 InFO-LSI（本地硅互连）。InFO-LSI 是 InFO-R，但在多个芯片下方有一块硅。这种局部硅互连将开始作为多个die之间的无源互连，但未来可以演变为有源（晶体管和各种 IP）。它最终也将缩小到 25 微米，但我们认为第一代不会出现这种情况。第一款采用这种类型封装的产品将在后面展示。

立即想到的比较最有可能是英特尔的 EMIB（嵌入式多芯片互连桥），但这并不是真正的最佳选择。它更像是 Intel 的 Foveros Omni 或 ASE 的 FOEB。让我们解释一下。



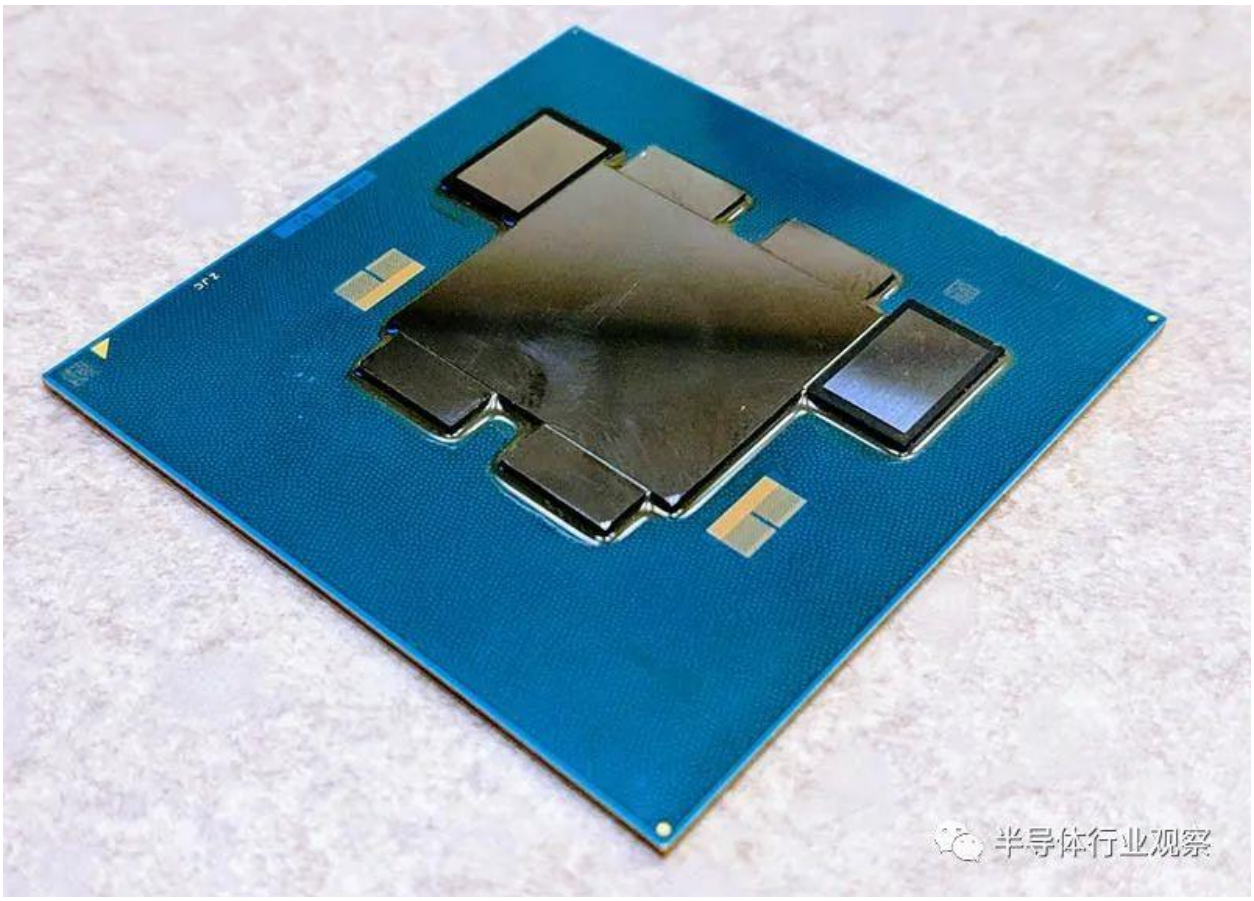
英特尔的嵌入式多芯片互连桥被放置在传统的有机基板腔体中。然后继续构建衬底。虽然这可

以由英特尔完成，但 EMIB 的放置和构建也可以由传统的有机基板供应商完成。由于 EMIB 芯片上的大焊盘以及沉积层压布线和通孔的方法，不需要在基板上非常准确地放置芯片。



通过继续使用现有的有机层压板和 ABF 供应链，英特尔放弃了更昂贵的硅基板材料和硅制造工艺。总的来说，这条供应链是商品化的，尽管目前由于短缺而相当紧张。自 2018 年以来，英特尔的 EMIB 一直在产品中发货，包括 Kaby Lake G、各种 FPGA、Xe HP GPU 和某些云服务器 CPU，包括 Sapphire Rapids。目前所有 EMIB 产品都使用 55 微米，但第二代是 45 微米，第三代是

