3D IC系统架构概述

An Overview of 3D IC System Architecture

陈昊/CHEN Hao[1,2]，谢业磊/XIE Yelei[1,2]，
庞健/PANG Jian[1,2]，欧阳可青/OUYANG Keqing[1,2,3]
（1. 移动网络和移动多媒体技术国家重点实验室，中国 深圳 518055；
2. 深圳市中兴微电子技术有限公司，中国 深圳 518081；
3. 射频异质异构集成全国重点实验室，中国 深圳 518061）
(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. Sanechips Technology Co., Ltd, Shenzhen 518081, China;
3. State Key Laboratory of Radio Frequency Heterogeneous Integration, Shenzhen 518061, China)

DOI：10.12142/ZTETJ.2024S1011
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20240909.1731.002.html
网络出版日期：2024-09-09
收稿日期：2023-11-25

摘要：随着芯片制造工艺接近物理极限，使用多Die堆叠的三维集成电路（3D IC）已经成为延续摩尔定律的最佳途径之一。利用3D IC将芯片垂直堆叠集成，可以极大程度降低互联长度，提升互联带宽。详细介绍了一些常见的3D IC系统架构方案，说明了使用不同3D架构对于整体芯片系统在性能、功耗等方面的优势，也列举了在物理实现、封装测试、工艺能力等方面的挑战。最后综述了一些业内使用3D IC的典型产品，并介绍了这些产品的系统架构、典型参数、适用领域，以及使用3D IC后给产品带来的竞争力提升情况。针对业界现状，认为应该把握机遇，不惧挑战，实现弯道超车。

关键词：三维集成电路；三维堆叠芯片；三维片上系统；存储堆叠逻辑；逻辑堆叠逻辑

Abstract: As the chip manufacturing process approaches its physical limits, multi-die stacking 3D integrated circuit (IC) technology has emerged as a promising approach to sustain Moore's law. Integrating chips vertically with 3D IC can significantly reduce interconnection length and improve interconnection bandwidth. This paper provides a detailed overview of common 3D IC system architecture solutions and discusses the advantages of using different 3D architectures in terms of performance, power, and area. It also outlines the challenges related to physical implementation, packaging, testing, and process capability. This paper summarizes some typical commercial products that utilize 3D IC technology and introduces their system architecture, typical parameters, applicable fields, and competitiveness improvement. Considering the current industry landscape, the paper suggests that China should comprehensively assess the current situation, capitalize on opportunities, confront challenges without fear, and strive for leadership in this domain.

Keywords: 3D IC; 3D stack integrated circuit; 3D system on chip; memory on logic; logic on logic

引用格式：陈昊, 谢业磊, 庞健, 等. 3D IC系统架构概述 [J]. 中兴通讯技术, 2024, 30(S1): 76-83. DOI: 10.12142/ZTETJ.2024S1011
Citation: CHEN H, XIE Y L, PANG J, et al. An overview of 3D IC system architecture [J]. ZTE technology journal, 2024, 30(S1): 76-83. DOI: 10.12142/ZTETJ.2024S1011

1985年，著名物理学家诺贝尔奖获得者理查德·费曼在日本发表“未来的计算机器”演讲时就预言，未来芯片发展的一大方向就是通过扩展平面芯片到三维层面来提升系统性能。2006年，三星电子CEO黄昌圭博士在国际电子器件年会（IDEM）上发表主题演讲“硅半导体业界新范式”时提出，电子技术新时代即将来临，可以将内存、逻辑、传感器、处理器等不同器件集合在一起的三维集成技术是电子技术新时代的核心。[1]

1 3D IC分类概览

2009年的国际半导体技术路线图（ITRS）从互联的不同层面对三维集成电路进行了规范。在板级的互联称三维封装，使用传统封装工艺，互联尺寸大，密度低，工艺简单，如封装上封装（PoP）、封装内封装（PiP）、多芯片模组（MCM）等。在封装级的互联称三维晶圆级封装，互联使用Bump-Pad和重布线层（RDL）工艺，如TSMC的CoWoS、InFO和Intel的EMIB等，又称封装内系统（SiP）。在片内的互联分为3种：三维芯片堆叠（3D SIC）为功能模块级堆叠，互联尺寸在10 μm左右，互联密度达一万个/mm^2；三维片上系统（3D SOC）为电路单元级堆叠，互联尺寸在1 μm左右，互联密度达一百万个/mm^2；三维集成电路（3D IC）为晶体管级三维堆叠，互联尺寸达100 nm级，互联密度高

达一亿个/mm^2，又称单片三维集成电路（Monolithic 3D IC）或顺序3D IC。从板级到封装级，再到片内高密度互联，3D IC工艺在不断成熟，三维互联也在微缩化。[2–3]

集成电路产业的发展与工艺技术演进密不可分。3D早期集中在板级和封装级互联，如PoP、PiP、SiP、2.5D等，使用打线或焊球进行片间互联，有源芯片间互联密度较低，互联速率、互联功耗、信号完整性等较差，需要芯片到芯片间互联模块（D2D IP）进行协议转换和驱动增强保证信号质量，而这会导致面积、功耗、延迟劣化。这些早期技术主要目的是突破工艺制造光罩尺寸限制，提升可拓展性，因此只能称为三维封装，而不是3D IC。现在3D IC已发展到3D SIC和3D SOC阶段，属芯片级或晶圆级互联，互联密度高，传输延迟低，可同步直连无需协议转换，晶体管间垂直互联对延迟、功耗开销小，与2D片内直联相容度高，弥合了晶圆制造和封装工艺间的鸿沟，让芯片互联走向新纪元。

