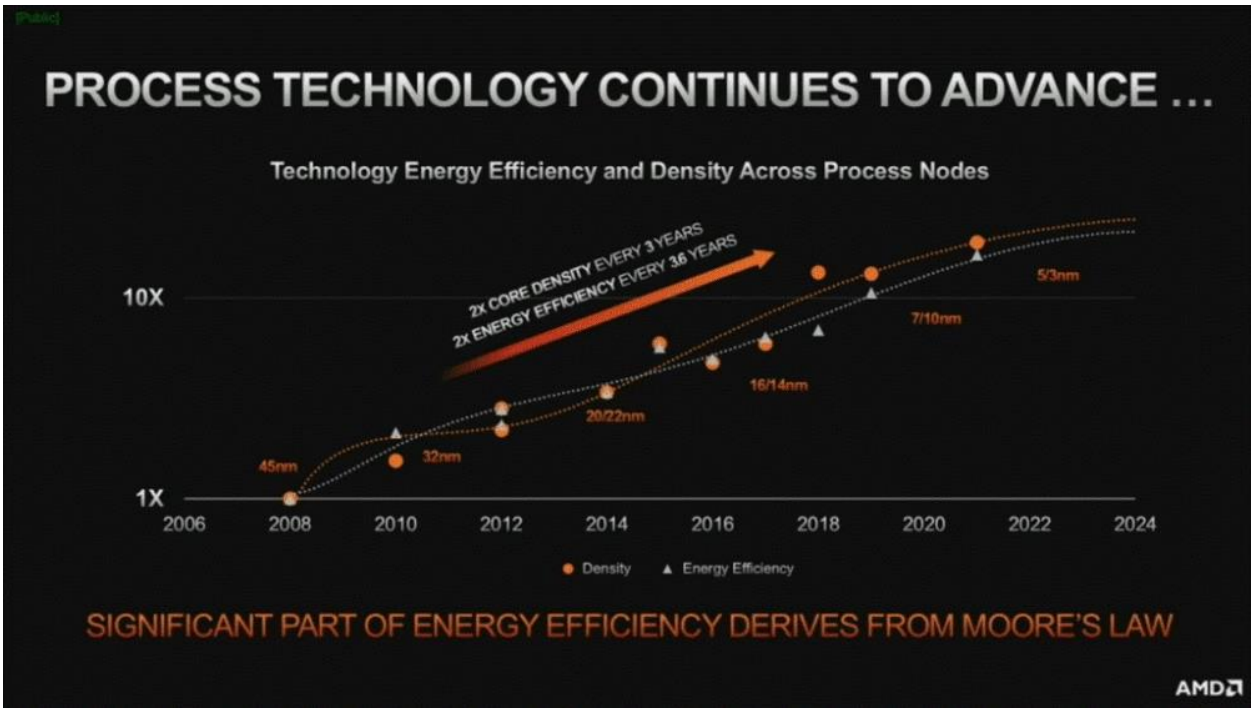
(China)

The strongest science of advanced packaging – January 31, 2022

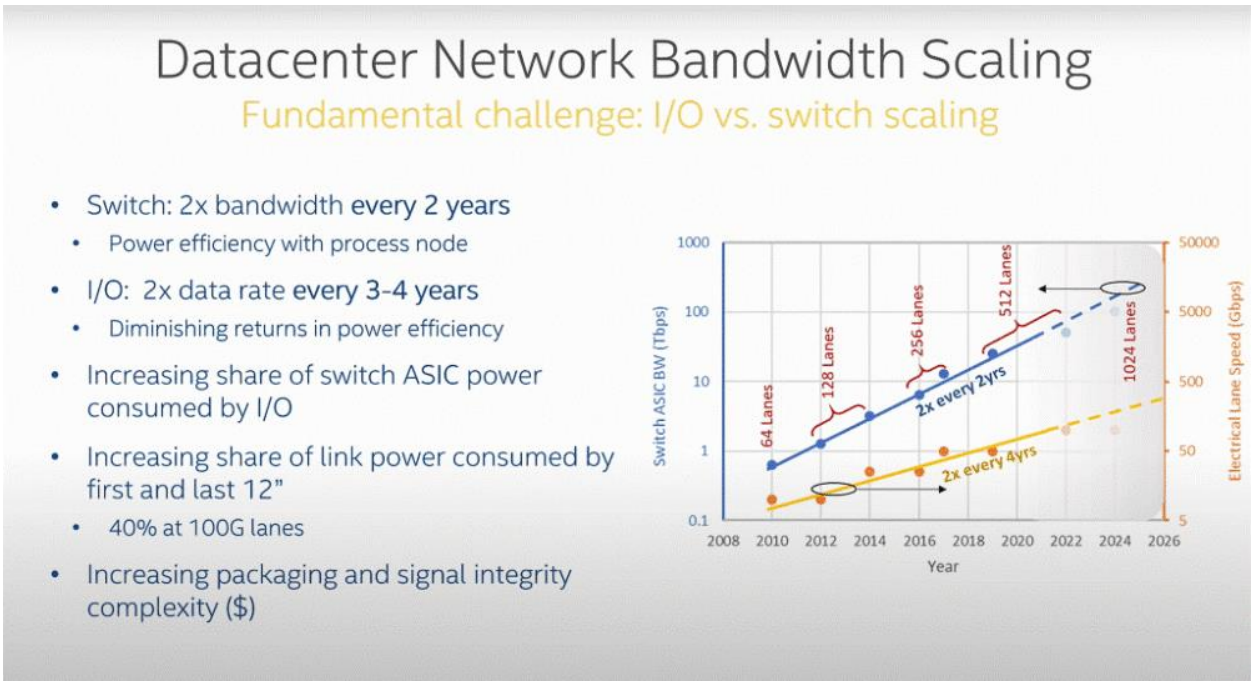
先进封装最强科普

在过去几年中，先进封装已成为半导体越来越普遍的主题。在这个由多个部分组成的系列中，SemiAnalysis将打破大趋势。我们将深入研究实现先进封装的技术，例如高精度倒装芯片、热压键合 (TCB) 和各种类型的混合键合 (HB)。

本次深入探讨将包括各种代工厂、IDM、OSAT和无晶圆厂设计公司的使用状况、设备采购以及技术选择的差异。它还将包含 Besi Semiconductor、ASM Pacific、Kulicke and Soffa、EV Group、Suss Microtec、SET、Shinkawa、Shibaura和 Xperi 等公司对设备和供应链的评论。