微米。



英特尔可以通过这个芯片将功率推送到上面的有源芯片。如果需要，英特尔还可以灵活地设计封装以在没有 EMIB 和某些小芯片的情况下运行。在英特尔 FPGA 的一些拆解发现，如果英特尔发货的 SKU 不需要它，英特尔将不会放置 EMIB 和有源芯片。这允许围绕某些细分市场的物料清单进行一些优化。

最后，英特尔还可以通过仅在需要的地方使用硅桥来节省制造成本。这与台积电的 CoWoS 形成鲜明对比，后者将所有芯片都放置在单个大型无源硅桥的顶部。稍后会详细介绍，但台积电的 InFO-LSI 和英特尔的 EMIB 之间的最大区别在于基板材料和制造工艺的选择。

更复杂的是，日月光还拥有自己的2.5D封装技术，与英特尔的EMIB和台积电的InFO-LSI截然不同。它被用于 AMD 的 MI200 GPU，该 GPU 将用于多台高性能计算机，包括美国能源部的 Frontier exascale 系统。ASE 的 FOEB 封装技术与台积电的 InFO-LSI 更相似，因为它也是扇出。TSMC 使用标准的硅制造技术来构建 RDL。一个主要区别是 ASE 使用玻璃基板面板而不是硅。这是一种更便宜的材料，但它还有一些其他好处，我们将在后面讨论。

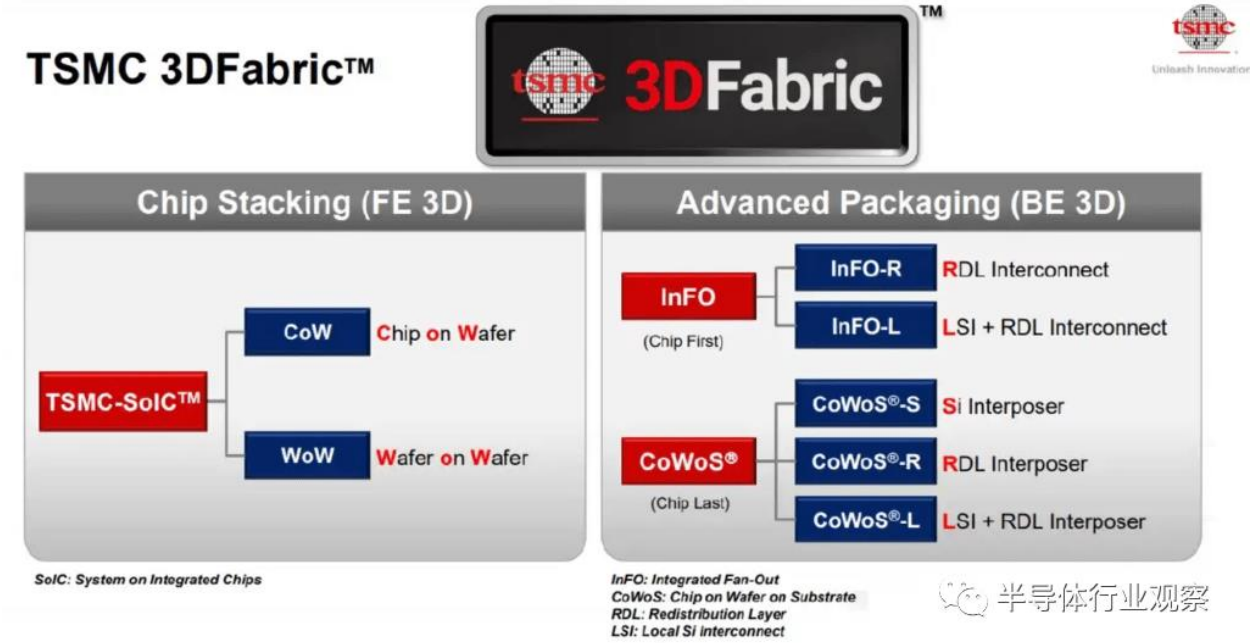
Platform (Ref. Flip Chip MCM)		FOEB	EMIB
Manufacturing values	*Supply Chain*	Simple	Complex
	Assembly Cycle time	High	Std Flip chip process (Exclude substrate fabrication time)
	Yield (%)	> 99	80 ~ 90% (Include substrate embedded bridge)
Chip	*Die QTY*	3 ~ 30	3 ~ 16
	Interposer	Organic/RDL	Organic Substrate
	I/O density	Very high	Std. Flip chip
	μ-Joint (pitch, μm)	25~40	-
	Cu Line (L/S, μm)	0.8/0.8 ~ 10/10 (Scalable)	> 5/5
	Chiplets integration	Excellent	Excellent
	Design Scalable	Good	Good
	Warpage	Low	Low
Package	*Size (mm^2)*	> 55*55	> 55*55
	Coplanarity	Comparable	Comparable
	C4 Stress	Low	Depend on die size