2 3D IC带来价值

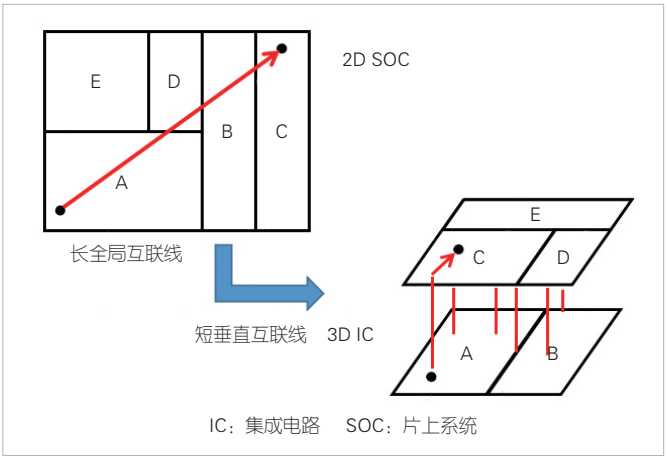
三维堆叠集成带来价值总体体现在4个层面：互联、内存、小型化和异构集成。

首先是互联层面。几十年来半导体先进工艺的发展带来了器件的微缩，可以在单位面积内集成更多器件，降低了器件延迟与功耗，但同时也带来了单位面积内绕线密度增加与栅极厚度减小的问题。芯片电路延迟由器件单元延迟和器件间互联绕线延迟两部分组成。绕线密度增加导致线宽线距降低，增加了互联RC延迟与互联功耗，器件互联线延迟与功耗占总延迟与总功耗的比例越来越高。另外，芯片功能复杂度提升增大了单芯片尺寸，故全局互联信号和时钟无法享受工艺微缩红利带来的延迟降低。此外，随着单位面积内器件集成度提高，更多互联绕线资源需求引起后道金属绕线密度提升，但受到制造工艺、应力、电源信号完整性等限制，后道金属密度和层数不能无限提高，绕线资源紧张导致拥塞与互联带宽的压缩。如图1所示，3D IC可通过模块切分并垂直堆叠，降低模块间全局互联长度，减少互联线延迟与互联信号和时钟网络上引入的功耗与面积，同时垂直相比水平互联拓宽了绕线资源，提升了模块间可容许的互联带宽。

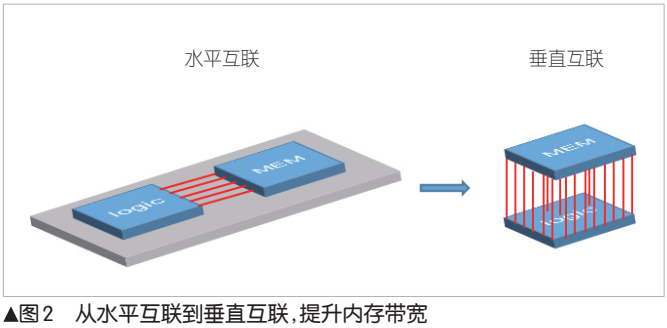
其次是内存层面。工艺技术、电路设计、系统架构的不断优化驱动处理器性能和主存容量成指数级提升。主存带宽每2年提升1.4～1.6倍，无法跟上处理器性能每2年提升约3倍的步伐，成为限制整体系统性能的短板，这就是内存墙[4]。在计算机体系架构中，使用静态随机存储器（SRAM）替代动态随机存储器（DRAM），采用高速缓存来构建多级内存层级，可以尽可能缓解内存墙影响。在先进工艺世代，SRAM微缩难度相比逻辑单元更大，其性能、功耗、面积等指标也滞后于算术逻辑单元（ALU）。另外，多级内存架构在物理上越靠近ALU的缓存，层级越低、容量越小、对带宽需求越高，多级缓存失效带来的延时惩罚就越严重。通过3D切分，将缓存、主存从原有的2D芯片互联中独立出来，与逻辑运算芯片垂直堆叠，物理上提升逻辑到内存互联带宽，打破2D场景下缓存容量限制，缓解内存墙问题。

再次是小型化。芯片制造工艺跟不上用户需求增长的步伐，为了满足更复杂功能实现与更高性能要求，在高性能计算领域使用多核多线程并行处理方式来构建超级计算机，在物联网领域集成计算和智能处理核心来构建边缘AI处理器。核数增多与功能集成导致面积扩增，会触及光罩极限，带来良率与成本劣化。2.5D水平互联可以解决单芯片切分问题，但无法满足带宽和封装尺寸要求。通过扩展Z维度，实现芯片垂直互联，可以提高集成密度，降低芯片外形尺寸，如图2所示。3D IC互联工艺还会对硅晶圆背面进行减薄，总体厚度相比于封装级三维堆叠显著降低。另外，利用堆叠前后的多步测试手段和使用芯片堆叠晶圆（D2W）或芯片堆叠芯片（D2D）工艺，可在提高单晶片良率的同时提升封装后芯片的良率和产率。

最后是异构集成。摩尔定律预测，每16～24个月，集



▲图1　从2D到3D，降低互联长度



▲图2　从水平互联到垂直互联，提升内存带宽

成电路上晶体管速度、集成度可以提升1倍。但随着器件尺寸不断接近物理极限，工艺制造时非理想效应凸显，这一步伐逐渐趋缓。不仅如此，数字电路、模拟电路放缓速率不同，模拟与内存电路晶体管速度在栅长22 nm时已经趋于饱和。数字电路尺寸微缩主要体现在垂直堆叠栅极构建多Fin结构。目前，N3节点鳍式场效应晶体管（FinFET）栅长约16 nm，继续降低通道长度会导致漏电流过大难以关断。综上可知，如果仍将数字、模拟、内存等使用同一种工艺制造在同一颗2D芯片上，将不利于整个系统的成本、面积和性能的优化[5]。使用不同工艺制造不同类型电路利用3D IC进行垂直互联，可以在满足性能的前提下更好地利用工艺甜点（Sweet Node），实现性能、成本和良率的最优化。不仅如此，在超越摩尔定律（Beyond CMOS）的范畴，一些新架构、新器件、新材料也可以利用3D IC进行异构集成。其中，三维图像传感器（3D CIS）就是典型的异构集成案例。此外，该技术还可以拓展到存内计算、射频、光互联等领域。