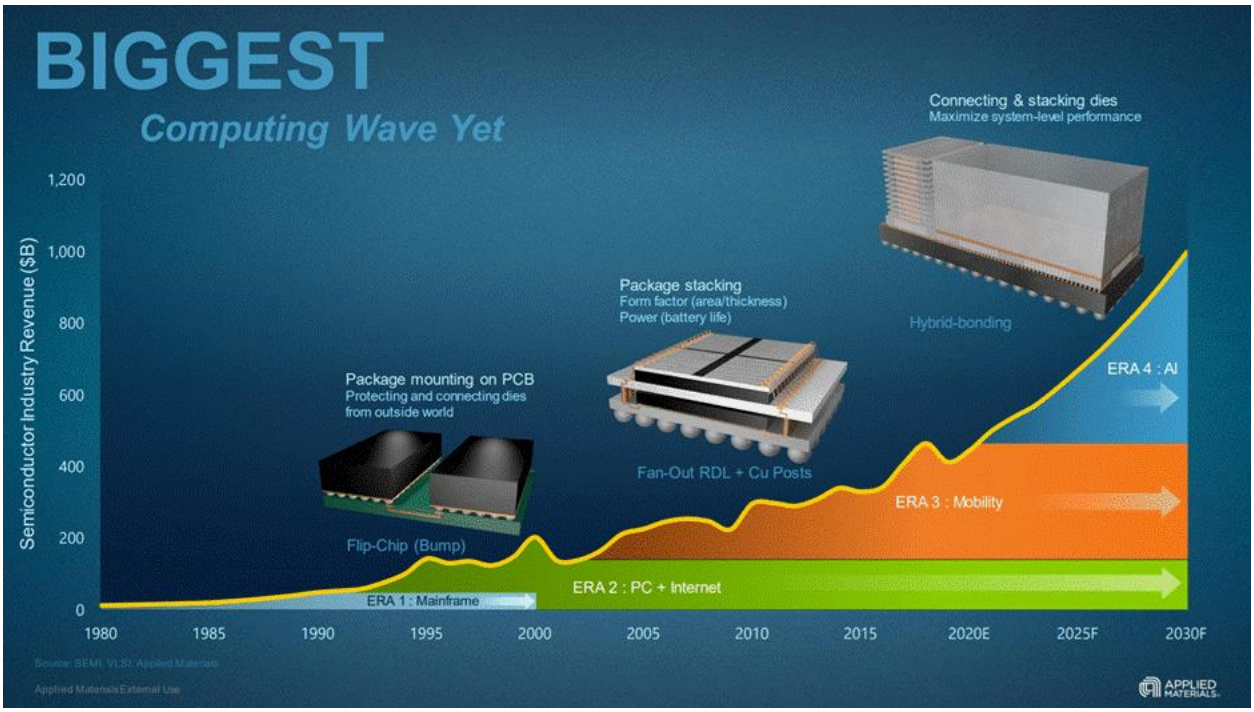


首先让我们讨论一下对先进封装的需求。摩尔定律以迅猛的速度发展。自台积电 32nm 失误以来，直到目前的 5nm 工艺节点，台积电的晶体管密度每年增长 2 倍。尽管如此，真实芯片的密度每 3 年增长约 2 倍。这种较慢的速度部分是由于 SRAM 缩放、功率传输和热密度的消亡，但大多数这些问题都与数据的输入和输出有关。



芯片上数据的输入和输出 (IO) 是计算的命脉。将内存置于芯片上有助于通过减少通信开销来减少 IO 需求，但归根结底，这是一种有限的扩展途径。处理器必须与外部世界进行交易以发送和接收数据。摩尔定律使业界的晶体管密度大约每 2 年增加 2 倍，但 IO 数据的速率每 4 年才增加 2 倍。几十年来，晶体管密度与 IO 数据速率的这种差异出现了巨大差异。共同封装的光学器件只是解决这个问题的一种方法，它并不是单独出现的。

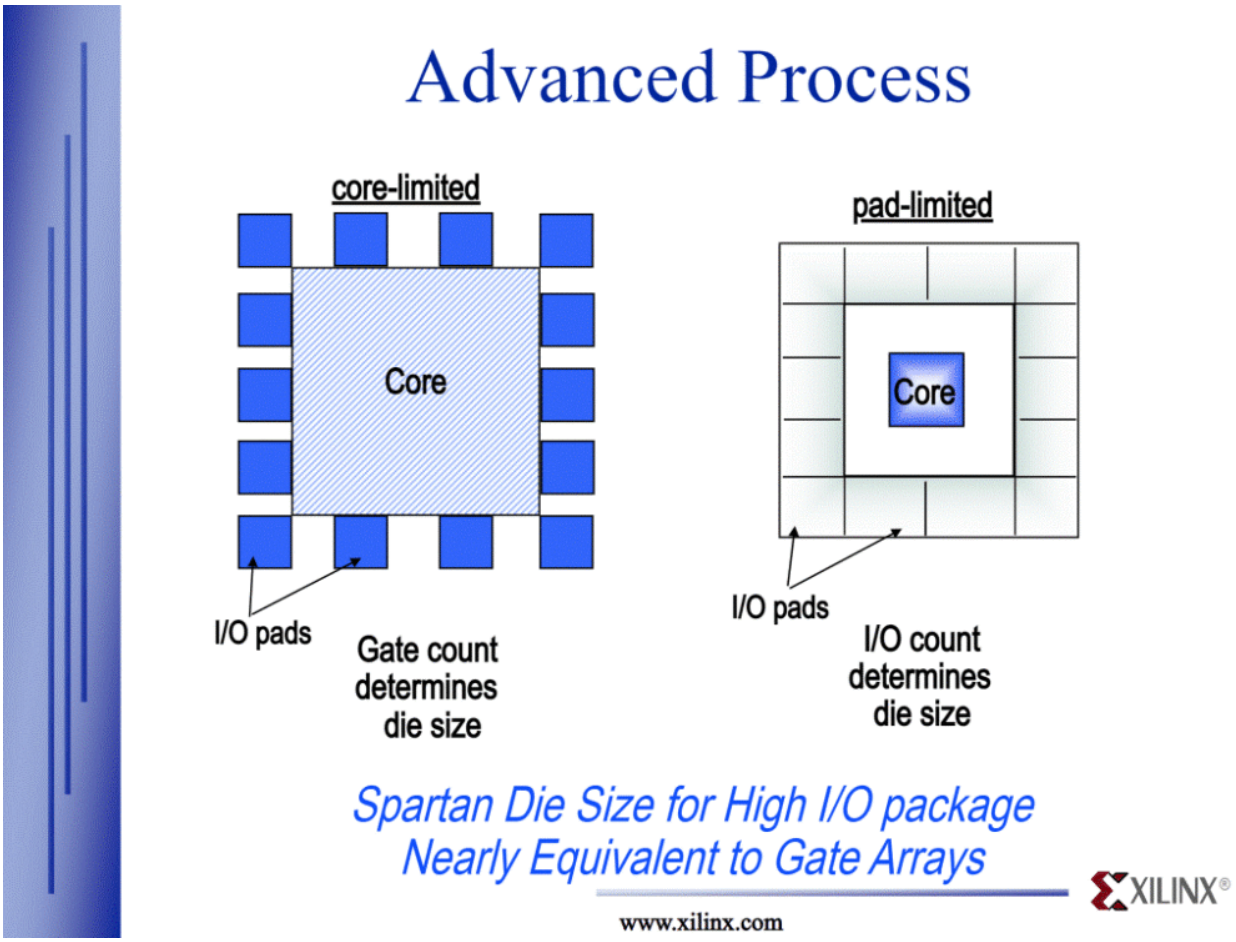
从根本上说，芯片需要容纳更多的通信或 IO 点才能跟上。不幸的是，这方面的最后一个主要步骤功能增加是在 90 年代转向倒装芯片封装。



传统的倒装芯片封装的凸点间距在 150 微米到 200 微米之间。这意味着每个 IO 单元在裸片的底侧相距 150 到 200 微米。台积电 N7将凸点间距降低到 130 微米，英特尔的 10nm 将凸点间距降低到 100 微米，这些进步被称为细间距倒装芯片。不要小看这些进步，因为它们极大地促进了更好的处理器，但 2000 年的封装技术与 2021 年的封装技术基本相同。

2000年的250mm?的芯片与2022年的250mm?芯片在晶体管数量、性能和成本方面有着难以置信的不同。摩尔定律每 2 年翻一番，表示晶体管数量增加了 2000 倍以上。显然，现实并不那么有利，但晶体管仍然增加了几个数量级。在硬币的另一面，封装没有享受同样水平的增长。

在台积电的 N7 节点上，AMD 的凸块间距从约 200 微米变为 130 微米，IO 仅增加了 2.35 倍。如前所述，英特尔在 10 纳米工艺上从200 微米的凸点间距变为 100 微米，从而实现了更大的缩放。这仍然只会使 IO 增加 4 倍。2.35倍或4倍的增加是相对于晶体管数量增加的舍入误差。