ASE 不是将无源互连芯片嵌入基板的空腔中，而是放置芯片，构建铜柱，然后构建整个 RDL。

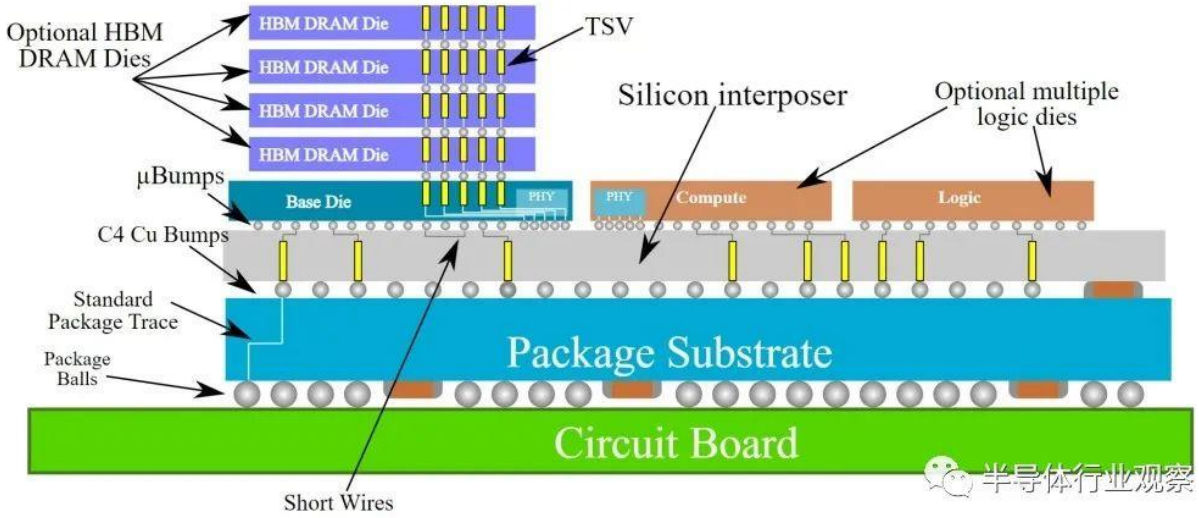
在 RDL 之上，有源硅 GPU die和 HBM die使用微凸块进行连接。然后使用激光脱模工艺将玻璃中介层从封装中移除，然后在使用标准倒装芯片工艺将其安装到有机基板上之前完成封装的另一面。

ASE 对 FOEB 与 EMIB 提出了许多声明，但有些是完全错误的。ASE 需要推销他们的解决方案是可以理解的，但让我们消除噪音。EMIB 收益率不在 80% 到 90% 的范围内。EMIB 的收益率接近 100%。第一代 EMIB 在芯片数量方面确实有缩放限制，但第二代没有。事实上，英特尔将发布有史以来最大封装的产品，一种采用第二代 EMIB 的92mm x 92mm BGA 封装的高级封装。通过在整个封装中使用扇出和光刻定义的 RDL，FOEB 确实保留了布线密度和芯片到封装凸点尺寸方面的优势，但这也更昂贵。

与台积电相比，最大的区别似乎是最初的玻璃基板材料与硅。部分原因可能是因为 ASE 的成本受到更多限制。ASE 必须以更低的价格提供出色的技术来赢得客户。台积电是芯片大师，专注于他们熟悉的技术，台积电有着将技术推向极致的文化，在这种推动下，他们最好选择硅。



现在回到台积电的其他高级封装选项，因为我们还有一些要做。CoWoS 平台还有 CoWoS-R 和 CoWoS-L 平台。它们与 InFO-R 和 InFO-L 几乎 1 比 1 对应。这两者之间的区别更多地与过程有关。InFO 是先芯片工艺，首先放置芯片，然后围绕它构建 RDL。使用 CoWoS，先建立 RDL，然后放置芯片。对于大多数试图了解高级封装的人来说，区别并不那么重要，所以今天我们将轻松地讨论这个话题。