根据具体应用，3D IC价值主要面向两大类场景：一类是高端、行业旗舰级应用，一类是中低端、用户消费级应用。在高端应用方面，3D IC可带来高性能与高功能集成度。3D IC可以缩短连线长度，提升通信带宽，异构集成多种小芯片来构建封装内系统，可用于人工智能模型、大型科学计算、生命科学等领域。在中低端应用方面，3D IC可实现设备小型化，降低互联功耗，能够更好地利用工艺甜点降低总成本等。因此，3D IC在桌面级与手持应用、可穿戴式设备和万物互联等领域也有较高实用价值。

3 3D IC系统架构

3D IC系统架构，可按顶层（Top die）和底层（Bottom die）堆叠功能类型进行区分，主要有内存堆叠内存（MOM）、内存堆叠逻辑（MOL）、逻辑堆叠逻辑（LOL）。

MOM架构形式可将多颗DRAM堆叠成为内存芯粒（Memory Chiplet）的形式，提升系统内存容量和内存带宽，如高带宽内存（HBM）和混合内存立方体（HMC）；也可SRAM堆叠成为缓存堆叠缓存（Cache on Cache）并集成到SOC内，做成可拓展系统级缓存（SLC），提升缓存容量，如AMD的3D V-cache；还可设计新型3D SRAM，通过拆分SRAM单元阵列优化访问延迟，如宾夕法尼亚州立大学（PSU）的谢源教授研究团队用字线和位线拆分构建3D SRAM[6]。

MOL架构形式可将SRAM高速缓存（L1/L2 cache）和逻辑单元堆叠，做成逻辑上缓存形式，如将中央处理器（CPU）中的L1/L2缓存拆分和ALU堆叠。比利时微电子中心（IMEC）的DRAGOMIR等使用openSPARC-T1核进行3D缓存堆叠，相比2D可以提升11.4%的核频率[7]。佐治亚理工大学（GT）和ARM的ZHU等基于CMN600和N1核进行了SLC和CPU堆叠，可以提升17%的单周期指令数（IPC）[8]；也可将DRAM主存（Main Memory）和SOC堆叠，用来替代L2/L3缓存，做成近存计算（PNM）架构，例如：圣母大学与惠普实验室的CHEN等建立CACTI-3DD模型分析了3D DRAM的优势[9]，阿里达摩院的NIU等用逻辑和DRAM三维堆叠实现了一款高带宽高能效AI处理器[10]。

LOL架构形式有以下几种：1）SOC on IP堆叠，将SOC中的外围接口IP放于底层，接口IP多为模拟模块，对计算性能要求不高，在低工艺节点实现，采用三维堆叠异质集成，降低尺寸与综合成本。2）Core on NOC堆叠（NOC指片上路由网络），把计算核心与NOC进行3D拆分堆叠，降低计算核心间网络路由物理距离。该方法对逻辑切分挑战较大，设计较复杂。3）Core on Core堆叠，将不同计算或处理核心相互堆叠。但Core一般是发热大户，可能导致散热方面问题。IMEC的CHEN等分析了3D CPU堆叠的散热相比2D结温会提高140%[11]。4）三维片上网络（3D NOC）的形式，这是一种新型片上互联方式。相比于传统的2D NOC，3D NOC可以显著降低平均网络访问延迟和时钟周期数，常见的有基于Tree结构和基于Mesh结构两种。ARM的FEERO和PANDE分析了3D NOC相比于2D NOC在带宽、功耗、吞吐量方面都有较大优势，结合NOC交换延迟的降低可以更好地提升性能[12]。5）处理器内部流水线级3D堆叠，3D NOC和流水线堆叠需要芯片互联架构层面做出适配性改动，对硅通孔（TSV）和芯片键合互联工艺微缩也提出较大挑战。

3D IC系统架构选择主要受需求侧和实现侧的影响。需求侧是动力和源泉，影响架构的可能方向；实现侧是限制与约束，影响架构的可行方向。两者的集中体现是3D IC系统设计。

4 3D IC系统设计

自2000年以来，美国国防部高级研究计划局先进微电子研究委员会（DARPA）就开始资助3D IC和异构集成相关项目。2000年，麻省理工学院（MIT）的RAIF教授等利用Rent定律理论论证使用3D IC可以显著降低互联延迟和芯片面积，但实现的最关键因素是垂直层间过孔的高密度互联，多层堆叠的性能瓶颈主要在于过孔的禁空区（KOZ）大小[13]。这为3D互联工艺微缩指明了方向。

3D IC研究早期由于工艺技术难以跟上架构发展，研究集中在架构的可行性论证和收益方面。2004年，Intel的BLACK等对深度流水线处理器——iA32进行了3D切分架构的尝试，通过三维布局规划，可以节约处理器流水线中的RC延迟，如时钟时延、浮点处理时延、寄存器堆访问时延等，可以降低约25%的流水线级数。综合来说，处理器级的3D实现可以提升15%性能的同时优化约15%的功耗[14]。

2006年，佐治亚理工学院的PUTTASWAMY等使用3D方式实现了256输入的物理寄存器堆，两层堆叠可以在优化58.5%的能耗同时达成24.1%的延迟优化，四层堆叠可以在优化58.2%的能耗同时达成36%的延迟优化。此外，他们还提出了寄存器堆的3种3D切分方案：寄存器切分、比特位切分、访问端口切分。3D切分堆叠带来的优势不局限于处理器微架构，对于不同的处理器配置，处理器中的关键延迟部件不同。不同的微架构需要调整3D堆叠策略，以达成最佳的时延降低[15]。寄存器堆是线延迟主导，因此借助3D实现来降低访问延迟和简化控制数据流来构建高性能微处理器将成为可能。