这带来了pad（硅片的管脚）受限设计的概念。将旧设计转移到新工艺节点时，设计本身可能会大幅缩小，但 IO 需求会阻止芯片尺寸缩小多少。由于需要 IO，裸片尺寸保持较大且留有空白空间。这些情况称为pad limited，并且非常频繁。

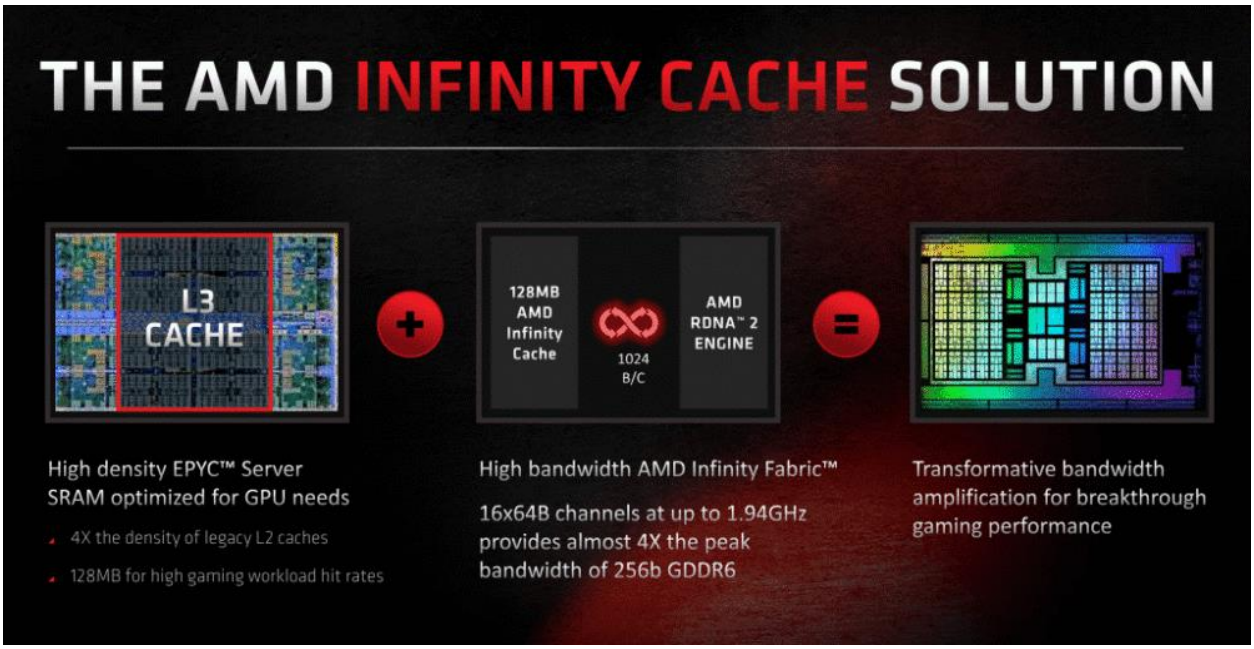
顺便说一句，这不仅与将使用先进封装的前沿有关，而且与围绕汽车芯片和一般半导体短缺的讨论有关。Intel 的首席执行官 Pat Gelsinger 认为，这些短缺的公司应该转向 Intel 16nm 代工服务。

Pat Gelsinger表示，今天，我们宣布在英特尔 16 和爱尔兰工厂的其他节点上提供欧洲代工服务，我们相信这有机会帮助加速结束供应短缺，我们正在与汽车和其他行业合作帮助建立这些能力。但我也想说有些人

可能会争辩说，好吧，让我们在旧节点上构建大部分汽车芯片。旧节点不需要一些旧晶圆厂吗？我们是想投资过去还是想投资未来？

一个新的晶圆厂需要 4 到 5 年的时间才能建成并具有生产价值。不是解决今天的危机，投资于未来，不要选择向后投资。相反，我们应该将所有设计迁移到新的现代节点，为未来增加供应和灵活性做好准备。

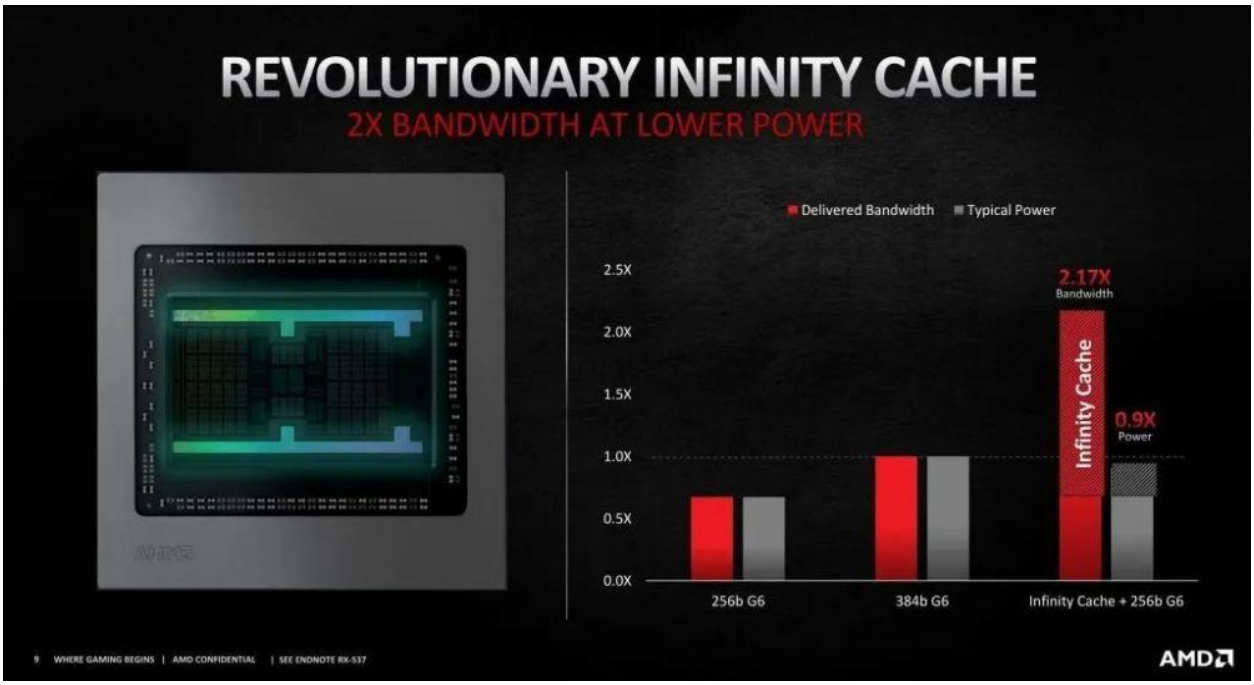
Intel的问题在于，当从古老的节点转移到相对现代的节点时，这些设计将受到pad限制。由于每 mm?的成本较高，单位成本经济学在这里不起作用，因为由于pad有限，芯片面积不能很好地缩放。除了这些成本之外，由于必须在较新的节点上重新设计旧芯片和整个重新认证过程，因此一次性成本也很高。将旧芯片移到新节点的解决方案是不可行的。