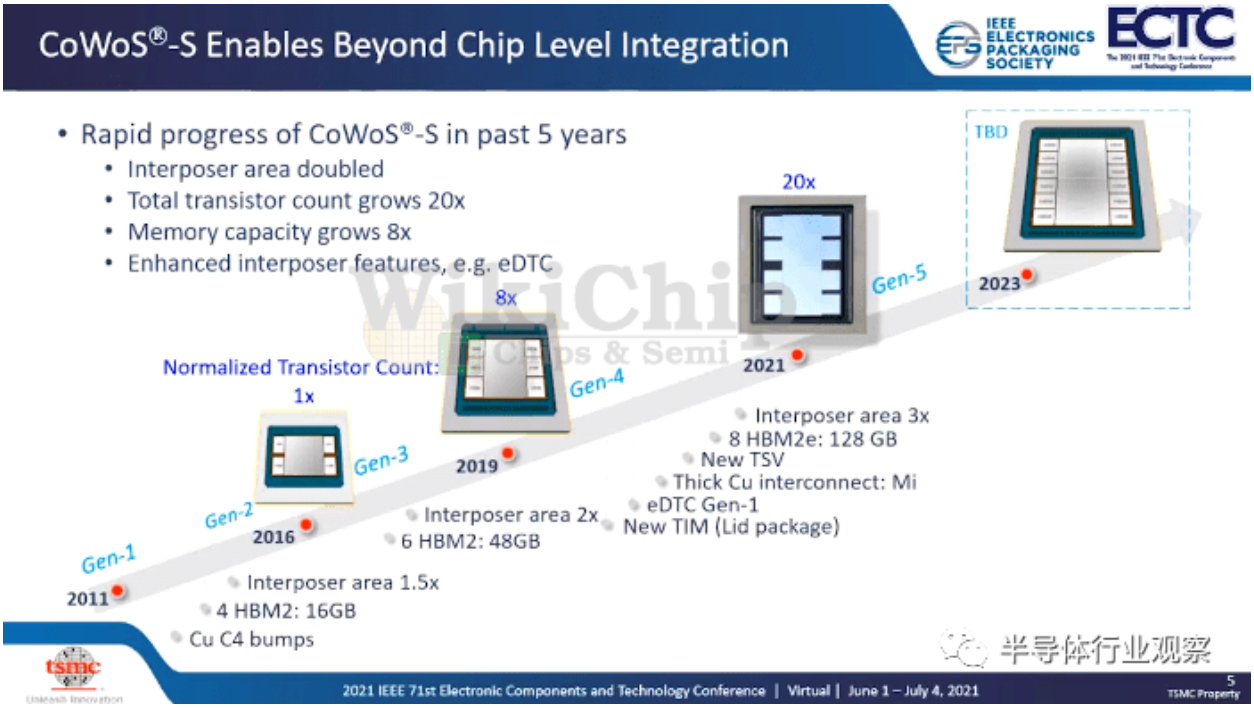


最大的亮点是 CoWoS-S（硅中介层）。它涉及采用已知良好的芯片，倒装芯片将其封装到无源晶圆上，该晶圆上具有图案化的导线。这就是 CoWoS 名称的来源，Chip on Wafer on Substrate。从长远来看，它是体积最大的 2.5D 封装平台。如第 1 部分所述，这是因为 P100、V100 和 A100 等 Nvidia 数据中心 GPU 使用 CoWoS-S。虽然 Nvidia 的销量最高，但 Broadcom、Google TPU、Amazon Trainium、NEC Aurora、Fujitsu A64FX、AMD Vega、Xillinx FPGA、Intel Spring Crest 和 Habana Labs Gaudi 只是 CoWoS 使用的几个值得注意的例子。大多数使用 HBM 计算的重型芯片，包括来自各种初创公司的 AI 训练芯片都使用 CoWoS。