2009年，MIT开发的3D工艺逐渐可以支持处理器微架构在硅上进行原型实现。宾夕法尼亚州立大学谢源课题组使用MIT Lincoln Labs的180 nm 3D FDSOI工艺制造了两种基本计算单元：3D加法器和3D乘法器。相比于2D实现，Kogge-Stone型3D加法器通过将12 bit计算带宽提升为72 bit，可以降低10.6%～34.3%的延迟和11.0%～46.1%的能耗，32×32的Wallace-Tree型3D乘法器可以降低14.4%延迟和6.8%能耗[16]。2010年，他们发布了一个用于H.264编解码的3D流处理器，该处理器可以提供多个内存通道，内存控制器和并行访问策略经过了重新设计，来充分利用3D DRAM的内存带宽提升[17]。这表明，在设计3D计算机系统架构时，应考虑如何将系统架构设计和3D堆叠集成带来的好处结合起来，例如：可以重新设计缓存层级和片上互联方案来充分利用3D带来的优势。

2009年，北卡罗莱纳州立大学（NSCU）FRANZON研究组报告了一种用于合成孔径雷达（SAR）的快速傅里叶变换（FFT）处理器。SAR FFT处理器架构有大量全共享全互联内存和计算单元，通过将一个大内存拆分成很多小的内存并实现内存逻辑垂直堆叠，可以提高计算单元访问内存的并行度，并降低FFT处理器60.3%的内存能耗。相比于2D实现，SAR FFT处理器可降低53%平均线长，提升24.6%计算频率，使内存带宽提升到原来的8.55倍，总硅面积降低25.3%[18]。

后来，人们看到了3D工艺制造上的局限性，开始在系统层面探索多核片上系统和3D IC之间的相容性。2012年，佐治亚理工学院SUNG课题组设计了一种基于3D堆叠内存的多核并行处理器架构——3D MAPS。这种架构的顶层是64个通用计算核心，底层是对应每个核心的4kB SRAM共256 kB。晶圆制造使用了格芯（GF）的130 nm技术，3D封装使用了安靠（Amkor）的Tezzaron TSV工艺。在3D MAPS单核设计方面，设计者详细考虑了流水线深度、寄存器堆容量、发射带宽、计算单元、指令集等的设计，并充分优化了内核微架构来满足3D IC技术提供的大内存带宽的优势。在多核间互联和核与内存互联方面，他们对核间路由的同步与协调和内存分块做了精细调整，让访存容量和核心处理能力相匹配，并可以让单个核心分别控制4个内存分区。除此之外，3D MAPS还引入了可测试性设计，并定制扫描链和测试控制器进行功能和性能基准测试。基准测试内容包括最大内存带宽、周期指令数和功耗等。基于Median Filter标准，3D MAPS可利用的内存带宽为63.8 GB/s，可达理论最大值的89.99%，这证明了微架构调整和3D IC结构之间的适配度[19-20]。

2021年，法国原子能委员会电子与信息技术实验室（CEA-Leti）与格勒诺布尔大学共同发布了一款基于有源中介层的6片芯粒组成的96核3D堆叠处理器——IntAct。有源中介层不同于2.5D情况下的中介层，除提供硅桥互联和电容电感等被动元件外，IntAct的有源中介层中还集成了电源管理模块SCVR、分布式3D插入式路由、传感器、串行解串器等模拟IP，还有用于可测试的设计与端口等。IntAct使用意法半导体（ST）的FDSOI 28 nm工艺堆叠在65 nm工艺的有源中介层上，可以达成分布式互联、3D异构集成、电源管理动态调压调频、芯粒重用、成本优化等诸多目标[21]。

2022年，苏黎世联邦理工（ETH）LUCA课题组提出了一种开源众核SOC架构——MemPool，其中高达256可编程核心集群可以共享大容量L1缓存，可以满足低延迟、低功耗和高吞吐量要求。基于MemPool，佐治亚理工学院和IMEC的研究人员结合3D IC技术，提出增强型MemPool-3D，有效将MemPool的体系架构设计和3D IC的优势结合起来，解决了常规MemPool在2D IC场景下的路由拥塞和全局传播延迟等问题[22]。3D IC系统设计汇总如表1所示。

可以看到，3D IC系统设计主要由架构设计、功能切分、工艺实现3个方面来决定。

架构设计方面，传统2D IC情况下仅有单个Die，通过2D的前道工艺（FEOL）和后道工艺（BEOL）来实现。2D的系统架构参数，如核数、主频、流水级别、缓存级别、缓存容量、访存带宽等结合2D IC的物理实现工艺做了最优

▼表1 3D IC系统设计汇总

	DAC–2009（NCSU）	3D IC–2009（PSU）	3D IC–2010（PSU）	ISSCC–2012（GT）	JSSC–2021（CEA–Leti）	DATE–2022（ETH,IMEC）
应用	3D SAR FFT processor	3D微架构–加法器，乘法器	3D DRAM H.264流处理器	3D MAPS–并行计算	IntAct–3D SOC	3D MemPool–多核并行计算
芯片工艺	MIT LL 180 nm 3DFDSOI	MIT LL 180 nm 3DFDSOI	GF 130 nm Tezzaron 3D	GF 130 nm Tezzaron 3D	ST 28 nm 65 nm Active interposer	28 nm(prototype)
3D拆分架构	3 Layers–MOL	3 Layers–LOL	5 Layers–LOL/MOM/MOL	3 Layers–MOM/MOL	2 Layers–LOL	2 Layers–MOL
3D pitch	3.9 μm	2.65 μm	4 μm	5 μm/2.5 μm	20 μm/40 μm	1 μm
3D堆叠架构	F2F/F2B	F2F/F2B	F2F/F2B	F2F/F2B	F2F	F2F
优势	53%线长降低 24.6%频率提升 8.55倍内存带宽提升 25.3% Si面积降低	10.6%～34.3%延迟降低 11.0%～46.1%能耗降低	并行存储访问 支持8个独立内存通道	64个通用计算核 比2D成本降低3% 63.8 GB/s内存带宽	集成片上电源调制器 可拓展式3D+2.5D集成 算力220 Gops超过7 nm同类芯片	256核共享L1缓存 9.1%性能提升 15%能耗降低