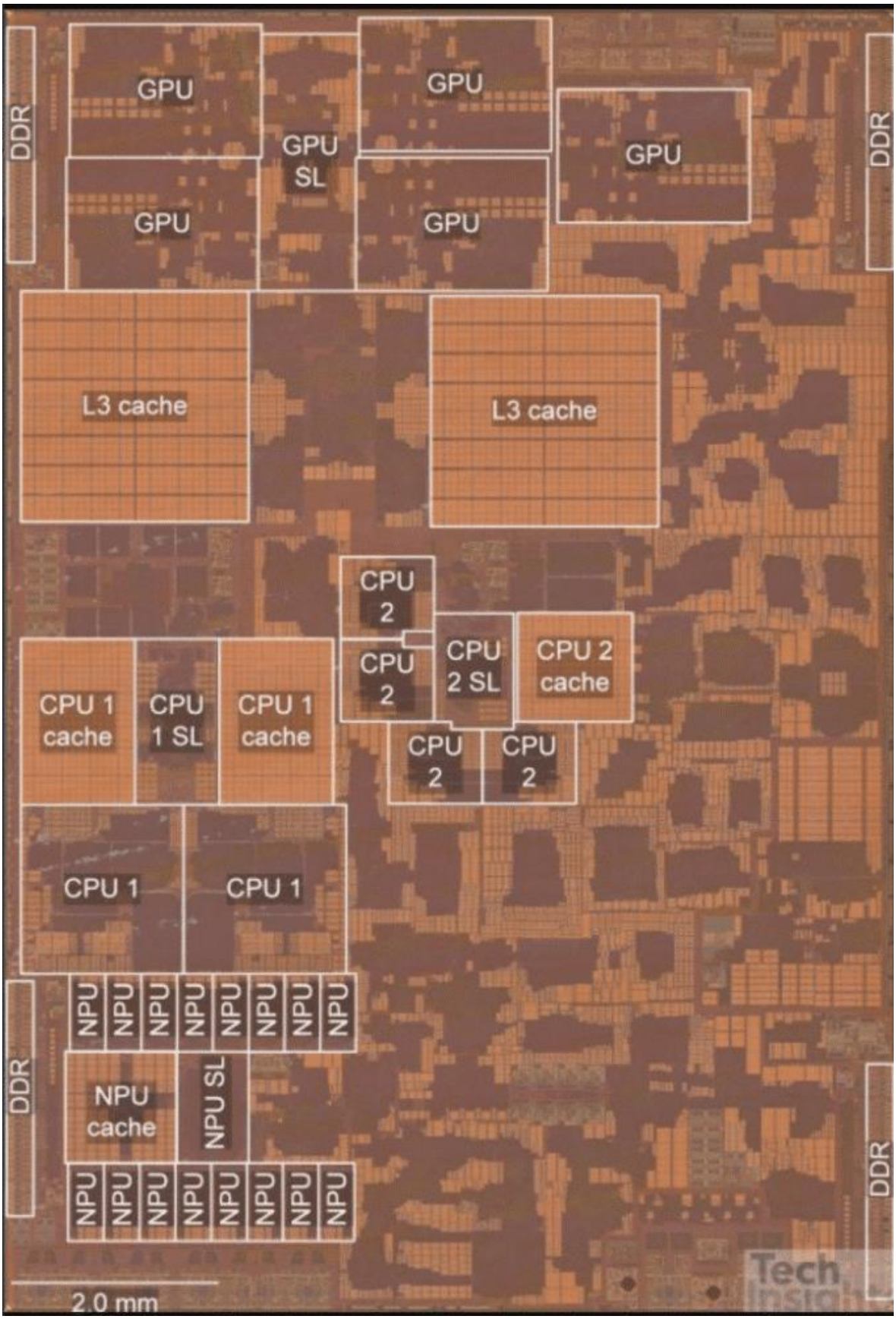
那么如何增加 IO 计数呢?

一种途径是寻找使芯片更大的方法。面积越大，IO的空间就越大。这不是最佳途径，但设计人员会经常增加芯片上的内存，以便在芯片上存储更多数据。这反过来又在一定程度上减少了 IO 需求。AMD 最近的架构就是一个很好的例子，因为它们在 CPU 和 GPU 上都有巨大的缓存。

AMD 将其命名为 InfinityCache（无线缓存）。解决方案是通过提供大量的片上 SRAM 来存储处理器中与计算最相关的数据，从而降低内存带宽要求。在 GPU 领域，AMD 明确表示他们能够通过添加无限缓存将 GDDR6 总线大小从 384 位减少到 256 位。苹果在这方面也很积极，在他们内部设计的处理器上塞满了大量的缓存。这些设计选择的一个组成部分与功率有关，但很大一部分也是由于pad限制。



另一种途径是添加各种专用电路以提高芯片效率。我们在大量的异构计算中看到了这一点。回到我们的 Apple A15 芯片分析，令人惊讶的是 CPU 或 GPU 的专用区域如此之少。这是人们谈论最多的两个方面。Apple 没有专注于这些营销方面，而是将大量区域用于其他功能。虽然没有标注，但右下角主要是图像信号处理器。这块巨大的部分正在做与拍照和视频相关的计算。还有另一个未标记的块与媒体编码和解码相关的计算有关。在 SoC 周围，可以找到这些相当小的统一矩形，它们是 SRAM 缓存，可将更多数据保存在芯片上，而不必进入内存。


GPU
GPU
GPU SL
GPU
GPU
GPU
DDR
DDR
L3 cache
L3 cache
CPU 2
CPU 2
CPU 2 SL
CPU 2 cache
CPU 1 cache
CPU 1 SL
CPU 1 cache
CPU 2
CPU 2
CPU 1
CPU 1
NPU
NPU cache
NPU SL
DDR
DDR
2.0 mm

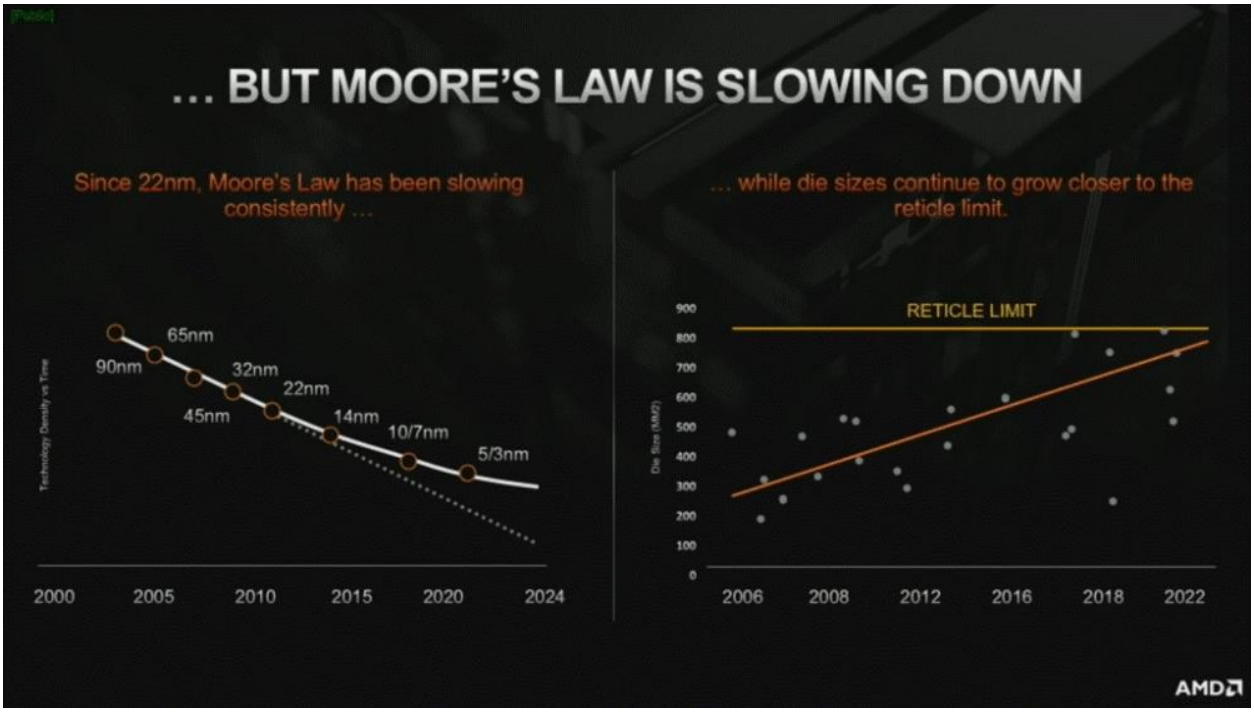
这些工作负载不能在传统CPU上运行。AI的模型越来越大，Facebook 的深度学习推荐系统模型有超过12万亿个参数。不断膨胀的模型尺寸致力于让你在应用上停留更长时间并点击更多广告。谷歌开发了自己的芯片，用于人工智能模型的训练和推理，称为 TPU。随着 VCU（一种新型处理器）的出现，他们扩大了在芯片工作的研究，如果专用于同一任务，它能够取代 1000 万个 CPU。