为了进一步说明 CoWoS 的普及程度，这里有一些来自 AIchip 的引述。AIchip是一家台湾设计和IP公司，主要利用台积电CoWoS平台协助与AI芯片相关的EDA、物理设计和产能工作。

台积电甚至没有参加与 CoWoS 容量相关的所有会议，因为台积电已经销售了他们生产的所有产品，而且要支持所有这些设计需要太多的工程时间。另一方面，台积电的客户集中度较高（英伟达），因此台积电希望与其他公司合作。AIchip 有点像中间人，即使 Tier 1 客户（Nvidia）预订了一切，AIchip 仍然获得一些容量。即便如此，他们也只能得到他们想要的 50%。

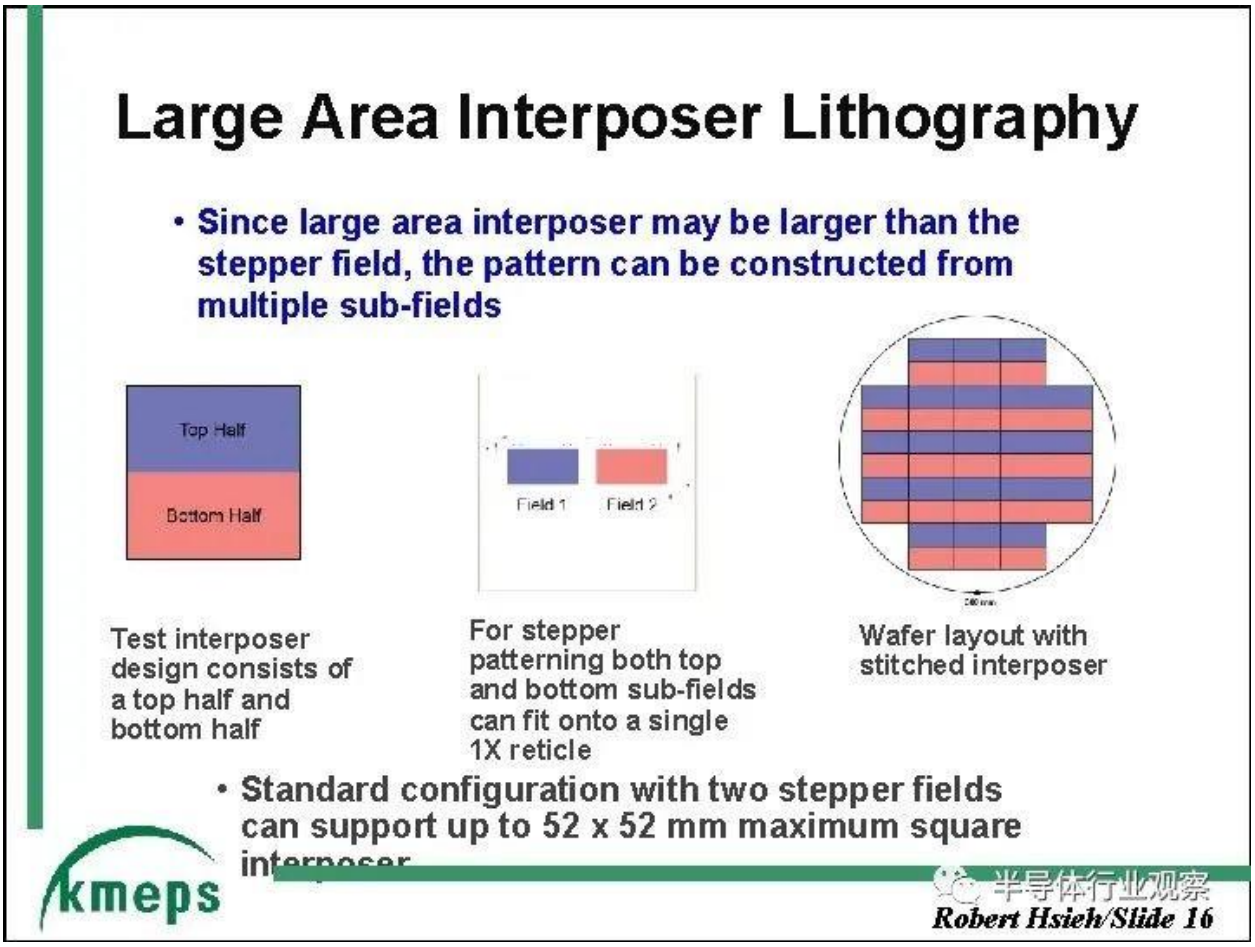
让我们转身看看英伟达在做什么。在第三季度，他们的长期供应义务跃升至 69亿美元，更重要的是，Nvidia 预付款16.4亿美元，并且未来将再预付款17.9亿美元。英伟达正在吞噬供应，特别是针对 CoWoS。