DAC：国际设计自动化会议
DATE：欧洲设计自动化与测试学术会议
DRAM：动态随机存储器
F2B：面对背堆叠
F2F：面对面堆叠
FDSOI：全耗尽绝缘体上硅工艺
FFT：快速傅里叶变换
GF：格芯
IC：集成电路
ISSCC：国际固态电路会议
JSSC：IEEE固态电路杂志
LOL：逻辑堆叠逻辑
MOL：内存堆叠逻辑
MOM：内存堆叠内存
SAR：合成孔径雷达
SOC：片上系统

化。过渡到3D情形，晶体管可以实现垂直堆叠并互联，增加了物理实现自由度。如果仍选择之前2D IC对应的系统参数与指标，则难以达成3D IC情况下的最优结果。因此，针对3D IC，要面向3D IC的工艺特点和价值取向，进行3D IC系统架构重构，充分利用3D IC优势。架构重构取决于具体产品的架构形式、做3D IC的目的、解决的关键问题，以及要突破的产品瓶颈。如果关注SOC多核性能，如延迟、吞吐量、核负载均匀性等，就需要考虑多核访存容量、带宽、路由性能等瓶颈，拓展缓存容量和带宽，进行NOC架构级修改或NOC节点配置修改。如果要提升具体IP单核性能，则要考虑IP核内限制性能提升的因素，如寄存器堆配置、流水线深度、L1/L2带宽、物理可实现性等。这些都可以利用3D IC来改善，具体需要结合架构与微架构修改和底层代码实现。如果要进行内存容量或带宽拓展，或将片外缓存拿到片内，或降低内存层次，实现近存架构，那么可采取MOM或MOL，并结合架构和代码具体情况将内存从2D SOC中切分出来，或将扩展内存挂载到原有SOC上。总体来说，考虑延迟敏感性，对互联带宽、内存容量有要求的场景，可以通过功能切分，在顶层和底层上实现垂直互联[23]。

功能切分与工艺可实现性需要同步考虑。哪些模块需要放到哪颗Die上？具体放到Die上的什么位置？解决这些问题的方法称为3D IC的功能切分。能够实现功能切分的层级和3D IC工艺微缩程度息息相关。其中，最重要的工艺指标是Die间互联（主要是TSV和混合键合）尺寸、间距、电气特性，这决定了Die间互联的带宽、速率、功耗。尺寸间距越小，电气性能越好，3D互联带宽、速率越高，功耗就越低。相对来说，只要打破带宽和速率的限制，3D IC架构实现就更灵活，可以满足的架构方式就更多样化。从2D SOC到3D SOC，不同的3D IC方案按功能切分层级进行分类。切分和堆叠的最小单元在芯片设计中的层次高低被定义为3D IC的切分粒度。基于SoC架构级别的切分是粗粒度的切分，互联密度低，工艺要求低，对应MOL主存堆叠逻辑和LOL中的Core on Core或SOC on IP形式。细一级的是基于IP功能模块的切分，对应MOM、MOL缓存堆叠逻辑和LOL中Core on NOC或3D NOC形式；再细一级的是功能模块内部逻辑级的切分，对应LOL中流水线级别拆分。最细粒度的拆分是门级的切分，通过LOL在总线位宽上拆分或网表内部基本单元级的拆分，需要依赖于超高密度互联的Monolithic 3D IC工艺[24]。可以看到，切分每细化一个级别，就会对3D工艺微缩提出更高要求，带宽可以做得更高，因此可以容许架构设计做出更大的改动。

我们看到，从2D到2.5D、3D SIC，再到3D SOC、Monolithic 3D IC，3D工艺的演进是架构革新的助推器，现阶段业内3D工艺普遍可以支持3D架构的第一、第二级别——架构级和功能级。也就是说，是3D互联工艺限制了Die间互联的性能和带宽，同时也限制了具体3D IC实现时的系统架构与切分方案的自由度。反之，更精细化的3D架构设计与切分方案也对3D工艺演进提出了更多的要求。随着3D

互联工艺的演进，3D架构设计与切分方案也可以更加多样化。

除此之外，3D IC系统设计也需要综合考量价值与开销，使整个芯片系统性能、功耗、面积、成本（PPAC）达到最优。

3D IC中最显著的问题是散热问题。芯片的三维堆叠带来功耗密度的多层叠加，但3D IC相比2D却没有引入额外的散热通路，与2D相比，3D的结温会更高，需要更全面的产热、散热系统解决方案。热源端需要在系统设计和逻辑实现时考虑热因素，采用低功耗的设计方法（如多电压域、时钟电源关断等）降低系统积热，需要在功能切分和物理实现时使用热感知的布局技术，并在设计签核时考虑热带来的时序和电源恶化的影响。散热端需要考虑功能切分与物理设计时尽可能使高发热和热敏感模块离散热器更近，添加额外的散热通路如Dummy die和Dummy TSV等辅助散热，还可以结合一些先进散热技术如微通道液冷、金刚石散热等作为芯片级冷却解决方案。