亚马逊有定制的网络芯片，也运行他们的管理程序和管理堆栈。他们拥有自己的芯片，专门用于AI 训练、AI 推理、存储控制和 CPU。当你查看 Marvell 和 Broadcom ASIC 服务的重点时，就会发现，硬件设计和架构的分散性只会增加。

就连英特尔，这家认为每个工作负载都应该在 CPU 上运行的公司，也认识到唯一的前进道路是异构设计。与针对每项任务的通用 CPU 硬件不同，该行业正在采用常见的工作负载并专门为它们构建芯片。这使架构师能够获得更高的每单位硅性能。

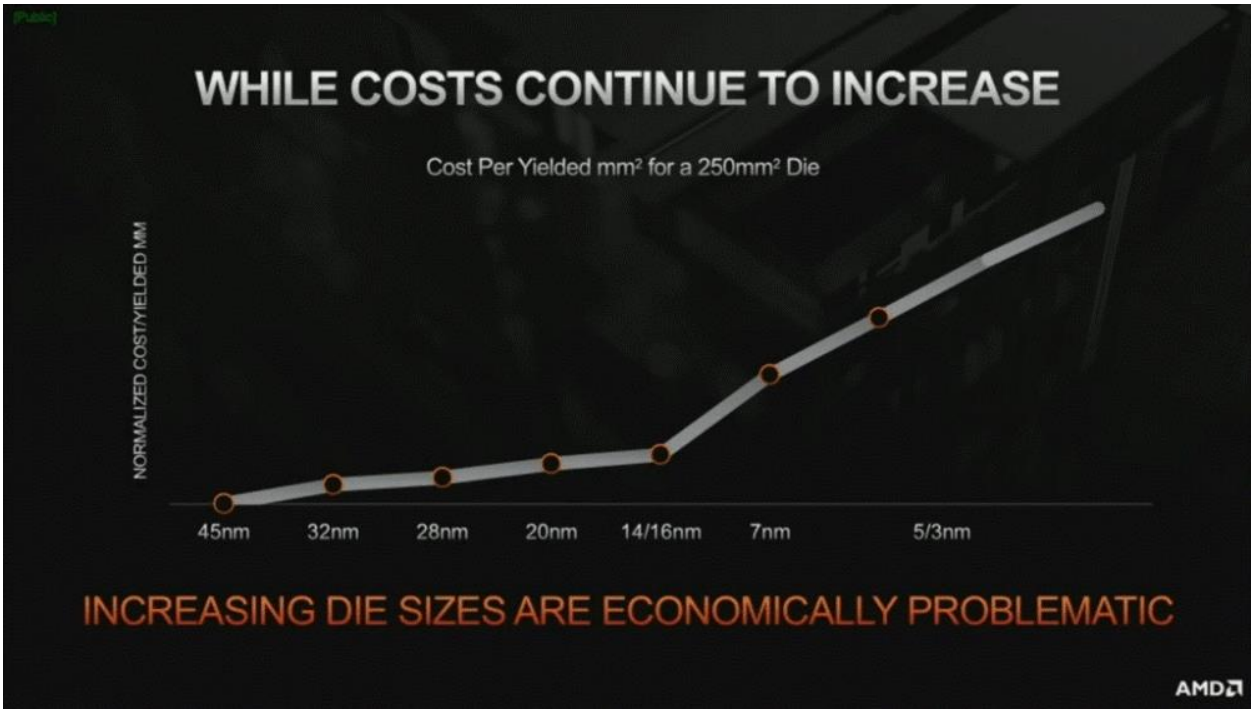
长话短说，除了 CPU 之外，专用集成电路的异构集成是至高无上的。不过，更多内存和更多异构计算并不是万能的。虽然通过增加内存和异构计算来增加芯片尺寸对于消除pad限制和提高能源效率非常有用，但这些都是要花钱的。很多钱。

更多的芯片面积意味着更多的引脚、更多的集成功能，但这也是成本失控的绝妙方法。并且芯片尺寸已经达到极限。例如，看看英伟达或英特尔的数据中心阵容。两者都接近“标线限制”超过 5 年。即使他们愿意，他们也无法继续制造更大的芯片。芯片收缩已经大幅放缓，助长了这个问题。



因此，收缩已经放缓，芯片尺寸无法增长得更大，设计也受到pad的限制，这些是唯一的问题吗?

不幸的是，不是。硅单元经济学也遇到了障碍。半导体行业及其下游企业单枪匹马地推动了整个经济的通缩环境，抵消了其他地方的通胀行动。没有它，80 年代以来的美国和欧洲将经历无休止的滞胀。不过，这种变革性的通货紧缩力量正在遇到障碍。半导体单位经济没有改善。事实上，将晶体管缩小到更小，它们甚至变得更糟。制造大芯片不仅昂贵，而且比之前的一代更昂贵。