回到技术上，CoWoS-S 多年来经历了一次演变。主要特点是中介层面积越来越大。由于 CoWoS 平台使用硅制造技术，因此它遵守称为光罩限制的原则。使用 193nm ArF 光刻工具可以印刷的最大尺寸为 33mm x 26mm (858mm 2)。硅中介层的主要用途也是光刻定义的，即连接位于其上的芯片的非常密集的电线。英伟达的芯片早已接近标线限制，但仍需要连接到封装的高带宽内存。

上图包含一个 Nvidia V100，这是 Nvidia四年前推出的 GPU，它的面积是 815平方毫米。一旦包含 HBM，它就会超出光刻工具可以打印的光罩限制，但台积电想出了如何连接它们。台积电通过做光罩拼接来实现这一点。台积电在此增强了他们的能力，可以为硅中介层提供 3 倍大

小的掩模版。鉴于标线拼接的局限性，英特尔 EMIB、TSMC LSI 和 ASE FOEB 方法具有优点。他们也不必处理与大型硅中介层一样多的费用。

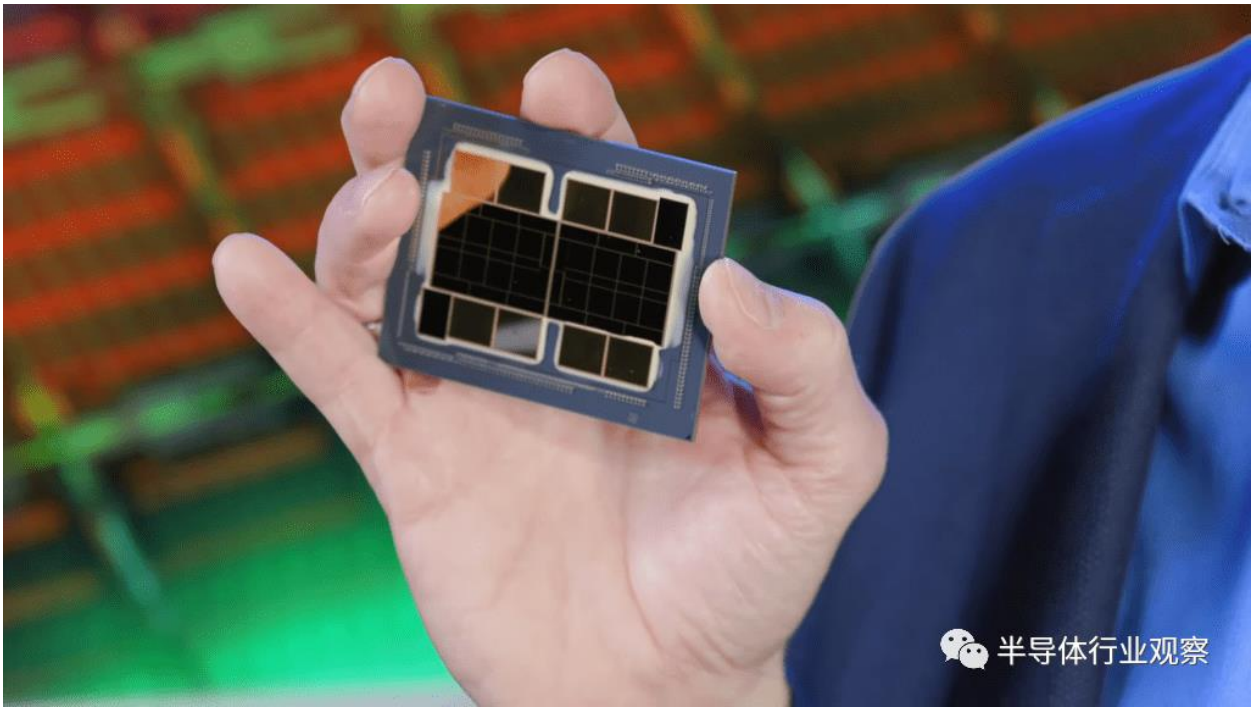


除了增加掩模版尺寸外，他们还进行了其他改进，例如将微凸块从焊料改为铜以提高性能/功率效率、iCap、新的 TIM/盖子封装等。

有一个关于 TIM/盖子包装的有趣故事。在Nvidia V100上，Nvidia 拥有一个无处不在的 HGX 平台，该平台可以运送到许多服务器 ODM，然后运送到数据中心。可以应用于冷却器螺钉以实现正确安装压力的扭矩非常具体。这些服务器 ODM 在这些价值 10,000 美元的 GPU 上过度拧紧了冷却器和芯片。Nvidia 的 A100 转移到在芯片上有盖子的封装，而不是直接冷却芯片。当 Nvidia 的 A100 和未来的 Hopper DC GPU 仍然需要散发大量热量时，这类封装的问题就会出现。为了解决这个问题，台积电和英伟达在封装上进行了很多优化。

三星也有类似于 CoWoS-S 的 I-Cube 技术。三星使用这种封装的唯一主要客户是百度的 AI 加速器。

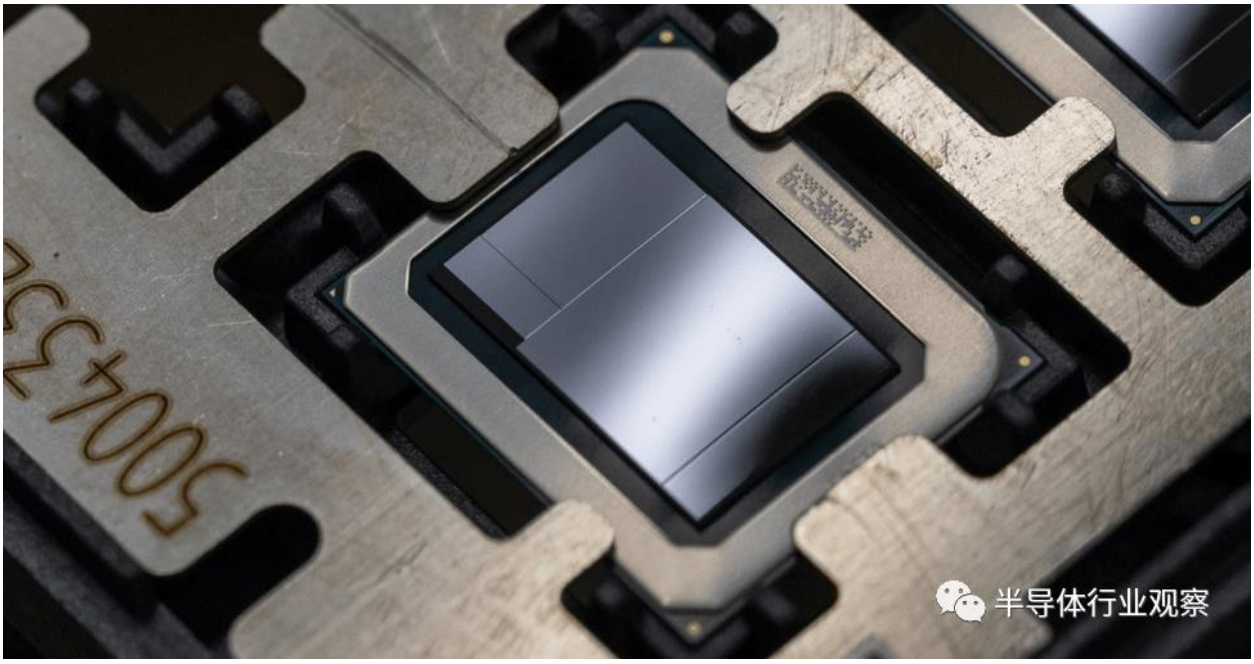
接下来我们有 Foveros。这就是英特尔的3D芯片堆叠技术。Foveros 不是一个裸片在另一个裸片的顶部活动，而后者本质上只是密集的导线，Foveros 涉及两个包含活动元素的裸片。有了这个，英特尔第一代 Foveros 于 2020 年 6 月在 Lakefield 混合 CPU SOC 中推出。该芯片不是特别大的容量或令人叹为观止的芯片，但它是英特尔的许多第一款芯片，包括 3D 封装和他们的第一个混合 CPU 内核具有大性能核心和小效率核心的架构。它采用了 55 微米的凸点间距。