除了散热之外，3D IC的工艺成熟度也备受关注。其中，芯片键合与硅通孔是3D IC最重要的两项技术。二者作为片间垂直互连的实际存在形式，从特征尺寸、电气特性、工艺良率3个方面决定了3D IC性能与质量。Bonding和TSV特征尺寸越小，互联密度越高，3D IC可以提供的片间带宽就越大。选取电气性能优异的互联材料，可以降低3D IC互联延迟与功耗。持续优化工艺细节和工艺流程，如改善晶圆键合时的平整度和对准问题，控制键合焊盘和工艺温度，改善铜凸问题等，对于提高3D IC整体良率、降低3D IC制造成本有重要意义。

此外，还需要考虑3D IC的成本问题。3D IC涉及多颗芯片堆叠，在成本核算方面更加复杂，包括设计成本、制造成本、封装成本、测试成本，以及不同供应商的物流成本。3D IC设计复杂度比2D更高，需要考虑的设计维度更多，但可以通过将高性能模块做成多种或多代产品可复用的3D Chiplet形式，以降低设计和制造成本。在制造层面，3D可以堆叠集成多种工艺，降低芯片面积，提升良率，但同时也需要加工更多光罩（Mask layer），提高一次性工程成本（NRE）。3D封装工艺对产线控制要求更高，良率曲线还在爬升，当前阶段比2D封装成本有显著提升。3D需要嵌入更多的测试与修复设计，尽可能保证筛片良率。此外，采用新的测试针卡、机台、测试用例等也会提高3D测试成本。

除前述之外，多物理场耦合、物理设计实现、多工艺签核、可测试性设计与测试流程等问题对3D IC整体可实现性、性能、成本折中也有重要影响。3D IC芯片系统设计应综合考虑功能、性能、可靠性、成本等多种因素，确定最终系统的参数选项，根据具体设计情况，迭代达成最佳实现目标。

5 3D IC产品概览

最近几年，产业界已推出较为成熟的3D IC产品。最早应用3D IC技术的产品是3D CIS和MOM的HBM。但由于3D CIS架构相对简单、应用领域单一，同理HBM不能脱离逻辑芯片独立功能存在。

通用3D IC产品在2019年产生。Intel发布了业内第一款商用3D IC的低功耗小尺寸SOC芯片——Lakefield处理器。这款芯片使用逻辑堆叠逻辑SOC on IP异质堆叠形式，顶层使用Intel 10 nm工艺的计算芯粒（Chiplet），包含CPU、GPU、图像处理器（IPU）和一些显示引擎。底层使用Intel 22 nm低功耗工艺的IO芯粒，包含通用串行总线（USB）、串行外设接口（SPI）等一些常用接口IP。顶层和底层之间使用Intel 3D Foveros Microbump互联工艺。3D同步互联速率可达500 MHz，3D互联功耗为0.2 pj/bit。Lakefield凭借低待机功耗和小尺寸，成为当时最小的桌面级CPU处理器，应用于Samsung Galaxy Book S、Lenovo X1 Fold等超薄笔记本中[25]。

2022年，Intel发布号称芯片航空母舰的Ponto Vecchio（PVC），这款芯片从SOC架构设计到实现工艺都代表Intel当前最先进的芯片水平。PVC结合了3D IC和2.5D技术优势，共有63个芯粒，其中47个是功能芯粒，另外16个芯粒用于支撑与散热，共由超过1 000亿个晶体管构成。PVC有两层堆叠，采用了混合SOC on IP和MOL 3D架构，顶层有16个计算芯粒和8个内存芯粒，内含128个Xe核和120 MB扩展L3缓存，底层有2个大的基础互联芯粒，内含片上集成电源模块（FIVR）、内存控制器、PCIe、CXL、L2缓存等。3D异步互联速率达2.97 GHz，互联功耗0.2 pj/bit。PVC是同时采用多家供应商不同工艺芯粒进行3D IC堆叠的第一款大规模商用产品，其计算芯粒为TSMC N5，内存芯粒与基础芯粒为Intel N7。这表明3D IC可在紧耦合互联层面弥合多家供应商界限，打造全新商业模式。PVC应用在阿贡实验室定制的面向人工智能与科学计算的超算服务器，预期算力可达2 ExaFlops[26]。

AMD的3D IC商用化脚步紧随Intel之后。2022年，AMD发布第一款使用3D V-cache技术的产品，底层是计算芯粒，内含8个CPU核和32 MB的L3缓存，顶层是内存芯粒，内含64 MB L3缓存。3D V-cache采用TSMC最新3D SoIC技术，利用间距9 μm的高密度TSV与混合键合（HB）进行互联。3D V-cache架构本质属于Cache on Cache的

MOM，其中3D堆叠互联区域只有L3缓存。3D V-cache目前有两代，第一代的内存芯粒与计算芯粒都是TSMC N7，第二代把计算芯粒升级为N5，复用之前N7的内存芯粒堆叠。3D V-cache同步登录AMD服务器与桌面处理器产品线。相较于2D版本，3D版本服务器性能可提升66%[27]。

2023年，AMD发布了新一代3D IC处理器架构——Instinct MI300。MI300的3D IC架构与PVC类似，也是两层堆叠SOC on IP和MOL混合3D架构。顶层是TSMC N5工艺计算芯粒，针对不同的产品线有两种分型：MI300A由3个Zen4架构CPU芯粒和6个CDNA3架构GPU芯粒组成混合架构；MI300X把3个CPU芯粒替换成2个GPU芯粒，共8颗GPU芯粒组成纯GPU架构。底层是4个TSMC N6工艺的基础芯粒，内含集成输入和输出（I/O）接口、路由仲裁与拓展缓存等模块。MI300主要面向大语言模型与高性能计算，用于劳伦斯-利弗莫尔国家安全实验室El Capitan超级计算机，算力可达2 ExaFlops[28]。