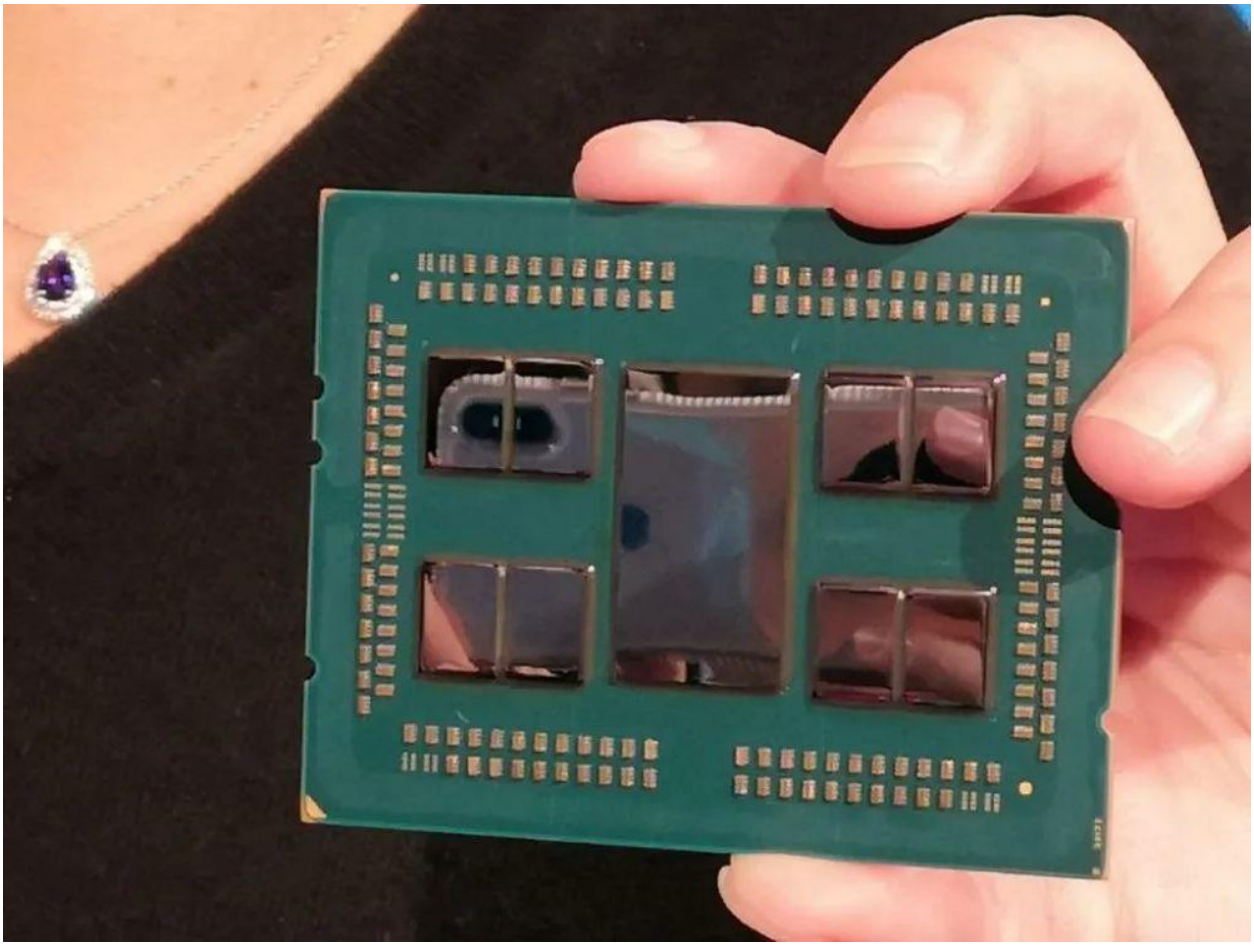


这张来自 AMD 的图表描绘了一幅非常病态的画面。虽然每个节点的转变并不相同，但很明显，在 7nm 和 5nm 处，该行业已经达到了拐点。每产出平方毫米的成本增加幅度不是很小，而是很大。尽管节点转换带来了类似的密度增益，或者可能由于 SRAM 缩放速度放缓而更糟，但成本的增加并没有跟上。与每晶体管成本相关的趋势逆转令业界震惊。这种逆转具有巨大的影响，甚至导致无知的银行家以此为理由，下调台积电的评级，称其估值过高。

Apple SOC Transistor Scaling							
SOC	Process Node	Transistor Count (Millions)	Die Size	Effective Density (MTs/mm^2)	Effective Density Scaling	Theoretical Density (WikiChip)	Effective vs Theoretical
A10	N12	3300	125.00	26.40		28.88	91.41%
A11	N10	4300	87.66	49.05	0.54	52.51	93.42%
A12	N7	6900	83.27	82.86	0.59	91.20	90.86%
A13	N7	8500	98.48	86.31	0.92	91.20	94.64%
A14	N5	11800	88.00	134.09	0.67	171.30	78.28%

摩根士丹利认为，由于摩尔定律正在放缓，晶体管成本缩放已经停止，台积电的定价压力将减弱。摩根士丹利通过一张可笑的图表来证明这一点，该图表显示 5nm 的晶体管成本低于 7nm，这与业内专家形成鲜明对比。随着 FinFET 节点的引入，每个晶体管的成本停滞不前，7nm 完全趋于稳定，而 5nm 则比以往任何时候都高。我们的读者可以算一算，N7 晶圆约为 9500 美元，N5 晶圆约为16000 美元。苹果的芯片尺寸几乎没有下降，但他们付了钱。

因此，每个晶体管的成本仍在增加，但对计算的需求比以往任何时候都增加。我们转向异构架构进行反击，但现在芯片设计过程要困难得多。该行业必须依靠许多拥有不同 IP 的团队按时交付并将其整合在一起。Synopsys 和 Cadence 等 EDA 供应商在协助方面做得非常出色，但这还不够。对于没有超过 1000 万个单元用例的任何人来说，一个可以购买特定应用 IP 或芯片并将其集成到硬件设计中的开放生态系统是必要的。即使对于这些公司，小芯片风格的系统架构也是答案。



随着我们继续收缩，预期收益率会缓慢下降。这是一个合乎逻辑的结论，因为每个连续的节点都会增加约 35% 的流程步骤。当前沿流程在数千个流程步骤中进行衡量时，错误开始迅速堆积。工业公司喜欢谈论“Six Sigma”，但这对半导体制造来说还不够。让我们假设一个有 2000 个工艺步骤的过程，每个步骤的每 cm?缺陷数为Six Sigma。那么D0（每 cm?缺陷率的行业术语）最终将是0.678。芯片越大，出现缺陷的可能性就越大。

如果这个假设的过程是构建英特尔的高端服务器 CPU，Ice Lake。这将导致每个晶片有 4 个良好的裸片和 76 个有缺陷的裸片。现在考虑这个分析是在 cm?水平上完成的，并且在前沿工艺节点上每 cm?有数十亿个晶体管。半导体行业比Six Sigma好得多。

除了尺寸的完美之外，还有什么解决方案?

Chiplets——小芯片！将大筹码分成许多小筹码。

AMD 是这方面最受欢迎的例子，但这是整个行业的趋势。AMD 可以设计 3 个芯片，一个CPU 核心小芯片和 2 个 IO 芯片。这 3 种设计覆盖了很大一部分市场。同时，英特尔设计了 2 个 Alder Lake 台式机芯片和 3 个 Ice Lake 服务器芯片，以服务于相同的潜在市场。因此，AMD 可以节省设计成本，制造比英特尔更多内核的 CPU，并节省收益成本。

要演示 yield 参数，请参见下表。AMD将 CPU 内核拆分为 8 个 CPU 内核小芯片。如果良率是 100%，英特尔将能够以比 AMD 更低的每个 CPU 内核的成本制造内核。但相反，英特尔必须在每个 CPU 内核上花

费更多，因为更大的芯片有更多的缺陷。下表有一些明显的警告，其中最大的假设是缺陷芯片的收获率为 0，并且英特尔和台积电具有相同的 D0。这些假设都不是真的，这个练习是为了演示目的。