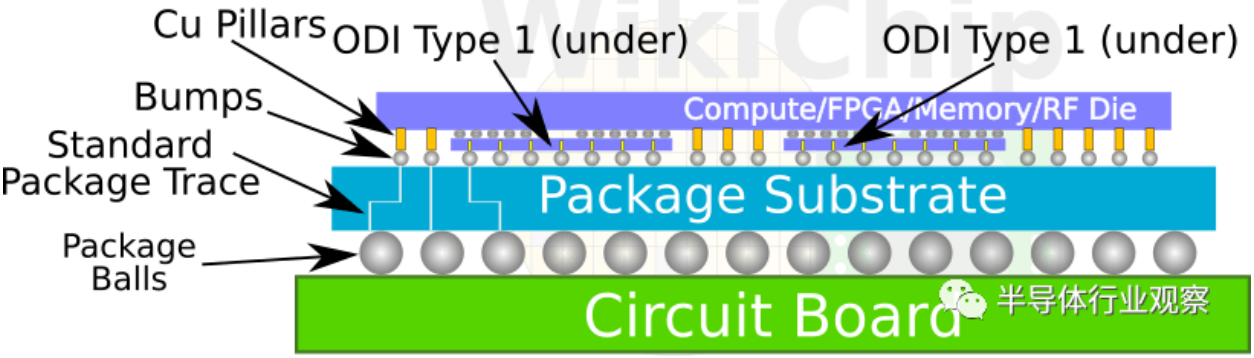
下一个 Foveros 产品是 Ponte Vecchio GPU，经过多次延迟，它应该在今年推出。它将包括与 EMIB 和 Foveros 一起封装的 47 个不同的有源小芯片。Foveros 芯片到芯片的连接采用

36 微米的凸点间距。

未来，英特尔的大部分客户端阵容都将采用3D堆栈技术，包括代号为Meteor Lake、Arrow Lake、Lunar Lake的客户端产品。Meteor Lake 将是首款采用 Foveros Omni 和 36 微米凸点间距的产品。第一个包含 3D 堆栈技术的数据中心 CPU 代号为 Diamond Rapids，其名称是 Granite Rapids。我们将在本文中讨论其中一些产品使用的节点以及英特尔与台积电的关系。



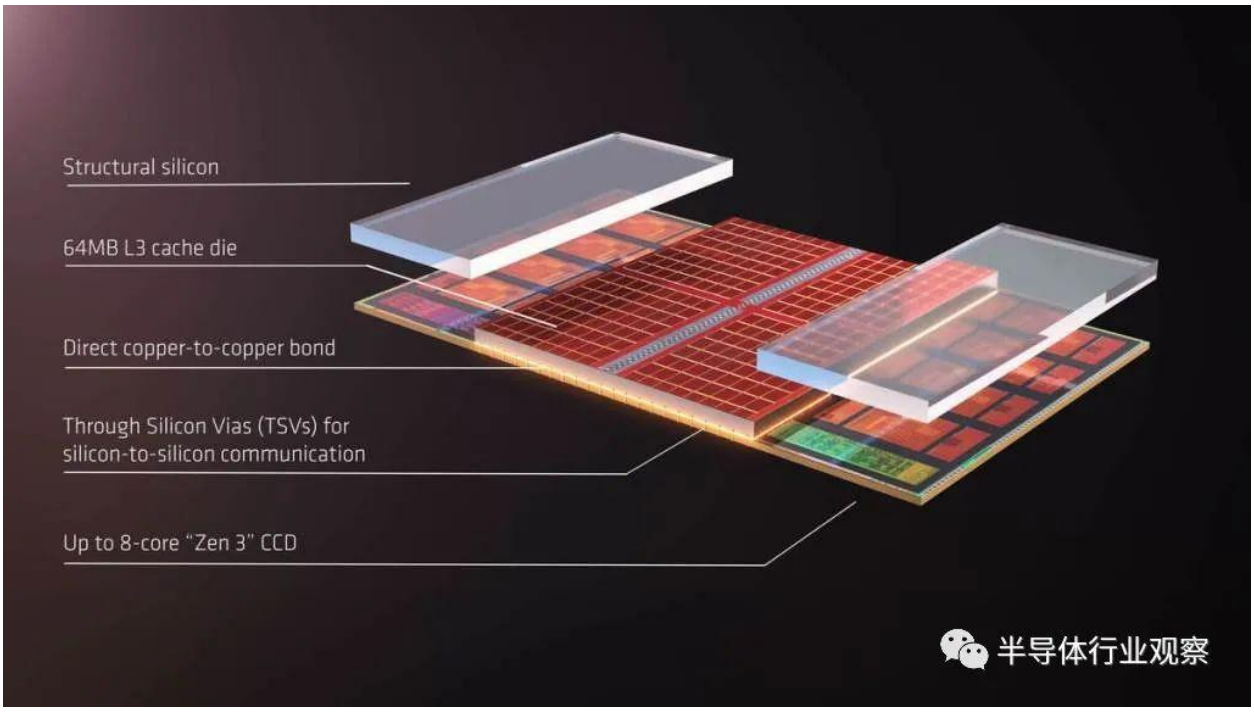
Foveros Omni 的全称是 Foveros Omni-Directional Interconnect (ODI)。它弥补了 EMIB 和 Foveros 之间的差距，同时还提供了一些新功能。Foveros Omni 可以作为两个其他芯片之间的有源桥接芯片，作为完全位于另一个芯片下方的有源芯片，或位于另一个芯片顶部但悬垂的芯片。



Foveros Omni 从未像 EMIB 那样嵌入基板内部，它在任何情况下都完全位于基板之上。堆叠类型会导致封装基板与位于其上的芯片的连接高度不同的问题。英特尔开发了一种铜柱技术，让他们可以将信号和电源传输到不同的 z 高度并通过芯片，这样芯片设计人员在设计 3D 异构芯片时可以有更多的自由。Foveros Omni 将从 36 微米的凸点间距开始，但在下一代将降低到 25 微米。