2022年，人工智能芯片公司Graphcore推出创新架构智能处理器——Bow IPU，使用TSMC 3D SoIC晶圆堆叠晶圆（WOW）方式，顶层是人工智能逻辑芯粒，底层是集成了深硅刻蚀电容（DTC）的电源管理芯粒，可以改善电源完整性并获取更高能耗效率[29]。

Meta公司在2024年国际固态电路会议（ISSCC 2024）上报告了一款用于穿戴式虚拟现实应用的3D IC原型芯片，基于TSMC SoIC W2W F2F堆叠，顶层是扩展SRAM与传感器模块，底层是芯片启动与控制系统。通过三维堆叠，这一产品至少可节约55%的内存访问功耗，提升40%的系统性能，超小的体积也为边缘式AI应用拓展了兼容性。这是3D IC在消费级应用方面的重大创新[30]。

在2022年ISSCC上，阿里达摩院发布了一款完全自主制造的基于3D IC的近存计算人工智能样片。该产品采用了3D WOW形式HB堆叠，顶层是25 nm工艺的DRAM芯粒，共36 Gbit DRAM，底层是55 nm工艺的逻辑芯粒，由神经网络引擎（NE）、匹配引擎（ME）等组成。虽然仅采用了远低于业界先进的工艺节点，但是3D跨Die互联速率可达150 MHz，3D互联功耗仅为0.88 pj/bit，并且吞吐量、带宽、能耗效率等相比其他采用更先进工艺的2D和2.5D芯片也有较大优势[10]。

6 机遇与挑战

可以看到，最近几年，3D IC芯片如井喷般涌现，根本原因是摩尔定律推进困难，先进工艺红利不再，存储墙问题凸显。伴随着3D制造工艺尤其是TSV和HB工艺的逐渐成熟，物理尺寸可与BEOL尺寸相比。此外，高性能计算、人工智能、大模型、数字孪生等场景应用对芯片大算力、大内存、大带宽的需求进一步加大。目前看来，3D IC的主要优势在于通过互联和内存性能提升产品高度，通过功耗与面积优化拓宽产品广度。随着工艺优化和产品迭代带来的成本与风险降低，3D IC也必将步入大规模商用化的道路。我们应该把握机遇，勇于做科技的领跑者。

致谢

感谢深圳市中兴微电子技术有限公司高级工程师武辰飞、李乐琪、黄彤彤、高静丽对本研究的帮助！

参考文献

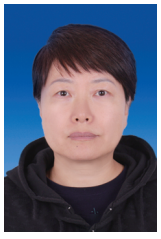
[1] GARROU P, BROWN C, RAMM P. Handbook of 3D integration, volume 1: technology and applications of 3D integrated circuits [M]. New York: John Wiley & Sons, 2011
[2] LAU J H. Chiplet design and heterogeneous integration packaging [M]. 2023
[3] LI Y, GOYAL D. 3D microelectronic packaging: from architectures to applications [M]. Second edition. Singapore: Springer, 2021
[4] GHOLAMI A, YAO Z, KIM S, et al. AI and memory wall [EB/OL]. (2024-03-21)[2024-08-08]. https://arxiv.org/pdf/2403.14123
[5] NAUTA B. 1.2 racing down the slopes of Moore's law [C]// Proceedings of IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2024: 16-23. DOI: 10.1109/ISSCC49657.2024.10454417
[6] TSAI Y F, XIE Y, VIJAYKRISHNAN N, et al. Three-dimensional cache design exploration using 3DCacti [C]//Proceedings of International Conference on Computer Design. IEEE, 2005: 519-524. DOI: 10.1109/ICCD.2005.108
[7] NAEIM M, YANG H Q, CHEN P H, et al. Design enablement of 3-dies stacked 3D-ICs using fine-pitch hybrid-bonding and TSVs [C]//Proceedings of IEEE International 3D Systems Integration Conference (3DIC). IEEE, 2023: 1-4. DOI: 10.1109/3DIC57175.2023.10155075
[8] ZHU L J, TA T, LIU R, et al. Power delivery and thermal-aware arm-based multi-tier 3D architecture [C]//Proceedings of IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED). IEEE, 2021: 1-6. DOI: 10.1109/ISLPED52811.2021.9502481
[9] CHEN K, LI S, MURALIMANOHAR N, et al. CACTI-3DD: architecture-level modeling for 3D die-stacked DRAM main memory [C]//Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2012: 33-38
[10] NIU D M, LI S C, WANG Y H, et al. 184QPS/W 64Mb/mm^{2}3D logic-to-DRAM hybrid bonding with process-near-memory engine for recommendation system [C]//Proceedings of IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2022: 1-3. DOI: 10.1109/ISSCC42614.2022.9731694
[11] CHEN R, LOFRANO M, MIRABELLI G, et al. Power, performance, area and thermal analysis of 2D and 3D ICs at A14 node designed with back-side power delivery network [C]// Proceedings of International Electron Devices Meeting (IEDM). IEEE, 2022: 23.4.1-23.4.4. DOI: 10.1109/IEDM45625.2022.10019349
[12] FEERO B S, PANDE P P. Networks-on-chip in a three-dimensional environment: a performance evaluation [J]. IEEE