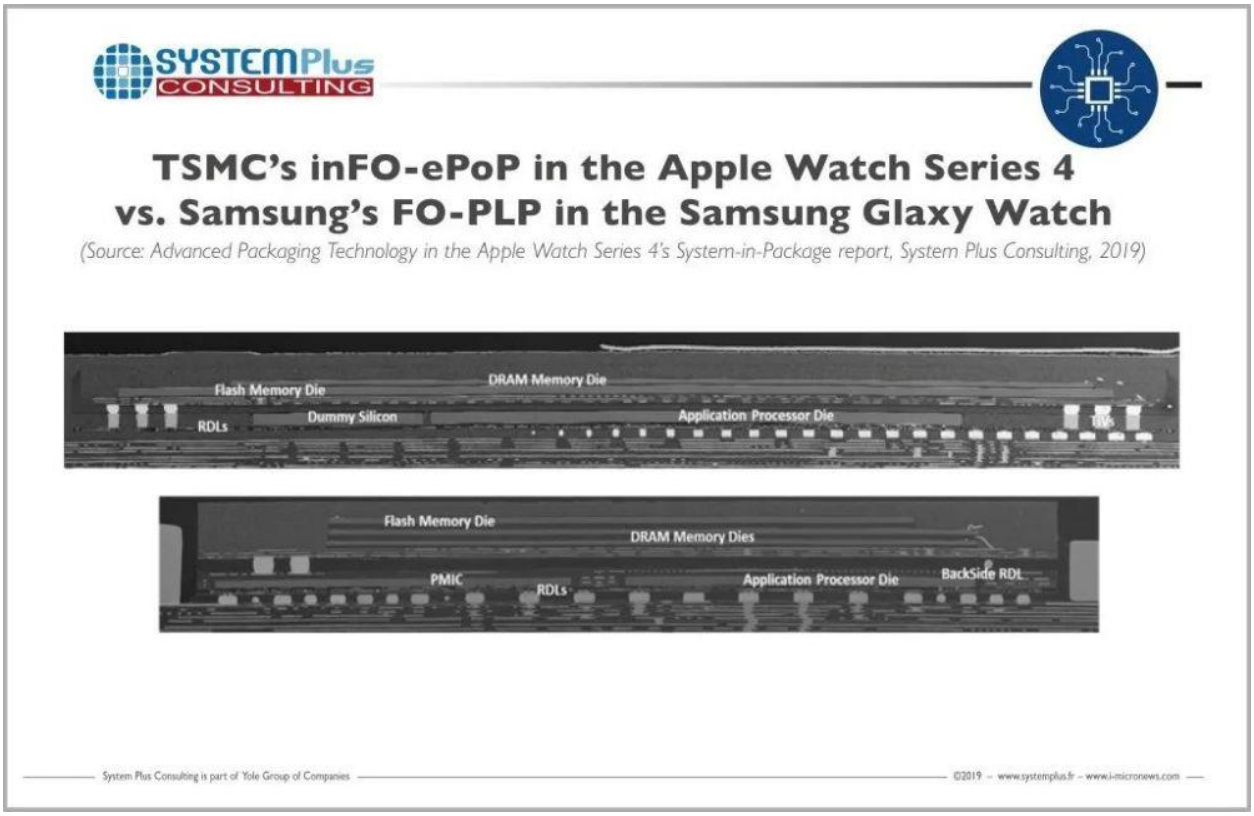
SemiAnalysis Server Chip Manufacturing Cost Analysis				
Company	Intel	AMD	Intel	AMD
Chip	Icelake SP XCC	Epyc Milan	Icelake SP XCC	Epyc Milan
Die Size (mm^2)	628	81	628	81
Dies Per Wafer (D0 = 0)	80	725	80	725
% Good Dies Per Wafer (D0 = 0/0.1)	100.0%	100.0%	55.0%	92.3%
Manufacturing Cost Per Wafer	$ 6,500	$ 9,500	$ 6,500	$ 9,500
Dies Per Server	1	8	1	8
Cores Per Server	40	64	40	64
Cost Per Server	$ 81.25	$ 104.83	$ 147.73	$ 113.60
I/O Die Cost Per Server	$ -	$ 45.00	$ -	$ 45.00
Total Chip Cost	81.25	149.83	147.73	158.60
Total Packaging Cost	4.5	10.7	4.5	10.7
Total Cost	85.75	160.53	152.23	169.30
Cost Per Core	2.14	2.51	3.81	2.65

小芯片（Chiplet）很棒，但它不是孤立的解决方案。我们仍然遇到许多相同的问题。每个晶体管的成本仍在上升，设计成本飙升，由于需要更多 IO 来与其他芯片接口，小芯片被pad限制。由于 IO 限制，部分芯片无法拆分，因此芯片尺寸仍在达到峰值。

解决办法是什么?

先进封装!

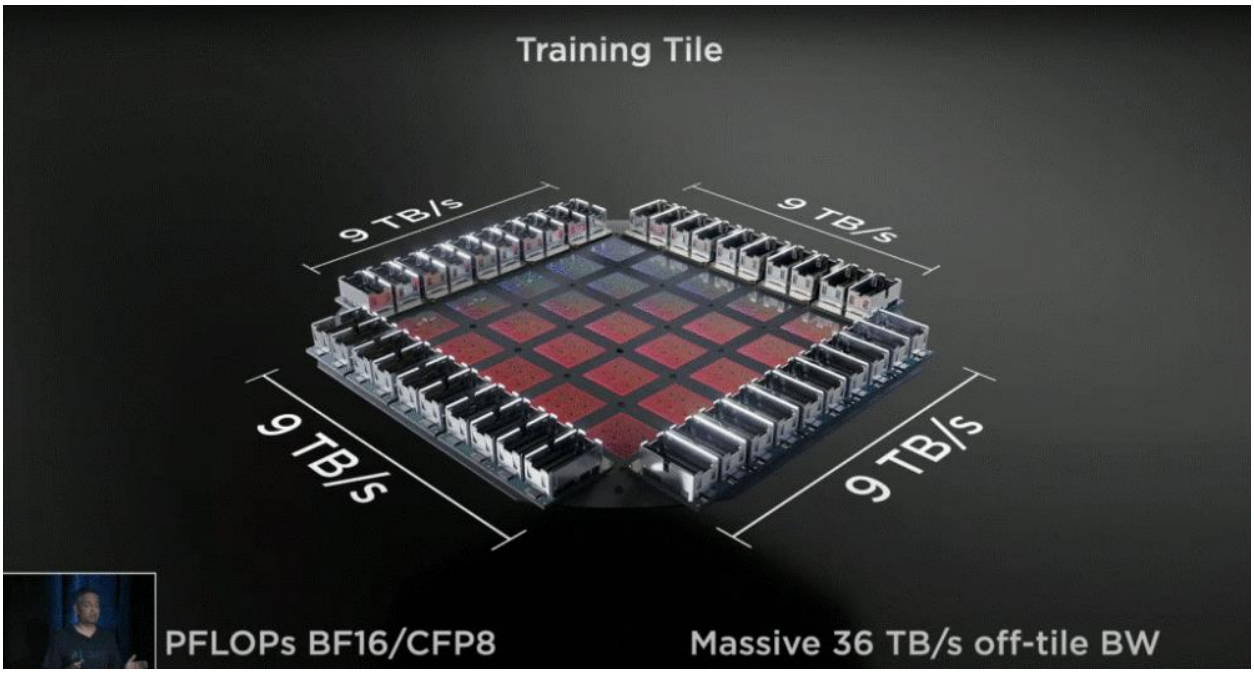
这就是我们要注意的地方，一些工具供应商将所有倒装芯片封装称为“先进封装”。SemiAnalysis 和大多数业内下游人士不会这么说。因此，我们将所有凸点尺寸小于 100 微米的封装称为“先进”。



最常见的先进封装类别称为扇出。有些人会争辩说它甚至不是先进的封装，但那些人大错特错。以Apple为例，他们将让台积电采用应用处理器芯片，并将其与 90 微米到 60 微米数量级的更密集凸块封装到重组或载体晶圆/面板上。与传统倒装芯片封装相比，凸点密度大约高出 8 倍。

这种重组或载体晶圆/面板然后进一步展开 IO，因此得名扇出。然后将扇出封装连接到主板。硅芯片的设计可以减少对pad受限的担忧，因为扇出处的pad较小。该封装还可以封装 DRAM 内存、NAND 存储和 PMIC 。集成扇出不仅有利于密度，而且它们还在封装上保留了大量的芯片间 IO。否则，该 IO 将不得不以更大的IO 间距尺寸通过主板进行接口。