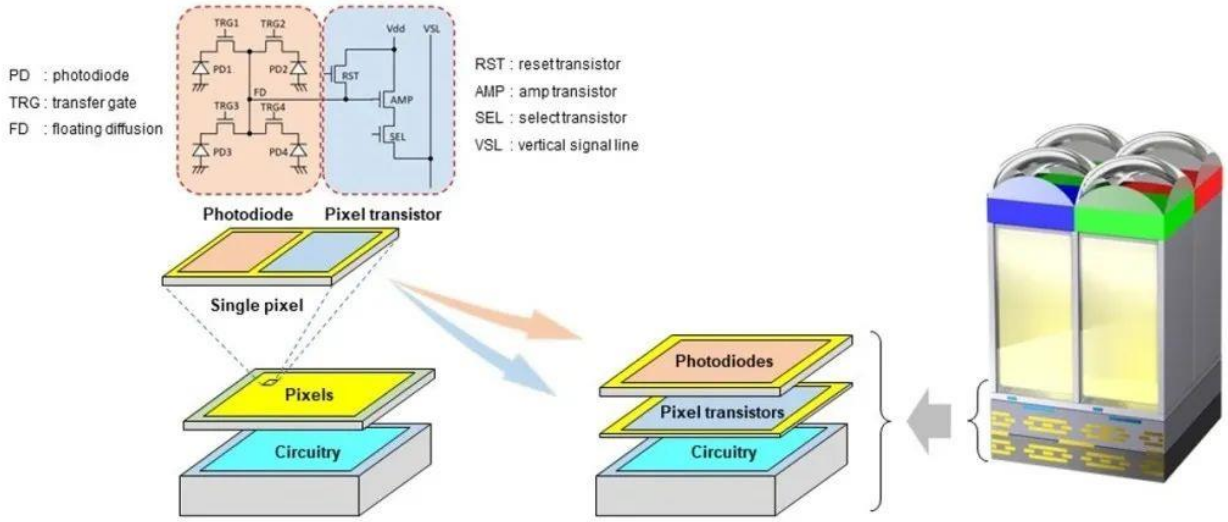
我们要注意的是，DRAM 还使用了先进的 3D 封装。HBM 多年来一直在三星、SK 海力士和美光使用先进封装。将制造存储单元并连接到暴露并形成微凸块的 TSV。最近，三星甚至开始推出 DDR5 和 LPDDR5X 堆栈，它们利用类似的堆栈技术来提高容量。SKHynix 正在其 HBM 3 中引入混合键合。SKHynix 将把 12 个芯片键合在一起，每个芯片的厚度约为 30 微米，并带有混合键合 TSV。

混合键合是一种技术，它不使用凸点，而是将芯片直接与硅通孔连接。如果我们回到倒装芯片工艺，没有凸块形成、助焊剂、回流或模下填充芯片之间的区域。铜直接遇到铜。实际过程非常困难，上面部分详述。在本系列的下一部分中，我们将深入研究工具生态系统和混合绑定类型。与之前描述的任何其他封装方法相比，混合键合能够实现更密集的集成。



最著名的混合键合芯片当然是最近宣布的 AMD 的 3D 堆叠缓存，它将于今年晚些时候发布。这利用了台积电的 SoIC 技术。英特尔的混合键合品牌称为 Foveros Direct，三星的版本称为 X-Cube。Global Foundries 公开了使用混合键合技术的 Arm 测试芯片。产量最高的混合键合半导体公司不是台积电，今年甚至明年也不会是台积电。出货最多的混合键合芯片的公司实际上

是拥有 CMOS 图像传感器的索尼。事实上，假设你有一部高端手机，你的口袋里可能有一个包含混合粘合 CMOS 图像传感器的设备。如第 1 部分所述，索尼已将间距缩小至 6.3 微米，而 AMD 的 V-cache 间距为 17 微米。



Conventional stacked CMOS image sensor

Stacked CMOS image sensor with newly developed 2-Layer Transistor Pixel technology

目前索尼提供 2 stack 和 3 stack 版本。在 2 堆栈中，像素位于电路的顶部。在 3 堆栈版本中，像素堆叠在电路顶部的 DRAM 缓冲区缓存的顶部。随着索尼希望将像素晶体管从电路中分离出来并创建具有多达 4 层硅的更先进的相机，进步仍在继续。由于其 CMOS 图像传感器业务，三星是混合键合芯片的第二大出货量出货商。

混合键合的另一个即将大批量应用是来自长江存储技术公司的 Xtacking。YMTC 使用晶圆到晶圆键合技术将 CMOS 外围堆叠在 NAND 门下方。我们在这里详细介绍了这项技术的好处，但简而言之，它允许 YMTC 在给定一定数量的 NAND 层数的情况下安装更多的 NAND 单元，而不是任何其他 NAND 制造商，包括三星、SK 海力士、美光、Kioxia 和西部数据。

关于各种类型的倒装芯片、热压键合和混合键合工具，有很多话要说，但我们将把这些留到下一篇。投资者对 Besi Semiconductor、ASM Pacific、Kulicke 和 Soffa、EV Group、Suss Microtec、SET、Shinkawa、Shibaura 和 Applied Materials 的共同认识是不正确的，这里的各种公司和封装类型使用工具的多样性非常广泛. 但赢家并不像看起来那么明显。

https://www.gelonghui.com/p/504713