transactions on computers, 2009, 58(1), 32–45
[13] RAHMAN A, REIF R. System–level performance evaluation of three–dimensional integrated circuits [J]. IEEE transactions on very large scale integration (VLSI) systems, 2000, 8(6): 671–678. DOI: 10.1109/92.902261
[14] BLACK B, NELSON D W, WEBB C, et al. 3D processing technology and its impact on iA32 microprocessors [C]// Proceedings of IEEE International Conference on Computer Design: VLSI in Computers and Processors, 2004. ICCD 2004. Proceedings. IEEE, 2004: 316–318. DOI: 10.1109/ICCD.2004.1347939
[15] PUTTASWAMY K, LOH G H. Implementing register files for high–performance microprocessors in a die–stacked (3D) technology [C]//Proceedings of IEEE Computer Society Annual Symposium on Emerging VLSI Technologies and Architectures (ISVLSI'06). IEEE, 2006: 6. DOI: 10.1109/ISVLSI.2006.56
[16] OUYANG J, SUN G, CHEN Y, et al. Arithmetic unit design using 180nm TSV–based 3D stacking technology [C]//Proceedings of IEEE International Conference on 3D System Integration. IEEE, 2009: 1–4. DOI: 10.1109/3DIC.2009.5306565
[17] ZHANG T, WANG K, FENG Y, et al. A 3D SoC design for H.264 application with on–chip DRAM stacking [C]//Proceedings of IEEE International 3D Systems Integration Conference (3DIC). IEEE, 2010: 1–6. DOI: 10.1109/3DIC.2010.5751446
[18] THOROLFSSON T, GONSALVES K, FRANZON P D. Design automation for a 3DIC FFT processor for synthetic aperture radar: a case study [C]//Proceedings of 46th ACM/IEEE Design Automation Conference. IEEE, 2009: 51–56
[19] KIM D H, ATHIKULWONGSE K, HEALY M, et al. 3D–MAPS: 3D massively parallel processor with stacked memory [C]// Proceedings of IEEE International Solid–State Circuits Conference. IEEE, 2012: 188–190. DOI: 10.1109/ISSCC.2012.6176969
[20] KIM D H, ATHIKULWONGSE K, HEALY M B, et al. Design and analysis of 3D–MAPS (3D massively parallel processor with stacked memory) [J]. IEEE transactions on computers, 2015, 64 (1): 112–125. DOI: 10.1109/TC.2013.192
[21] VIVET P, GUTHMULLER E, THONNART Y, et al. IntAct: a 96–core processor with six chiplets 3D–stacked on an active interposer with distributed interconnects and integrated power management [J]. IEEE journal of solid–state circuits, 2021, 56(1): 79–97. DOI: 10.1109/JSSC.2020.3036341
[22] CAVALCANTE M, AGNESINA A, RIEDEL S, et al. MemPool–3D: boosting performance and efficiency of shared–L1 memory many–core clusters with 3D integration [C]//Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2022: 394–399
[23] LOH G H, XIE Y, BLACK B. Processor design in 3D die–stacking technologies [J]. IEEE micro, 2007, 27(3): 31–48. DOI: 10.1109/MM.2007.59
[24] Batude P, Ernst T, Arcamone J, et al. 3–D sequential integration: a key enabling technology for heterogeneous co–integration of new function with CMOS [J]. IEEE journal on emerging and selected topics in circuits and systems, 2012, 2(4): 714–722
[25] GOMES W, KHUSHU S, INGERLY D B, et al. 8.1 lakefield and mobility compute: a 3D stacked 10 nm and 22FFL hybrid processor system in 12 × 12 mm^2, 1 mm package–on–package [C]//Proceedings of IEEE International Solid–State Circuits Conference – (ISSCC). IEEE, 2020: 144–146. DOI: 10.1109/ISSCC19947.2020.9062957
[26] GOMES W, KOKER A, STOVER P, et al. Ponte vecchio: a multi–tile 3D stacked processor for exascale computing [C]// Proceedings of IEEE International Solid–State Circuits Conference (ISSCC). IEEE, 2022: 42–44. DOI: 10.1109/ISSCC42614.2022.9731673
[27] BURD T, LI W, PISTOLE J, et al. Zen3: the AMD 2nd–generation 7nm x86–64 microprocessor core [C]//Proceedings of IEEE International Solid–State Circuits Conference (ISSCC). IEEE, 2022: 1–3. DOI: 10.1109/ISSCC42614.2022.9731678
[28] SMITH A, LOH G H, SCHULTE M J, et al. Realizing the AMD exascale heterogeneous processor vision: industry product [C]// Proceedings of ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA). IEEE, 2024: 876–889. DOI: 10.1109/ISCA59077.2024.00068
[29] MOORE S. Graphcore uses TSMC 3D chip tech to speed AI by 40% [EB/OL]. (2022–03–03) [2024–08–08]. https://spectrum.ieee.org/graphcore–ai–processor
[30] WU T F, LIU H C, SUMBUL H E, et al. 11.2 A 3D integrated Prototype System–on–Chip for Augmented Reality Applications Using Face–to–Face Wafer Bonded 7 nm Logic at <2 μm Pitch with up to 40% Energy Reduction at Iso–Area Footprint [C]// Proceedings of IEEE International Solid–State Circuits Conference (ISSCC). IEEE, 2024: 210–212. DOI: 10.1109/ISSCC49657.2024.10454529

作 者 简 介

陈昊，移动网络和移动多媒体技术国家重点实验室研究员、深圳市中兴微电子技术有限公司后端设计部后端设计工程师；主要研究方向为三维集成电路物理实现，负责先进技术规划和研发工作。

谢业磊，移动网络和移动多媒体技术国家重点实验室研究员、深圳市中兴微电子技术有限公司封测工程部封装设计专家；拥有9年以上封装设计研究经验，负责多个先进封装预研项目的开发工作。



庞健，移动网络和移动多媒体技术国家重点实验室研究员、深圳市中兴微电子技术有限公司先进封装技术总工；负责封装技术规划和研发，完成多款先进封装交付。

欧阳可青，深圳市中兴微电子技术有限公司副总经理、IC平台研发中心主任，并担任射频异质异构集成全国重点实验室副主任、移动网络和移动多媒体技术国家重点实验室研究中心主任；长期从事复杂SOC芯片的设计方法学研究工作，在先进工艺、数模混合设计、2.5D/3D先进封装、高可靠性设计等领域取得多项关键技术突破。