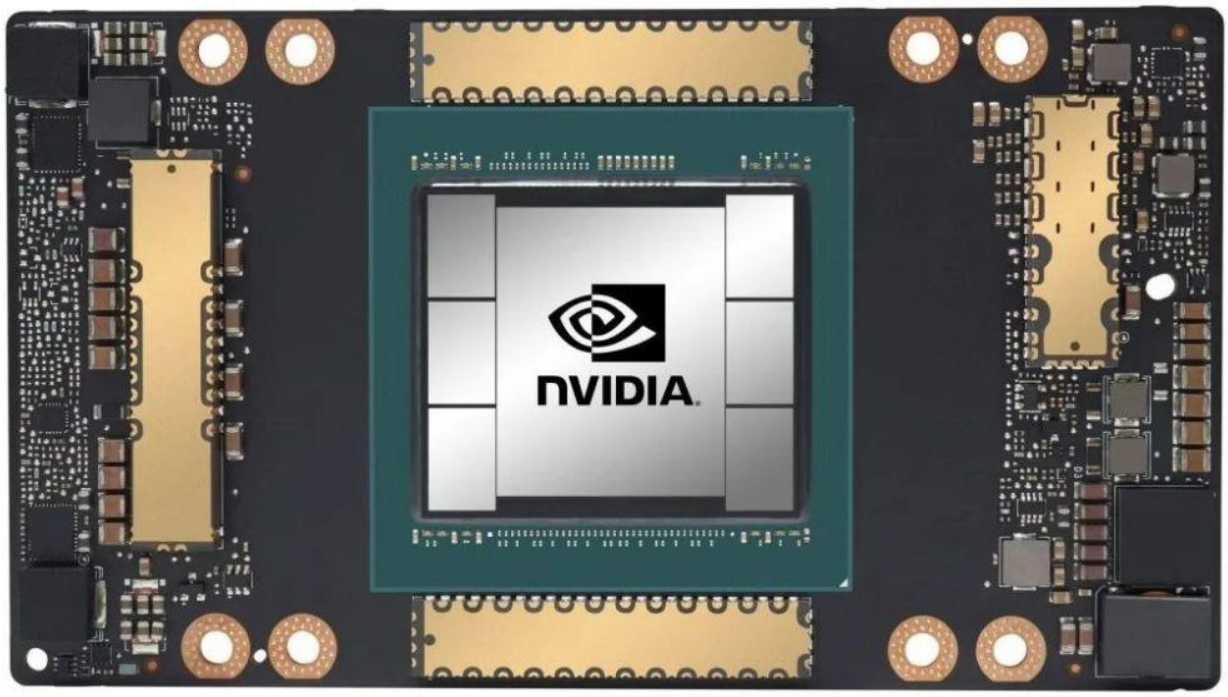
集成扇出对于高性能应用程序变得越来越普遍，不仅仅是移动应用程序。增长最快的用例是在十多年来设计一直受到限制的事物的网络方面。AMD 将在其服务器 CPU 和 GPU中非常积极地采用扇出。Tesla Dojo 1是集成扇出封装的另一个引人注目的例子，但在晶圆级。SemiAnalysis透露，特斯拉将在发布公告前使用这种包装类型。



在先进封装中，有 2.5D 和 3D 封装。2.5D 涉及封装在其他硅片上的硅片，但较低的硅片专用于布线，没有有源晶体管。这通常以55 微米到 50 微米的间距完成，因此凸点密度高出约 16 倍。最常见和最高容量的用例是具有 TSMC CoWoS（基板上晶圆上芯片）的 Nvidia 数据中心 GPU。台积电将有源芯片封装在只有互连和微凸点的晶圆上。然后使用传统方法将这叠芯片封装到基板上。

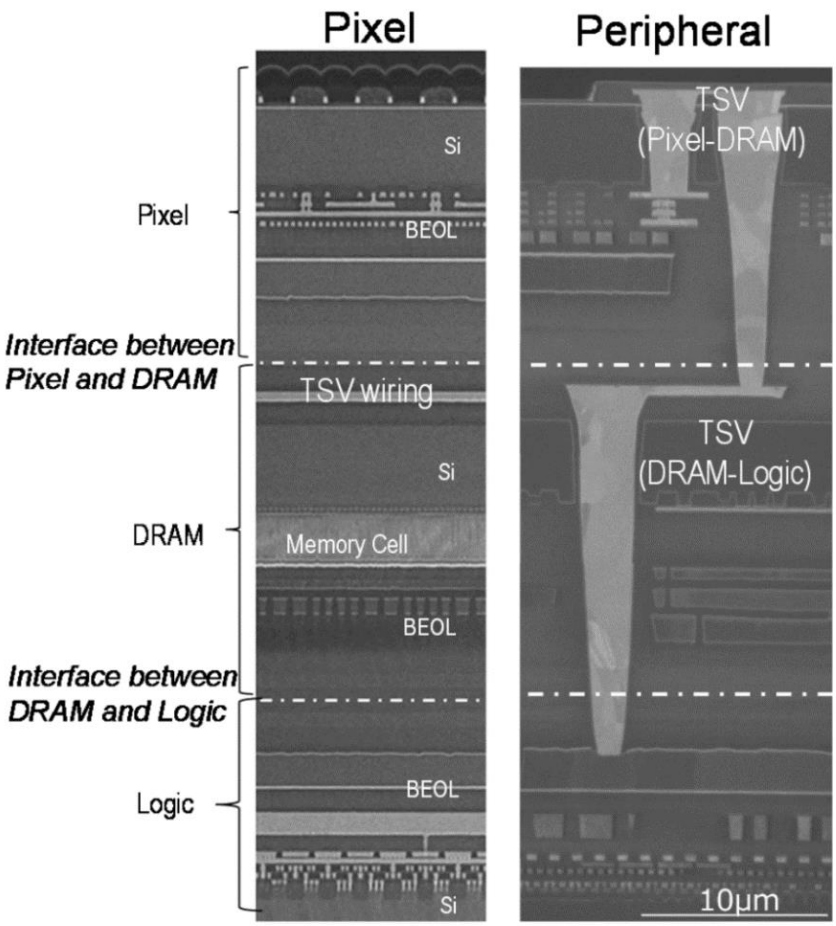
其他示例基本上包括每个带有 HBM 的处理器。HBM 是作为一种阶梯函数增加内存带宽的方法而建立的，这种方法高于传统形式的 DRAM。它通过使用更宽的内存总线来实现这一点。这些宽总线会产生与 IO 计数相关的问题，但 HBM 是从头开始设计的，以便在同一包内共存。这颠覆了 IO 问题，同时也允许更紧密的集成。

2.5D 的更多示例包括基于Intel EMIB 的产品、Xilinx FPGA、AMD 最新的数据中心 GPU 和Amazon Graviton 3。



3D 封装是将一个有源芯片封装在另一个有源芯片之上。这最初由英特尔以 55 微米间距的逻辑硅一起发货，但批量用例将在 36 微米及以下。台积电和 AMD 将推出 17 微米间距的 3D堆叠 V-cache。该技术从凸块过渡到硅通孔 (TSV)，并且具有更大的扩展空间。

其他应用，例如索尼制造的 CMOS 图像传感器，其间距已经达到 6.3 微米。为了保持比较，36 微米间距的凸点密度高出 31 倍，17 微米间距实施的铜TSV 的 IO 密度将提高 138 倍，而索尼的6.3微米间距的CMOS图像传感器的IO密度比标准翻转芯片高567倍。



这只是对主要封装类型的基本解释，但我们将深入研究本系列中的不同类型的封装。对未来的封装类型、工具以及工具供应商有很多不同的赌注。设备和 IP 方面比人们乍一看想象的要兴奋得多，但在我们深入研究之前，需要先解释基础知识。

对于即将到来的创新海洋，有很多可投资的想法和角度。摩尔定律的放缓正在推动根本性的变化。我们正处于先进封装推动的半导体设计复兴之中。

转半导体行业观察

----全文到此为止，如果喜欢，请点下“在看”或分享到朋友圈。

https://www.bilibili.com/read/cv15080380